THE COMPLETE TECHNICAL PAPER PROCEEDINGS

FROM:

NCTA

Technical Papers

A Comparison of Economic and Operational Tradeoffs for the Deployment of Broadcast, Multicast, and Unicast Infrastructures within an IP Video Environment

Carol Ansley, Jim Allen, Tom Cloonan
ARRIS

Abstract

As the Cable Industry evaluates the incorporation of IP Video as the next stage of video delivery, an important consideration is the need for analogues of the current Broadcast, Switched, and Unicast protocols within the IP Video deployments. Initially, it was assumed that the new IP Video world would look much like the current Legacy Video world, with its own architecture based on a triplet of protocols- one for Broadcasted (always-on Multicast) video, one for Switched (Multicast) Video, and one for Unicast Video. As the industry has continued to traverse the complex learning curve, this fundamental understanding has come into question.

Arguments have been made for the elimination of Broadcast, based on the idea that a Multicast deployment would provide increased network efficiencies. An opposing viewpoint is that a small Broadcast tier coupled together with a Unicast tier might provide greater network simplicity by eliminating the need for (and complexities of) a Multicast tier. This paper will use simulations based upon subscriber behavior to explore design approaches for several possible deployment scenarios. The analysis would consider network efficiency, possible economic factors, and possible feature interactions in an effort to help guide MSO decisions as they move forward towards future IP Video deployments.

ON VIDEO EVOLUTION

Legacy video services have long been the core of the basic cable service offering. In the distant past, these services were offered using NTSC-based analog programming, with one program per 6 MHz channel (in North America). In the past twenty years, this service has been augmented (and in some cases, replaced) with the arrival of MPEG-based digital video services transporting digital program streams over Quadrature-Amplitude Modulated (QAM) 6 MHz channels. This new service capitalized on advanced coding and compression techniques that permitted ten or more standard-definition program streams to be temporally multiplexed into a single 6 MHz channel (in North America).

In a legacy video environment, there are typically two distinct video service types offered to subscribers:

a) Linear video services
b) Video on Demand (VoD) services

Linear Video Services

Linear video services have been a part of the cable networks since their inception. Linear video services provide the “normally scheduled” program line-up to subscribers, with transitions between programs usually occurring at half-hour increments throughout the day. A program has a pre-assigned, scheduled time-slot when it is transmitted, and as many viewers as are interested can watch the broadcast program feed at the same time. Over the years a multicasting technology called Switched Digital Video (SDV) also evolved, reducing bandwidth demands by enabling program delivery only when a subset of one or more subscribers wanted to watch a program. Thus, we can define the two common methods used for the delivery of Linear video services:

a) Broadcast - each of the Linear program streams is transmitted from the head-end over the HFC plant to all of the subscribers. As a result, all programs consume bandwidth at all times, whether being viewed or not. However, Broadcast offers the benefit that only a single copy of the program needs to be transmitted into a particular Service Group when multiple users are viewing the program and a single headend signal can be split to accommodate any number of service groups within the same ad zone.
b) Switched Digital Video - only the Linear programs that are currently being viewed by one or more subscribers within a Service Group are transmitted from the head-end to that Service Group. Linear programs that are not being viewed are not transmitted, so bandwidth savings result relative to Broadcast techniques of transmission. These bandwidth savings do not come for free, because they do require a two-way protocol to exist between the client devices and a head-end management system. The actual magnitude of the bandwidth savings over straight broadcast depends on many factors, which are discussed later in this paper. Like Broadcast, SDV offers the benefit that only a single copy of the program needs to be transmitted into a particular Service Group regardless of the number of users viewing the program.

__Video On Demand Services__

VoD services permit offerings such as standard Video on Demand, Network Digital Video Recording (nDVR), and Start-Over. VoD services offer subscribers access to an extended library of stored video content. These “extra” programs are traditionally provided as a free or fee-based service. Viewers can select a program from the Video on Demand content library at their convenience. They can start and stop the program as they wish, and often trick modes such as fast-forward, rewind, and pause are available. Since it is unlikely that two subscribers will choose to watch the same VoD program at exactly the same time, no effort is made to broadcast or multicast VoD content. It is simply unicast to the single user who has requested the content. Unlike Broadcast and SDV feeds, each new VoD selection must be sent individually to each new viewer, so there is a one-for-one utilization of bandwidth for each new stream.

__IP Video Services__

Just as MPEG-based digital video services were used to augment and in some cases replace analog video services over the past twenty years, a new technology is now being viewed by the cable industry as a potential augmentation (or eventual replacement) for MPEG-based digital video services. This new entrant capitalizes on the recent advances in DOCSIS technology and advances in video delivery over IP.

IP Video delivers encoded and compressed video program content from origin servers to client devices by inserting the audio and video information into the payloads of Internet Protocol (IP) packets that are then passed over

IP networks. IP Video architectures have the potential to enable support of new end devices and new revenue opportunities based on personalized advertising or expanded video services.

As MSOs begin to architect new video delivery systems to take advantage of IP Video techniques, the video delivery models that have been successfully utilized in the legacy video delivery world come first to mind. As such, one would expect that MSOs might consider IP Video as a delivery system for all of the following service types:

1. IP Video VoD services
    a. Standard VoD
    b. nDVR
    c. Start-Over
2. IP Video Linear services
    a. IP Video Linear Always-On services (similar to legacy Linear Broadcast services)
    b. IP Video Linear Switched services (similar to legacy Linear SDV services)

It should be clear that the various IP Video VoD services will likely be delivered using point-to-point IP Video unicast delivery. These services will likely be based on the latest IP/TCP/HTTP transport technologies, the dominant protocol stack used for unicast IP Video Streaming. The increasing use of non-television devices to access this content also suggests that reuse of popular Internet technology would be advantageous, when it fits with the unique cable industry infrastructure.

What is less clear is the most efficient method or methods to implement the delivery of traditional Linear Video services over IP. There are many ways to emulate Linear Video delivery within an IP environment so that, in the end, the video is ultimately delivered via IP packets to client devices within the subscriber's home. Some of the possible techniques that are currently under consideration are enumerated here:

a) Point-to-point, unicast IP/TCP/HTTP packet streaming from head-end origin servers (or caching servers) over DOCSIS to each individual subscriber client device, requiring lots of point-to-point connections to be established for popular programs
b) Dynamic, point-to-multipoint, multicast IP/UDP packet delivery from head-end origin servers (or caching servers) over DOCSIS to any subscriber clients that join the multicast stream
c) Always-On, point-to-multipoint, multicast IP/UDP packet delivery from head-end origin servers (or caching servers) over DOCSIS to any subscriber clients that join the multicast stream
d) Legacy Linear MPEG-TS transmission on the HFC plant, with IP Encapsulation in a residential Media Gateway and unicast IP/TCP/HTTP Video delivery within the home network, re-uses the existing MPEG-based delivery infrastructure and reduces IP Video architectural complexity

It should be noted that of the various techniques listed above, the first three utilize IP delivery techniques to the home. The final technique utilizes a unique hybrid approach, where the content is sent via traditional MPEG-TS delivery over the HFC network, but then uses an IP unicast stream for final delivery over the home network. This last technique, while valid for consideration as a method for IP video delivery overall, will not be discussed further in this paper. We will concentrate on discussion of IP Video architectures that use IP in the transport arena.

Another layer of complexity that is not addressed in this paper is the Quality of Service (QoS) architecture for IP Video. Within DOCSIS, there is a substantial QoS infrastructure that can preserve or improve the performance of individual IP Video flows with respect to the overall volume of IP traffic. Subscribers today are accustomed to always-on TV service from their perspective, SDV is typically engineered to be indistinguishable from legacy broadcast delivery. An important part of the IP video architecture will involve the decision to preserve, or not, the current levels of video service reliability and the implementation of that decision. While it has some relevance to the protocol topics discussed in this paper, the QoS topic is complex enough to warrant another discussion focused on that topic.

With IP technologies, each technique for transport (unicast, multicast, broadcast) has its own set of advantages and disadvantages. For example, Unicast is relatively simple to deploy since it is based on variants of basic HTTP transactions. It currently being used by several MSOs to provide some IP-based subscribers with the a limited equivalent of Linear services, but a unicast approach can be wasteful of the limited HFC bandwidth if any unicast program is actually sent to more than one subscriber at the same time. Unicast delivery also suffers from a "simulcast effect" that may exist if early deployments of IP video begin while legacy video distribution to legacy STBs is also in place, as the same programs will need to be simulcast across both distribution systems.

Multicast, in dynamic or static varieties, can be more complex to deploy since it may require an additional headend server to support bandwidth management, similar to SDV. Depending upon the current configuration of a headend's routers, switches and CMTSs, they may also require upgrades to support multicast protocols. If these multicast or broadcast techniques are eventually utilized, they will yield bandwidth savings on the HFC plant due to the fact that multiple viewers of a single stream within a particular Service Group will not require extra replications. A dynamic multicast approach is more bandwidth efficient than a static Always-On approach. The multicast approaches may also suffer from the "simulcast tax" mentioned above, but the overall bandwidth cost of that simulcast may be reduced by the inherent advantages of dynamically switching a program in only when it is actually to be viewed.

The rest of this paper attempts to quantify and explore the many tradeoffs associated with these technologies.

UNICAST IP VIDEO DELIVERY

It is important that we clearly define the protocols that we have analyzed in the paper for delivery of unicast IP Video. If it is mapped into the layers of the Open System Interface (OSI) model, IP Video clearly uses IP as its Layer 3 (Network Layer) protocol. However, it can use any one of two different Layer 4 (Transport Layer) protocols: Transmission Control Protocol (TCP) or User Datagram Protocol (UDP). TCP is a connection-oriented protocol that provides guaranteed packet delivery, flow control, and congestion control for the data transport, whereas UDP is a simpler, connectionless protocol that provides none of the advanced services of TCP.

During the early days of IP Video (in the early-1990s), the content was initially delivered to the home using Ethernet/IP/UDP/RTP encapsulations. Custom players and custom servers were usually utilized. It worked fairly well, but it did run into issues with in-home NAT boxes and congested networks.

The original IP Video Download protocols were used for over a decade, and they are still used (to some extent) today. However, the latest improvements in IP Video delivery began to be utilized in the middle of the 2000 decade. This new approach is oftentimes called HTTP-based Adaptive Streaming.

HTTP-based Adaptive Streaming has come to be used quite extensively in most applications that require a unicast IP Video stream to be delivered from a single origin server to a single client device. The application program typically uses TCP transport services for downloading fragments of the video content file by invoking Hypertext Transfer Protocol (HTTP).

HTTP-based Adaptive Streaming replaced the single HTTP GET message of the first Downloading protocols with a series of repeated HTTP GET messages, with each HTTP GET message requesting a different, small chunk (or fragment) of the video content file. As a result, only the video content that is to be viewed is actually requested, so the problems associated with wasted bandwidth are minimized. In addition, since the video fragments tended to be fairly short in duration (2-10 seconds was typical), it was easy to efficiently support simple trick modes. The short-duration fragments also made it possible for the clients to rapidly identify network congestion and adjust their HTTP GET messages to request higher or lower resolution fragments that could be accommodated by the available network bandwidth at any instant in time. These rapid adjustments in the resolution (and bit-rate) of successively-requested video fragments came to be known generically as HTTP-based Adaptive Streaming.

Current Unicast IP Video is essentially a TCP-based, HTTP pull model, with unicast packets only being sent from the source to the destination whenever the destination requests the content (with HTTP GETs). Other than the routing tables that help to steer the packets, no other state information is required within the intermediate network elements to ensure correct transmission of the packets between the source and the destination. The typical control plane protocols and data plane exchanges for Unicast IP Video are illustrated in **Fig. 1** and **Fig. 2**.

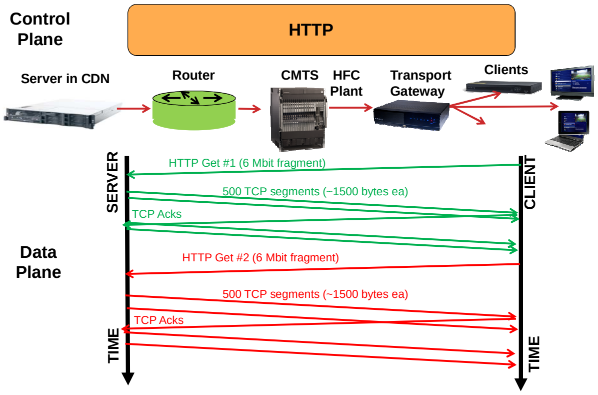

Control Plane
HTTP
Server in CDN
Router
CMTS
HFC Plant
Transport Gateway
Clients
SERVER
CLIENT
HTTP Get #1 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
Data Plane
HTTP Get #2 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
TIME
TIME

Figure 1 - Unicast over Transport Gateway

Fig. 1 illustrates an example unicast architecture where clients send HTTP GETs directly to the head-end video server, and a Transport Gateway merely passes the upstream and downstream packets between the HFC plant and the Home Network.

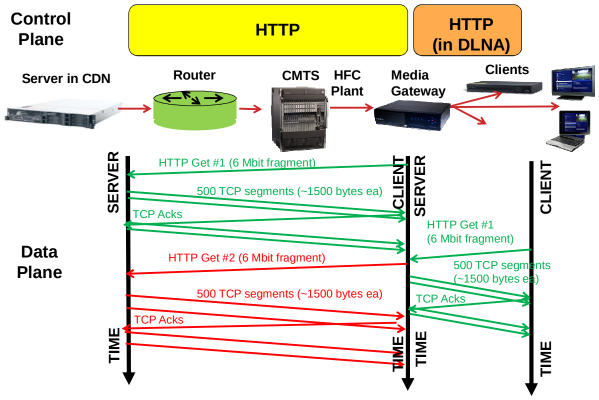

Control Plane
HTTP
HTTP (in DLNA)
Server in CDN
Router
CMTS
HFC Plant
Media Gateway
Clients
SERVER
CLIENT
SERVER
CLIENT
HTTP Get #1 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
HTTP Get #1 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
Data Plane
HTTP Get #2 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
TIME
TIME
TIME
TIME

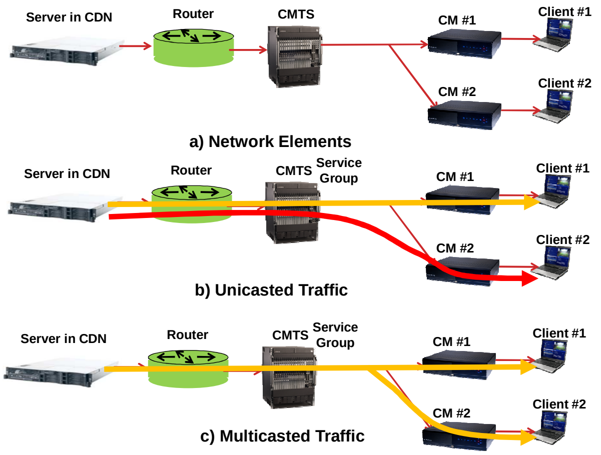
Figure 2 - Unicast over Caching Media Gateway

Fig. 2 illustrates an example unicast architecture where clients send HTTP GETs through a DLNA network within the home to a server application running on the Media Gateway within the home, and the Media Gateway also has an HTTP client application running on it that would have (hopefully) previously used HTTP GETs to request and cache fragments from the head-end video server. If the content was not cached, then the Media Gateway would simply relay the HTTP GET upward towards the head-end video server.

Caching operations in the network may be added to reduce traffic on the back bone network, but do not appreciably add to overall architectural complexity.

MULTICAST IP VIDEO DELIVERY

The use of Multicast for IP Video delivery improves bandwidth efficiencies over Unicast video delivery on the HFC plant as well as on the MSO's back-office network. This fact is simply illustrated in **Fig. 3**, where we have assumed that two users (Client #1 and Client #2) are both accessing the same linear video content at the same time. For comparative purposes, we will assume that the bandwidth associated with this video content is 7 Mbps. If delivered using Unicast, then the resulting bandwidth consumed in both the HFC plant and the MSO back-office network is 14 Mbps, since two separate streams containing the video content must be propagated through the network. If delivered using Multicast, then the resulting bandwidth consumed in both the HFC plant and the MSO back-office network is only 7 Mbps, since only a single stream containing the video content is propagated through the network- both CM #1 and CM #2 receive and pass the stream on to their respective clients, resulting in the inherent "replication" of the stream near the stream destinations.



a) Network Elements

b) Unicasted Traffic

c) Multicasted Traffic

Figure 3 - Unicast vs. Multicast

While Multicast IP Video delivery is more bandwidth efficient for streams that are simultaneously viewed by more than one recipient, it is also more complex to manage than unicast IP Video delivery. This added complexity is primarily due to the fact that multicast IP Video requires additional protocol support in the intermediate network elements and the client devices.

Multicast IP Video is quite different from Unicast IP Video. Since there are multiple destinations receiving the Multicast IP Video feed, the TCP-based, HTTP pull model used in

Unicast IP Video cannot be utilized for Multicast IP Video.

As a result, a UDP-based push model is used for IP Multicast, with packets being sent from the source to the multiple destinations without HTTP GETs or TCP ACKs being required. While one could (in theory) send the IP Multicast to all possible destinations, that approach would be quite wasteful of both bandwidth within the network and processing power within all of the destinations. As a result, standard IP Multicast solutions limit the scope of destinations to which the multicast streams are sent. In particular, the multicast streams (which are identified by a particular Multicast Group IP Address as the Destination Address within the IP packet header) are only sent to destinations that have formally requested that the stream be transmitted to them. This formal request is typically made within a LAN using the Internet Group Multicast Protocol (IGMP) for IPv4 systems and using the Multicast Listener Discovery (MLD) protocol for IPv6 systems.

In both cases (IGMP and MLD), the destination desiring access to the content within a multicast stream would typically use the appropriate protocol to send a “Join Message” (a.k.a. a Membership Report or a Multicast Listener Report) that would be broadcast to the router(s) in its LAN. If/when the destination desires to no longer receive that particular multicast stream, it can optionally send a “Leave Message” (a.k.a. a Leave Group or a Multicast Listener Done). In order for routers to stimulate destinations to report that they are joined to a particular group, they would typically send a “Query Message” (a.k.a. a Membership Query or a Multicast Listener Query). Routers must maintain state information indicating which of their ports have listeners that have indicated a desire (via Join Messages) to receive each multicast stream. The routers then must forward packets associated with each particular multicast streams to the ports that have listeners associated with that multicast stream. Routers typically communicate their desire to receive a multicast stream from other routers using one of several possible multicast routing protocols, including PIM-SSM, PIM-SM, PIM-DM, DVMRP, MOSPF, MBGP, and CBT. The routers involved in multicast address exchanges must be capable of communicating using a common multicast routing protocol.

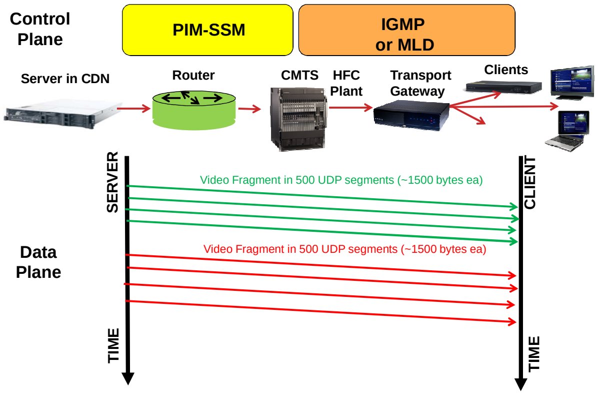


Figure 4 - Multicast with UDP

There are many different architectures that one can envision for the deployment of a Multicast IP Video delivery system- several of them are illustrated below. **Fig. 4** illustrates an example architecture with a data plane that uses UDP transport of multicast IP Video packets from the head-end multicast server to the clients in the home, with the packets passing through a Transport Gateway within the home. The control plane within **Fig. 4** uses IGMP or MLD to establish the multicast path between the clients and the CMTS, and it uses PIM-SSM to establish the multicast path between the CMTS and back-office routers and back-office multicast server.

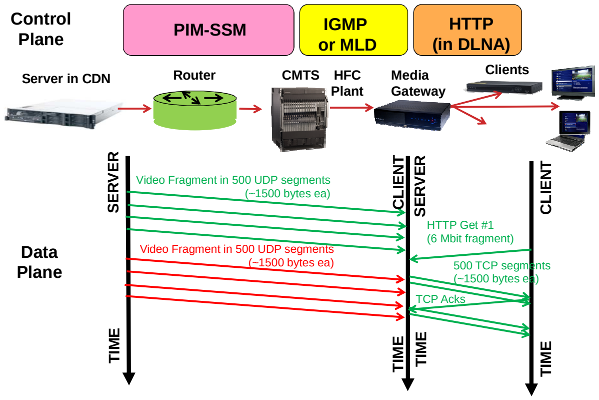

Control Plane
PIM-SSM
IGMP or MLD
HTTP (in DLNA)
Server in CDN
Router
CMTS
HFC Plant
Media Gateway
Clients
SERVER
CLIENT
SERVER
CLIENT
Video Fragment in 500 UDP segments (~1500 bytes ea)
Data Plane
Video Fragment in 500 UDP segments (~1500 bytes ea)
HTTP Get #1 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
TIME
TIME
TIME
TIME

Figure 5 - Multicast with UDP & Conversion to Unicast TCP

Fig. 5 illustrates an example architecture with a data plane that uses UDP transport of multicast IP Video packets from the head-end multicast server to the Media Gateways in the home, with the video content file being re-constituted by the Media Gateway. The Media Gateway then acts as an HTTP server to distribute the video content over a unicast HTTP/DLNA connection to the HTTP client within the Home Network. The control plane within Fig. 5 uses HTTP/DLNA within the Home Network, it uses IGMP or MLD to establish the multicast path between the Media Gateway and the CMTS, and it uses PIM-SSM to establish the multicast path between the CMTS and back-office routers and back-office multicast server.

The widespread deployment of SDV has also proven in many parts of the multicast technology that are directly applicable to multicast IP Video distribution. Many optimizations that were developed to make SDV robust and efficient can be extended to the IP Video world to enable IP Multicast to be successful. One example technique would be the automatic joining of all available SDV multicast streams by an Edge QAM even before any specific end device has chosen one of those programs. This pre-join speeds up the acquisition of a new program by a client, as the Edge QAM merely needs to be instructed by an SDV server which stream to activate on which QAM and PIDs. This feature is not a part of traditional multicast as practiced by IT professionals, but it an obvious improvement directly applicable to a CATV IP Video architecture. The analogous IP Video feature would instruct the CMTS to join all IP Video multicasts, which would let a device activate a new stream with just a transaction with its CMTS. Since the CMTS manages its own bandwidth constraints, the SDV Server's bandwidth allocation might be transferred entirely to the CMTS, resulting in no new network elements for multicast.

COMPARING UNICAST AND MULTICAST IP VIDEO

The primary benefit of Multicast IP Video delivery is its basic ability to reduce the bandwidth required to deliver video content to multiple destinations when two or more of those destinations are viewing the same content at the same time. Many MSOs already treat bandwidth on their HFC networks as a critical and precious resource as multiple services compete for that bandwidth and the situation can only become more contentious as HSD continues to increase its requirements and video any time anywhere continues as well. If these trends continue, then the primary benefit of Multicast IP Video may prove to be very important.

Unicast IP Video delivery also has a place, even with its bandwidth usage, since it can provide a simple deployment model for early stages of IP video deployments, when the concentration of IP video users in any one service group is low. Trends within the universe of any time any place video distribution may also tend to accelerate the usage of network DVR and other unicast services, which will increase the amount of natively-unicast traffic.

Depending upon the stage of the IP Video deployment and the deployment choices connected with other related areas, such as network DVR, the answer of what may be the most efficient may vary depending upon whether one considers network bandwidth, operational/deployment costs, and service flexibility. When considering a real world deployment, the answer may even be that a mix of technologies will be required to ensure that MSOs can obtain an optimal efficiency from their HFC plant for video delivery.

Some of the issues to be considered are listed below.

1. Common protocols for multicast IP Video and any optimizations over DOCSIS should be available in an open forum, similar to the TWC ISA or Comcast NGOD SDV specifications
2. Any optimizations for Unicast IP Video that allow robust performance for first screen viewing should be provided in an open forum for maximum benefit
3. Current encoding methods will require multiple choices for unicast and/or multicast stream delivery, new encoding choices, such as SVC, could improve multicast efficiency for multicast and stream management for unicast
4. Reliability concerns in the IP Video packet delivery
5. Distributed Denial Of Service attacks by hackers on head-end equipment
6. Multicast must be tied into a Connection Admission Control algorithm to identify overload conditions when a new Multicast stream cannot be set up

While the issues listed above should be carefully considered, it is important to note that many technical and architectural proposals have already been created to mitigate most of the issues. Granted, some of these proposals require more complexity to be added to the equipment, but they nevertheless provide solutions to the problems.

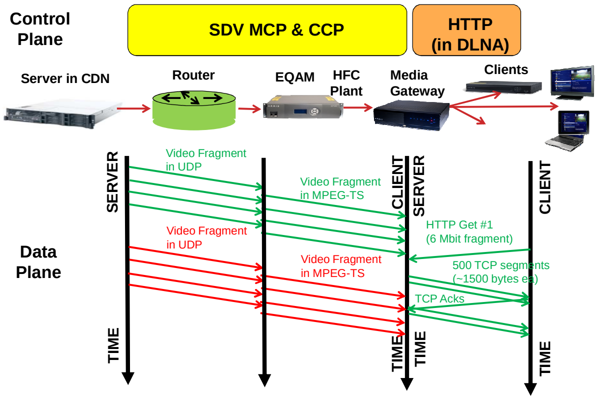


Figure 6 - Multicast with UDP & MPEG-TS & Conversion to Unicast TCP

Fig. 6 illustrates an example hybrid architecture with a data plane that uses UDP transport of multicast IP Video packets from the head-end multicast server to the head-end EdgeQAM, MPEG-TS transport from the head-end EdgeQAM to the Media Gateways in the home. The Media Gateways re-constitute the video content file. The Media Gateway then acts as an HTTP server to distribute the video content over a unicast HTTP/DLNA connection to the HTTP client within the Home Network. The control plane within **Fig. 6** uses HTTP/DLNA within the Home Network, and it uses SDV-oriented protocols like the Channel Change Protocol (CCP) and the Mini-Carousel Protocol (MCP) to establish a video stream flow between the back-office server and the Media Gateway.

SIMULATION RESULTS

The fact that Broadcast, Unicast and Multicast are closely related protocols allows us to simulate their respective behavior using a common simulation base – modeling both Broadcast and Unicast as special cases of the more general Multicast model. Of these three protocols, however, only Unicast natively permits trick modes such as pause and replay – a quality that, though it may be quite valuable to viewers, has no correspondence in the other two protocols. We have focused, therefore, on modeling only properties (listed below) that can be used to describe all three protocols.

Broadcast can support an arbitrarily large number of viewers (when the downstream program capacity is sufficient to carry every program in the lineup). Unicast, on the other hand, can support an arbitrarily large number of offered programs (when the downstream capacity is sufficient to dedicate a separate program channel for every viewer). Multicast, however, possesses both of these properties and is also able to provide bandwidth-efficient service even when either of the two above constraints on the downstream program capacity cannot be met.

Viewer Modeling Parameters

Two attributes of a video delivery network lie largely outside the control of the MSO. These properties can be measured but not controlled by the MSO. Numerical values for these parameters are best attained through careful analysis of actual viewer tuning behavior. These attributes are:

1. Acceptable Tuning Blockage Probability
2. Program Viewership Popularity

Customers will ultimately decide with their feet how often (relative to competing providers) they are willing to tolerate being denied a program that they have requested. Video service providers, however, are forced to make a reasonable guess at exactly what this limit of viewer tolerance might be, as we know of no applicable field study in this area. Throughout this paper we have assumed viewers will be satisfied if they are denied a program selection request no more than 0.1% of the time (or once per 1000 tuning requests).

It is also the customer population that determines the relative popularity of each of the programs offered in the lineup. Modeling this property of the viewer population can present a significant challenge since relative program popularity varies substantially with time-of-day, day-of-week and with the demographics of the neighborhood served by the service group.

While neither Broadcast nor Unicast services are sensitive to program popularity, the relative popularity of programs in the offered lineup plays a significant role in Multicast by determining how many programs can be expected to be multiplexed onto a limited amount of downstream bandwidth.

Fortunately the dynamic nature of a Multicast protocol causes it to automatically adapt to changes in relative program popularity (both temporally and also between service groups). This means that it is not so important to know exactly which programs are most popular – only that we know in a general sort of way.

A number of studies have suggested that if we first sort a program lineup by market share, from most to least popular, then the popularity or market share of a program (n) can be approximated using a Power Law Distribution, shown here:

$$P_n = n^{-\alpha} / \sum_{n=1}^{N} n^{-\alpha}$$

In this equation $\boldsymbol{P_n}$ represents the probability that a randomly chosen viewer is currently watching program n (from a lineup with a total of N possible choices). The parameter, alpha (α), can take a value only between 0 and 1 and should be chosen to provide the best fit to actual field data. Like any Probability Density Function (PDF) the total area under the curve must always be zero. Figure 7 shows the shape of typical Power Law Distributions for various population sizes. Although the Power Law is at best an approximation of an actual program lineup popularity, we have found that an alpha value around 0.8 provides a fairly reasonable first order approximation of many actual field measurements. Except where explicitly stated otherwise, we have used this value in this paper.

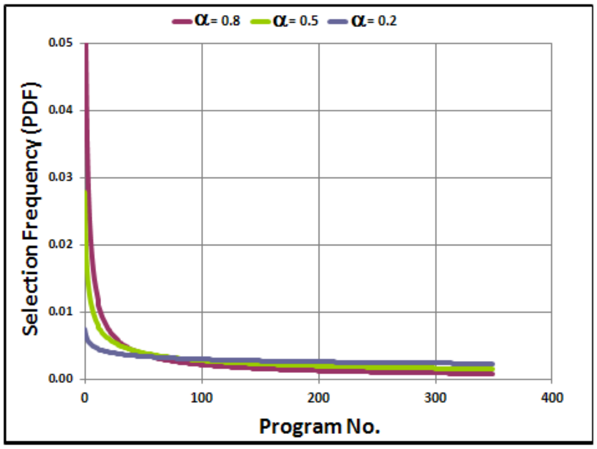


Figure 7 - Simulated Popularity Curves

The next figure illustrates normalized program popularity curves from 4 sample Service Groups. Each service group had about 400 programs available and had between 300 and 500 settop boxes. The curve was developed by accounting for all channel dwell times across 1 week.

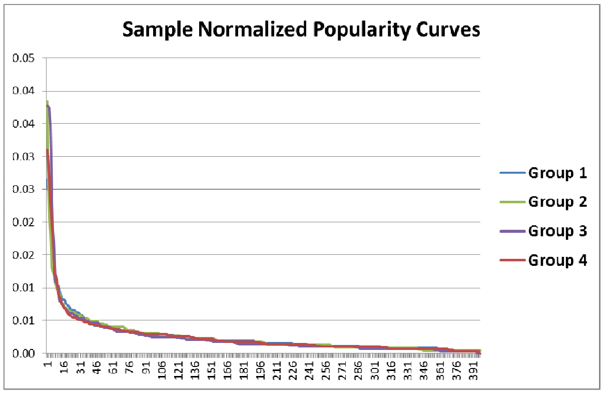


Figure 8 - Sample Popularity Curves

These curves illustrate that the Power law approximation holds up fairly well across a range of service group sizes.

<u>Network Modeling Parameters</u>

In addition to the viewer modeling parameters discussed above, three more attributes are required to model characteristics of the video network that are very much under the control of the MSO. These are:

1. Number of Offered Programs
2. Downstream Program Capacity
3. Number of Viewers in a Service Group

The challenge for network designers is to optimize these parameters to provide the maximum level of service to the viewers at an affordable equipment cost. These are not, however, three independent variables. Once any two of these three variables are chosen the value of the remaining parameter is dictated by the values chosen for the first two under the constraints imposed by the level of blocking deemed acceptable and the popularity profile of the offered program lineup.

Of these three attributes, the service group size will normally be the property that varies the most between network nodes and is least likely to be precisely determined at network design time.

This paper uses software simulations, employing Monte Carlo techniques, to model and chart the relationships among these attributes. Results of these simulations are shown in the following sections.

Downstream Program Capacity

Figure 9 shows simulation predictions for the downstream program capacity (as a function of the size of the service group) required to provide viewers with a lineup of 200 programs using each of the three video delivery protocols. For the purposes of this simulation, all programs are assumed to require the same amount of bandwidth. The chart contains Multicast curves consistent with a 0.1% blocking probability for three different values of alpha – showing the sensitivity of Multicast performance prediction over quite a wide range of values.

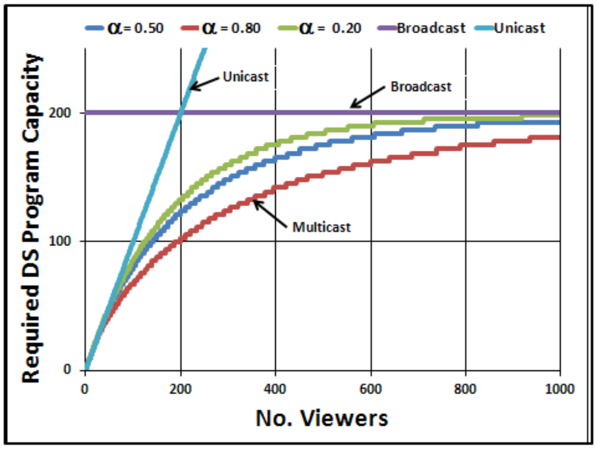


Figure 9 - Required Capacity vs. No. Viewers

Broadcast, of course, always requires a constant amount of downstream capacity (sufficient to carry a single copy of every offered program). Unicast, on the other hand, requires a separate downstream program channel for each individual viewer. The curve for the Multicast service is asymptotic to Unicast for very small service group sizes (very small numbers of viewers are likely to each select a different program). As the service group size gets very large the Multicast curve becomes asymptotic to the Broadcast service (since every program in the lineup will likely be selected by at least one of the very large number of viewers).

It is in the intermediate service group sizes that Multicast can be seen to require less downstream capacity than either of the other protocols. The vertical distance between the Multicast curve and either of the other protocols represents the downstream channel capacity that can be saved by using Multicast rather than the other protocol.

Program Lineup Size

A chart like the one in Figure 9 can tell us the relationship between downstream capacity and service group size for a known program lineup. Often, however, it may be that downstream program capacity is constrained a priori and we would like to know the relationship between the service group size and the number of programs that we could provide in the program lineup.

The next figure assumes that downstream capacity is available for only 100 simultaneous video programs with the resulting relationship between the service group size and the number of programs that could be offered.

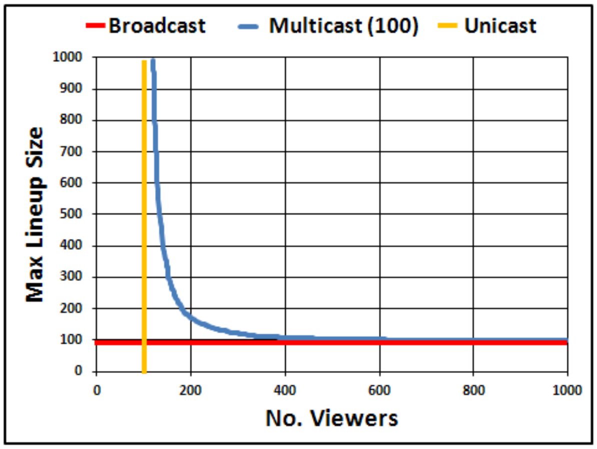


Figure 10 - Maximum Lineup Size

Again we see that the Multicast curve is asymptotic to Broadcast service for very large service group sizes and to Unicast for small service group sizes, still assuming the same 0.1% blocking probability. The straight horizontal line corresponding to Broadcast service shows that Broadcasting always requires a separate program channel for each offered program, but can support an arbitrarily large service group. The straight vertical line corresponding to Unicast reveals that Unicast can support an infinite number of offered programs (when the downstream program capacity is greater than the number of viewers in the service group) but cannot support even a single viewer more without failing to meet the required blocking probability.

The vertical distance between the Multicast and Broadcast curves shows how many more programs Multicast could support in the program lineup (as a function of the service group size). The horizontal distance between the Multicast and Unicast curves, on the other hand, shows how many more viewers could be in a Multicast service group (as a function of the number of offered program choices).

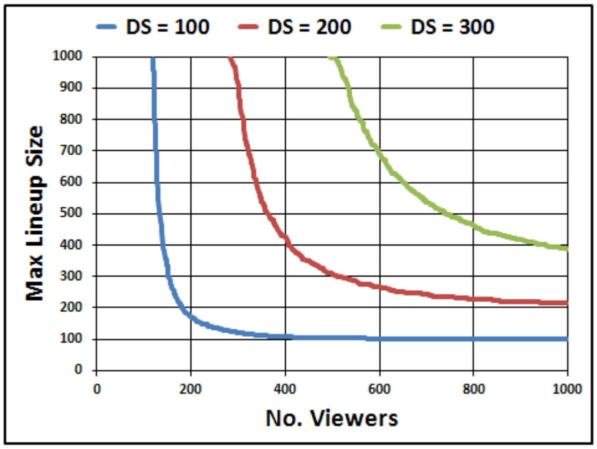


Figure 11 - Lineup Size vs. Downstream Program Capacity

Figure 11 shows the same curve (for 100 downstream Multicast program channels) but adds two more curves – for 200 and 300 Multicast downstream program channels. These curves seem to indicate that the power of Multicast service (i.e., the distance from the Multicast curve to either of the asymptotes) increases significantly for larger numbers of downstream program channels and for larger program lineup sizes. Curves for smaller numbers of downstream program channels (like Figure 9) closely hug both asymptotes with only a fairly narrow range of service group sizes in which Multicast shines relative to the other protocols.

This behavior suggests that a network evolution plan that begins by transferring a small number of Broadcast programs onto a small amount of downstream Multicast bandwidth may not immediately experience the full advantage that might come later when a larger program lineup is offered via Multicast. This finding also has significance for the importance of improvements in coding efficiency. As the number of programs that can be efficiently carried within a given network bandwidth increases, this analysis suggests that the increase in multicasting gain will be non-linear. For example, if the number of programs carried in a given bandwidth can be doubled, taking a service group from a ceiling of 100 streaming programs to 200 streaming programs, the actual offered lineup could increase from 150 linear programs to 900 linear programs for 300 viewers while still maintaining the same blocking ratio.

Extension with Actual Data

Because of the extensive deployment of SDV in some markets, there is a large body of data that can allow a comparison of simulated results with real-world behavior. The information in the section comes from SDV deployments in several different regions. Because of the variations due to local conditions, it is not always possible to find perfect matches to the simulations. The data in this paper was chosen to represent average conditions, and may represent data that was averaged over many service groups.

As was observed in the previous sections, simulations predict that the size of a service group and the number of offered programs can significantly influence the program popularity behavior which is directly related to the efficiency of various multicast/unicast/broadcast implementations. In actual deployments, there is a limited dispersion in the sizes of service groups. Service groups that are very large or very small are difficult to gather significant amounts of data on. The next figure, Figure 12, compares a small group of service groups and generally confirms the logical assumption that the size of a service group has an effect on the number of programs that it will consume in the aggregate. These groups do show, however, that the effect is not linear; there is not a 50% decrease in the number of active programs when the service group is 50% smaller. This finding agrees with the simulation shown in Figure 9.

Based on some actual viewership data that included peak tuning activity across many hundred service groups, an interesting dichotomy was observed. The relationship between the size of the program lineup and the percentage of programs that had at most one viewer was strongly correlated, implying that for a given size of service group the addition of more programs to a channel lineup will tend to add mostly unicast instances. For the same study, larger program lineups had a weaker, though still positive, correlation with multicast viewing instances. In contrast,

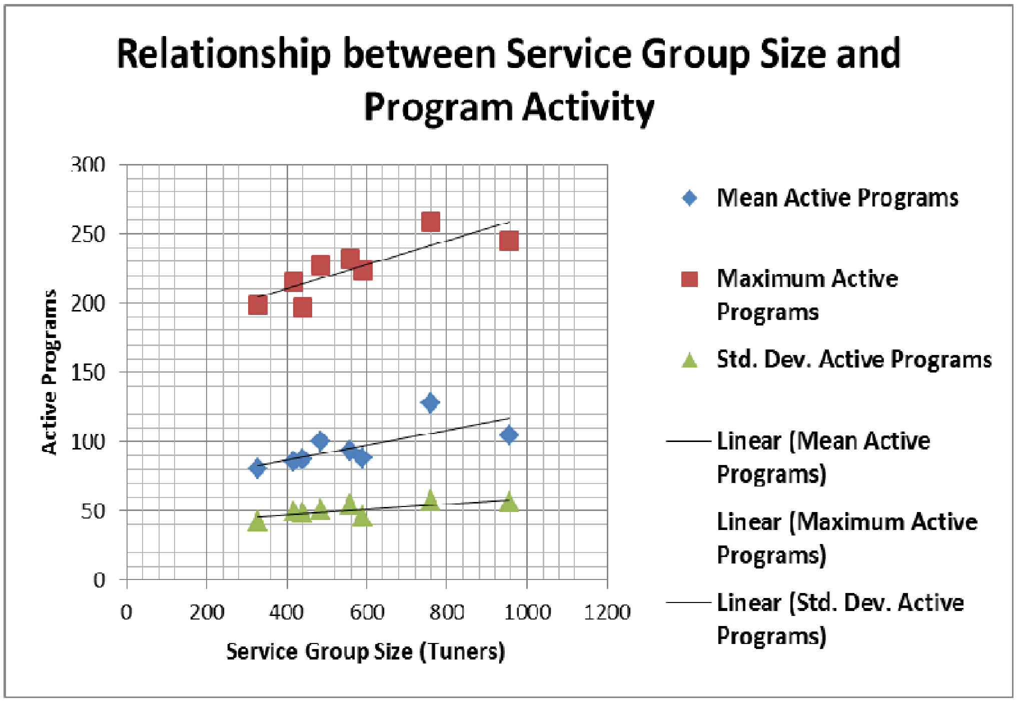


Figure 12 - Comparison of Service Group Size and Viewership

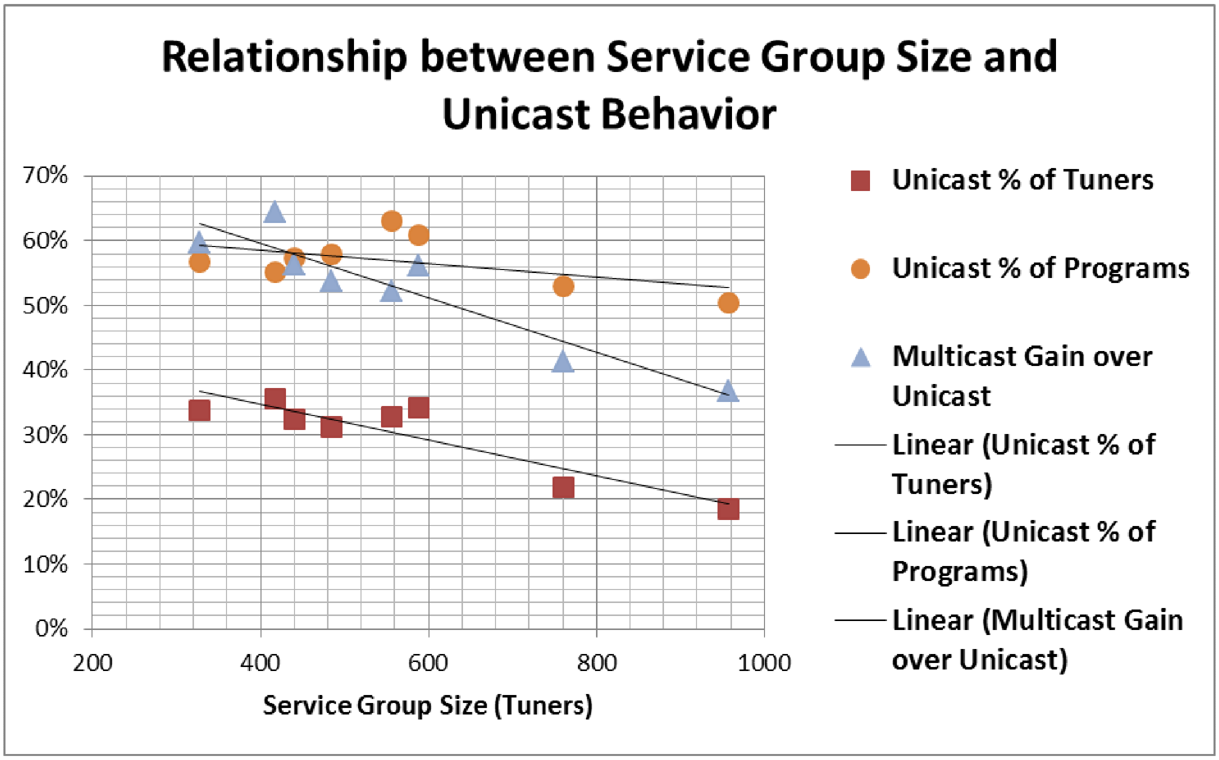


Figure 13 - Relationship of Service Group Size to Unicast Behavior

comparing increased numbers of viewers per service group for the same size of program lineup showed a strongly negative correlation with the number of unicast instances. In other words, as the number of viewers in a service group increased, they tended to watch similar programming to the other subscribers, which reduces the overall unicast percentage. This observation was compared against the test block of service groups and a similar pattern was seen in Figure 13. The service groups all had similar program lineups, and the percentage of unicast traffic declined as the number of subscribers grew in the service groups.

These observations, taken together, suggest that there is an optimum service group range that balances the number of viewers and the program content available to them.

Comparison of Deployment Scenarios

Another important area in which real world data can provide important information is the relative network impact in real time of implementations of the various protocols we have been discussing. The diagrams in this section were taken from a detailed analysis of the channel change logs of 8 service groups chosen at random.

A week’s worth of channel change logs from 8 different service groups were analyzed and used to drive various network simulations. The service groups were chosen to be roughly representative of common configurations. The wide variety of network and node configurations means that any extrapolations must be taken with a grain of salt, but they may still prove useful illustrations of the performance of different proposed systems.

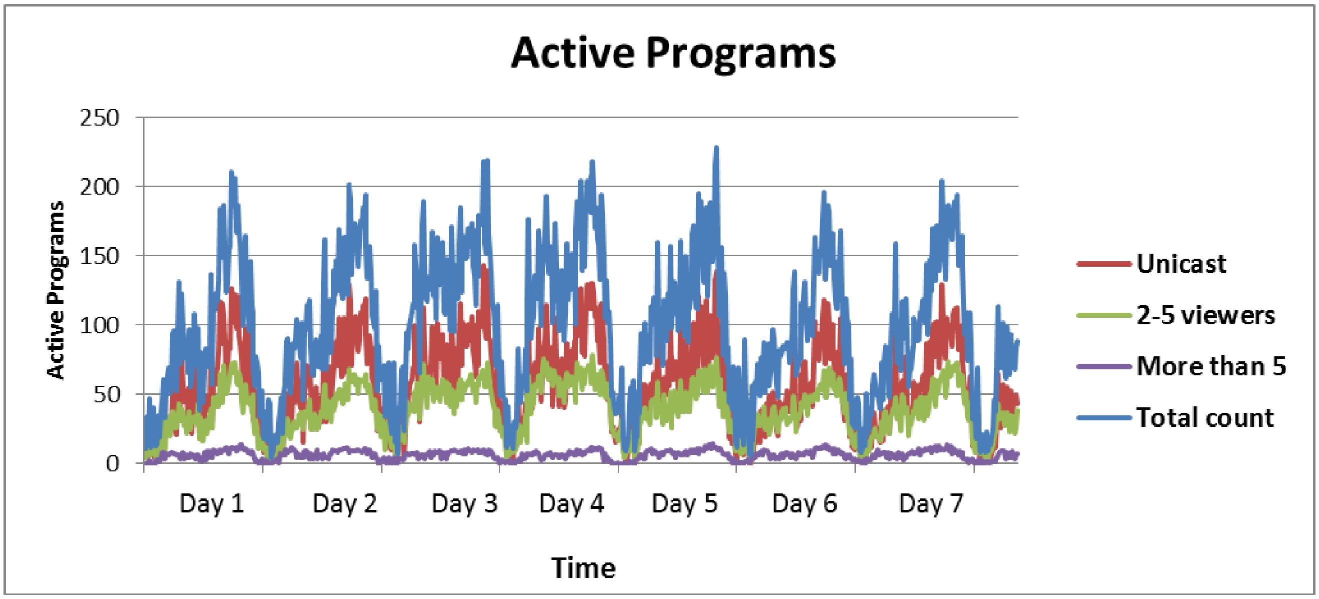


Figure 14 - Classifying Programs by Viewership

The channel change logs allowed the simulation to play out a week's worth of channel change events in various scenarios to see if the resulting network would be practical.

First to be considered is the question of the practicality of an all-unicast solution compared to an all-multicast solution. A reasonable way to study this problem is to study the distribution of viewers to programs. The 8 service groups referenced behaved similarly. One service group's results are used for illustration below, but the other service groups showed very similar results.

When one considers the distribution of viewers per program, the programs watched by only one viewer constitute the majority as shown in Figure 14. Across the SGs studied the percentage of programming viewed by only a single tuner peaked between 50% and 63% as shown in Figure 13, Unicast % of Programs.

But the dominance of Unicast in the program view is a bit misleading if one is considering an actual unicast deployment. If one considers the same period with the same

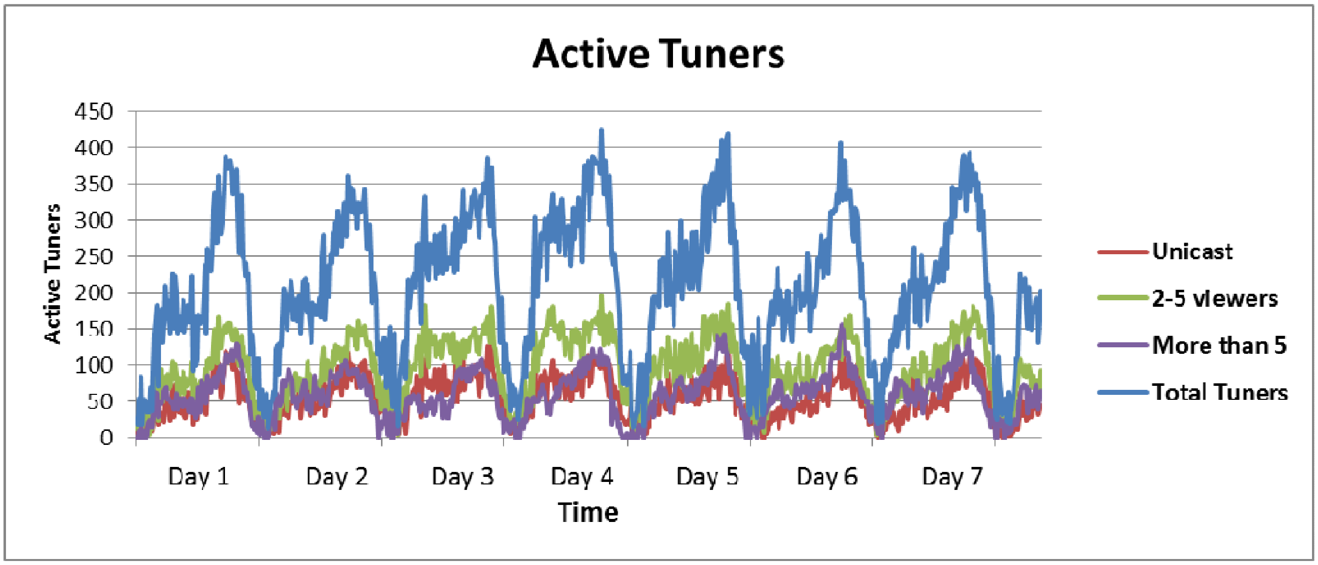


Figure 15 - Distribution of Tuners versus Other Tuners

service group, but instead studies the actual number of tuners attached to each program, a different picture emerges. The actual viewers, tuners really, are split fairly evenly across the different categories used in the graphs. From the larger group of SGs, the peak percentage of tuners that were alone in viewing a program ranged between 18% and 34%, as shown in Figure 13, Unicast % of Tuners.

Turning back to the sample service group, Figure 15 clearly shows that while the majority of streams, particularly during primetime, only have a single viewer, the majority of the viewers are actually on channels with more than one viewer.

This result implies that to move to an all unicast model for IP video requires substantially more bandwidth than a model using multicast. On average, for the service groups used as examples, an all unicast model would require 55% more bandwidth than an all multicast model. Using our example service group again, in Figure 16 the difference between an All Unicast and All Multicast model can be seen.

One other option that deserves consideration is a model that combines a static multicast tier, emulating broadcast, with a unicast tier. This combination could allow a reduction in the complexity of an IP Video deployment by simplifying the network engineering required since the static tier could be processed to improve its compression statistics, and possibly that scenario would require less protocol support that would be unique to CATV.

Using the sample service groups and the tiering shown before, the programs that had only been unicast were identified, and it was assumed that the rest of the program lineup was broadcast. That scenario was 27% more efficient, on average, than a full broadcast model. A full multicast model would have allowed 77% bandwidth reduction over broadcast. Another scenario was considered where any channels that had had at most 2 viewers were left as unicast, with the rest broadcast. This scenario offered a 50% bandwidth reduction over broadcast with performance close to that of multicast during primetime.

In Figure 17, several scenarios are compared

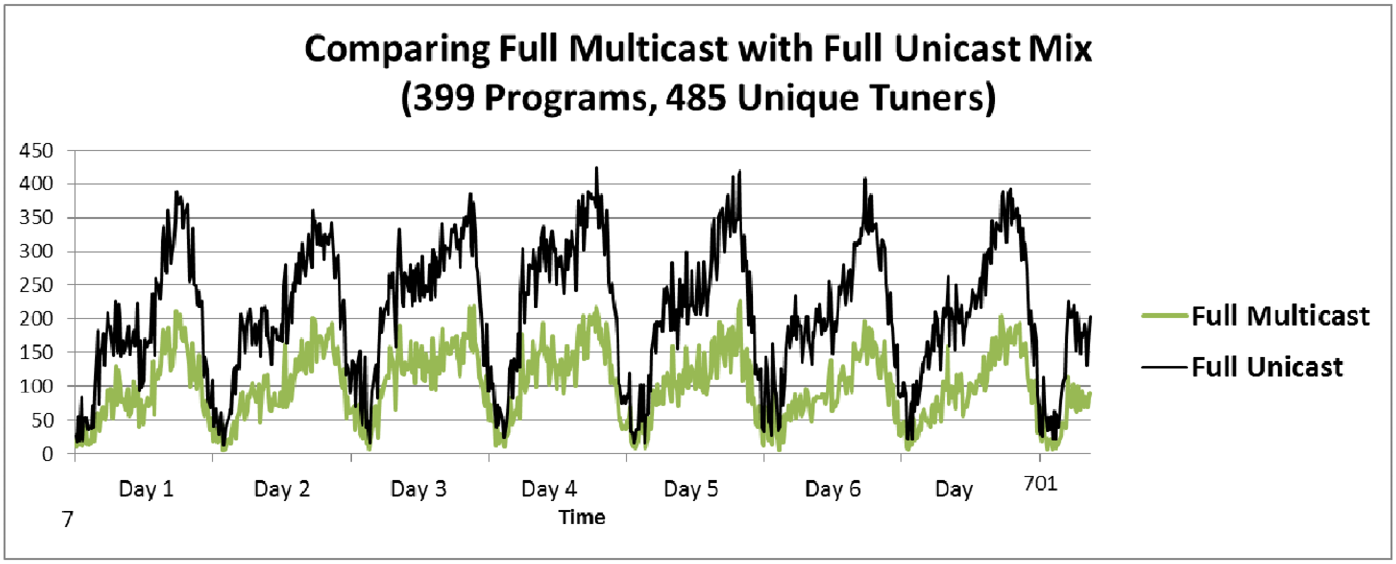


Figure 16 - Comparison of Multicast and Unicast

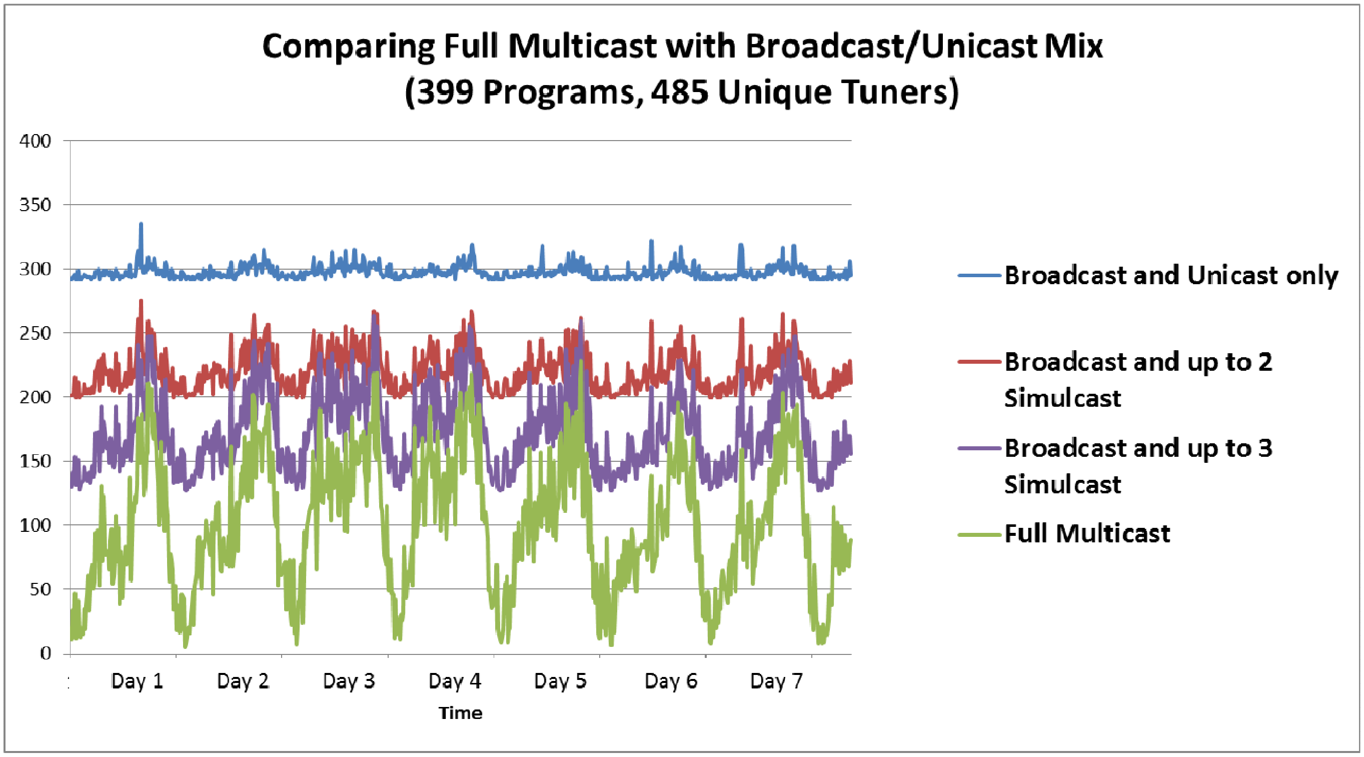


Figure 17 - Comparison of Multicast with Hybrid Broadcast/Unicast Model

using the example service group again. A more nuanced picture emerges from consideration of this figure. During primetime, the most popular channels are almost always playing, so utilizing broadcast or multicast has little effect during that time, so long as the most popular channels are correctly identified for broadcast. Allowing the least popular channels to be unicast does improve bandwidth utilization over broadcast, and as the channels committed to unicast increase the efficiency of this scheme approaches that of full multicast. Multicast offers a lower average bandwidth utilization, but its benefits are most apparent outside of primetime, traditionally the most congested time of the day in residential areas.

The value of the tradeoffs within the choice of IP Video distribution protocols is difficult to quantify. Pure broadcast's simplicity is counter-balanced by its total lack of network bandwidth efficiency. Pure unicast delivery is more complex than broadcast, but with only a small increase in DS bandwidth efficiency over broadcast. The two-way nature of the popular unicast video delivery protocols also uses more upstream bandwidth than either broadcast or multicast. Full multicast distribution offers the best bandwidth efficiency to reduce outside plant expenditures, but has not been extensively deployed past the headend and may pose unknown challenges.

Other Considerations

Some concerns have been raised about the practical limits of channel change times using multicast. DOCSIS3.0 multicast specifications involve fairly complex scenarios wherein a CM/STB must send a request to the CMTS to join a multicast group, and the CMTS must attempt to join the multicast group, then respond to the CM. An IP Video CM could conceivably have to change its DS bonding group and worst-case even reset to reach a new multicast stream.

While these scenarios are possible within the specification's limits, a sensible IP Video architecture can make many simplifications and improvements by observing the choices the successful SDV architecture has made to improve its performance. For instance, the time it takes to join a multicast group cold, so to speak, was recognized as a potential problem within SDV. The solution that was developed within the SDV architecture was to have the EQAM join as many multicasts as it would potentially source over its channels. The CMTS, occupying the same network position as the EQAM for IP Video, is equally capable of joining multiple multicast groups, thus eliminating potential router latency from the aggregate channel change time.

The analyses in the foregoing sections have assumed that subscribers will continue to behave mostly as they do today. A critical part of that assumption relates to the behavior of content providers and the regulatory landscape. If the content providers were to change from their current course and deemphasize linear programming and promote a more VOD-style consumption of their content, similar to that provided by most over-the-top providers today, then any assumptions made based on extrapolations from the behavior of today's subscribers would become moot. Most analysts have not predicted that sort of change any time soon due to primarily commercial factors, but a radical change is always a possibility driven by a new application or possible new regulations.

The Path Forward

As the operators move toward incorporating IP video into their day-to-day operations, the availability of both unicast and multicast protocols within the IPTV 'toolbox' may prove to be quite valuable.

For early low-volume deployments, unicast delivery offers a simple first step. It enables experimentation with alternative user interfaces, and hybrid STB/cloud architectures, without the complications of a volume deployment. For some networks with very small effective service groups this technology may continue to be cost-effective even as the network approaches saturation.

As IP Video deployment moves out of limited trials and into larger deployments in more traditional larger service groups, multicast can be employed to enable a cost-effective deployment of services that still fall into today's linear model. Depending upon the tradeoffs possible between network bandwidth, service group size and program popularity mix, as well as the popularity of new services such as network DVR, there is not a single answer as to the most efficient IP Video distribution model. An MSO that has moved to small service group sizes for other reasons may be able to utilize a mix of unicast and multicast with good results. An MSO that has not lowered the size of its average service groups may well decide to make more extensive use of multicast to get the most efficiency out of its network. An MSO with many commercial customers that could use the non-prime-time bandwidth freed up by multicast may also choose to implement a full multicast solution.

As the content distribution model evolves past the traditional linear program distribution, unicast may return to prominence if few users tend to watch the same thing at the same time. Some events, like sports or breaking news, may still attract enough viewers to leave multicast a place on the table even then.

In Summary

IP Video delivery over unicast protocols has flourished on the Web, but for the CATV application of bulk delivery of programming over a pipe with limited bandwidth, the unicast model tends to break down due to the sheer volume of users.

Broadcast has been great for CATV for many years, but as the number of programs has proliferated and the required variety of resolutions for those programs has grown as well, the sheer volume of programming selections has tended to exhaust the available bandwidth.

Multicast, perhaps in combination with unicast, may offer a robust solution, similar to the use of SDV in conjunction with VOD in current MPEG distribution network. This combination of technologies can offer an expansive list of programming suitable for many different device types, while still fitting within practical constraints of the available bandwidth envelope.

A Software Friendly DOCSIS Control Plane

Alon Bernstein
Cisco Systems

Abstract

It has been 15 years since the initial set of DOCSIS specs have been authored. In those 15 years software engineering has seen an explosion in productivity at the same time that the DOCSIS control plane has remained fairly unchanged. Can we apply these productivity tools to the DOCSIS control plane to facilitate greater simplicity and feature velocity?

This paper will outline both software trends and protocol design trends that are relevant to the above discussion and how they can be applied to DOCSIS.

OVERVIEW

DOCSIS is primarily an interface protocol between a CM and a CMTS. There is a wide palette of options for a protocol designer and all are relevant to DOCSIS design. Each option has its tradeoffs and the role of the protocol designer is to choose the option that fits the system requirements and constraints the best. The list of options include:

- Generalized interface vs. mission specific interface
- Legacy protocol vs. mission specific protocol
- Stateless vs. stateful
- Client-Server vs. Peer-to-peer

And more…In many cases there are no simple rights and wrongs and a choice that might have made sense at the time of the protocol design turns out to be sub-optimal as systems often end up getting deployed in a manner that is different then what they were designed for. All of these choices have an impact on software. Its not always the case the choice that is optimal for software is the ideal for meeting the system requirements, still its unfortunate that in many cases the software implementation ease is considered as a relatively low priority item. This observation is patricianly in place for DOCSIS since the amount of software resources needed to support the DOCSIS set of protocols is significant.

Before proceeding, a word of caution: in cases where there are requirements and constraints that supersede software requirements then clearly the guidelines explained in this document will not apply. The challenge is to identify these requirements and constrains correctly and not to pre-optimize at the expense of "software friendliness". To quote the author of the "Art Of Computer Programming":

> *"We should forget about small efficiencies, say about 97% of the time: premature optimization is the root of all evil"* (Donald Knuth)

SOFTWARE CONSIDERATIONS

Software engineering and protocol design share the same approach to simplifying complex system requirements:

- Modularization: sub-divide a large and complex system into simple and easy to test modules with well-defined inputs and outputs.
- Layering: Define the hierarchy of modules, what services each component provides to another.
- Abstraction: identifying common services that can be shared across modules

Software design methodologies have evolved around the same timelines as the Internet revolution and the creation of networking protocols. But while the timelines are similar the amount of change software went through is much larger the amount of change in the suite of networking protocols that drive the Internet.

Software has evolved from the “C” programming language to object oriented languages such as C++ and Java which allowed for further modularization/layering and abstraction of code to web technologies that brought amazing scale, speed and flexibility. At the same time network protocols stuck with the OSI 7 layer model [6] as possibly the only attempt to apply modularization/layering and abstraction to networking. Case in point: many of the routing protocol RFCs have pages upon pages of interface specifications and message formants. If written from with a “software friendliness” point of view they could have had a well-defined separation of the methods to distribute data across a group of routers (which is similar in many of the routing protocols) and the actual routing algorithms.

Obviously the issue outline above has a wider scope then DOCSIS, so to keep the discussion focused here are a couple of examples of why the current DOCSIS specifications does not follow basic software implementation guidelines along with a high-level proposal on how to fix it (going into fine details is outside the scope of this paper):

Example 1: DOCSIS registration.

The DOCSIS registration process starts with bringing up the physical layer, jumps to authentication and IP bring up, then to service provisioning then back to physical layer bring up.

Initially services are created in the registration process. Additional services are added with a different mechanism then the one used in registration (DSx).

Why is it not software friendly? The fact that the cable modem bring-up has a dependency on the IP layer bring-up makes it difficult to independently develop (aka “feature velocity”) and independently test (aka “product quality”), the registration process.

This is a good example where an idea that seemed to offer:

1. Simplification: because the same mechanism used to provision services is also used for the modem bootstrap
2. Optimization: fewer messages since all the various layers are squeezed into the same

Turns out to be not-such-a-good-idea when it comes to software implementation. This becomes painfully obvious when the system is physically distributed. Imagine an implementation where DOCSIS functionality is segregated into processor A and Layer 3 functionality into processor B. Because of the way registration is handled the DOCSIS processor needs to know a little about the IP layer and the IP processor needs to know a bit about DOCSIS. Clearly these are solvable problems and “anything can be done in software” but as mentioned above there is a price to pay in speed of implementation, testability and debug of system issues.

How would a software friendly registration protocol look? A software friendly specification would have clear and independent stages as depicted in the figure below:

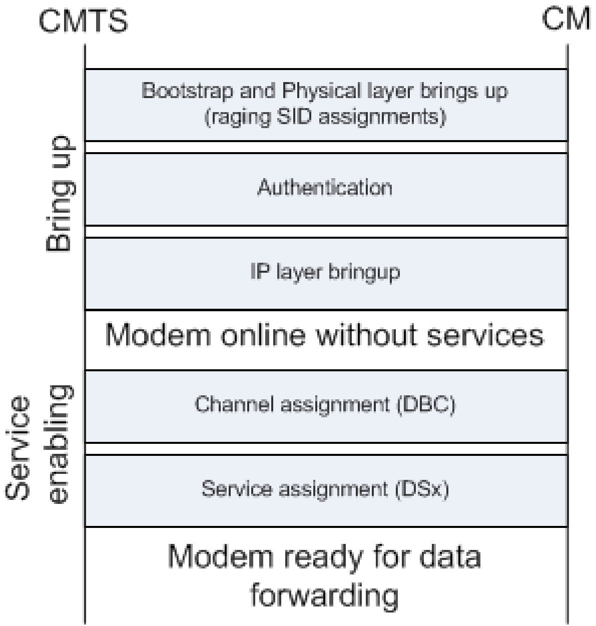

CMTS
CM
Bring up
Bootstrap and Physical layer brings up (raging SID assignments)
Authentication
IP layer bringup
Modem online without services
Service enabling
Channel assignment (DBC)
Service assignment (DSx)
Modem ready for data forwarding

Figure 1 SW friendly registration

1. bootstrap: initial physical layer bring up

2. Authentication: validation of the cable modem for network access
3. L3 bring up: DHCP processing and IP address assignment

Each one of the above steps would be treated as an independent transaction and the three of them would be the workflow needed to bring a cable modem online.

For these 3 steps the major deviation from the current registration is that we don't rely on TFTP for service provisioning. There are two reasons to skip TFTP; the first being that the IP layer is not even up so we can't access anything beyond the CMTS and the second being that we want to postpone the service providing part to a later stage anyway. Having said all that, the CM still needs to communicate with the CMTS and it still needs some form of a service flow to do it. If we don't have any services provisioned how do the CMTS and CM communicate? The "temporary flow" that DOCSIS creates anyway for registration would just leave on until after the IP bring-up phase and only after that would be replaced by the "real" in the service enablement phase

As far as the next steps go we fortunately have clear transactions to handle:

- Service provisioning using DSX
- Changing physical layer parameters with the DBC

These can be used to change services and channel assignments after the modem is online. Note that the CMTS can still use TFTP to retrieve service parameters and those will be parsed into a DSA message.

Naturally there are tradeoffs to this proposal; the number of messages has increased and a new form of service enablement has been added, however the payoff is significant in terms of software modularity.

Example 2: The Mac Domain Descriptor MDD message is a dumping ground for information about plant topology, IPv6, error message report throttling, security, physical layer parameters and more. A software friendly specification would create independent messages for each of the functional areas. Though one can argue that MDD is "just a transport" for data that can be managed by independent modules in the software, however the inclusion of all of them in the same message creates dependencies that are easier to avoid were the MDD to be broken into separate messages.

END-TO-END PRINCIPAL

Some link-level protocols (such as DOCSIS) assume reliability is required and come up with their own set of timers to assure delivery at the link layer. This might be justified if the link layer is highly unreliable, and even in that case the timeouts set for retransmission must be an order of magnitude shorter then the timeouts of the end-to-end application. If retransmission timers are too long then all sorts of odd corner cases might occur. For example: the DSx-RSP timeout is about 1 sec and there can be 3 of them. If an end-to-end signaling protocol, such as SIP [4] has a message re-transmission time of the same order of magnitude then a DOCSIS implementation might release a SIP message when the application level has already timed out and re-transmitted its own copy. This will not cause the system to break since a robust implementation knows how to deal with messages that are duplicated or out of order, but it will clutter the error counters and fault logs with a "duplicate message" event which would have been avoided if the DOCSIS link layer counted (as it should have) on an end-to-end session establishment protocol. The reader might ask, "what if the end-to-end protocol is designed to be unreliable"? Even in that case it's not the role of the DOCSIS link-layer to assure delivery if the higher lever application does not require assurance in order to operate correctly.

The DOCSIS software may trigger a timeout for a DSx-RSP, however the expiration of this timeout would be only used for recording a failure and releasing system resources allocated for the DSx, not for triggering a retransmission.

If the media is highly unreliable and failures are a common occurrence then they might be room for link-level error repair but in that case the timeouts need to be an order of magnitude shorter then the application timeouts - a suggested range would be 100ms or so.

A further simplification based on the end-to-end principal is to remove the DSx-ACK phase. DOCSIS uses a 3-way handshake for DSx. The rough outline of the conversation at the service activation phase goes like this:

1. CMTS -> CM: please start a service (DSA-REQ)
2. CM -> CMTS: ok, I started the service (DSA-RSP)
3. CMTS -> CM: cool, my CMTS resources are ready you can start sending data (DSA-ACK)

But this third step is not really needed for the same end-to-end argument. For example consider this zoon-in of a PCMM message sequence (figure 9 in ref [2])

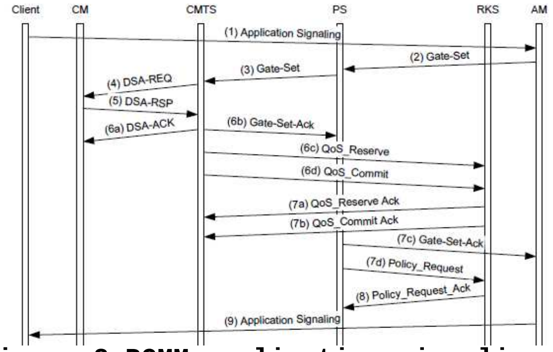


Figure 2 PCMM application signaling

It's obvious from this diagram that the DSA-ACK is not fulfilling any useful function. For one it's a "dead-end" not resulting in any further action, and since it is sent at the same time as the Gate-Set-Ack it is useless in guaranteeing any sequencing of events.

As a side note, some have suggested that the DSA-ACK is needed for extra reliability but this would be an even worst violation of protocol rules since the DSA-RSP is already an acknowledgement and a protocol should not acknowledge and acknowledgment.

ENCODING

DOCSIS uses TLVs to serialize information however TLVs are not common in modern networking stacks and not supported in many of the productivity tools and code generation tools used today. Non-TLV types of encodings include JSON/HTTP/XML/google "protocol buffers" and others. The advantages of the above mentioned tools are:

1. They come with code generation tools that relive the software developer from the burden of parsing messages into native data structures.

2. Most of them encode information in human readable strings that makes debugging easier.

TLVs are a more compact form of serializing data but as bandwidth available on the cable media increases this is becoming a non-issue.

TLVs might also be easier to parse, but CPU power is much less of an issue then it was at the time the DOCSIS specification were written. In fact, the modern cable modems have more powerful CPUs then early CMTS products!

OPEN SOURCE

Another software trend that has been going strong is the movement to open source. As a development methodology it has proven to deliver on wide scale and highly complex software projects. How can open source apply to cable? The CMTS/CM interaction is not likely to be of interest to the open source committee since its so domain specific and for product differentiation reasons it's highly unlikely that CMTS/CM vendors will open their source code.

This document proposes to use source a companion to the CableLabs standard documentation process. For example, if a new registration process was to be pursued then high-level function calls and JSON encodings could be published as open source. This would hopefully promote better interoperability and shorter ATP cycles as it removes a lot of ambiguity in the interface design.

DATABASE TECHNOLGIES

A CMTS implementation needs to manage a database of cable modems, plant topology and more. In many cases this information needs to be shared with a CM and so one view of a DOCSIS system could be that it's a distributed database of CM state and resources. With that observation it's clear that the only type of data sharing that DOCSIS allows for is the transactional type. That used to be the only model for data sharing in the database industry in general but the scale that companies such as google and facebook had to grow to gave rise to a new model, one that priorities performance over accuracy. Clearly for some types of data this model will not work well (financial transactions for example) while for others it makes sense (searching through web pages).

An interesting observation made by the Internet community is captured in what's called the "CAP theorem" [5]. In a nutshell what the CAP theorem states is that when a database designer is requested to support a distributed database that provides [1]:

1. Performance
2. Consistency of data across components of the distributed database
3. Resiliency in cases for system malfunctions, for example, packet drops.

Only two out of these three requirements can be met. The designer still has to choice of which two are fulfilled, but it is not possible to meet all three.

As mentioned above DOCSIS supports a transactional sharing of data that represents a choice of consistency and resiliency over performance, and as long as performance requirements are met (for example, number of voice call created per-second) it is a win-win situation. But as new applications become available and the load on the control plane increases it may make sense to consider other choices. The proposal in the previous section to avoid re-transmissions represents the option of demoting resiliency. Another option is to assume an "optimistic model" where the CMTS can allocate and activate resources on a DSA-REQ, assuming that a positive DSA-RSP will follow and intentionally allowing for short period of times of inconsistency if cases where the DSA-RSP was not successful.

Another useful tool from the database world is the concept of "data normalization". Its outside the scope of this paper to go into the detail of data normalization (see ref [3]), but in a nutshell it's a set of guidelines on how to break complex data into a list of tables with rows and columns where each row is fairly atomic. When inspecting some of the DOCSIS MIBs and MAC messages its obvious that some break at least one of the normal forms. For example, the inclusion of a "service flow reference" in the same table as the "service flow id" violates the normal form that prohibits dependencies between columns of a table. Without getting into too many details this paper only makes the observation that management constructs that are "normal" are easier to implement in software.

SECURITY

An obvious security hole in DOCSIS is letting the cable modem parse the configuration in order to reflect it back to the CMTS. It's worth mentioning it in this document because (ironically) this might have been the single attempt in DOCSIS to help software by offloading the task of parsing the cable modem configuration to the cable modem. However, in order to plug this security hole the CMTS needs to parse the configuration anyway and overall it's a great example of how premature optimization can create more harm then good.

[1] [ab] CAP stands for "consistency, availability, and fault tolerance". I took the liberty of translating the above to terms familiar to the cable community since the original terminology might be confused with existing cable terms.

CONCLUSION

DOCSIS has obviously been a very successful protocol. The DOCSIS provisioning and back-office system is part of this success, especially when comparing it to its DSL counterparts where a strong standard for provisioning and service enablement does not exist. Nevertheless a 15-year critical review, and possible updates would certainly help DOCSIS to become even better in facing the challenges ahead.

This paper suggests that software implementation ease and modern software tools need to play a bigger role in the design of future DOCSIS protocols and while some of the proposals made here are of academic and demonstrative value only, others can be relevant to future versions and enhancements of DOCSIS.

REFERANCES

1. DOCSIS MULPI : http://www.cablelabs.com/specifications/CM-SP-MULPIv3.0-I18-120329.pdf
2. PCMM: http://www.cablelabs.com/specifications/PKT-SP-MM-I06-110629.pdf
3. Codd, E.F. (June 1970). "A Relational Model of Data for Large Shared Data Banks". Communications of the ACM 13 (6):377–387.doi:10.1145/362384.362685.
4. Session Imitation Protocol, RFC 3261
5. CAP Theorm : http://www.cs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf
6. ITU-T, X.200 series recommendations: http://www.itu.int/rec/T-REC-X/en

ARCHITECTING THE DOCSIS® NETWORK TO OFFER SYMMETRIC 1GBPS SERVICE OVER THE NEXT TWO DECADES

Ayham Al-Banna
ARRIS Group, Inc.

Abstract

The paper analyzes various options to increase the capacity of HFC networks in order to meet the capacity demands over the next two decades. A smooth migration plan is proposed to enable MSOs offering beyond than 1Gbps US service. A High-split prototype system is built and initial results are introduced.

1. INTRODUCTION

The current architecture of Hybrid Fiber Coaxial cable (HFC) networks along with the exponential growth in bandwidth demand are placing the cable Multiple Service Operators (MSOs) at competitive disadvantage as they face capacity limitations. These limitations may preclude the MSOs from satisfying the customers' demands if not properly addressed.

In order for the MSOs to maintain their business and offer more services at faster speeds (e.g., services to business customers, IPTV fans, gamers, etc.), they need to immediately start brainstorming, architecting, and upgrading their networks in ways that will meet the pressing bandwidth demands. This process requires taking smart and gradual steps toward the goal system architecture, which will support beyond than symmetrical 1Gbps service.

Multiple factors need to be considered while going through the system and plant migration: cost, network architecture, spectrum allocation, operational issues, technical challenges, headend equipment (e.g., Converged Cable Access Platform (CCAP) compatible?, servers scale?, etc.), customers Quality of Experience (QoE), etc. The list goes on! Not only do MSOs have to think about the above factors as they prepare their networks for future services, they also need to think thoroughly about the appropriate sequence of steps to take such that an optimal architecture is achieved. The optimal architecture can be defined as a flexible network topology that results in maximum capacity and minimum cost over extended periods of time.

This paper is organized as follows. Section 2 describes the traffic growth trends based on recent real data. Several multiple factors that play heavily in the decision process of network migrations are briefly described in Section 3. Section 4 lists and analyzes the available options to extend the US BW to offer 1Gbps service. A sample plan that offers *smooth* migration steps to result in an optimal network architecture, which offers symmetric 1Gbps architecture and multi-gigabit system in the future, is described in Section 5. Section 6 concludes the paper.

2. RECENT TRENDS IN BW DEMAND

The traffic demand has been growing exponentially for the last 30 years. Different applications and services appeared at different times over the last three decades to ensure that the traffic growth stays on track! Among many, business services, gaming, and IPTV make today's motivation for guaranteed traffic growth for the next few years. The constant traffic growth over the past three decades is shown in Fig. 1, which shows the maximum DS rate per subscriber over cable networks [1]. This curve is sometimes referred to as the Nielsen curve for Cable networks.

Recent data obtained from different MSOs shows similar growth pattern for the average traffic on their networks. In particular, Fig. 2 shows the DS BW Average Cumulative Growth Rate (CAGR) per subscriber for three different MSOs over the past couple of years. Note that the CAGR value for all MSOs is more than 50% per year. Figure 3, on the other hand, depicts the US BW CAGR per subscriber over the past two years for two different MSOs. Observe from the figure that the CAGR averaged over both MSOs results in an US BW growth rate of about 30% per year.

The data in Fig. 2 and Fig. 3 shows that while some MSOs may observe slow growth rate on their networks, other MSOs observe larger growth rates. Additionally, the cumulative traffic growth averaged over all MSOs for the past two years agrees with the traffic growth trend observed for the past thirty years as was shown in Fig. 1.

From Figs. 1 through 3, the average DS and US BW per subscriber CAGR is shown to be >50% and 30%, respectively, for the past thirty years. Therefore, it might be reasonable to assume that the traffic growth will maintain the same trend in the future. In subsequent analyses in this paper, where we focus on the US BW problem in HFC networks, we assume that the US CAGR is at 30% on average.

Given that the US CAGR is at 30%, the question at hand is: when will the current 5-42MHz spectrum be totally consumed and therefore an upgrade of some sort is necessary? The answer to that question not only depends on the US CAGR, but also on the value of the maximum offered subscriber rate (Tmax) today. Between the US CAGR and Tmax values offered today, it will be straightforward to predict when the current US spectrum runs out of capacity, which will be shown later in this section.

The maximum offered rates have been published recently [2]. Table 1 shows DS and US Tmax values currently offered in North America. Note that the table lists Tmax values offered by different industries (Cable and others). Tables 2 and 3, show DS Tmax values offered by different MSOs in Europe. Observe that some European MSOs offer higher rates than their counterparts in North America. In particular, the maximum DS Tmax currently offered in Europe is 360Mbps by Zon Multimedia (See Table 3). The current offering of Zon for DS Tmax and US Tmax shows a constant ratio of 15 between the rates. Therefore, the US Tmax value offered by Zon is assumed to be around 24Mbps. Note that this is close to the 20Mbps US Tmax being offered by Videotron in North America.

One important point to observe from Table 1 is that the maximum Tmax service is offered by Verizon, which is not a cable MSO! Therefore, in addition to customer traffic demand, Table 1 clearly shows the other side of the equation that pushes cable MSOs to add capacity to their networks: Competition!

With the assumptions that the US CAGR is 30% and the current offered US Tmax value is 24Mbps, the next step is to calculate the time when the current US spectrum runs out of capacity. Given a certain US Tmax value per subscriber, some MSOs might consider providing a total capacity of 1.5*Tmax in order to offer service with adequate Quality of Experience (QoE) to their subscribers. Other MSOs might choose other factors that are different from 1.5*Tmax (e.g., 2*Tmax). For the analysis in this paper, we assume that a capacity of 1.5*Tmax is required in order to offer good QoE service for customers with Tmax as the maximum rate per subscriber.

Figure 4 shows the extrapolated growth of US Tmax per subscriber using the above assumptions. Note that the current US spectrum (5-42MHz) capacity is assumed to be around 133Mbps. This is because the total BW of 37MHz may not be completely usable at the highest possible modulation order (some channels can potentially run at QAM256 while others will run at QAM16). Also, strong FEC is assumed for the same reason (some parts of the spectrum are very clean while others are really challenging). The combination of moderate order modulation order (QAM64) and strong FEC (code rate = 0.75) compensates for noisy channels, unusable spectrum, and spectrum that used for services other than data). The total capacity of 133Mbps might be close to what MSOs can achieve in the real-world from the 5-42MHz spectrum. You may notice that this number is a little higher than the more conservative estimates that have been published earlier by the ARRIS' team (the author included) [3], which assumed a total capacity for the 5-42MHz to be around 118Mbps. Upcoming sections in this paper, where comparisons between the capacities of different split options is provided, will refer to the past capacity work and will point out that the estimates might be a little conservative and therefore can be slightly increased. In all cases, the total capacity always depends on the plant condition and MSO's usage plan for the spectrum. Observe, however, that the difference in capacity numbers (15Mbps) due to different assumptions is not significant given the Tmax CAGR growth rate shown earlier.

Observe in Fig. 4 that the current US spectrum runs out of Tmax capacity just before year 2017. This corresponds to service offering of Tmax~90Mbps. Note that 1Gbps US Tmax service will be required around year 2026, if not earlier. One may realize that it not too early to start planning for network architecture updates and migration strategies in order to offer capacities that satisfy the projected traffic demands over the upcoming years.

3. PLAYING FACTORS IN HFC NETWORKS MIGRATION

This section briefly describes the various factors to be taken into consideration while going through the system and plant migration process in order to meet the capacity demands over the next two decades. Not only these elements need to be studied thoroughly, but also the interaction between them needs to be analyzed carefully. The interaction happens because some elements depend on others, where the choice of some elements affects the choice of others. There might be no one ideal solution for all MSOs. However, different MSOs may have different optimal solutions depending on their position from the factors listed below.

3.1. Network Architecture

Both the components composing the network and network topology affect the performance heavily. The number and characteristics of amplifiers, line extenders, bridgers, taps, and other passive devices affect both signal loss and noise. The characteristics of some of these equipment also define the operational BW where signals can be transmitted in the DS or US direction. The type and length of coaxial cables (trunk and drop) affect the signal loss too. The length and type of fiber links as well as the features of the optical transmitter and receiver also affect the performance.

How deep the fiber node in the plant affects the performance. For example, longer cascades results in more attenuation, noise, and worse filters roll-offs, which impair the signals transmitted around band edges. Shorter cascades on the other hand, result in better network performance.

Networks topology needs to be analyzed frequently because the plant topology changes over time as MSOs update their network to expand the capacity of their networks. The change in network topology may affect various customers differently depending on the location of the customer relative to the network update. Specifically, Fig. 4 shows an example of N+5 network topology. After node segmentation and splitting, the network topology becomes as shown in Fig. 5. Note that it is sometimes difficult to balance the number of subscribers between new fiber nodes as apparent from Fig. 5, which affect the capacity per subscriber. Also, the example in Fig. 5 is a good illustration to the node splitting and segmentation process whose output does not guarantee that new nodes have the same cascade length. Figure 5 shows that the resultant nodes possess different lengths (i.e., different number of cascaded amplifiers behind the fiber nodes). Again, this affects the attenuation, noise, and therefore capacity.

The number of cascaded amplifiers behind a fiber node has declined over the years. Some MSOs estimate the current average of their networks to be at N+5 (to N+6)[1]. The current network topologies along with the limited US spectrum (5-42MHz in the USA, 5-65MHz in Europe) place a tight limit on the capacity that can be offered by today's networks and therefore gradual sequential upgrades will be necessary to cover the demand as well as competition over the next two decades!

3.2. Spectrum Allocation

This is a critical topic because it touches many areas. The choice of which split to choose for the US spectrum (mid-split, high-split, top-split) comes as a result of studies of technical feasibility, which analyzes the technical challenges and offered capacity associated with the implementation of each split option. Besides cost, operational aspects are affected depending on the chosen split. For example, affected operational parts include: reclaiming/reallocating analog TV channels, moving DS spectrum, capping DS BW, transition bands (guard bands between DS and US), addressing the Out-Of-Band (OOB) signaling of Set-Top Boxes (STB), etc. This factor (spectrum allocation) is studied in more details in later sections of this paper.

3.3. Operational Issues

Various Operational issues are to be addressed when network migration occurs. Depending on the network architecture and the update to occur, operational aspect that can be affected include: Analog channels reclamation and reassignment, specifying spectrum for DOCSIS and digital channels, addressing STB OOB signaling, DOCSIS and Video management, network maintenance process (depending on equipment being in the headend or in the headend and FN together), network reliability and availability (again, related to equipment being in headend or in headend and FN). Observe that placing more intelligent equipment in the FN introduces higher risk in terms of network availability and reliability. Some of these operational aspects will be addressed in later sections of this paper.

3.4. Technical Challenges

The technical aspects of any solution or proposed network update must be studied thoroughly. The technical study results in recommendations regarding feasibility, cost, capacity estimates, and implementation requirements. For example, the feasibility of certain US spectrum split is a function of the signal attenuation experienced on that split. Another example of how technical studies are important is that understanding the different noise and channel impairments, which exist

[1] The total number of actives behind a single FN is currently estimated to be around 30.

on different parts of the spectrum and how they can be mitigated via different PHY and MAC technologies, will affect the proposed solution requirements, capacity, efficiency, and cost. A technical evaluation of different capacity-expanding options is included later in this paper, where a migration plan is proposed.

3.5. Headend Equipment

While network topology affects the system performance and offered services, headend equipment also plays a major role into that. The MSOs needs to make sure they specify requirements for products that can scale very well with the projected service offerings. This scale is related to number of channels as well as number of service groups, service group size, servers scale, management and scale of IP addressing scheme (IPv4 & IPv6), etc.

Additionally, not only scale is important, but also the architecture of the headend equipment should be chosen to minimize cost and maximize capacity. Available architectures include Integrated and modular. The MSOs need to make sure they place requirements that result in optimal system architecture in terms of capacity and cost.

On a side note, the Cable industry already started the effort of specifying the scale and requirements of the next generation network architecture, where different requirements were listed in the Converged Cable Access Platform (CCAP) specifications.

3.6. Quality of Experience (QoE)

Quality of Experience is one of the most challenging topics to be addressed. The problem with this topic is that it deals with the customer's perception about the service. The MSO has to collect various system and traffic parameters in order to analyze how the service offering is rated in the customer's eye. The MSOs normally works with system vendors on developing different algorithms and performance metrics that measure the satisfaction of the customers. In this kind of analysis, good questions to be addressed include: For how many seconds can the subscriber wait for a webpage to download? What is the webpage size that the customer is trying to download? How often is he online? How often does he jump between pages while online? What about games latency? What is the pattern of the traffic of a certain game? Does that apply to all games? Does statistical multiplexing help? If so, how does it interact with the number of bonded channels?, etc. The list can go forever!

In order to make sure that the customer has good QoE, the MSOs also need to understand how networks availability affects QoE. Additionally, the effect of the FN size, SG size, offered Tmax needs to be analyzed and understood. Then, the MSO may need to work with system vendors to create algorithms that manage latency and service flows priorities to result in best potential customer QoE.

3.7. Cost

This is the most important factor to consider when planning networks migrations. It is a function of all of the above factors. The goal of network migration is to offer adequate capacity at minimum cost. In many scenarios, the MSOs use the cost per unit of BW as a metric to decide between different proposed solutions. The cost of a certain proposal should take into consideration the investment protection provided by different solutions. It is instructive here to mention that backward compatibility can offer large cost savings, for most of the time, as it capitalizes on using the established base. In many cases, the savings exceed the added cost and complexity which occur when requiring that the new solution be backward compatible with the existing technology.

3.8. Next Steps & Sequence of Steps

There are many network topology options to consider when it comes to the plant migration. These options include: utilizing the available spectrum efficiently, expand the US spectrum, introduce new techniques for better spectral efficiency (like more efficient Forward Error Correction (FEC)), introduce new robust and more efficient PHY technologies (like Orthogonal Frequency Division Multiplexing (OFDM)), require backward compatibility for added enhancements, Go deeper with fiber, etc.

The decision of choosing particular options and the sequence of implementing the options depend on all of the above factors that need to be analyzed thoroughly. In particular, the available options listed above need to be evaluated technically, operationally, and financially. The purpose of this paper, in the next few sections, is to analyze these proposals from the technical point view to provide recommendations to the MSOs as they brainstorm about their network. The technical analysis will provide implications regarding the cost of different solutions. Some options will also be evaluated from the operational point view.

4. OPTIONS TO ACHIEVE 1GBPS IN THE UPSTREAM

This section lists and analyzes the different options, from which the MSOs can choose when planning networks updates in order to produce the goal network architecture. Along with the analysis, technical and operational challenges that may appear throughout the migration process will be exposed and addressed.

4.1. Utilizing the Available BW Efficiently

The utilization of the current 5-42MHz spectrum is far from efficient. In particular, there are portions of the spectrum that are not used at all, while other parts are used inefficiently such that the obtained capacity is way less than what can be potentially offered by that part of the spectrum.

The DOCSIS3.0 has many tools and features in order to help the MSOs achieve the best capacity out of their US spectrum [4] [5] [6] [12]. Some of these parameters include:

- Multiple access technologies (e.g., Advanced Time Division Multiple Access (ATDMA) and Synchronous Code Division Multiple Access (SCDMA)). SCDMA can be very helpful in fighting impulse noise in the lower part of the 5-42MHz spectrum.
- Center frequency selection
- Symbol rate range (0.16 – 5.12 Msymbol/sec)
- Modulation orders (QPSK, 8QAM, 16QAM, 32QAM, or 64QAM)
- Reed-Solomon Forward Error Correction (RS-FEC) to correct up to 16 bytes
- Codeword size selection
- 24-tap pre-equalization
- Long preambles up to 1536 bits
- Ability to adjust to longer/more powerful Preambles
- Proprietary noise mitigation techniques
  - Ex: Ingress Noise Cancellation
- ATDMA Interleaving…
- SCDMA Interleaving
- SCDMA de-spreading
- SCDMA spreading
- SCDMA Trellis Coded Modulation (TCM)
- SCDMA Maximum Scheduled Codes (MSC) feature
- SCDMA Selective Active Codes (SAC) feature
- Channel bonding (MAC layer feature used for PHY layer noise mitigation)
- & Many Many others (Last Codeword Shortened (LCS), max burst size,

scramble seed, differential encoding, etc.)

Detailed analysis of utilizing the above tools and optimizing the spectrum usage can be found in [4] [5] [6]. The abundance of parameters and the flexibility in choosing their values makes it a challenge to optimize them to result in the best spectral efficiency. Therefore, automated tools can be used to measure the different types of noise and also search the solution space of all the parameters and choose the optimal ones that result in the best spectral efficiency. For example, Fig.7 shows that the spectrum can have different types of noise in different portions. Therefore, the automated algorithm shown in Fig. 7 captures the noise in the channel and specifies the best modulation profile and channel parameters that result in the best spectral efficiency. Any automated algorithm needs to have the flexibility to specify constraints for the optimal solution. This is highly desired especially if the MSO does not want to use certain parameters or want to specify certain range for specific parameters. An example of that is shown in Fig. 8, where the algorithm can accept multiple constraints and then searches the constrained solution space to find the optimal parameters that result in the best spectral efficiency.

4.2. Segmenting and Splitting Nodes

Examples of node splits and segmentations were provided in previous sections. The process of node split and segmentation helps in many ways:

- Less Noise funneling as a result of reducing the number of subscribers per node or service group. Lower noise translates to higher SNR and therefore increased capacity.
- Less attenuation because: the deeper the node is, the shorter the coaxial cable becomes, and therefore less signal attenuation is introduced. The lower attenuation translates to higher SNR and therefore increased capacity.
- More average capacity per subscriber. This comes as a natural result of reducing the number of subscribers per node or service group.
- Less contention for BW. Again, this is a natural result of reducing the number of subscribers per node or service group. The reduction in BW contention makes the assumption of requiring 1.5Tmax (or 2Tmax) of capacity to offer Tmax service more reasonable.

Since node splits and segmentations offer all of the above benefits and increased capacity, one may think of performing this process infinitely as demand increases. This, in fact, can be a good approach! However, the cost of node splits rise exponentially every times they are to be performed because the number of resultant nodes doubles after every node split operation. Therefore, there will be a time, when performing the next node split operation will cost more than changing the US spectrum split or laying fibers all the way to the homes or and therefore the natural step after those many node split operations becomes Fiber To The Home (FTTH). This will then make the most reasonable decision from cost point view and also offers multiple times of capacity that may actually be needed by that time.

4.3. Adding More US Spectrum

At some point in the future, the MSOs will need to add more US spectrum to their networks to provide enough capacity to meet the traffic demands. Adding more US spectrum can take many forms: mid-split (5-85MHz), High-split (5-200MHz, 5-238MHz, 5-300MHz, etc.), and top-split (placing US spectrum above the current DS BW). This is shown in Fig. 9 and Fig. 10.

The above splits can be classified into two categories: diplex category (mid-split and high-split), and triplex category (top-split). In particular, in the diplex category, there is only one transition band in the spectrum which separates the US spectrum below the transition band and the DS spectrum above the transition band as shown in Fig. 9. The triplex category, on the other hand, contains two transition bands separating the US and DS spectra as shown in Fig. 10. Specifically, in the triplex architecture, the lower part of the spectrum is used by US traffic, which is followed by the first transition band that is followed by the DS spectrum. The second transition band sits above the DS spectrum and separates it from the US spectrum at the top.

In order for the MSOs to have enough capacity to offer 1Gbps Tmax service and beyond, they will need to move to either high-split or top-split as a goal architecture. This is because mid-split does not offer enough capacity and also MSOs may choose to move from sub-split to high-split directly (instead of going through mid-split) in order to save on the cost of plan upgrade. In particular, the move from sub-split to high-split directly avoids the need to touch the plan multiple times. Other MSOs, however, might choose to go through the mid-split step in order to avoid addressing the OOB STB signaling issue for few years, which allows them to phase out these STBs before moving to high-split architecture.

There are multiple advantages and disadvantage for both the top-split and high-split options. Some of the advantages of the top-split option are:

1. It does not interfere with the OOB STB signaling (frequency range is 70-130MHz).
2. The DS spectrum layout does not need to change. No video channels are affected.

On the other hand, there are several disadvantages for the top-split option including:

1. High signal attenuation, which results in reduced total capacity and inefficient spectrum usage (analysis shown later).
2. More expensive than the high-split option [3].
3. Requires two transition bands which translate to wasted capacity.
4. Requires large bandwidth for the top transition band. In general, the bandwidth of the transition band depends on the frequency of the band. Since the top transition band occurs at high frequency, the transition band bandwidth will be large and this translates to more wasted capacity.
5. Places a cap on the growth of DS spectrum. Once the US spectrum is placed on the top of the DS spectrum, there will be no room to expand the BW of the DS spectrum. Any future growth for the DS will be very challenging because it has to be on the top of the US spectrum and therefore results in these exact disadvantages of wasted capacity (if that option is ever feasible).
6. Requires high modem transmit power for reliable transmission (still at lower capacity).
7. Requires changing all actives to introduce the second transition band.

The high-split architecture, on the other hand, has various advantages including:

1. Offers the highest system capacity (analysis shown later).
2. Less signal attenuation.
3. Single transition band is required.
4. The transition band is narrow because it happens at low frequency.
5. Offers the cheapest solution [3].
6. Does not place a limit on the growth of the DS spectrum.

7. Leverages some of the existing HFC components like laser transmitters and receivers as some of them do support the high-split BW.
8. Offers some backward compatibility because the current DOCSIS3.0 specifications have the US DOCSIS defined from 5-85MHz. This capability already exists in the hardware of various CMTS and modem equipment.

Some of the disadvantages of the high-split option are:

1. It interferes with the OOB STB signaling.
3. It affects the layout of the DS spectrum because the bottom part of the DS spectrum is chewed by the new US spectrum. Some modifications to the DS spectrum layout and channel assignments need to occur.
2. Requires changing all actives to move the current transition band to a higher frequency.

As mentioned above, one of the challenges introduced by the high-split architecture is addressing the OOB STB signaling scheme. There are different scenarios for addressing this issue including:

1. Some MSOs do not have this issue because they have IP or DOCSIS STBs deployed as opposed to legacy STBs which require the signaling in the frequency range 70-130MHz.
2. Phase-out legacy STBs out of the plant. Some MSOs use 9 years as turn-over time for their STBs. Therefore, if the MSOs plan to move to the high-split option in the future and start planning accordingly, the legacy STB problem may not be an issue. The MSOs still have at least 5 years before they need to make any change with the spectrum from a Tmax perspective. This was illustrated in Fig. 4, where the offered Tmax capacity by the current 5-42MHz spectrum runs out of steam around 2017, when the MSO can offer about 90Mbps (assuming that a required channel capacity of 1.5Tmax to offer Tmax service). Note, however, that if the MSOs assume 2Tmax capacity is needed to offer Tmax service, the 5-42MHz spectrum will be consumed (from Tmax point view) one year elarier, namely in 2016, enabling the MSOs to offer Tmax service of ~70Mbps by then. The date, when the capacity of the 5-42MHz spectrum is consumed, can be pushed further in the future if spectrum is used more efficiently via optimizing modulations profiles parameters (shown in earlier sections) and introducing DOCSIS enhancements (will be explained in later sections).
3. Use up-conversion and down-conversion techniques to move the STB signals to higher frequencies beyond the high-split limit. Several approaches are available to perform this, where each approach has its own advantages and disadvantages. The discussion of these solutions is outside the scope of this paper.

Extensive analysis has been done by the ARRIS' team (the author included) to compare different split options from cost and capacity point view [3]. The detailed analysis in [3] is summarized here for convenience. This analysis shows that the high-split option is the most economical solution that offers the highest capacity.

The assumptions used in this analysis are kind of conservative because it was assumed that parts of the spectrum are completely unusable (which may not be the case in most plants). Also, the analysis defines the capacity to be the available DOCSIS3.0 bonding capacity offered by the spectrum. In other words, the analysis does not assume channels used for legacy devices or spectrum monitoring to be part of the available capacity. Specifically, only 22.4MHz was assumed to generate the capacity numbers for the 5-42MHz spectrum. This was rationalized by the different items listed in Table 4. Others assumptions used for this analysis are shown in Tables 5 and 6, while the analysis results are shown in Fig. 11. As mentioned earlier, these numbers can be slightly increased because the assumptions were a little conservative. However, this may not change the course of actions that the MSOs need to do to augment their networks because the difference is insignificant compared to the CAGR of US Tmax.

As can be seen from the above analysis, the high-split option makes the best potential choice for the US spectrum as MSOs plan to upgrade their networks to offer adequate capacity for the required Tmax offerings. Therefore, ARRIS has built a high-split prototype system to mimic the example real-world N+3 network architecture shown in Fig. 12. The real prototype setup is show in Fig. 13. In Fig. 13, all of the active HFC components are ARRIS-made and modified and support 200MHz high-split operation.

The purpose of this effort is to characterize the system and identify any potential limitations or hurdles that may appear as a result of transmitting US signals using the high-split system. The ultimate goal of this experiment is to develop and propose solutions to any identified challenges well before the time of real network migration has come. System analysis for the high-split setup in Fig. 13 has already started. Fig. 14 shows an initial Noise Power Ratio (NPR) curve measured at early stages of the experiment. Further analyses and experiments are still pending and the obtained results will be shared in future papers.

4.4. Introducing PHY Enhancements (Higher Order Modulations) for Better Spectral Efficiency

Introducing higher order modulation options for US transmissions can be a smart move to increase the offered capacity. Currently, the US part of DOCSIS3.0 can support up to QAM64 (or QAM128 with Trellis Coded Modulation (TCM)). Introducing higher order modulations like QAM256, QAM1024, and QAM4096[2] can help in achieving higher spectral efficiencies if/when the plants can support them. For the above modulation orders, QAM256 offers 33% more spectral efficiency than QAM64. QAM1024 offers 25% more capacity than QAM256, and QAM4096 offers 20% more capacity than QAM1024.

As mentioned earlier, node splits and segmentations can result in reduced signal attenuation and noise funneling. Both of these result in higher SNR values that enable the operation of higher order modulation profiles. DOCSIS3.0 noise mitigation toolkit can also help enable the use of higher order modulation orders. Additionally, the next two sections will explain few enhancements that can be added to the DOCSIS, which result in SNR gains that can enable the operation of high order modulation orders.

4.5. Introducing PHY Enhancements (New PHY) for Better Spectral Efficiency

Enhancements to the DOCSIS standard can go beyond offering higher order modulations. Adding modern transmission technologies to DOCSIS toolkit can increase the spectral efficiency. For example, the multi-carrier Orthogonal Frequency Division Multiplexing (OFDM) technology is one of the common PHY techniques used in many of the modern applications including the

[2] These are even-order modulations. Odd modulation orders can be proposed too for higher granularity.

European standard Digital Video Broadcast standard (DVB-C2) [7].

OFDM can be implemented efficiently using the Fast Fourier Transform (FFT) algorithm. Therefore, it requires less chip resources when compared to other transmission technologies with comparable noise immunity, which is one attractive feature that enabled OFDM to be used by different applications. OFDM is also known to have good immunity to various types of noise and channel impairments, which is enabled by the use of subcarriers that also results in long symbol duration, which helps the performance in the presence of impulse noise. The good noise immunity is another attractive feature of OFDM. The proposal to use OFDM (to be exact, Orthogonal Frequency Division Multiple Access (OFDMA)) for US DOCSIS is not a new concept in this paper. In particular, the author analyzed the performance of multi-carrier signals in the presence of HFC noise in 2009 [8] and also proposed the use of OFDM technology for US transmissions in DOCSIS back in 2010 [9].

This section analyzes the gain obtained from using OFDM for US transmissions in DOCSIS networks. The gain obviously depends on the assumptions and input parameters to the model. The analysis assumes an Additive White Gaussian Noise (AWGN) channel. Therefore, the analysis shown here is not an extensive or comprehensive analysis but only shows the gain obtained for one example scenario. More detailed analysis for the benefits of using multi-carrier signals can be found in [9]. Fig. 15 shows capacity estimates for an US single carrier DOCSIS signal and Fig. 16 shows an analysis for the capacity of 200MHz high-split system that uses OFDM. Comparing the results in Fig. 16 to those in Fig. 15, the gain resulting from using OFDM instead of Single carrier is about 2.6% or 0.129 bps/Hz of capacity improvement.

Observe that the increased capacity obtained from introducing OFDM as a new PHY technology for US transmissions is 2.6% when calculated at QAM256. Note that this value is highly dependent on the choice of the OFDM parameters, particularly the cyclic prefix code length and the preamble-to-burst-length ratio. The above improvement at QAM256 is equivalent to an additional 0.214 bits, which translates to 0.63dB of SNR gain. Although the gain may not be very large, some MSOs may choose to use OFDM for US transmissions in order to utilize the US spectrum in the most efficient way and also to use the noise and impairment immunity of OFDM to provide reliable transmissions in harsh plant conditions. In fact, the gain provided by the use OFDM can increase significantly when other parameters are used and also when different noise types (other than AWGN) and channel impairments exist on the channel [9].

Apart from the insignificant capacity improvement provided by OFDM when used with US DOCSIS transmissions shown in the above example, there are many benefits that can be drawn from using OFDM for US DOCSIS including:

1. Backward compatibility with US Channel Bonding: The MSOs can consider bonding across two different PHY technologies and therefore achieve the best possible spectrum utilization. This concept was originally introduced in [9].
2. Easy coexistence and smooth migration: The ability to turn on/off OFDM subcarriers makes it straightforward to accommodate legacy channels within the BW used for the new technology. The reader may be referred to [9] for more details.
3. Low Cost and Optimized Implementation [9]: The OFDM is based on the efficient FFT algorithm and is believed to result in simpler

implementation, which translates to lower cost.

4. Robust to noise and channel impairments: the OFDM is one of the most powerful PHY technologies in terms of its ability to fight different noise types and also mitigate interference [8] [10]. In fact, OFDM is used for wireless channels which are more challenging than the DOCSIS US channels because of multipath fading.
5. More Efficient US Bandwidth Utilization: the analysis above shows that OFDM can result in better spectral efficiency. The analysis assumed AWGN channel, where the results showed minor gain. The gain can be much larger when different noise scenarios and channel impairments exist on the plant [8] [9].
6. Load-Balancing: MSOs can choose to load-balance the traffic on the US between two different PHY technologies. This concept was originally introduced in [9] and also helps with backward compatibility.

One *potential* drawback of OFDM is increased latency. This can appear if the subcarriers width is selected to be very small, which results in increased symbol duration and therefore extended latency. If the subcarriers width is chosen in such a way that the OFDM symbols durations are similar or shorter than the SCDMA symbol durations used in DOCSIS, there will be no extra latency.

4.6. Introducing PHY Enhancements (New FEC) for Better Spectral Efficiency

Another enhancement that can be added to DOCSIS, which results in highly efficient spectral efficiency, is the use of modern Forwarded Error Correction Techniques (FEC). For example, Low Density parity Check (LDPC) codes are known to be much more efficient that the traditional Reed-Solomon codes (RS) codes that are currently being used in DOCSIS. The LDPC scheme was invented many years ago (in 1960's) by Gallager who, at the time, was working on his PhD thesis in MIT on this topic [11]. The LDPC error correction scheme was abandoned for many years because of its implementation complexity that needs high processing power. Recently, LDPC codes have been used in many applications including the DVB-C2 standard [7], which was enabled by the advances in processing platforms.

In order to evaluate the gain offered by the LDPC coding scheme, computer simulations were performed for a QAM signal with un-concatenated Reed Solomon (RS) to represent the current DOCSIS signals [12]. The results of these computer simulations (packet size = 250Bytes) are plotted in Fig. 17 along with other performance numbers for QAM LDPC FEC that are obtained from the published DVB-C2 standard [7]. Note that the above simulated numbers for RS FEC are close to the numbers derived from the J.83 Annex A, where RS FEC and not concatenated RS (RS with convolutional codes) [13] is used, and also similar to the DOCSIS US signals that use vanilla RS FEC. If comparison is to be made against concatenated RS FEC, one will find that the gain achieved by adding LDPC FEC is less because concatenated RS FEC performs better than vanilla RS FEC. Vanilla RS FEC was used in this analysis because it is what currently being used in DODCSIS US transmissions.

Observe that the gains in the QAM256 for the three plotted data points are 4.4dB, 5.1dB, and 7dB, depending on the code rate. Similarly, the gain ranges between 4.2dB and 5.5dB for the QAM64 case depending on the code rate. Therefore, the *average* gain between the LDPC numbers (from DVB-C2) and the RS S numbers (simulated DOCSIS RS FEC) is found to be 5.5dB and 4.85dB for the

QAM256 and QAM64 modulations, respectively. These average SNR gains of 5.5dB or 4.85dB translate to 1.83 bits and 1.62 bits of capacity improvement, respectively.

The above analysis used the *average* SNR gain obtained from using LDPC (the gain is function of the code rate). Therefore, one can be extra conservative and assumes a minimum gain or generous and assumes maximum gain depending on code rate usage on the target network. This paper uses the average gain in the analysis as a reasonable assumption.

4.7. Protecting the Established Base via Backward Compatibility and/or Coexistence

Backward compatibility and coexistence are critical tools to attain investment protection for the established base. As mentioned above in the new PHY proposal, backward compatibility and coexistence can be achieved easily using the OFDM PHY technology if selected as a new PHY for future DOCSIS US transmissions. Several aspects of backward compatible features are offered by OFDM: backward compatibility with US channel bonding across different PHYs, coexistence via the ability to turn on/off subcarriers of OFDM, and load balancing between the legacy and new PHY channels [9].

4.8. Going Deep with Fiber

FTTH is still way in the future! With the current offered capacities and the various available options for MSOs to augment their networks to result in increased capacity, there will be so many years before the MSOs will need to go down the FTTH path.

In fact, gradual migration steps that the MSOs do normally get them smoothly toward FTTH. For example, node splits and segmentations process gets the node closer to the subscribers' homes, which makes it easy and more economical to jump to FTTH at some point in the future. By then, the required capacity will be high (multi-gigabits) and the move to the FTTH will come in the right time. This is one of the beauties of cable networks that they offer the opportunity for timely investments, where spent money and resources are actually used. This is opposed to investing in FTTH, where a large amount of money and resource is spent to offer capacities that are not needed yet.

4.9. Capacity Analysis Summary

This section summarizes the capacity analyses that were introduced in previous sections of this paper. We will start with the analysis from section 4.3, where expanding the US spectrum was proposed. We will use the estimates from that section [3]. Assuming QAM256, the net offered capacity by 200MHz high-split was found to be 855.6Mbps, while the net offered capacity by 238MHz high-split was found to be 999.5Mbps (1Gbps).

Section 4.6 showed an average SNR gain of up to 5.5dB using LDPC alone for the QAM256 case. Additionally, section 4.5 showed additional SNR gain of 0.63dB as a result of using OFDM. Therefore, the total gain introduced by using OFDM and LDPC, compared to the current US DOCIS technology, can be 6.13dB. This gain is equivalent to 2.04 bits. Therefore, the capacity of the 200MHz and 238MHz high-split systems will be as follows:

1. 200MHz High-split: 855.6/(200-5) = 4.3877 bps/Hz. Adding 2.04 bits will increase the above spectral efficiency (calculated at QAM256) to: 4.3877*(8+2.04)/8= 5.51bps/Hz (net capacity is 1.073Gbps).
2. 238MHz High-split: 999.5/(238-5) = 4.2897 bps/Hz. Adding 2.04 bits will increase the above spectral efficiency (calculated at QAM256) to:

4.2897*(8+2.04)/8= 5.38 bps/Hz (net capacity is 1.254 Gbps).

Assume that reduction in noise and signal attenuation that results from multiple node splits and segmentations, as well as optimizations of modulations and channel parameters, result in conservative gain estimate of 3dB (equivalent to one additional bit). Therefore, the capacity of the 200MHz and 238MHz high-split systems is increased as follows:

1. 200MHz-High-split: 5.51*(10.04+1)/10.04 = 6.05bps/Hz (net capacity is 1.18 Gbps).
2. 238MHz-High-split: 5.38*(10.04+1)/10.04 = 5.92bps/Hz (net capacity is 1.378 Gbps).

Since the offered channel capacity is well above 1Gbps in both of the above high-split architecture, one may argue that 200MHz high-split (with offered capacity of 1.18Gbps) is enough to offer a service with Tmax= 1Gbps. Although we assumed earlier that MSOs might choose to require 1.5Tmax of channel capacity to offer a Tmax service, one may suggest that a channel capacity of 15% more than the 1Gbps Tmax value is enough. The rationale behind that is that:

1. After so many node splits and segmentations, the number of subscribers per service groups drops exponentially. This reduces the chances that two subscribers will ask for BW at the same time.
2. When the Tmax value is really large (Tmax = 1Gbps), US bursts from subscribers consume very little time and therefore contention drops significantly. In particular, data transmissions from any single subscriber may not take an extended period of time and therefore will not likely affect other customers that are about to transmit their content. It is, therefore, likely that all customers attain the desired Tmax rates for their service.

Some MSOs may choose to be more cautious and decide to use 238MHz high-split option as a target US spectrum. After all, it is expected that either of the high-split options 200MHz (net capacity of 1.18Gbps) or 238MHz (net capacity of 1.254Gbps) will be able to offer a service with Tmax=1Gbps and beyond.

5. SAMPLE MIGRATION PLAN TO REALIZE SYMMETRIC 1GBPS SYSTEM AND BEYOND!

DOCSIS scales very well! It offers just-in-time steps for plant upgrades, where money and resources that are spent will actually be used. This is opposed to investing in FTTH before it is needed; following that path may lead to a large amount of money and resource being spent to offer capacities and capabilities that are not needed yet.

This section proposes smooth migration steps that MSOs might consider taking when upgrading their networks as they move into the future. These steps offer just-in-time investments that are necessary to offer the needed capacity that meets traffic demands. A natural consequence of these gradual steps is that they will likely occur over many years, with the end goal of migrating to a FTTH architecture when it is truly required. By migrating to FTTH at the right time, this approach will avoid upfront investments that will not be actually used until much later. Based on traffic engineering studies, the need for a FTTH architecture appears to be needed only when traffic demands require much more bandwidth than is provided by DOCSIS or DOCSIS variants. This condition appears to be many years down the road, so the economics of the upgrade process to FTTH can probably be deferred until that time.

As explained earlier, there are many steps and options to take in the process of network migration. One proposed sequence of these steps is given below:

1. **Step 0: Use the available spectrum efficiently.** Section 4.1 addressed this topic. For more details, refer to [4] [5] [6].

2. **Step 1: Node segmentations and splits.** This was covered in Section 4.2.

3. **Step 2: Add more BW.** This is divided into two categories:
   a. **CATEGORY 1 of STEP 2: Expand the US spectrum using High split as goal architecture.** This can be done in a single step to save on upgrade costs or via passing through Mid-split to gain more time to avoid the OOB legacy STB signaling problem. This topic was covered in Section 4.3.
   b. **CATEGORY 2 of STEP 2: Enhancements to DOCSIS.**
      i. <u>Higher order modulations</u>. This is viable because less noise and attenuation as a result of multiple noise segmentations / splits as well as other DOCSIS enhancements mentioned below. Section 4.4 covered this topic.
      ii. <u>New FEC (e.g., LDPC)</u>. This provides several dBs of SNR gain over RS. Section 4.6 addressed this topic.
      iii. <u>New PHY (e.g., OFDM)</u>. OFDM is an easy to implement technology that is robust against different types of noise. Section 4.5 covered this subject. For more details, refer to [9] [10]. Note that a new PHY may not be required because the capacity gain may be marginal as shown earlier. However, if MSOs would like to get the most out of the plant and use noise-robust technology, OFDM makes a good choice.
      iv. <u>Backward Compatibility</u>. This is a key item to maintain the increased offered capacity. Example is bonding across new and legacy channels. This was covered in Section 4.7. For more details, refer to [9].
   c. **NOTE:** The above categories of step 2 (items a & b) can be done in any order or even concurrently. This is a key feature to this proposal. In fact, some MSOs may choose not to go beyond category 1 if they think that it provides enough capacity. Others may jump to category 2 as it may line up better with the timing of their plans to expand the US spectrum. Others may go to both options concurrently (or consecutively) with a bold move to get the most capacity out of the plant.

4. **Step 3: FTTH.** way in the future. Natural step after many node segmentations/splits, which will enable MSOs to offer multi-gigabit service in DS and US. This was covered in section 4.8.

6. CONCLUSIONS

The paper studied different options available to the MSOs as they brainstorm to augment their networks for added capacity. The paper proposed using the current spectrum efficiently, performing node segmentations/splits, adding more US spectrum (Mid-Split/High-Split), and adding enhancements to DOCSIS (Higher order modulations/LDPC/OFDMA/Backward Compatibility for added features). A proposed sequence of gradual migration steps was included, which is deemed to carry the MSOs deep into the future with adequate offered capacity according to the provided analysis.

Since the high-split architecture was shown to make the best technical option for US transmissions, a description a high-split prototype system built by ARRIS was included in the paper. The prototype is aimed at studying and analyzing any potential challenges with the high-split proposal, which enables vendors to offer solutions for any problems or issues well before any mass deployment. Initial results for the prototype system were provided. Future papers are planned to share the results as more experiments are done and data is collected.

ACKNOWLEDGEMENTS

The author would like thank the following ARRIS colleagues for their contributions to this paper: Dr. Frank O'Keeffe, Dr. Tom Cloonan, and Mike Emmendorfer. Additionally, the author would like to thank the following ARRIS colleagues for their support in building the High-split prototype: Brent Arnold, Bill Dawson, Ron Miller, Doc Cummings, Bill Ward, Ameen Khalil, David Warner, and others.

REFERENCES

[1] Tom Cloonan, "On the Evolution of the HFC Network and the DOCSIS CMTS: A Roadmap for the 2012-2016 Era," Proceedings, SCTE 2008 Cable Tec-Expo (June, 2008).

[2] Alan Breznick, "Introduction: The Broadband Outlook", Light Reading Conference on Cable Next-Gen Broadband Strategies 2012 (March, 2012).

[3] Mike Emmendorfer, et. al., "Next Generation - Cable Access Network (NG-CAN): Examination of the Business Drivers and Network Approaches to Enable a Multi-Gigabit Downstream and Gigabit Upstream DOCSIS Service over Coaxial Networks", SCTE Canadian Summit, (March, 2012).

[4] Ayham Al-Banna, "DOCSIS3.0® Performance in the Presence of US HFC Noise", International Technical Seminar, SCTE-South America, (March, 2012).

[5] Tom Cloonan, et. al., "Novel CMTS-based Bandwidth Management Schemes Employing Congestion and Capacity Measurements with Throughput-Maximizing Adjustments for DOCSIS 2.0 Operation", SCTE Conference on Emerging Technologies, (January, 2005).

[6] Ayham Al-Banna, et. al., "DOCSIS® 3.0 Upstream Channel Bonding: Performance Analysis in the Presence of HFC Noise", SCTE-ET NCTA Conference, (April, 2009).

[7] Dirk Jaeger and Christoph Schaaf, "DVB-C2 High Performance Data Transmission on Cable – Technology, Implementation, Networks", (2010).

[8] Ayham Al-Banna and Tom Cloonan, "Performance Analysis of Multi-Carrier Systems when Applied to HFC Networks", SCTE-ET NCTA Conference, (April, 2009).

[9] Ayham Al-Banna, "WiMAX Links and OFDM Overlay for HFC Networks: Mobility and Higher US Capacity", 2010 Spring Technical Forum, NCTA-SCTE, (May, 2010).

[10] Ayham Al-Banna, “Multiple US PHY Technologies: Which Way to Take in Future HFC Networks?”, ANGA Cable Conference, (May, 2011).

[11] Robret Gallager, “Low Density Parity-Check Codes”, MIT Press, Cambridge, MA, (1963).

[12] CableLabs – “Data Over Cable Service Interface Specifications DOCSIS 3.0: Physical Layer Specification”, (October, 2010)

[13] Telecommunication Standardization Sector of ITU, “J.83: Series J: Transmission of television, Sound, programme and other multimedia Signals – Digital transmission of television Signals”, (April, 1997)

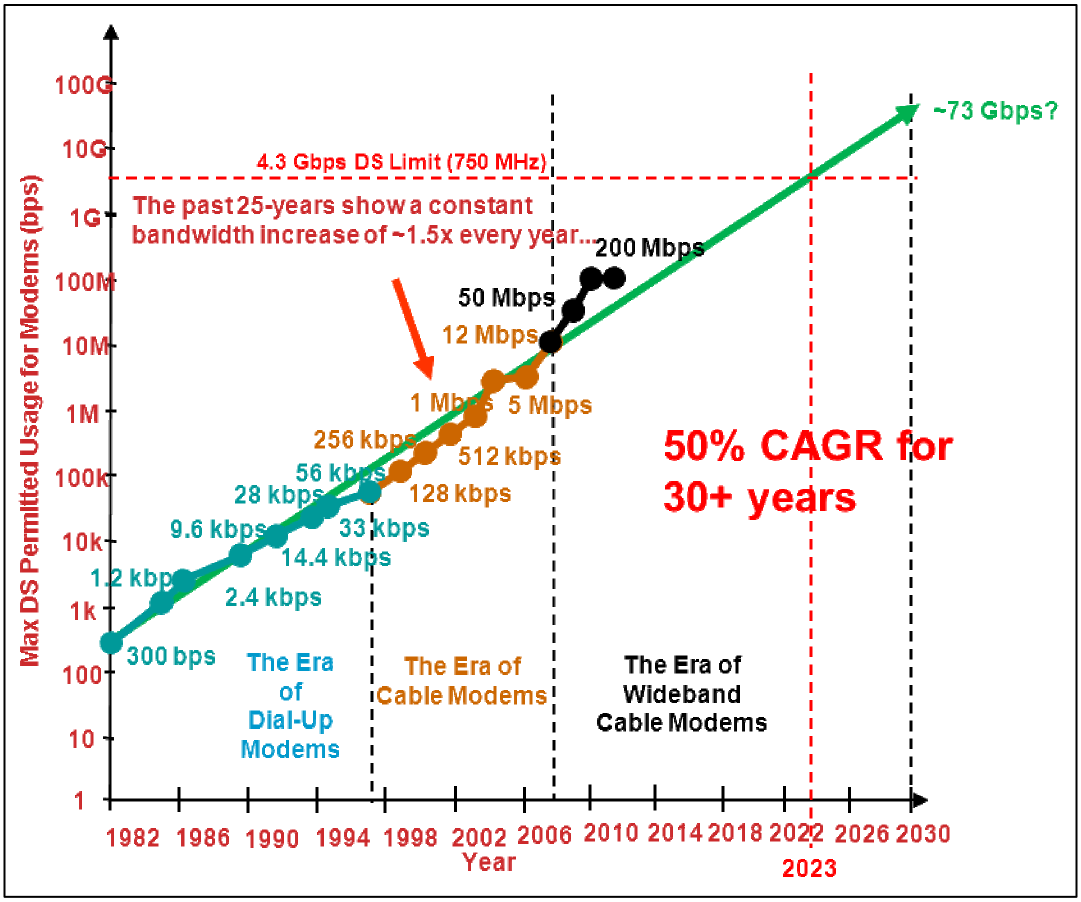


Fig. 1. The Nielson Curve for traffic growth over cable networks (Max. DS Usage/subscriber)

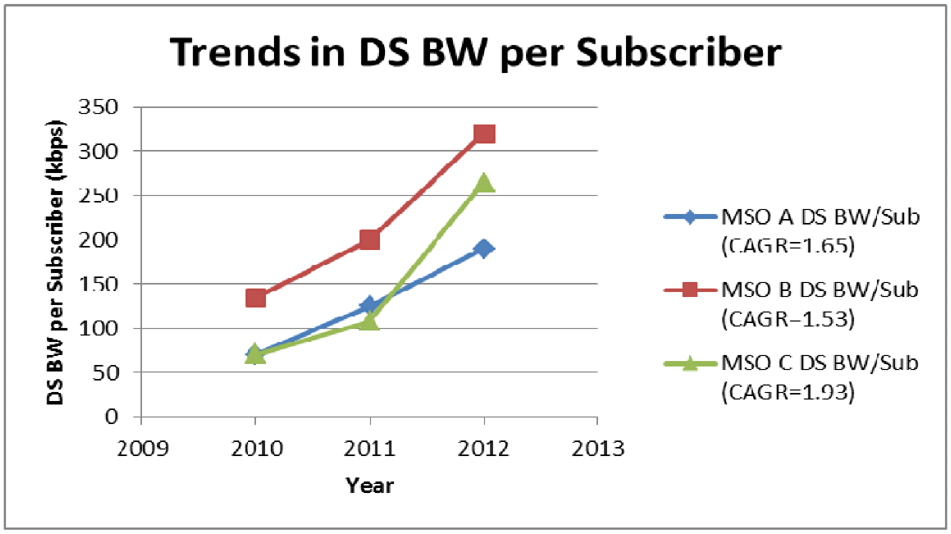


Fig. 2. CAGR of average DS BW per subscriber for three different MSOs over the past two years (>50% DS CAGR on average)

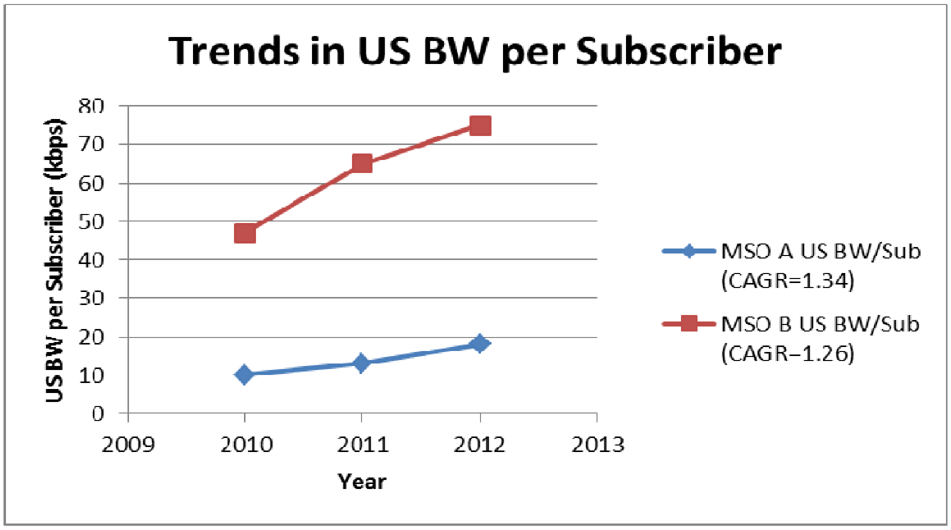


Fig. 3. CAGR of average US BW per subscriber for two different MSOs over the past two years (~30% US CAGR on average)

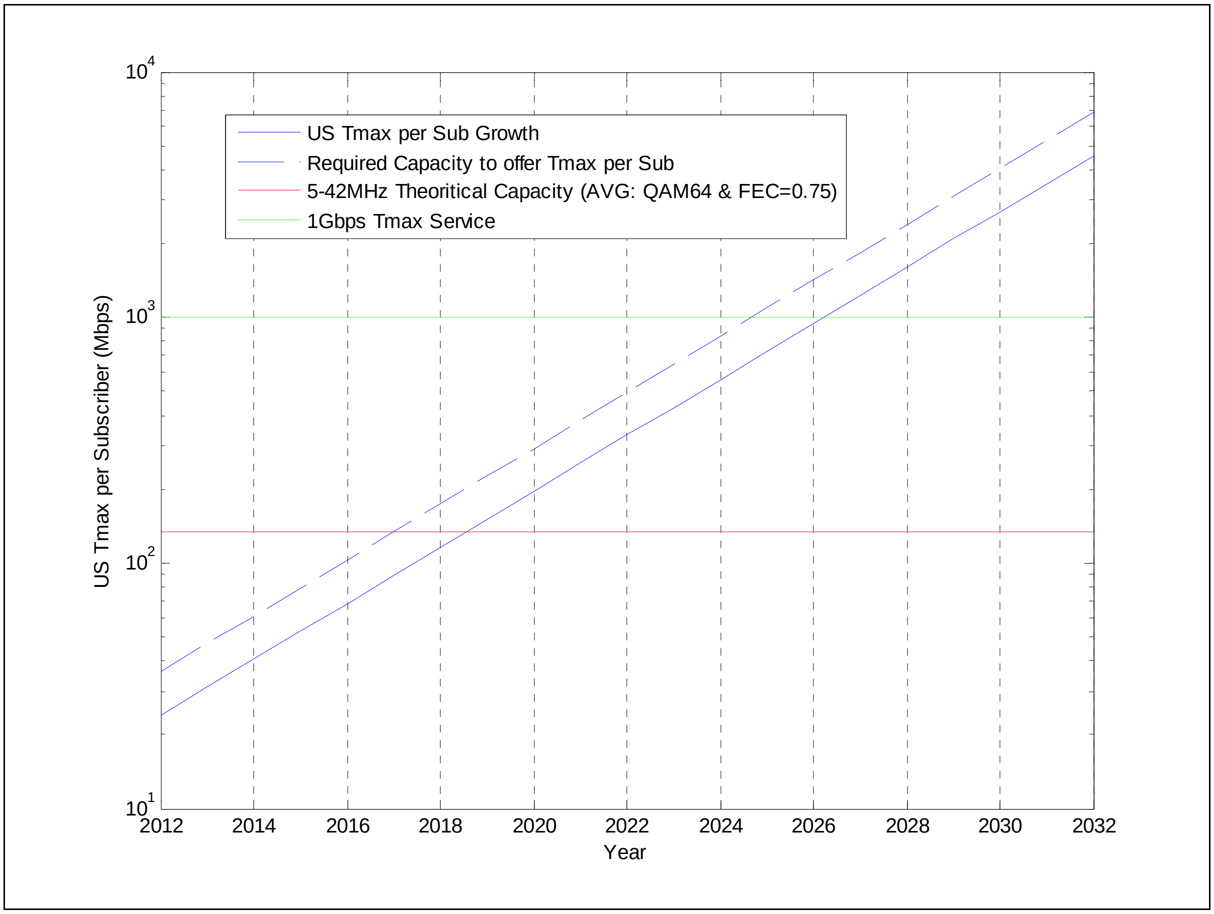


Fig. 4. US Tmax per subscriber growth over the next two decades (assuming CAGR = 30% & starting Tmax = 24Mbps per Subscriber in 2012)

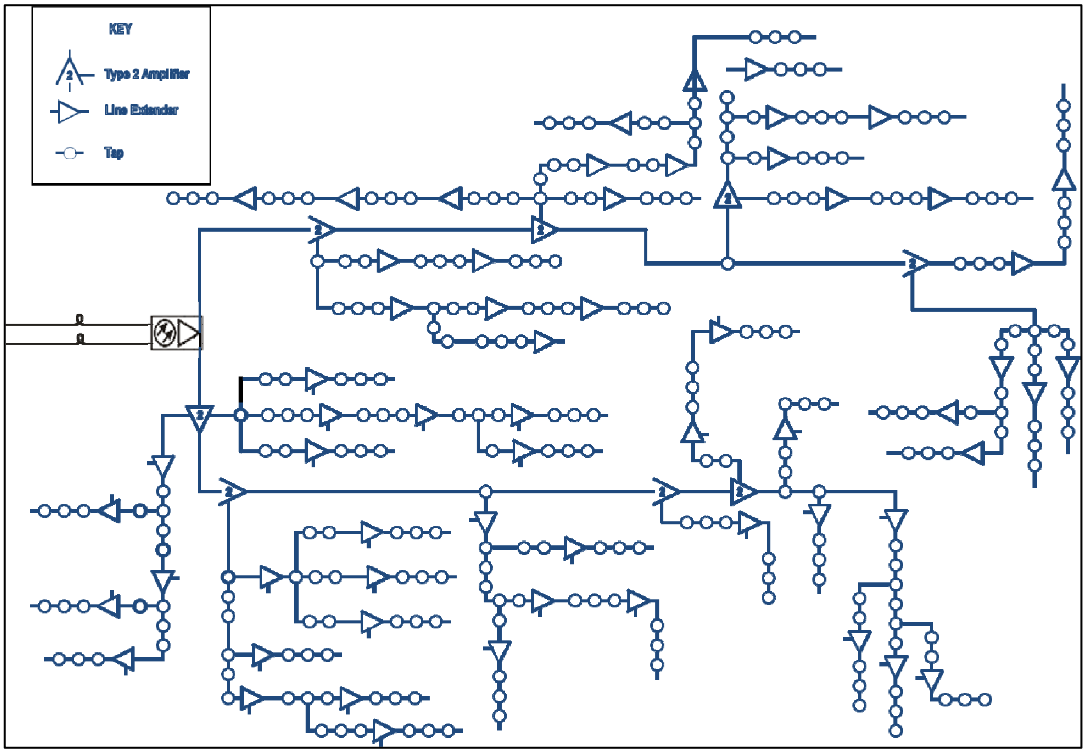


Fig. 5. Example of N+5 Network topology.

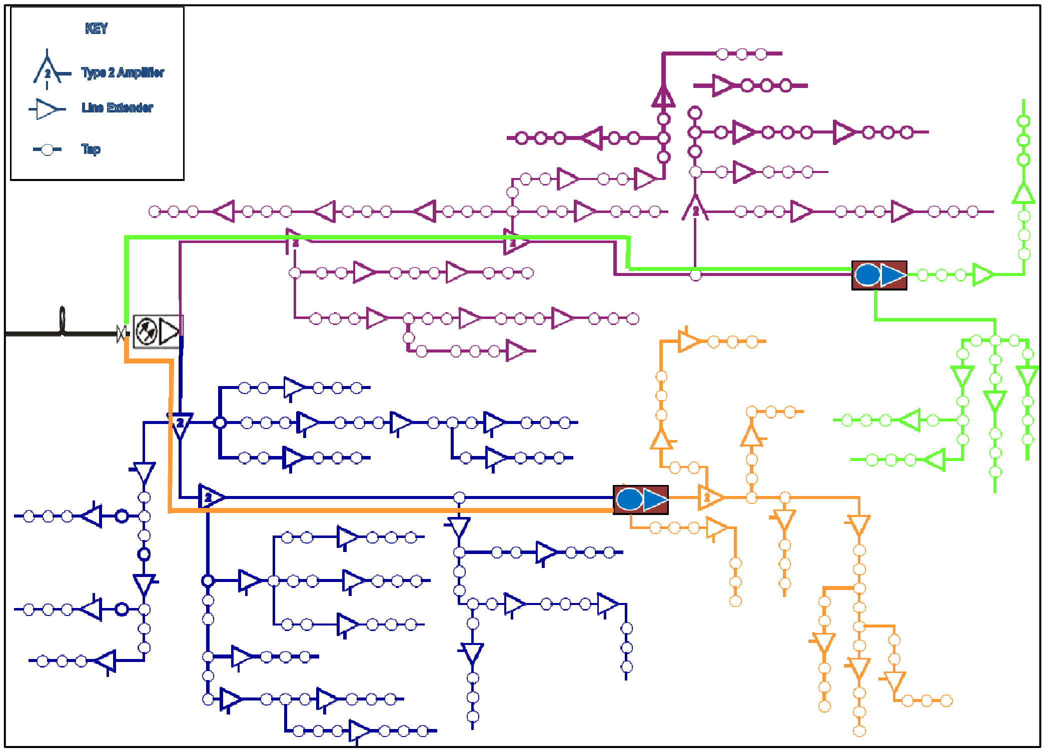
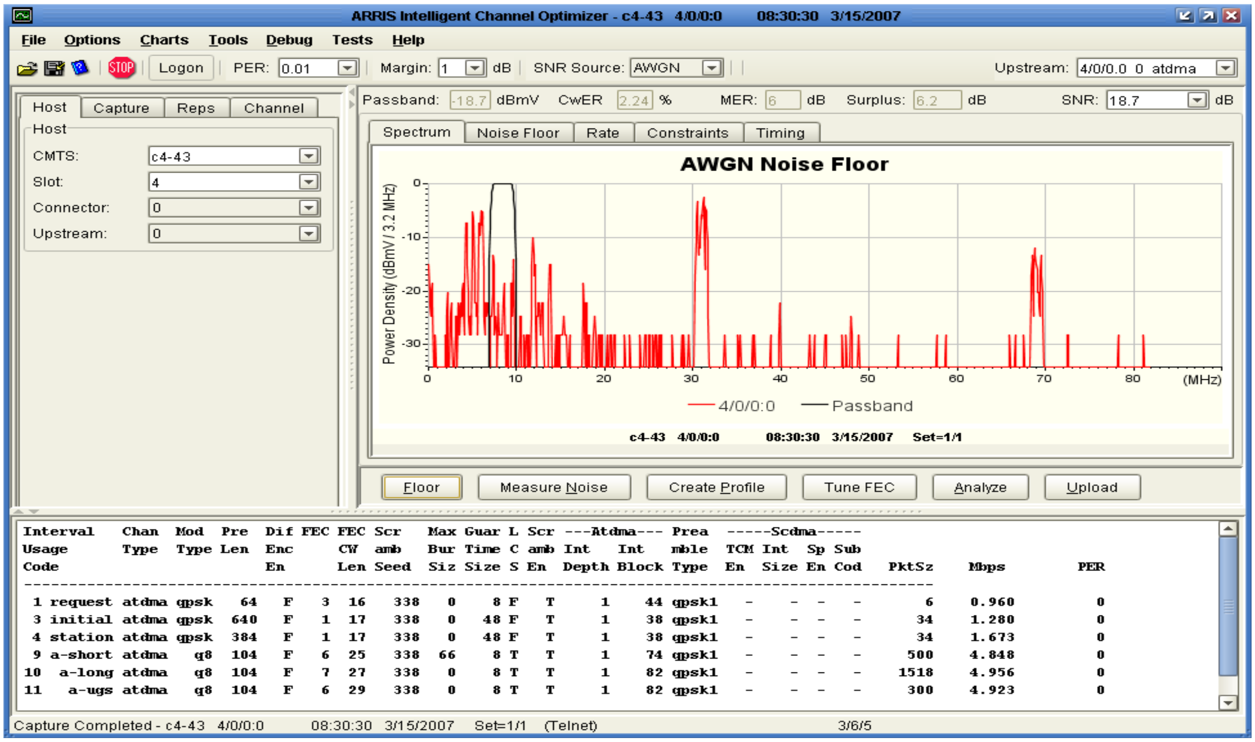


Fig. 6. Segmenting and Splitting the FN in the network example shown in Fig. 5.



PER: 0.01 | Margin: 1 dB | SNR Source: AWGN | Upstream: 4/0/0.0 0 atdma

Passband: -18.7 dBmV | CwER 2.24 % | MER: 6 dB | Surplus: 6.2 dB | SNR: 18.7 dB

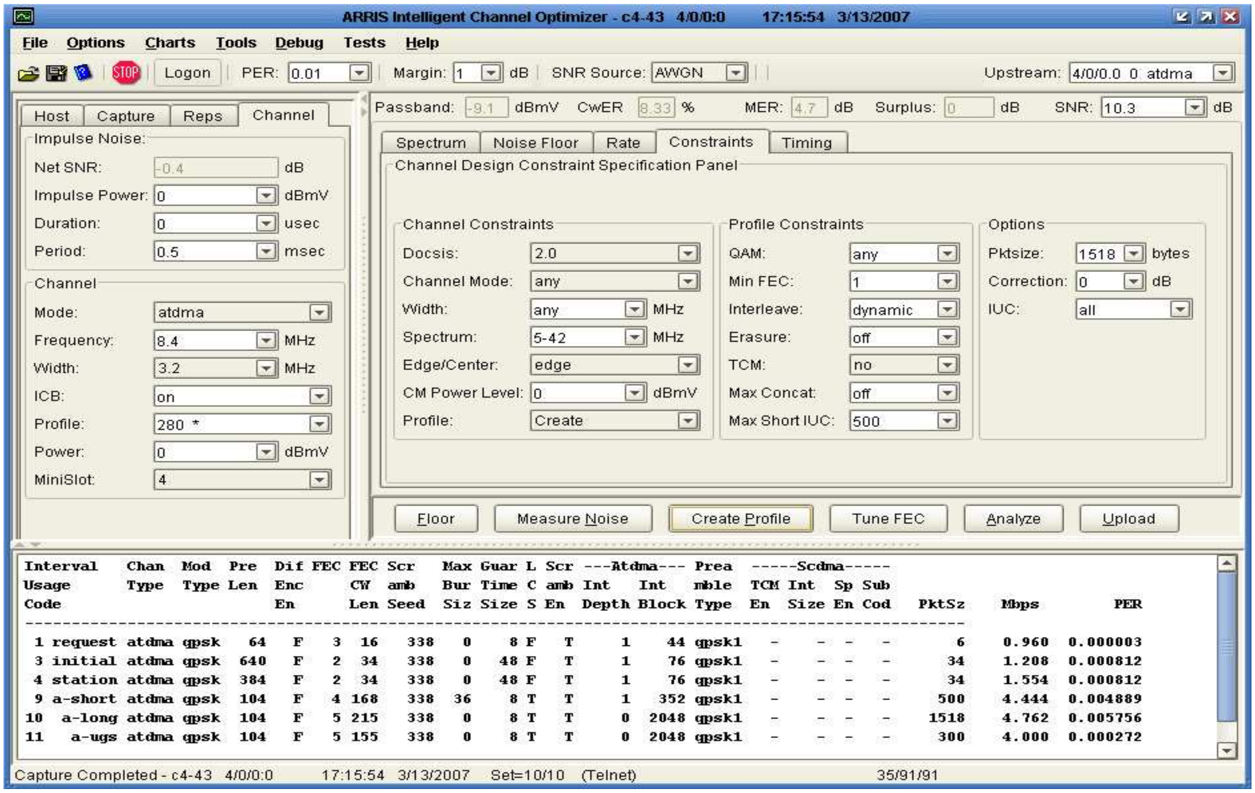
Host

CMTS: c4-43
Slot: 4
Connector: 0
Upstream: 0

```
Interval    Chan  Mod  Pre  Dif FEC FEC Scr   Max Guar L Scr ---Atdma--- Prea   -----Scdma-----
Usage       Type  Type Len  Enc     CW  amb   Bur Time C amb Int    Int  mble   TCM Int  Sp Sub
Code                        En      Len Seed  Siz Size S En  Depth Block Type   En  Size En Cod   PktSz     Mbps        PER
---------------------------------------------------------------------------------------------------------
  1 request atdma qpsk   64   F    3  16    338   0    8 F    T      1    44 qpsk1   -    -  -    -       6    0.960         0
  3 initial atdma qpsk  640   F    1  17    338   0   48 F    T      1    38 qpsk1   -    -  -    -      34    1.280         0
  4 station atdma qpsk  384   F    1  17    338   0   48 F    T      1    38 qpsk1   -    -  -    -      34    1.673         0
  9 a-short atdma   q8  104   F    6  25    338  66    8 T    T      1    74 qpsk1   -    -  -    -     500    4.848         0
 10  a-long atdma   q8  104   F    7  27    338   0    8 T    T      1    82 qpsk1   -    -  -    -    1518    4.956         0
 11   a-ugs atdma   q8  104   F    6  29    338   0    8 T    T      1    82 qpsk1   -    -  -    -     300    4.923         0
```

Capture Completed - c4-43 4/0/0:0 08:30:30 3/15/2007 Set=1/1 (Telnet) 3/6/5

Fig. 7. Automation of optimizing the upstream modulation profile and channel parameters (choosing the best parameters for the noise that exists on the plant)



PER: 0.01 | Margin: 1 dB | SNR Source: AWGN | Upstream: 4/0/0.0 0 atdma

Passband: -9.1 dBmV | CwER 8.33 % | MER: 4.7 dB | Surplus: 0 dB | SNR: 10.3 dB

Impulse Noise:

Net SNR: -0.4 dB
Impulse Power: 0 dBmV
Duration: 0 usec
Period: 0.5 msec

Channel

Mode: atdma
Frequency: 8.4 MHz
Width: 3.2 MHz
ICB: on
Profile: 280 *
Power: 0 dBmV
MiniSlot: 4

Channel Design Constraint Specification Panel

Channel Constraints

Docsis: 2.0
Channel Mode: any
Width: any MHz
Spectrum: 5-42 MHz
Edge/Center: edge
CM Power Level: 0 dBmV
Profile: Create

Profile Constraints

QAM: any
Min FEC: 1
Interleave: dynamic
Erasure: off
TCM: no
Max Concat: off
Max Short IUC: 500

Options

Pktsize: 1518 bytes
Correction: 0 dB
IUC: all

```
Interval    Chan  Mod  Pre  Dif FEC FEC Scr   Max Guar L Scr ---Atdma--- Prea   -----Scdma-----
Usage       Type  Type Len  Enc     CW  amb   Bur Time C amb Int    Int  mble   TCM Int  Sp Sub
Code                        En      Len Seed  Siz Size S En  Depth Block Type   En  Size En Cod   PktSz     Mbps        PER
-------------------------------------------------------------------------------------------------------------
  1 request atdma qpsk   64   F    3  16    338   0    8 F    T      1    44 qpsk1   -    -  -    -       6    0.960  0.000003
  3 initial atdma qpsk  640   F    2  34    338   0   48 F    T      1    76 qpsk1   -    -  -    -      34    1.208  0.000812
  4 station atdma qpsk  384   F    2  34    338   0   48 F    T      1    76 qpsk1   -    -  -    -      34    1.554  0.000812
  9 a-short atdma qpsk  104   F    4 168    338  36    8 T    T      1   352 qpsk1   -    -  -    -     500    4.444  0.004889
 10  a-long atdma qpsk  104   F    5 215    338   0    8 T    T      0  2048 qpsk1   -    -  -    -    1518    4.762  0.005756
 11   a-ugs atdma qpsk  104   F    5 155    338   0    8 T    T      0  2048 qpsk1   -    -  -    -     300    4.000  0.000272
```

Capture Completed - c4-43 4/0/0:0 17:15:54 3/13/2007 Set=10/10 (Telnet) 35/91/91

Fig. 8. Automation of optimizing the upstream modulation profile and channel parameters (specifying constraints)

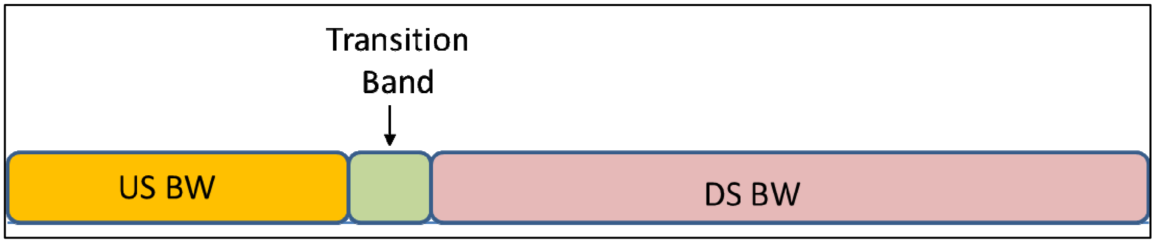


Fig. 9. Mid-Split and High-Split options for US spectrum usage

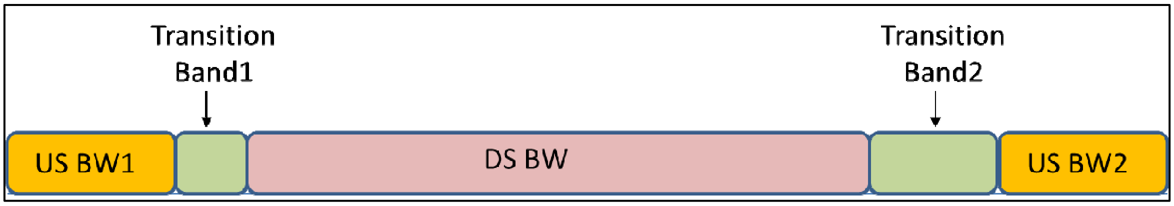
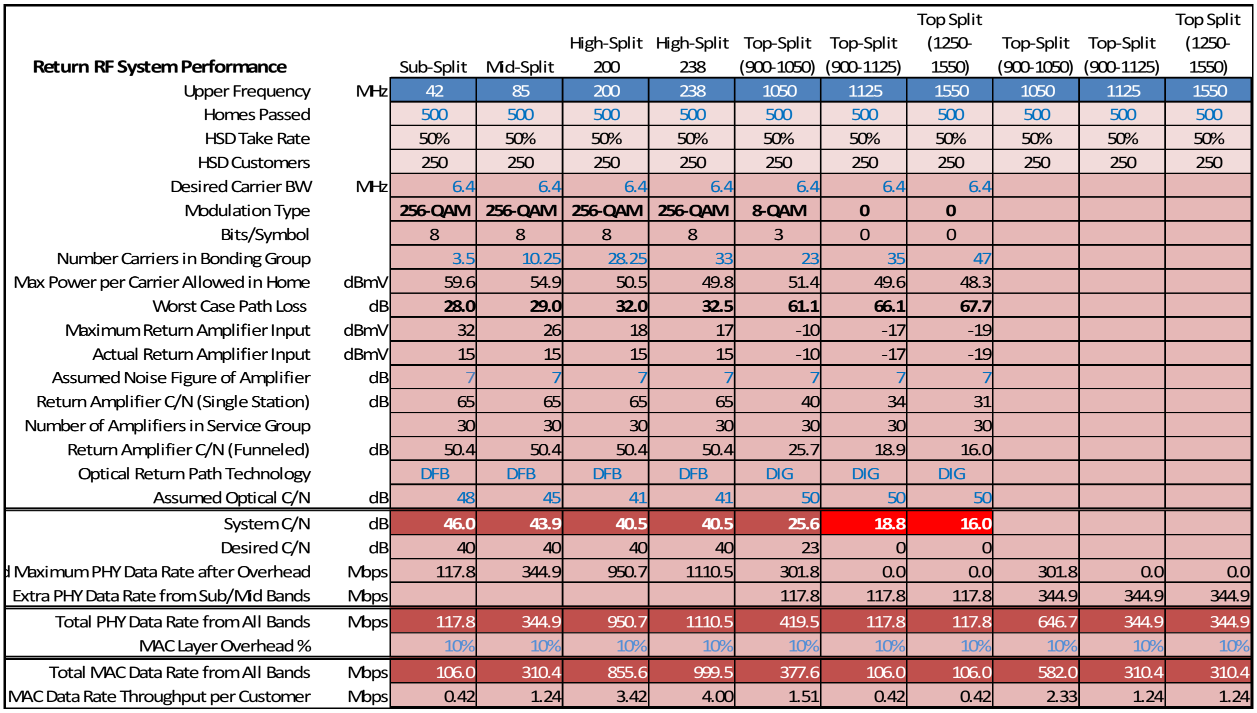


Fig. 10. Top-Split option for US spectrum usage



								Top Split			Top Split
				High-Split	High-Split	Top-Split	Top-Split	(1250-	Top-Split	Top-Split	(1250-
Return RF System Performance		Sub-Split	Mid-Split	200	238	(900-1050)	(900-1125)	1550)	(900-1050)	(900-1125)	1550)
Upper Frequency	MHz	42	85	200	238	1050	1125	1550	1050	1125	1550
Homes Passed		500	500	500	500	500	500	500	500	500	500
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	250	250	250	250	250	250	250
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4			
Modulation Type		**256-QAM**	**256-QAM**	**256-QAM**	**256-QAM**	**8-QAM**	**0**	**0**			
Bits/Symbol		8	8	8	8	3	0	0			
Number Carriers in Bonding Group		3.5	10.25	28.25	33	23	35	47			
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	50.5	49.8	51.4	49.6	48.3			
Worst Case Path Loss	dB	**28.0**	**29.0**	**32.0**	**32.5**	**61.1**	**66.1**	**67.7**			
Maximum Return Amplifier Input	dBmV	32	26	18	17	-10	-17	-19			
Actual Return Amplifier Input	dBmV	15	15	15	15	-10	-17	-19			
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7			
Return Amplifier C/N (Single Station)	dB	65	65	65	65	40	34	31			
Number of Amplifiers in Service Group		30	30	30	30	30	30	30			
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	50.4	25.7	18.9	16.0			
Optical Return Path Technology		DFB	DFB	DFB	DFB	DIG	DIG	DIG			
Assumed Optical C/N	dB	48	45	41	41	50	50	50			
System C/N	dB	**46.0**	**43.9**	**40.5**	**40.5**	**25.6**	**18.8**	**16.0**			
Desired C/N	dB	40	40	40	40	23	0	0			
d Maximum PHY Data Rate after Overhead	Mbps	117.8	344.9	950.7	1110.5	301.8	0.0	0.0	301.8	0.0	0.0
Extra PHY Data Rate from Sub/Mid Bands	Mbps					117.8	117.8	117.8	344.9	344.9	344.9
Total PHY Data Rate from All Bands	Mbps	117.8	344.9	950.7	1110.5	419.5	117.8	117.8	646.7	344.9	344.9
MAC Layer Overhead %		10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Total MAC Data Rate from All Bands	Mbps	106.0	310.4	855.6	999.5	377.6	106.0	106.0	582.0	310.4	310.4
MAC Data Rate Throughput per Customer	Mbps	0.42	1.24	3.42	4.00	1.51	0.42	0.42	2.33	1.24	1.24

Fig. 11. Analysis of different split options for the US spectrum in DOCSIS networks [3]

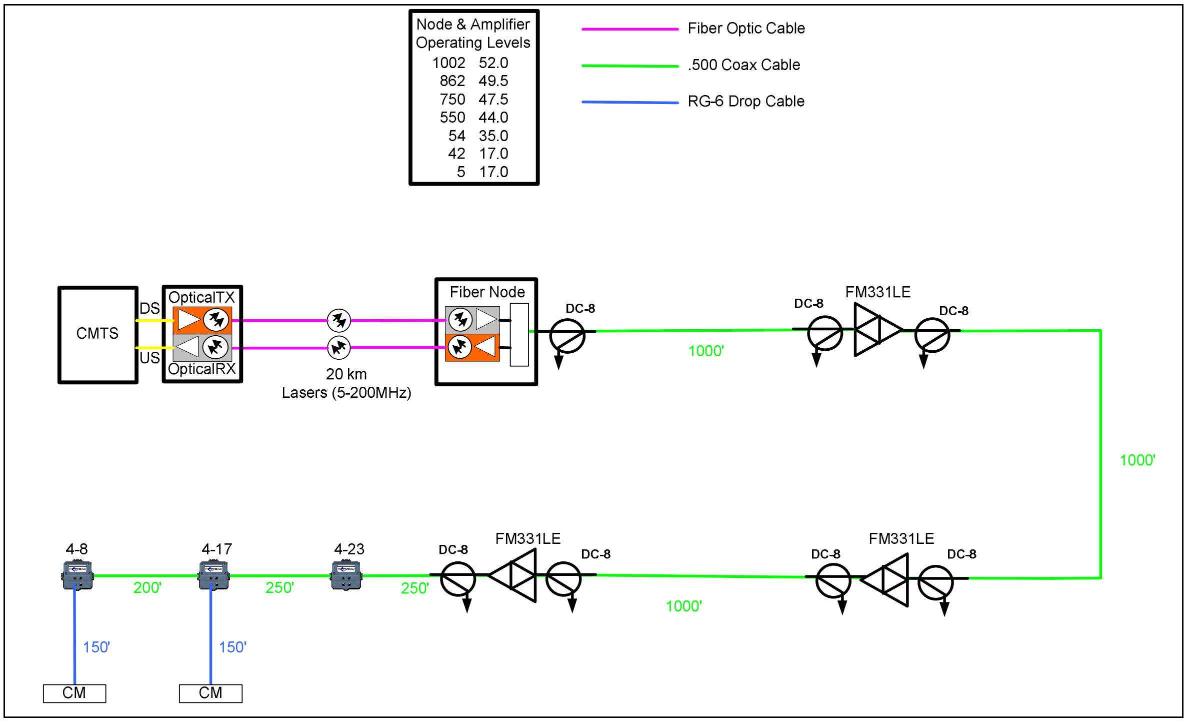


Fig. 12. Example setup for Real-world N+3 network architecture

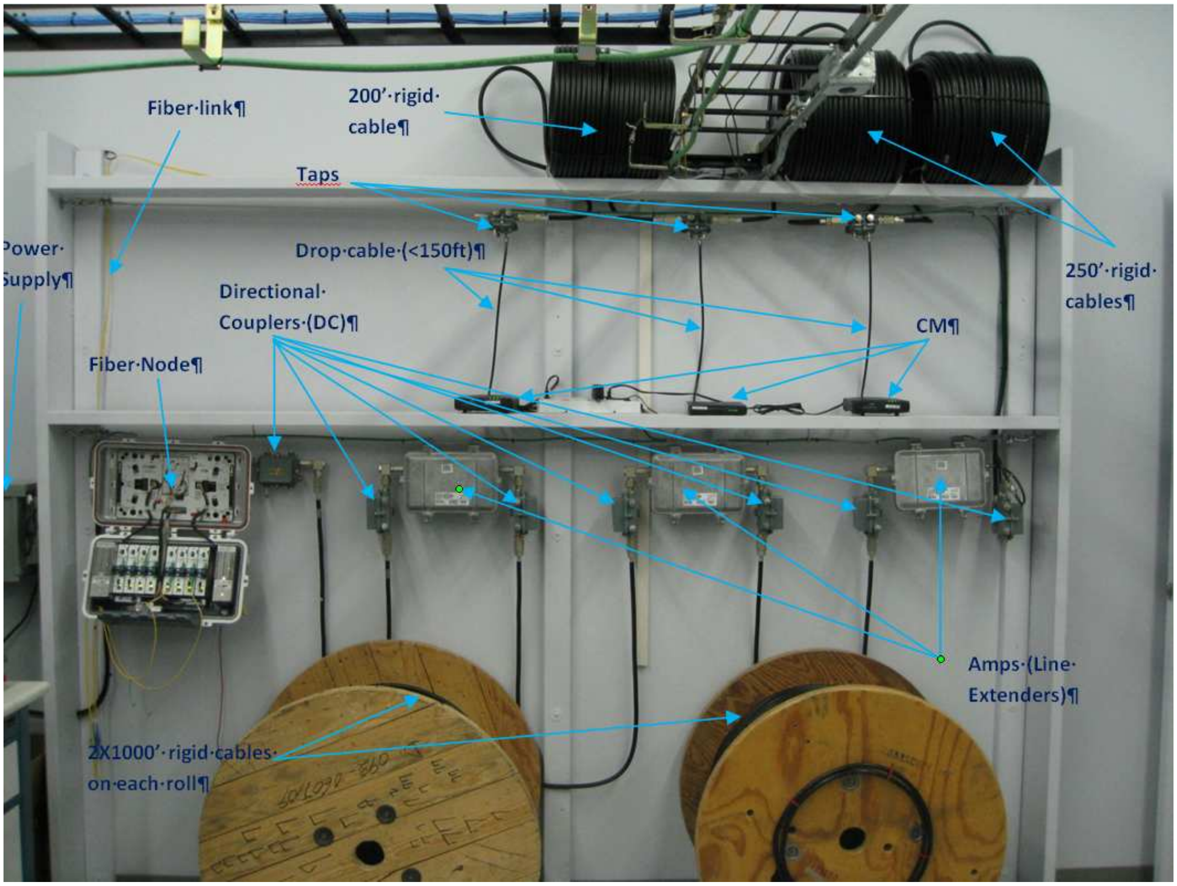


Fig. 13. ARRIS Implementation of high-split prototype architecture network to mimic the setup in Fig. 12 (laser Tx/Rx in the headend is not shown in the figure).

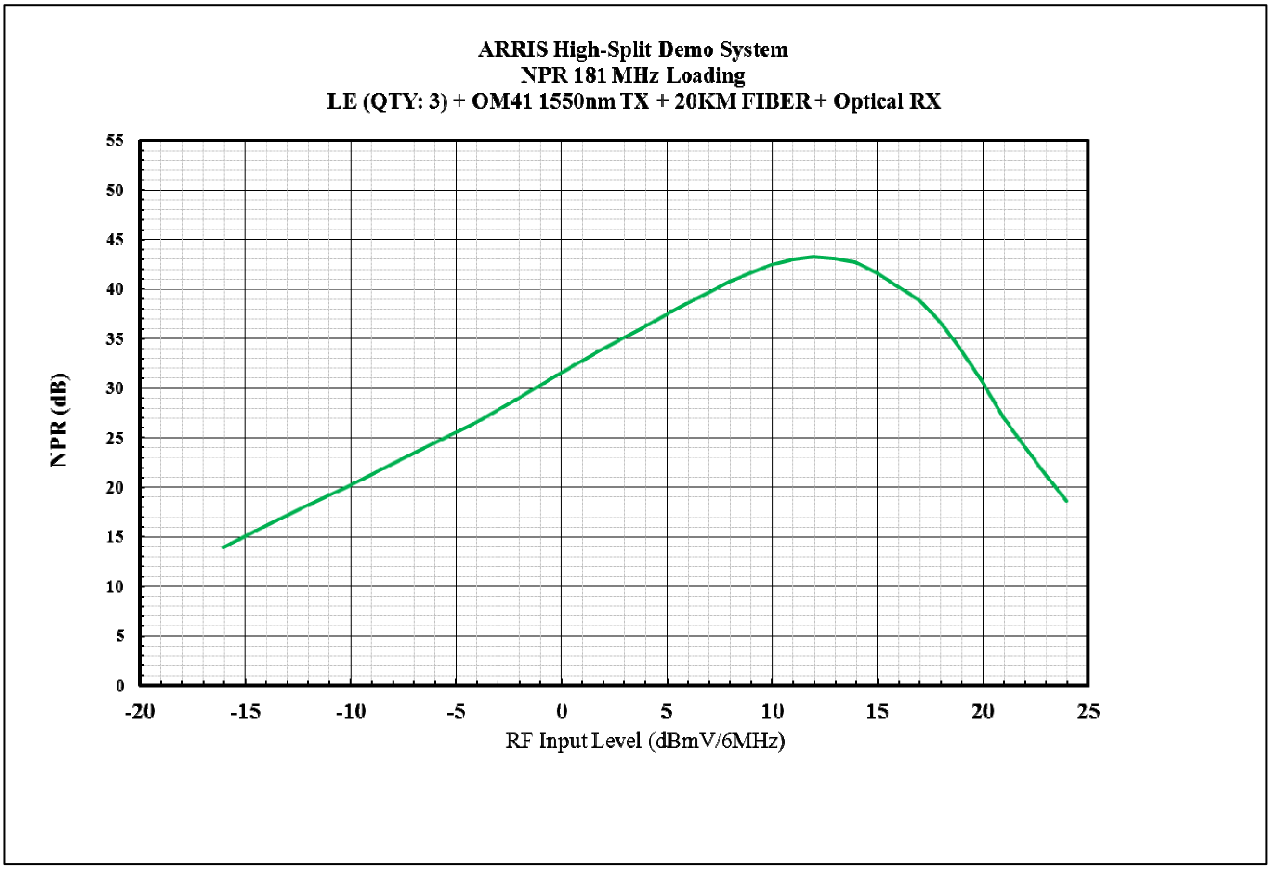
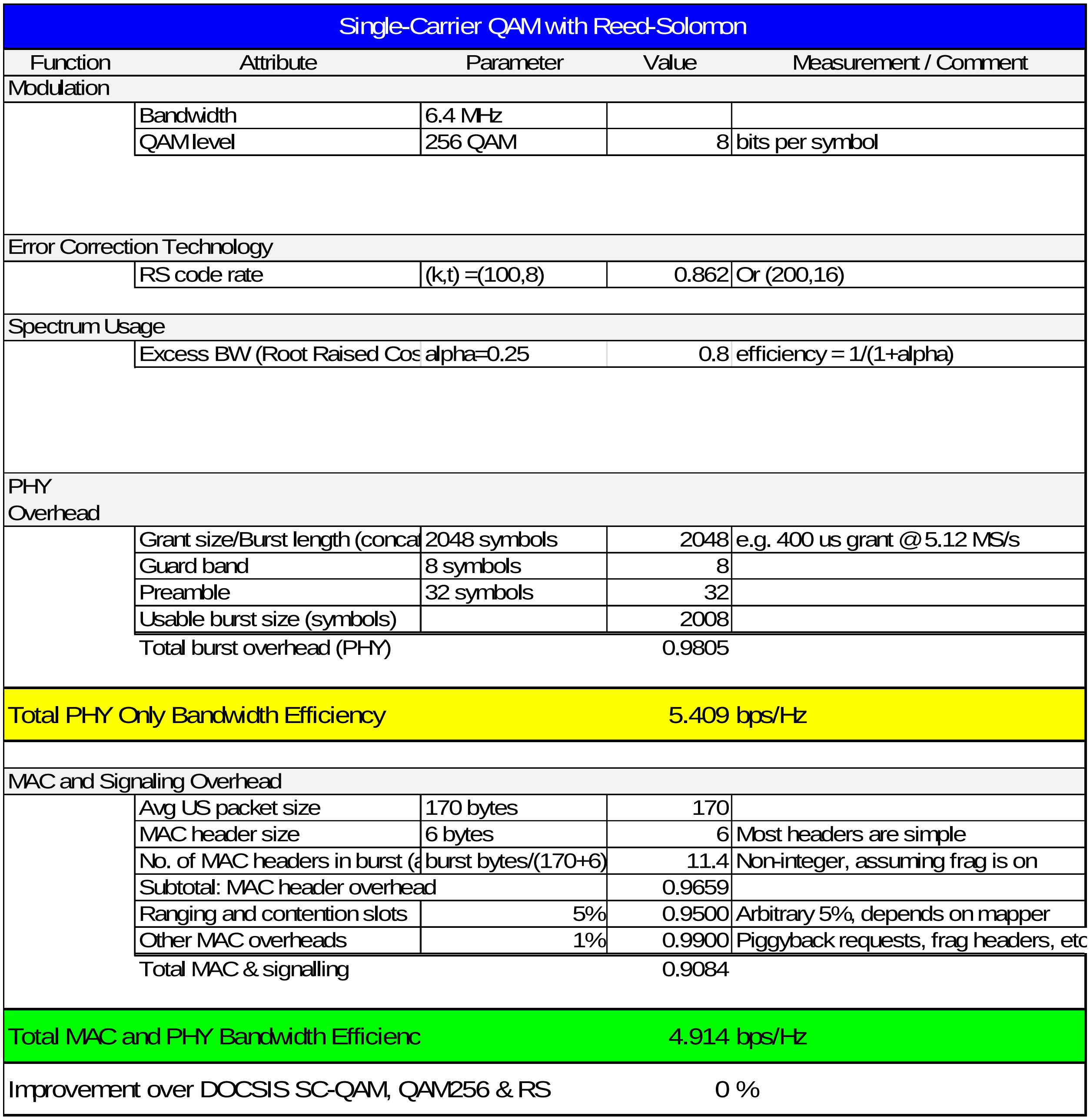
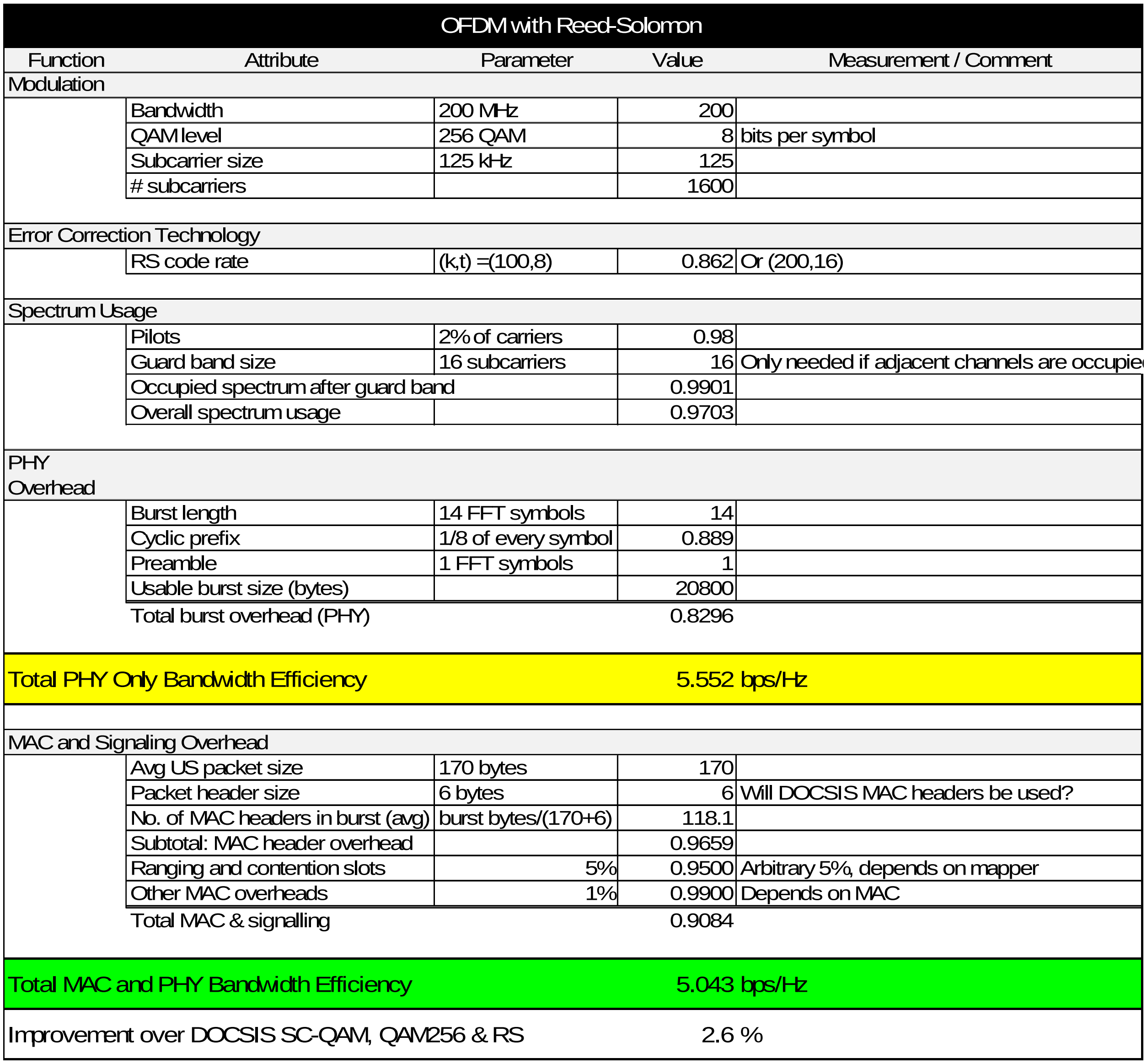


Fig. 14. An initial NPR curve for the plant setup shown in Fig. 13.



Single-Carrier QAM with Reed-Solomon				
Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	6.4 MHz		
	QAM level	256 QAM	8	bits per symbol
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Excess BW (Root Raised Cos	alpha=0.25	0.8	efficiency = 1/(1+alpha)
PHY Overhead				
	Grant size/Burst length (concat	2048 symbols	2048	e.g. 400 us grant @ 5.12 MS/s
	Guard band	8 symbols	8	
	Preamble	32 symbols	32	
	Usable burst size (symbols)		2008	
	Total burst overhead (PHY)		0.9805	
Total PHY Only Bandwidth Efficiency			5.409	bps/Hz
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	MAC header size	6 bytes	6	Most headers are simple
	No. of MAC headers in burst (a	burst bytes/(170+6)	11.4	Non-integer, assuming frag is on
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Piggyback requests, frag headers, etc
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficienc			4.914	bps/Hz
Improvement over DOCSIS SC-QAM, QAM256 & RS			0	%

Fig. 15. Capacity analysis for Single carrier DOCSIS signal



OFDM with Reed-Solomon

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	200 MHz	200	
	QAM level	256 QAM	8	bits per symbol
	Subcarrier size	125 kHz	125	
	# subcarriers		1600	
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Pilots	2% of carriers	0.98	
	Guard band size	16 subcarriers	16	Only needed if adjacent channels are occupie
	Occupied spectrum after guard band		0.9901	
	Overall spectrum usage		0.9703	
PHY Overhead				
	Burst length	14 FFT symbols	14	
	Cyclic prefix	1/8 of every symbol	0.889	
	Preamble	1 FFT symbols	1	
	Usable burst size (bytes)		20800	
	Total burst overhead (PHY)		0.8296	
Total PHY Only Bandwidth Efficiency			5.552 bps/Hz	
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	Packet header size	6 bytes	6	Will DOCSIS MAC headers be used?
	No. of MAC headers in burst (avg)	burst bytes/(170+6)	118.1	
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Depends on MAC
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			5.043 bps/Hz	
Improvement over DOCSIS SC-QAM, QAM256 & RS			2.6 %	

Fig. 16. Capacity analysis for OFDM signals when used for DOCSIS US transmissions

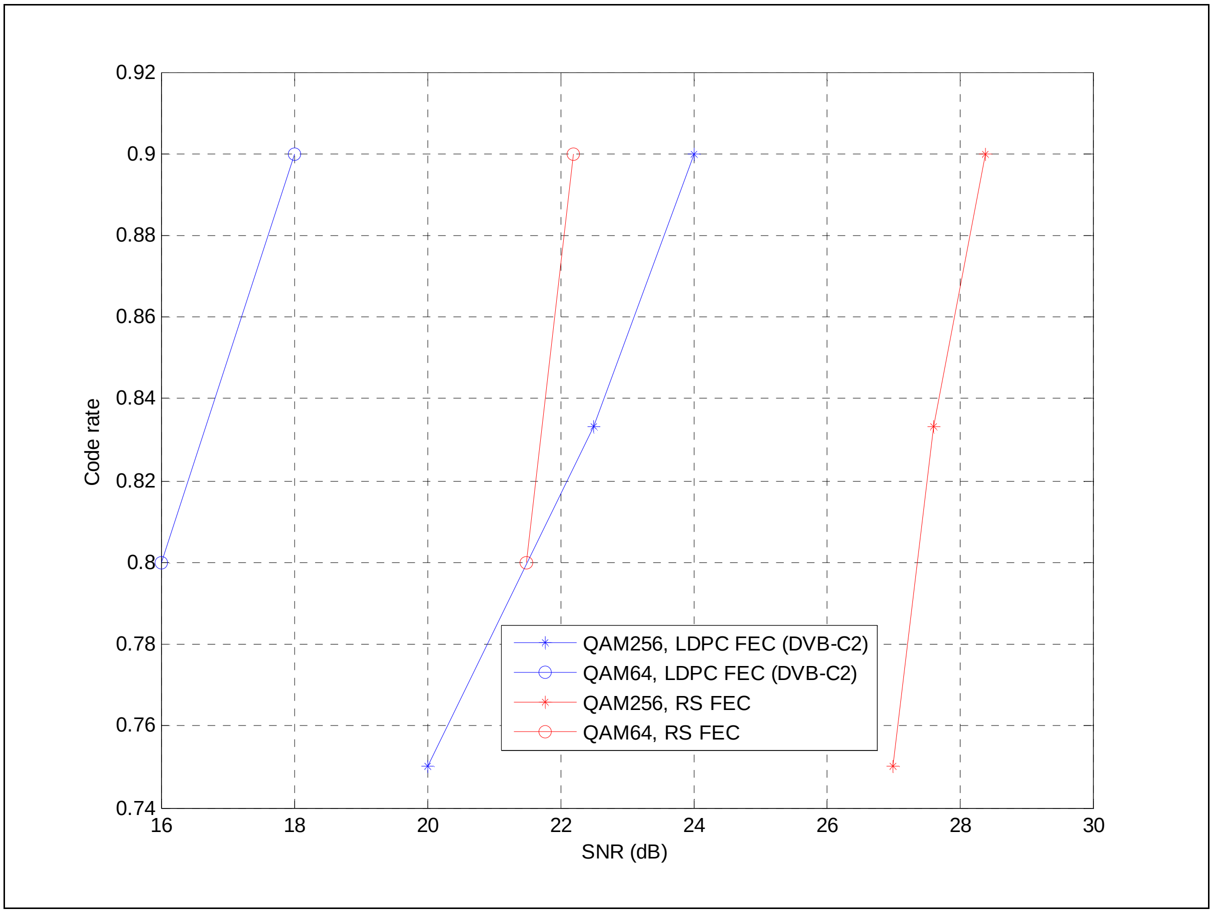
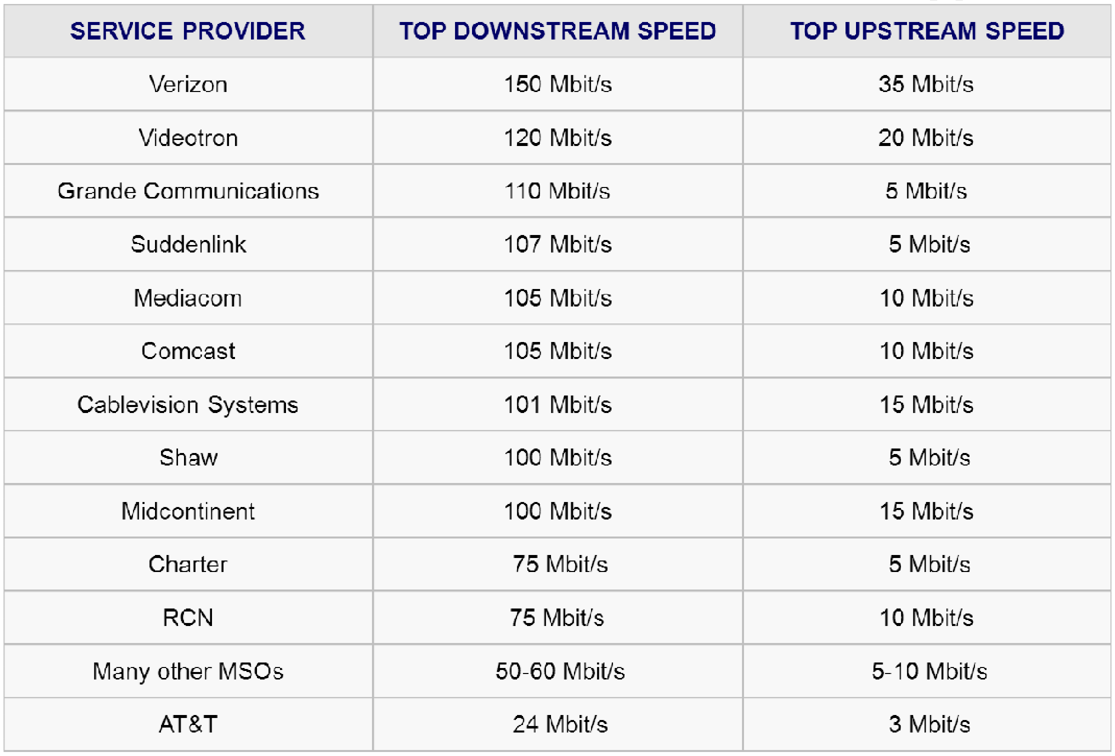


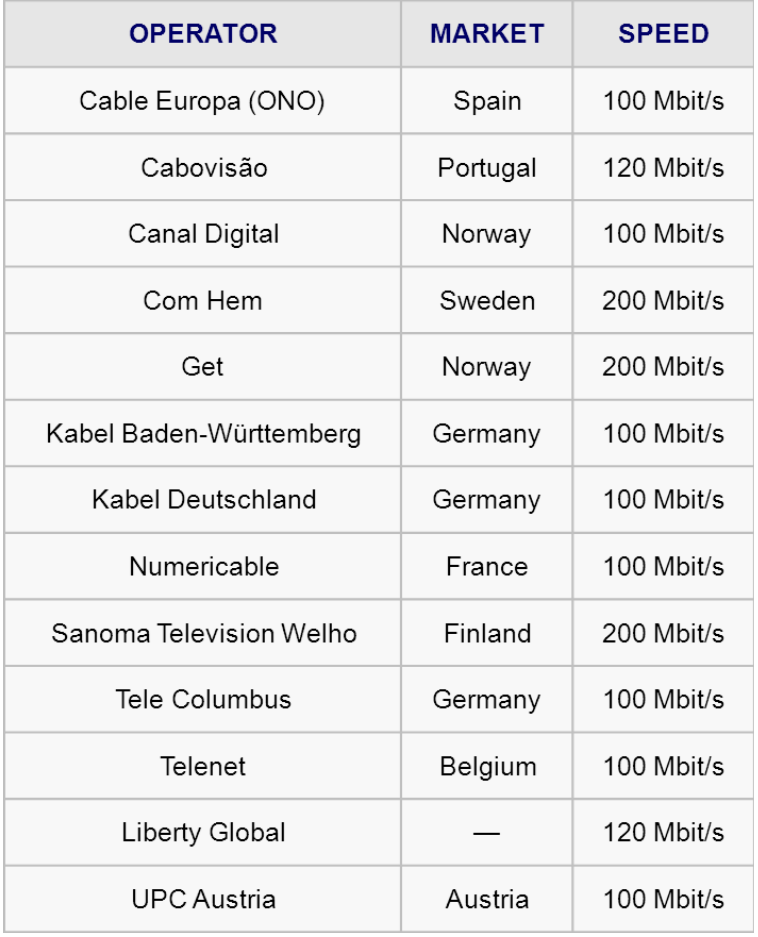
Fig. 17. Comparison between RS FEC (computer simulations) and LDPC FEC (from DVB-C2)

Table 1. Offered DS and US Tmax values in North America [2]



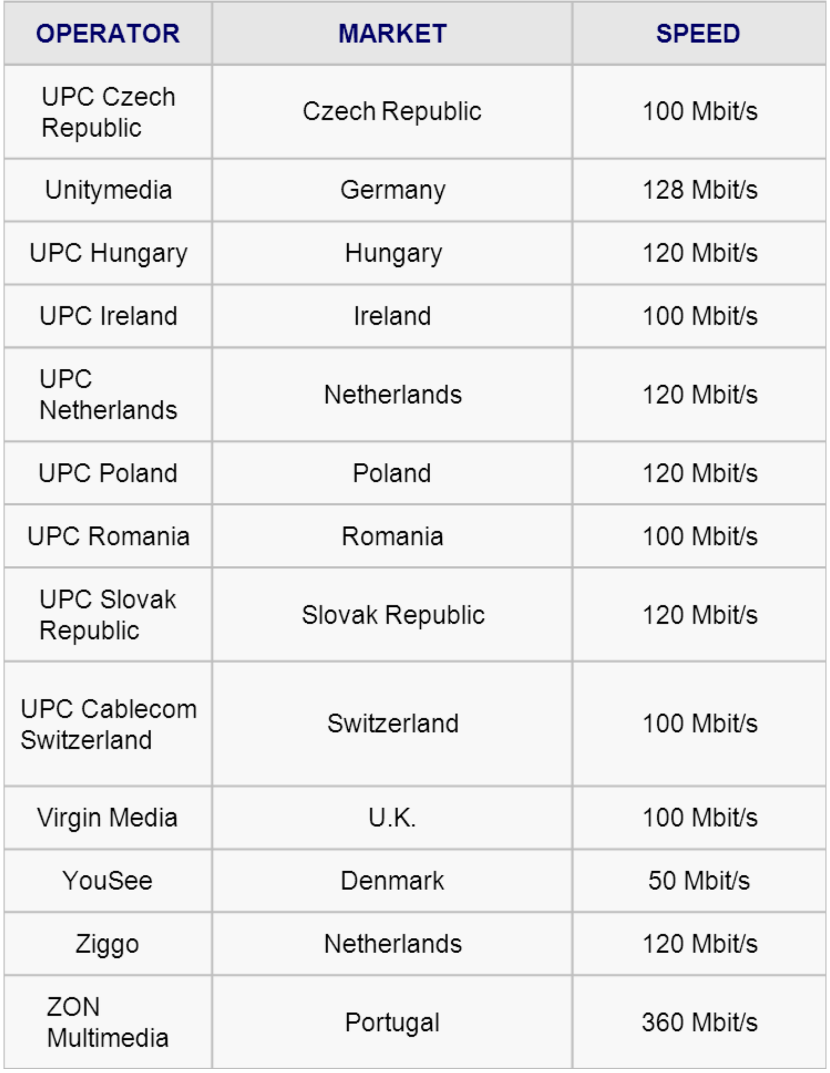
SERVICE PROVIDER	TOP DOWNSTREAM SPEED	TOP UPSTREAM SPEED
Verizon	150 Mbit/s	35 Mbit/s
Videotron	120 Mbit/s	20 Mbit/s
Grande Communications	110 Mbit/s	5 Mbit/s
Suddenlink	107 Mbit/s	5 Mbit/s
Mediacom	105 Mbit/s	10 Mbit/s
Comcast	105 Mbit/s	10 Mbit/s
Cablevision Systems	101 Mbit/s	15 Mbit/s
Shaw	100 Mbit/s	5 Mbit/s
Midcontinent	100 Mbit/s	15 Mbit/s
Charter	75 Mbit/s	5 Mbit/s
RCN	75 Mbit/s	10 Mbit/s
Many other MSOs	50-60 Mbit/s	5-10 Mbit/s
AT&T	24 Mbit/s	3 Mbit/s

Table 2. Offered DS Tmax values in Europe [2]



OPERATOR	MARKET	SPEED
Cable Europa (ONO)	Spain	100 Mbit/s
Cabovisão	Portugal	120 Mbit/s
Canal Digital	Norway	100 Mbit/s
Com Hem	Sweden	200 Mbit/s
Get	Norway	200 Mbit/s
Kabel Baden-Württemberg	Germany	100 Mbit/s
Kabel Deutschland	Germany	100 Mbit/s
Numericable	France	100 Mbit/s
Sanoma Television Welho	Finland	200 Mbit/s
Tele Columbus	Germany	100 Mbit/s
Telenet	Belgium	100 Mbit/s
Liberty Global	—	120 Mbit/s
UPC Austria	Austria	100 Mbit/s

Table 3. Offered DS Tmax values in Europe (Continued) [2]



OPERATOR	MARKET	SPEED
UPC Czech Republic	Czech Republic	100 Mbit/s
Unitymedia	Germany	128 Mbit/s
UPC Hungary	Hungary	120 Mbit/s
UPC Ireland	Ireland	100 Mbit/s
UPC Netherlands	Netherlands	120 Mbit/s
UPC Poland	Poland	120 Mbit/s
UPC Romania	Romania	100 Mbit/s
UPC Slovak Republic	Slovak Republic	120 Mbit/s
UPC Cablecom Switzerland	Switzerland	100 Mbit/s
Virgin Media	U.K.	100 Mbit/s
YouSee	Denmark	50 Mbit/s
Ziggo	Netherlands	120 Mbit/s
ZON Multimedia	Portugal	360 Mbit/s

Table 4. Assumptions about spectrum usage used in analyzing the capacity of 5-42MHz spectrum in [3]

Bandwidth	Description
37	Sup-split Upstream spectrum (5-42MHz)
-2	Assumed 2MHz as roll off (40-42MHz) being unusable
-5	Assumed that the noisy spectrum (5-MHz) to be unusable
-2	Legacy STBs
-2	Legacy Status Monitoring
-3.2	3.2MHz channel for legacy QAM16 DOCSIS
22.8	Possible spectrum for DOCSIS3.0 US channel bonding
22.4	Assumed value for capacity analysis

Table 5. Typical Fiber node assumptions used to compare different split options [3]

Item	Value	Unit
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	1000	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

Table 6. Express/distribution segments assumptions used to compare different split options [3]

"Express" (untapped) Segment Characterization	Unit	Sub-Split	Mid-Split	High-Split 200	High-Split 238	Top-Split (900-1050)	Top-Split (900-1125)	Top Split (1250-1550)
Upper Frequency	MHz	42	85	200	238	1050	1125	1550
Typical Maximum Cable Loss (Amp to Amp 70 deg F)	dB	6.5	9.2	14.1	14.8	35.7	36.9	43.3
Additional Gain Required for Thermal Control (0 to 140 deg F)	+/-dB	0.5	0.6	1.0	1.0	2.5	2.6	3.0
Total Reverse Amplifier Gain Required	dB	**6.9**	**9.8**	**15.1**	**15.8**	**38.2**	**39.5**	**46.4**
"Distribution" (tapped) Segment Characterization		Sub-Split	Mid-Split	High-Split 200	High-Split 238	Top-Split (900-1050)	Top-Split (900-1125)	Top Split (1250-1550)
Upper Frequency	MHz	42	85	200	238	1050	1125	1550
Worst Case Path Loss	dB	**27.9**	**28.9**	**33.1**	**33.5**	**63.0**	**68.0**	**69.9**
Path Loss from First Tap	dB	*27.9*	*28.9*	*31.0*	*31.0*	*42.2*	*44.6*	*44.8*
Distribution Cable Loss	dB	0.4	0.6	0.9	0.9	2.1	2.2	2.6
Tap Port Loss	dB	21.9	21.9	22.0	22.0	25.4	27.2	24.5
Drop Cable Loss	dB	2.1	2.9	4.6	4.6	10.1	10.4	12.2
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5
Path Loss from Last Tap	dB	*24.4*	*26.9*	*33.1*	*33.5*	*63.0*	*68.0*	*69.9*
Distribution Cable Loss	dB	4.0	5.7	8.8	9.2	21.2	22.0	25.8
Tap Insertion Loss	dB	7.9	7.9	9.2	9.2	16.7	18.7	17.9
Tap Port Loss	dB	6.9	6.9	7.0	7.0	10.4	12.2	8.5
Drop Cable Loss	dB	2.1	2.9	4.6	4.6	10.1	10.4	12.2
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5

ARCHITECTUAL APPROACHES FOR INTEGRATING SP Wi-Fi IN CABLE MSO NETWORKS

Rajiv Asati, Distinguished Engineer, rajiva@cisco.com
Sangeeta Ramakrishnan, Principal Engineer, rsangeet@cisco.com
Rajesh Pazhyannur, Technical Leader, rpazhyan@cisco.com

Abstract

Cable MSOs have an enticing opportunity with Wi-Fi residential and business services.

In this paper, we discuss the common requirements, challenges (that Cable MSOs face) and necessary architecture (that MSOs could use) for integrating SP Wi-Fi in Cable MSO networks to support both residential and hotspots use-cases. This paper also qualifies various architectural approaches for network transport in the context of DOCSIS access along with the time-to-market perspective, so as to enable MSOs to quickly capitalize on this opportunity.

1. INTRODUCTION

Wi-Fi is a pervasive & proven access technology that is commonly used by Homes and Enterprises around the world, and its usage by Service Providers (SPs) is gaining traction as well. SPs can use Wi-Fi to deliver one or more of the triple-play services (e.g Video, Voice, Data) to the customers indoor and outdoor, and enhance the customer/user experience (by allowing mobile consumption of content as well as access to data).

In fact, SPs, particularly, Mobile SPs have been leveraging Wi-Fi for better cost-efficiency and QoE. As the number of mobile devices keeps growing exponentially, it is expected that the Mobile network traffic would keep growing exponentially as well (studies have predicted a 18-fold increase in mobile data traffic in the next 5 years, as illustrated in Figure 1).

There will be 5B mobile devices and 2B M2M nodes
Global mobile traffic wil
grow 26X to 6.3 EB/mo
Video will be 66% of all mobile traffic by 2015
Source: Cisco Visual Networking Index (VNI) Global Mobile Data Forecast, 201
© 2011 Cisco and/or its affiliates. All rights reserved.
Cisco Cor

Unfortunately, most mobile SPs do not have enough licensed radio spectrum to accommodate this increase. Given that a large amount of traffic is consumed indoors (in homes, offices, public-spaces like hotels, café's, etc), where Wi-Fi connectivity is much more widely available than cellular, the usage & focus on Wi-Fi to offload traffic from cellular networks has greatly increased. In fact, 'Mobile Data Offload & Onload Video Whitepaper (published by Juniper Research in April 2011) predicts that Wi-Fi usage for mobile traffic offload could exceed ~1EB / month by 2015. This is illustrated in Figure 2.

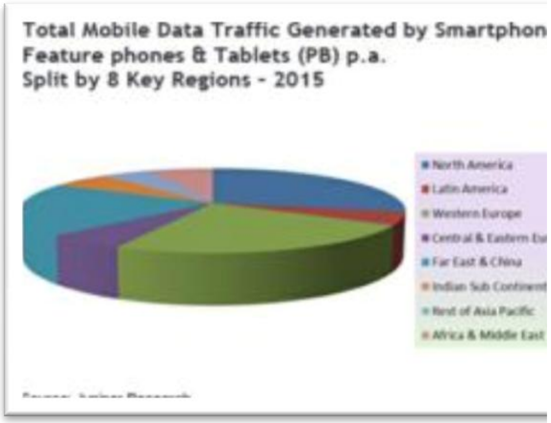


Figure 2 Mobile Traffic offload Prediction

Needless to say, Mobile SPs would need to acquire sites for installing Wi-Fi based macro-cells, and hence, mobile SPs are increasingly motivated to rely on other SPs/Providers offering the Wi-Fi based solutions.

Cable MSOs have a fantastic opportunity with Wi-Fi. In this paper, we discuss the common challenges (that Cable MSOs face) and necessary architecture (that MSOs could use) for integrating SP Wi-Fi in Cable MSO networks to support both residential and hotspots use-cases. We also qualify various architectural approaches for network transport in the context of DOCSIS access along with time-to-market perspective, so as to enable MSOs to quickly capitalize on this opportunity.

2. SP Wi-Fi: MSO REQUIREMENTS / CHALLENGES

SP Wi-Fi primarily refers to an 802.11 Wi-Fi system deployed and managed by a Service Provider (SP) for public access (aka community access) to its network for services such as High Speed Data Internet service. Public Access means that Wi-Fi is available to the customers of the SP and/or partner SPs and/or any customers. SPs may provide managed (and sometimes hosted) Wi-Fi services to other service providers (e.g. Mobile SPs).

SP Wi-Fi differs from general Wi-Fi e.g. Enterprise Wi-Fi (or Residential Wi-Fi) in three key aspects:

1. **Scale** – The number of APs and user clients tends to be very large – thousands to millions.
2. **Carrier Grade** – The high-availability and manageability aspects tends to be of carrier class (e.g. 5 9's)
3. **Multi-Vendor** – The existence of multiple vendor devices is expected – warranting the usage of standards based end-to-end architecture.

2.1 Use-Cases

SP WiFi architecture should be flexible enough to enable Cable MSO to serve one or more the following deployment use-cases:

1. **Residential** (Indoor) –re-use the Wi-Fi APs that are integrated with the (SP managed) residential gateways to provide public access Wi-Fi. In this case, the AP is located indoor (in a residential customer home).
2. **Metro** (Outdoor) –deploy Wi-Fi APs outdoor in public places to provide public access Wi-Fi. In this case, the APs are typically mounted on aerial cable strands, street-poles, roof-tops etc.
3. **HotSpot / SMB** (Indoor) –re-use the managed Wi-Fi service to SMBs such as coffee shops, bookstores, retail-stores etc., having 10s or 100s of employees, for both private and public access WiFi.
4. **HotSpot** (Outdoor) –deploy large concentration of APs in a relatively small area such as stadium, amphitheaters, parks etc. having large number of users in that area. The APs are usually located outdoor to offer public access Wi-Fi.

5. **Wholesale / offload** – allow partners' customers to access the Wi-Fi services, and/or backhaul mobile operators' customers traffic over the MSO infrastructure. In this case, the APs are located indoor and outdoor.

2.2 Access Point / 802.11 Radio

Access Point (AP) is the most fundamental element in the SP WiFi architecture. Hence, the AP requirements must be carefully assessed. The following are some of the key considerations for the Wi-Fi AP:

1. Coverage: refers to AP's range to = what throughput upto what distance. Coverage determines the number of APs required to cover a certain area. Naturally, 802.11n radio on AP is preferred for optimal coverage.
2. Capacity: refers to the maximum number of clients that AP can concurrently support/associate. Some prefer to define capacity in terms of maximum number of active users that can be supported with each user guaranteed a minimum throughput. Capacity directly influences the number of APs required to cover a certain area (e.g. the number of APs are determined by capacity requirements rather than coverage).
3. Interference Management: refers to AP's capability to continuously select the best radio channel (through constant monitoring since startup) while managing the radio interference so as to get the best radio performance. The interference could be generated by other Wi-Fi APs or by non Wi-Fi sources such as Bluetooth, DECT phones, Microwave etc. Naturally, techniques such as Beamforming to improve the signal strength received by the client, interference identification for reporting etc. become important.
4. Dual radio– refers to AP supporting simultaneous usage of 2.4GHz and 5GHz. This is particularly important for APs that are used for creating private and public WLANs. This should be controllable by the MSOs.

2.3 Security

Security is one of the most-pressing issues, as security threats such as snooping, Eavesdropping, session hi-jacking, session side-jacking, evil twin attack etc. expose the insecurity in WiFi networks that rely on open SSID. Hence, it is important to have secure SSID/WLAN.

Note that most SP Wi-Fi deployments have not used secured SSID because of lack of support on clients for EAP methods and/or complexity in distributing and managing user-security credentials. Hopefully, this will change with Hotspot2.0 recommendations. Please see more details on this here [Hotspot2.0].

Additionally, in case of residential SP WiFi, the AP must support at least one private WLAN/SSID for the residential customer's usage, and at least one public WLAN/SSID for public usage, for security reasons.

In summary, SP WiFi architecture should include user authentication and cryptography (e.g. WPA-2 Enterprise), as well as separate control and management of public and private WLANs so as to pave the way for 'Secure WLANs'.

2.4 Inter-Operator Roaming

It would be desirable to let the users use other MSOs' or SPs' Wi-Fi networks to get one or more services (such as high speed data connectivity to the Internet) when the users are roaming [Wi-Fi-Roam]. However, how would the customer's device know the right SSID (assuming more than one SSIDs) on the partner Wi-Fi network? If the users knew the right SSID, they may have to manually login and get authenticated so as to use partner Wi-Fi network. This is deemed not only inconvenient to the user, but also as a lost opportunity for the MSOs to influence users' network selection.

Once authenticated, then depending on the mobility requirement, home network or the partner network should assign the IP address to the user client device. If the roaming users managed to use partner Wi-Fi network, then they may get limited time before they are asked to re-authenticate, causing them another source of inconvenience. Lastly, as MSOs allow the roaming users, appropriate billing ruleset, Lawful Intercept etc. have to be enforced. Of course, this all assumes the MSOs to have struck the roaming agreements with other MSOs & SPs.
To address this challenge, IEEE 802.11u could be necessitated. Please see more details on this here [Hotspot2.0].

2.5 Mobility

Mobility is defined in many different ways, resulting in many different requirements. However, MSOs may not find all the mobility requirements to be important and/or applicable. A brief summary of mobility requirements is provided below:

- Fast Roaming: enables AP-to-AP handover user re-authenticate the user. Specifically, the re-association procedures are performed in parallel with key negotiation procedures, as per IEEE 802.1r.
- Micro-Mobility: In deployments with a small number of APs in a site (such as bookstore, restaurant) there is need to support mobility to reduce adverse impact on end user experience as they roam within the site. In most scenarios, when user walks out of the site, they will lose Wi-Fi coverage. Reconnecting to Wi-Fi in another location/site would typically result in users getting a different IP address.
- Macro-Mobility: In deployments where there is large contiguous area covered by Wi-Fi (such as outdoor APs) there is need for end users to maintain IP address as they roam between Wi-Fi APs. In such cases, the solution may need tunnels between centralized Wi-Fi aggregators (WLC, CMTS, MAG, etc) to provide this form of mobility
- Inter-Vendor Mobility: As mentioned earlier, SP Wi-Fi deployments tend to comprise network elements e.g. APs from different vendors, hence, it is important to ensure that mobility works between different vendors' APs. Further, in some scenarios, the vendors may provide overlapping Wi-Fi coverage.
- Inter-Technology Mobility: A significant portion of Wi-Fi devices are likely to have a cellular (3G/4G radio) as well. In some cases, it may be desirable to provide mobility as users roam between radio-technologies (between Wi-Fi and Cellular). Such mobility can be provided by using client based mobility mechanisms (Mobile IP, DSMIPv6) or network based mobility mechanisms (such as PMIPv6)..

While many of the above requirements may be reasonable, it is worth noting that continuous Wi-Fi coverage is a prerequisite of any form of mobility. Hence, mobility may not be possible everywhere or applicable, requiring careful justification.

2.6 Traffic Separation

As SP WiFi traffic is transported over the MSOs network infrastructure, traffic separation capabilities in the network especially on the access (e.g. DOCSIS) side will become critical.

2.6.1 Separation of HSD subscriber's traffic from SP Wi-Fi traffic

Most operators have bandwidth caps and tiers of service deployed whereby each subscribers' traffic is separately measured (for bandwidth cap purposes) and QoS is applied to ensure the traffic complies to the tier of service the user has subscribed to (example, 6Mbps down, 1Mbps up). Once the cable modem deployed at a business or home, is enabled for SP Wi-Fi, operators will want to ensure that the SP Wi-Fi users' traffic does not count towards the HSD subscriber's limits. Given that in DOCSIS the Service Flow is the unit on which accounting and QoS is applied, the architecture needs to ensure that the SP Wi-Fi traffic is mapped to a different service flow than that of the subscriber's HSD service flow. This mapping needs to be done both in the Upstream and Downstream directions.

An implicit challenge here is that needing specific US and DS classifiers may result in having unique CM config file for each modem. The chosen architecture must address this challenge.

2.6.2 Separation of Services per Fiber Node

The previous section discussed the separation of a single HSD subscriber's traffic from the SP Wi-Fi users attached to the same CM/AP. Additionally operators may want to ensure that a certain amount of bandwidth is set aside for HSD use versus SP Wi-Fi use across the entire Service Group. This would ensure that one service on an aggregate doesn't crowd out the other service on a Service Group. It would also be beneficial if any unused bandwidth provisioned for one service was made available for the other service to use as needed.

DOCSIS provides the Bonding Group construct which can be used to provide such a service separation between the two services. By using overlapping bonding groups across a set of RF channels, and steering HSD service flows to one Bonding Group and the SP Wi-Fi service flows to the other bonding group, operators can achieve such separation. Depending on how much bandwidth an operator wishes to set aside for each service, they can configure the bonding groups appropriately to achieve their goals.

2.7 Network Transport

SP WiFi services may need to be deployed over various types of access networks e.g. DOCSIS/HFC, EPON/Fiber etc. that are present in MSO networks. For example some operators are considering offering business services over EPON. The overall architecture chosen for deployment will need to be such that they are easily deployable across different access technologies. Hence the Access Point itself will need to support various backhaul technologies such as DOCSIS, EPON etc.

For utmost cost-effectiveness, it would be desirable to leverage the IP or MPLS (or 802.1 based carrier Ethernet) network transport that is already used by MSOs for other services. In fact, many MSOs have converged their networks (or on the path to do so) and been using MPLS technology for various services. The key is to choose the network transport that yields the simplification of SP WiFi architecture while satisfying other SP WiFi requirements that are important to the MSO.

2.8 Provisioning & Management

In particular, the WiFi APs should be automatically configured without needing any manual intervention for utmost cost-effectiveness (given the expected scale).

Thankfully, both DOCSIS cable modem and eDOCSIS[1] device already allows auto-

[1] An eDOCSIS device consists of an embedded DOCSIS cable modem (eCM) and one or more embedded Service/Application Functional Entities (eSAFEs) such as eAP, eRouter, eSTB, eMTA etc. There are already various vendors' eDOCSIS devices

configuration of cable modem (and DPOE allows auto-configuration of ONU) and integrated AP. Moreover, eDOCSIS device, by definition, has a single software image for the entire device.

However, if the chosen SP WiFi architecture requires each modem to rely on a unique config file, then it could become a provisioning challenge (as MSOs generally use a few cable modem config files across tens of thousands or millions of modems. This challenge can be solved if template based cable modem config file generation method is used.

For residential SP Wi-Fi deployments in particular the number of APs may well be as high as the number of deployed cable modems. Hence being able to provision at scale is critically important.

Needless to say that CMTS provisioning should not be needed on a per modem basis.

In summary, seamless integration of the WiFi provisioning (e.g. AP provisioning) into the existing provisioning infrastructure is going to be required for possible auto-provisioning of APs.

2.9 Subscriber Management

Like other services, SP WiFi services will also require subscriber management. This may include capabilities such as bandwidth accounting, quality of service, legal intercept etc. Such services will require a policy enforcement engine that is subscriber aware and learns the policies to be applied from a policy management system. All SP WiFi traffic will have to be routed through such a policy enforcement engine in order to provide the above-mentioned subscriber services.

Subscriber management could occur centrally in which case all traffic needs to be routed to the Subscriber Management Gateway. Different options are available to achieve this, and are discussed in more detail in the Transport Network section 4.1.

It is worth noting that for HSD services, such subscriber management capabilities are applied at the CMTS, hence no requirements to route HSD traffic to any other central entity really exist in MSO networks.

2.10 IPv6

Given the IPv4 address exhaustion becoming a reality for many MSOs & SPs sooner or later and given that SP Wi-Fi would involve 10,000s of APs and millions of users, it is imperative to have IPv6 in SP Wi-Fi usage from day 1. This means that IPv6 should be used not only for addressing users, but also for the underlying infrastructure (e.g. APs, CMTSs, PEs, etc.) irrespective of any IP tunneling is used or not. In other words, both user and AP addressing should be done using IPv6.

While using IPv4 is an option, MSOs would end up requiring many more bandaids (e.g. Carrier Grade NATs) to make it work in a large-scale environment, thereby negatively impacting CAPEX and OPEX associated with SP Wi-Fi.

2.10 Monetization

Once the basic SP Wi-Fi services (e.g. high speed data) get rolled out for the purposes such as customer retention, MSOs may increase the focus on monetization. This would require the architecture to be flexible enough to allow intelligent network to help with advanced services such as advertising, remote monitoring/security etc.

3. SP Wi-Fi ARCHITECTURE

(including 802.11n Wi-Fi Cable Gateway devices [Wi-Fi-GW]) in MSO deployments.

The SP Wi-Fi architecture needs to be flexible enough to satisfy some or all of the requirements (described in section 3) in an incremental & modular way. Such a flexibility would be an important trait to MSOs, since not every MSO would deem every requirements applicable to them day 1.

The SP Wi-Fi architecture needs to be flexible enough to satisfy some or all of the requirements (described in section 3) in an incremental & modular way. Such a flexibility would be an important trait to MSOs, since not every MSO would deem every requirements applicable to them day 1.

This section provides a simplified overview of SP Wi-Fi architecture, and focuses on the architectural approaches for transporting SP Wi-Fi traffic through the transport network while hinting at their flexibility. The Figure 3 below illustrates a high-level SP Wi-Fi architecture:

A SP Wi-Fi architecture illustrated above contains one or more of the following elements:

1. Wi-Fi Access Points: The Wi-Fi Access Points may be either embedded with a cable modem (as in outdoor or residential) i.e. eDOCSIS device (also referred to as Cable Wi-Fi Gateway) or deployed separately from the cable modem (as in many indoor hotspots).

2. Access Network: This is the DOCSIS based HFC network (or EPON or Ethernet based Fiber network) comprising CMTS or CCAP, Fiber Nodes, and CMs (or ONUs) providing network connectivity to/from the AP. The CMTS terminates DOCSIS connections from the cable modems as well as connects to the metro/aggregation Network.

3. Metro/Aggregation Network: The

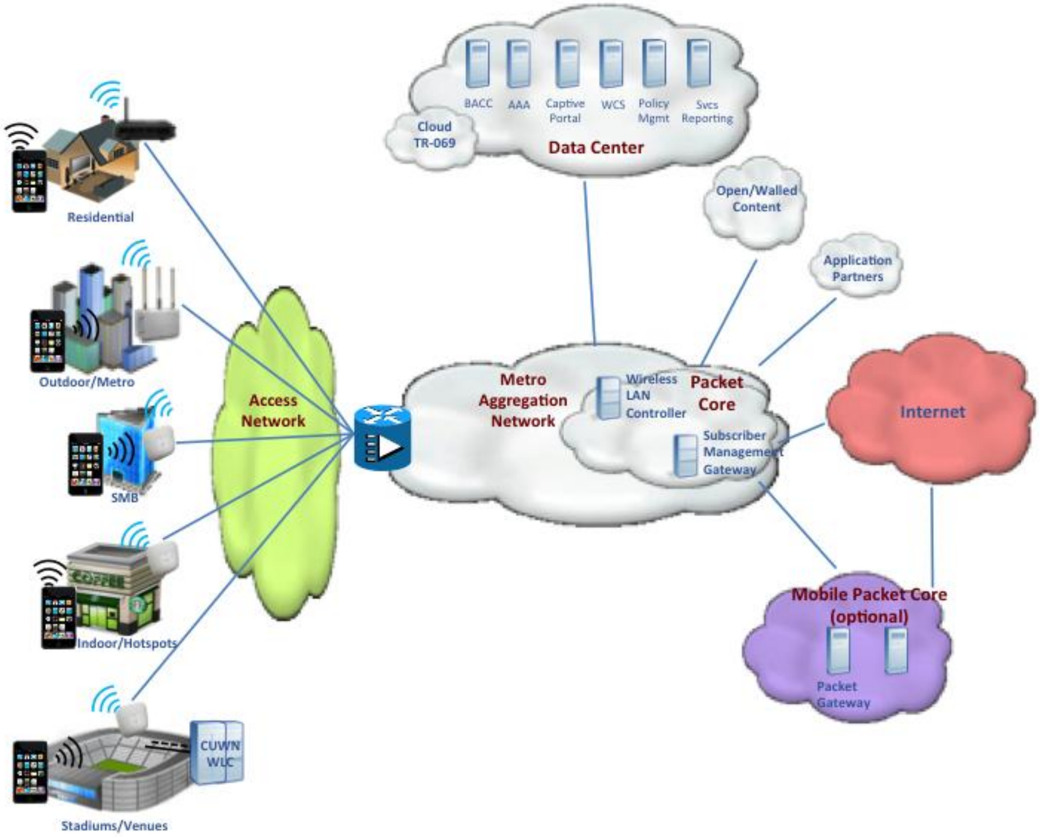


Figure 3 SP Wi-Fi Architecture (simplified)

network that CMTS uses to ultimately connect the users to the internet or partner networks or the open/walled-garden content. There may also be a regional and/or backbone network (not shown in the figure) between the metro network and internet. Metro network is usually an IP or IP/MPLS network (or sometimes a layer2 Ethernet/bridged network).

4. Wireless LAN Controller (WLC): The WLC is a centralized point of control and management of Wi-Fi APs using CAPWAP protocol (IETF RFC 5415). It tunnels data plane (user) traffic to/from the AP using the CAPWAP data plane tunnel (Please see section 3.1.1.2). It is part of the Wi-Fi packet core. It is worth pointing out that not all Wi-Fi APs are based on CAPWAP. Specifically, residential APs (i.e. eDOCSIS device) are not based on CAPWAP. This is better illustrated in the next section.

5. Subscriber Management Gateway: The Subscriber Management Gateway (dubbed as the centralized entity in this paper) is an IP point of attachment that functions as a Policy Enforcement Point (PEP). Specifically, the gateway is responsible to maintain user awareness and enforce of the relevant QoS settings, bandwidth limits, accounting, DPI, etc. The gateway is also referred to as Intelligent Services Gateway (ISG). It is part of the Wi-Fi packet core.

   It is worth pointing out that the Subscriber Management Gateway function could be implemented on the CMTS.

6. Data Center: The Service Network containing elements such as BAC, AAA, DNS, DHCP, Policy Servers and OSS/BSS elements providing network management and service management

7. Mobile Packet Core: This is optional, but it is needed for ensuring inter-technology (3G to Wi-Fi, say) or inter-domain mobility. This includes 3GPP specific elements such as PDN Gateway etc. pertaining to cellular network.

3.1 Network Transport Architecture

Wi-Fi AP connects wireless user devices to each other and/or to a wired network. In general[2], Wi-Fi AP is a layer2 bridge device that bridges Wi-Fi user devices' Ethernet frames between 802.11 wireless network (WLAN) and wired network (LAN). (One could relate AP to a Cable Modem, which is also a layer2 bridge device, but it bridges wired user devices' Ethernet frames between Ethernet network (LAN) and DOCSIS network).

> Due to subscriber management requirements described in section 2.9, the traffic from the Wi-Fi Access Points will need to be routed to a centralized entity located on the wired network for subscriber management. The subscriber management capability may reside on the WLC, ISG, MAG or even the CMTS depending on the chosen architecture.

This means that the Wi-Fi user device must have layer2 connectivity upto that centralized entity through the AP, even if AP and the centralized entity are multiple hops away from each other and reachable via the underlying network. If the underlying network

[2] A non-bridging AP will allow the association of wireless user clients, but will not allow connecting to a wired network.

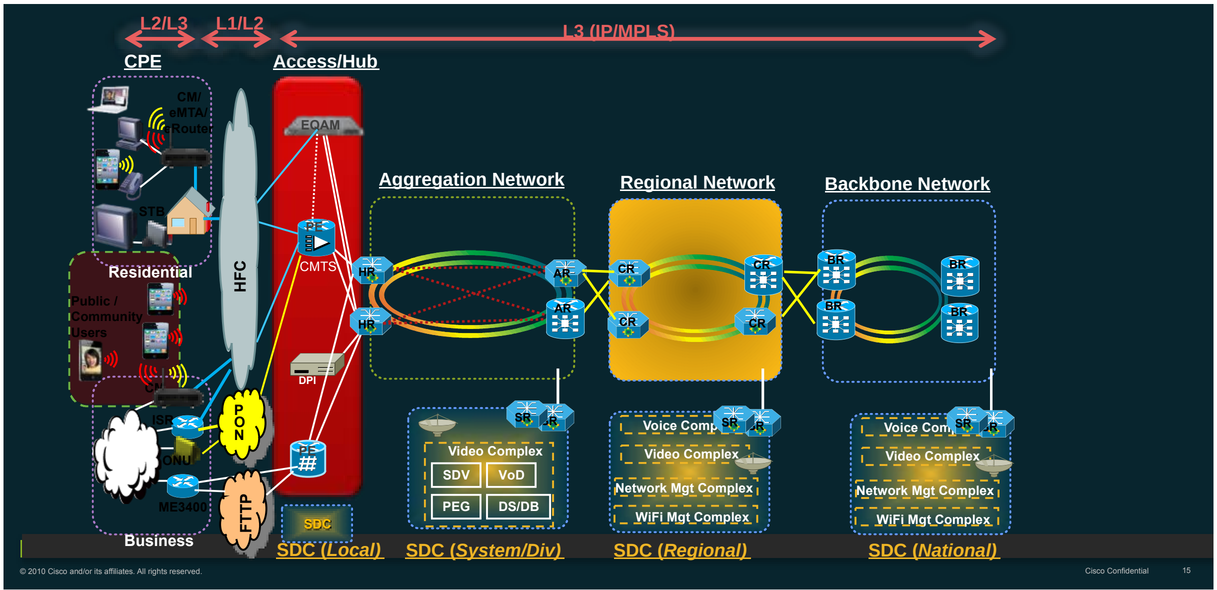
is a layer2 network (i.e. Ethernet bridged network), then it is relatively simpler to ensure the needed layer2 connectivity between user devices and the centralized entity. However, if the underlying network is a layer3 network (e.g. IP or IP/MPLS network comprising routers), then it can get complicated, depending one the chosen architectural approach (there are number of architecture approaches, as discussed later in this section).

Before we discuss various architectural approaches, it is important to put the MSO network in the perspective. The underlying network in the context of a cable MSOs is commonly a layer3 network in which CMTS (or CCAP) presents itself as the layer3 next-hop (as well as layer2 next-hop) to the user devices behind the standalone modems (e.g. CM, ONU) or embedded modems [eDOCSIS] (i.e. eCM) acting as the bridge.

the underlying network infrastructure must facilitate the bidirectional connectivity between the Wi-Fi user device and the centralized entity acting as the first IP next-hop, wherever that entity is located. This can be done in number of ways, based on the chosen architecture and requirements.

This section discusses such architectural options while keeping Cable MSOs' network infrastructure in mind. While this section focuses on DOCSIS access, it is well applicable to EPON access given the DPoE relevance. The following network transport architectural approaches are qualified for backhauling SP Wi-Fi traffic:

1. IP tunneling from AP
2. BSoD L2VPN
3. BSoD L3VPN



A reference Cable MSO network high-level diagram (not showing SP Wi-Fi elements) is shown in `Figure 4`.

As discussed earlier, if the underlying network is a layer3 network (e.g. IP or IP/MPLS network comprising routers), then

While each of the above architectural approaches are described in detail in the subsequent sections, the `Figure 5` below briefly illustrates them with their data plane specifics and how they relate to one of key AP capabilities:

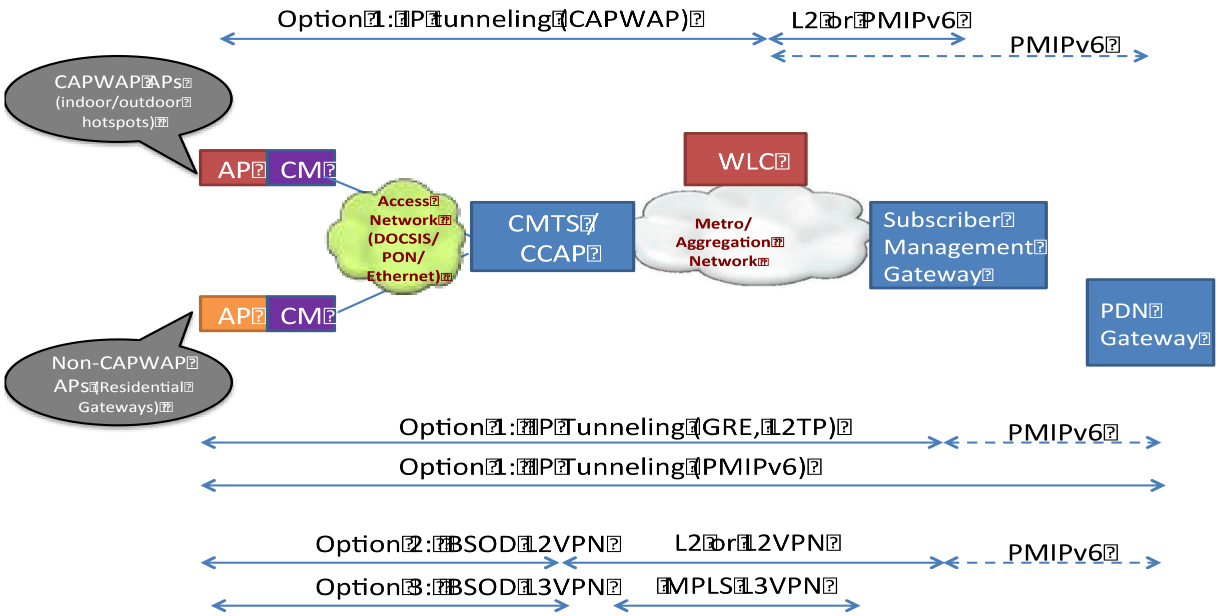

Option 1: IP tunneling (CAPWAP)
L2 or PMIPv6
PMIPv6
CAPWAP APs (indoor/outdoor hotspots)
AP
CM
WLC
Access Network (DOCSIS/ PON/ Ethernet)
CMTS/ CCAP
Metro/ Aggregation Network
Subscriber Management Gateway
AP
CM
PDN Gateway
Non-CAPWAP APs (Residential Gateways)
Option 1: IP Tunneling (GRE, L2TP)
PMIPv6
Option 1: IP Tunneling (PMIPv6)
Option 2: BSOD L2VPN
L2 or L2VPN
PMIPv6
Option 3: BSOD L3VPN
MPLS L3VPN

Figure 5 Network Transport Architectural Approaches - Data Plane

- CAPWAP APs: The traffic to/from AP is IP tunneled to the WLC using CAPWAP. The traffic between the WLC and Subscriber Gateway is a L2/802.1Q. PMIPv6 usage is optional.
- Non-CAPWAP APs: The traffic to/from AP is either tunneled over the network (option 1) or forwarded natively (option 2 or 3). PMIPv6 usage is optional.

The next section discusses each of the above network transport architectural options in details.

3.1.1 IP Tunneling from AP

There are number of options within this particular architectural approach that leverages IP tunneling from AP itself so as to tunnel the Wi-Fi traffic (either at layer2 or layer3) through the network.

3.1.1.1 PMIPv6

The architectural approach here is to build an over-the-top IP tunnel between AP and a remotely located centralized entity, using GRE over IP. In this approach, the data plane comprises “IPv4|v6 over GRE over IPv4|v6 over Ethernet [over DOCSIS (or PON)]” in the last-mile access and “IPv4|v6 over GRE over IPv4|v6” (over MPLS, if existed) in rest of the network (upto that centralized entity).

PMIPv6 is well standardized at the IETF [RFC5213] and [RFC5844]. PMIPv6 involves Mobility Access Gateway (MAG) and Local Mobility Anchor (LMA). LMA is defined to be the topological anchor point i.e. home agent for the Mobile Node’s (e.g. Wi-Fi user device's) IP prefix(es) and manages MN’s binding state via MAG. MAG manages mobility-related signaling for the MN that is attached to its access link. It is responsible for tracking the MN's movements to and from the access link and for signaling to the LMA.

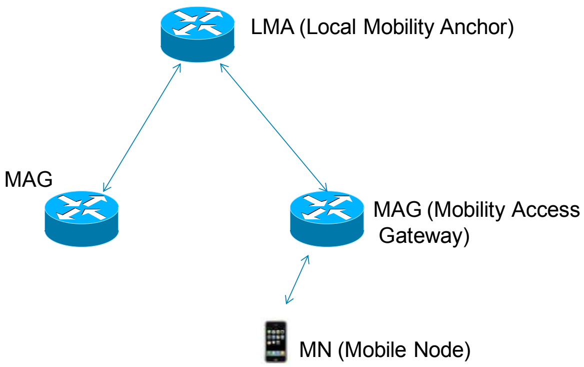
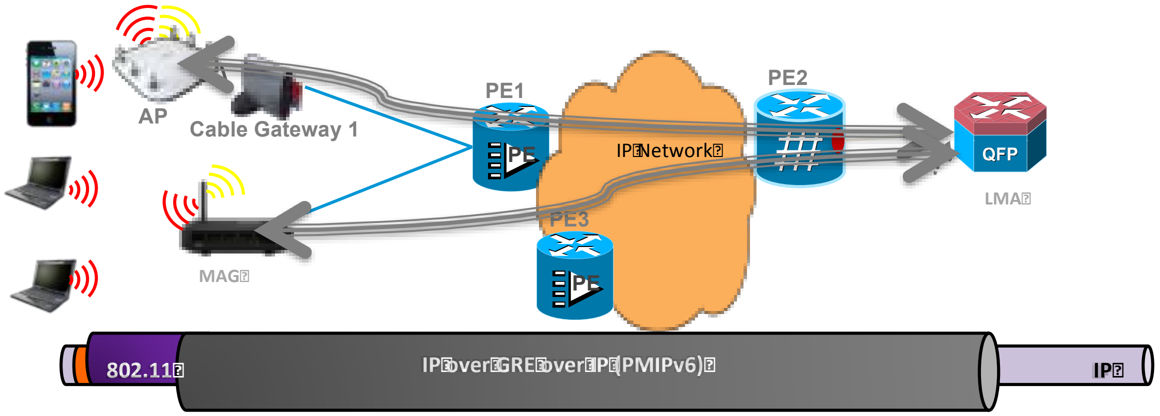


Figure 6 PMIPv6 Components

Error! Reference source not found. above illustrates PMIPv6 components.

While GRE over IP is commonly used tunnel mode, PMIPv6 also allows for other tunnel modes such as 'Ethernet over IPv6 over IPv6', Ethernet over UDP over IPv4 etc.



PMIPv6 is the only protocol that is claimed to qualify SP WiFi (with 802.1x/EAP) as the 'trusted non-3GPP access' and ensure mobility in every scenario.

Using PMIPv6 based architectural approach, an AP (acting as the MAG) uses PMIPv6 protocol messages to inform the LMA about the Wi-Fi user device (e.g. Mobile Node) getting attached. This allows AP/MAG and LMA to install (or update) the corresponding forwarding entries for the IP address assigned to the Wi-Fi user device. AP/MAG terminates user's layer2 and sends/receives user's IP traffic over the PMIPv6 tunnel. In other words, AP/MAG acts as the IP next-hop/gateway for the Wi-Fi user. While the Wi-Fi user is connected to AP/MAG at layer2, its IP address is anchored the LMA. This allows IP mobility, when the Wi-Fi user roams and changes AP/MAG attachments.

Figure 7 illustrates PMIPv6 tunneling applicability for SP Wi-Fi in sample MSO network topology.
It is important to highlight that instead of enabling PMIPv6 (MAG function) at the AP (as shown in this particular approach), it can

be instead enabled on ISG, WLC or CMTS (as shown in other architectural approaches e.g. BSoD L2VPN) in an incremental manner for mobility.

The advantages of this approach are – (a) scales extremely well, (b) provides IP mobility for all scenarios, (c) integrates with 3GPP based cellular network

The disadvantages of this approach are – (a) requires MAG function as well as user management/control on AP/Modem – increased complexity on residential modems/gateways, (b) requires unique config file per modem for DS classification, (c) subjected to fragmentation and reassembly on last-mile access, (d) prohibits 5-tuple classification for QoS in the network, (e) results in sub-optimal multicast replication (e.g. network capacity wastage) if multiple user devices consume the multicast content

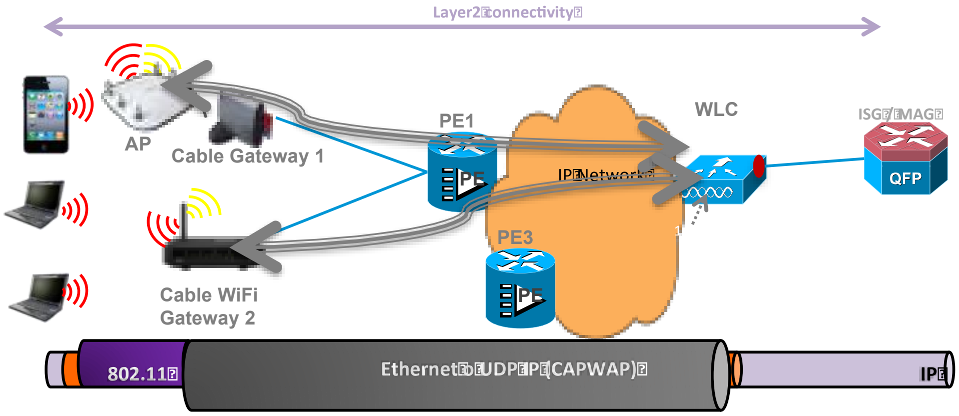
3.1.1.2 CAPWAP

The architectural approach here is to deliver the WiFi 802.11 traffic to a remotely located centralized entity e.g. Wireless LAN Controller (WLC), using UDP over IP. In this approach, the data plane comprises users' "Ethernet over UDP over IPv4|v6 over Ethernet [over DOCSIS (or PON)]" in the last-mile access and "Ethernet over UDP over IPv4|v6" (over MPLS, if existed) in rest of the network (upto WLC). UDP port is a well-known port 5247.

CAPWAP is well standardized at the IETF [RFC5415] and [RFC5416].

CAPWAP is a de facto protocol for Control and Provisioning of APs, and extensively used in most SP Wi-Fi deployments use-cases.

Figure 8 illustrates CAPWAP tunneling applicability for SP Wi-Fi in sample MSO network topology:



Using this approach, an AP establishes a CAPWAP tunnel (i.e. UDP over IP tunnel) with WLC (e.g. centralized entity). The 802.11 frames sent by the user device are forwarded by AP over the CAPWAP tunnel to WLC, which decapsulates the CAPWAP header and forwards the user device' IP packet using IP forwarding lookup. If the IP destination of the packet is another WiFi user device, then the IP packet is encapsulated in the 802.11 header and placed on the CAPWAP tunnel towards the appropriate AP. If the IP destination of the packet is on the wired network, then the IP packet is forwarded as usual.

The returning traffic gets subjected to the IP forwarding lookup, and gets placed on the appropriate CAPWAP tunnel, which is terminated at the AP. AP then delivers the 802.11 frames to the WiFi user device.

CAPWAP provides fragmentation and reassembly as per the path MTU discovery done by both AP and WLC, and allows for optional encryption using DTLS. CAPWAP also allows for PMIPv6 integration, as/if/when desired. This means that PMIPv6 elements (e.g. MAG and LMA) can incrementally be introduced, in which the MAG function can be enabled at the WLC.

The advantages of this approach are – (a) provides network administrators with a structured and hierarchical model to control & configure the APs, (b) controls hand-offs between AP during user roaming = foundation for mobility (c) works with layer2 or layer3 network, (d) allows 802.11 link-layer control, (e) works with NAT

The disadvantages of this approach are – (a) CAPWAP is not deemed useful for the residential APs, (b) network capacity wastage due to unnecessary multicast replication at WLC may happen if multiple user devices consume the multicast content

3.1.1.3 GRE

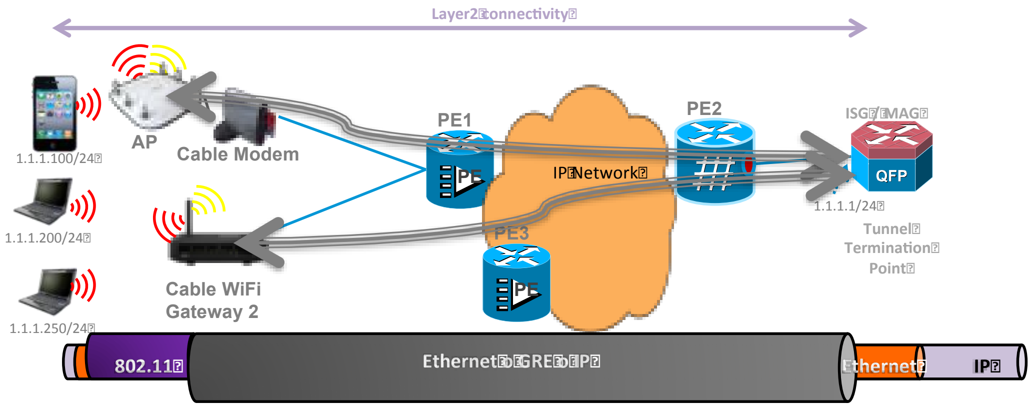
The architectural approach here is to build an over-the-top IP tunnel to deliver the Wi-Fi user device's Ethernet traffic between AP and a remotely located centralized entity (i.e. tunnel termination entity), using GRE. This approach requires IP connectivity between AP and the centralized entity. In this approach, the data plane comprises users' "Ethernet over GRE over IPv4|v6 over Ethernet [over DOCSIS (or PON)]" in the last-mile access and "Ethernet over GRE over IPv4|v6" (over MPLS, if existed) in rest of the network (upto that centralized entity).

> While Ethernet over GRE over IP usage is not well known or used, it is standardized at the IETF [RFC1771].

`Figure 9` illustrates GRE tunneling applicability for SP Wi-Fi in sample MSO network topology.

Using this approach, an AP establishes a GRE tunnel with the remote L2TP tunnel concentrator (e.g. centralized entity) and sends/receives Wi-Fi user device's Ethernet frames, over GRE (over IP) tunnel. It is important to note that GRE doesn't require a control channel and can be set up in a stateless manner without requiring any tunnel configuration.

The advantages of this approach are – (a)



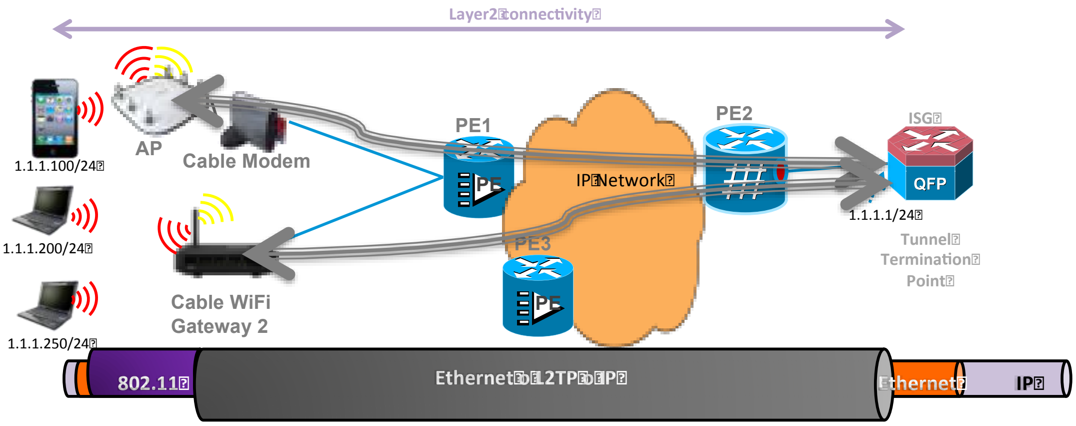
maintains simplicity on AP or Gateways (b) scales well (if stateless tunneling is used, (c) maintains subscriber management/control at the remotely located centralized entity (e.g. tunnel termination point) based on IP, (d) provides IP mobility natively within the Layer 2 domain, (e) can integrate with PMIPv6 (by having the MAG function on the tunnel termination point) to provide macro-mobility.

The disadvantages of this approach are – (a) does not integrate with 3GPP and doesn't provide mobility in all scenarios, (b) requires unique config file per modem for DS classification, (c) relies on IP tunneling, (d) subjected to fragmentation and reassembly on last-mile access, (e) prohibits 5-tuple classification for QoS in the network, (f) results in sub-optimal multicast replication (e.g. network capacity wastage) if multiple user devices consume the multicast content

3.1.1.4 L2TP

The architectural approach here is to build an over-the-top Layer 2 circuit (over IP network) to deliver the Wi-Fi traffic (e.g. Ethernet frames) between AP and a remotely located centralized entity (i.e. tunnel termination entity), using Layer 2 Tunneling Protocol (L2TP). This approach requires IP connectivity between AP and the centralized entity. In this approach, the data plane comprises users' "Ethernet over L2TP over IPv4|v6 over Ethernet [over DOCSIS (or PON)]" in the last-mile access and "Ethernet over L2TP over IPv4|v6" (over MPLS, if existed) in rest of the network (upto that centralized entity).

L2TPv2 is standardized at the IETF [RFC2661], whereas L2TPv3 is standardized at the IETF [RFC3931]. Figure 10 illustrates L2TP tunneling applicability for SP Wi-Fi in sample MSO network topology.



Using this approach, an AP establishes an L2TP tunnel with the remote L2TP tunnel concentrator (e.g. centralized entity) and sends/receives Wi-Fi user device's Ethernet frames, over L2TP (over IP) tunnel. It is important to note that L2TP requires a control channel to establish the tunnel.

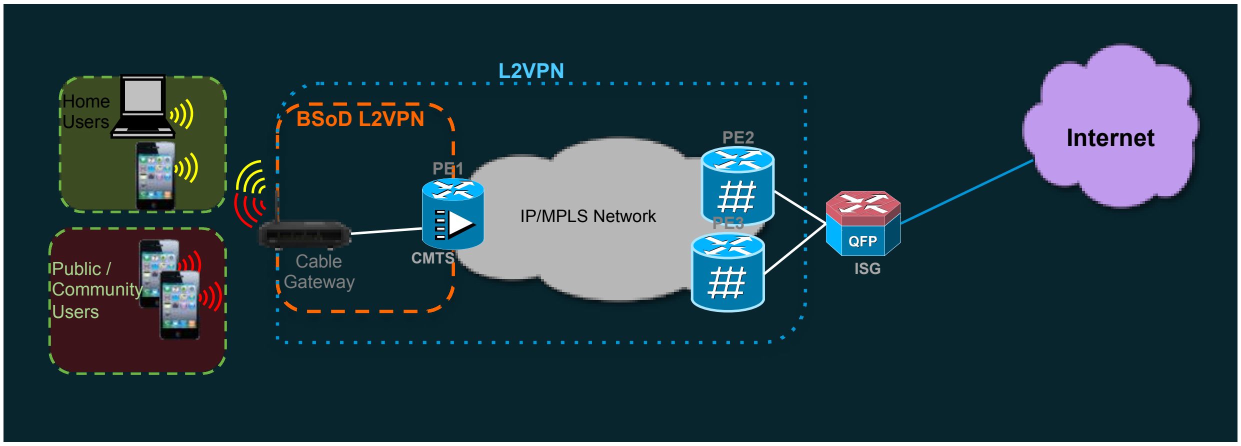
This architectural approach allows for PMIPv6 integration, as/if/when desired by the MSO to achieve mobility between Wi-Fi and Wi-Fi as well as cellular and Wi-Fi. This means that PMIPv6 elements (e.g. MAG and LMA) can incrementally be introduced in the

MSO network, in which the MAG function can be enabled at the L2TP tunnel concentrator.

The advantages of this approach are – (a) has its own control channel, (b) can make use of a cookie for added security

The disadvantages of this approach are – (a) does not scale (beyond few thousand tunnels), (b) requires unique config file per CM for proper DS classification, (c) does not integrate with 3GPP and doesn't provide mobility in all scenarios by itself,

3.1.2 BSoD L2VPN



The idea in this architectural approach is very simple – use Layer 2 VPN to deliver the Wi-Fi traffic to a remotely located centralized entity at layer2 (without requiring any IP lookup). In this approach, the data plane comprises Ethernet [over DOCSIS (or PON)] in the last-mile access and Ethernet over MPLS (or just Ethernet) in rest of the network (upto the centralized entity).

Thankfully, Layer 2 VPN is a well known and well used option in many MSO deployments already, given that CableLabs standardized the Layer 2 VPN over DOCSIS in form of BSoD L2VPN [BSODL2VPN] and enabled many MSOs to use Layer 2 VPN to serve business customers with Metro Ethernet services (e.g. MEF (E-LINE, E-LAN, E-TREE), TLS etc.) when the VPN sites are attached to the HFC access. It is becoming quite useful for other purposes such as traffic separation for different services.

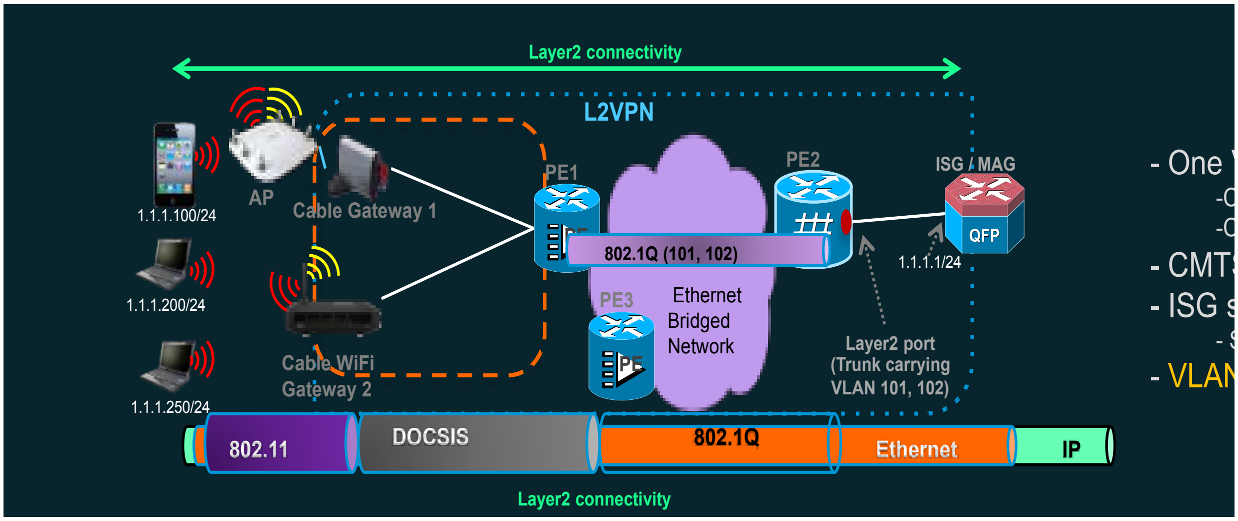
`Figure 11` illustrates L2VPN applicability in sample MSO network topology. It is important to note that the service-flows used for SP Wi-Fi (e.g. Public/Community users) are different from the ones used by the residential users. This automatically allows for traffic separation and IP prefix/address assignment separation between SP Wi-Fi users and residential users (throughout the network).

Using BSoD L2VPN, a CM is able to classify the upstream traffic (received from the AP) using SSID (in case of embedded CM) or VLAN (in case of standalone CM) present in the Ethernet frames, and forward the traffic over a particular DOCSIS service-flow (e.g. impose DOCSIS Header on the received Ethernet frame) to the CMTS. A CM is also able to forward the downstream traffic (received from the CMTS on a particular DOCSIS service-flow) to the AP (e.g. remove DOCSIS header and retrieve Ethernet frame).

Using BSoD L2VPN, a CMTS is able to forward the upstream traffic (received from the CM) on its uplink e.g. NSI towards the centralized entity, after removing the DOCSIS header and imposing an 802.1Q or 802.1AD or MPLS header, as per what MSO chose (and set in the config file). CMTS is also able to forward the downstream traffic (received from the network/centralized entity) after removing the 802.1Q or 802.1AD or MPLS header, to the Cable Modem on a particular DOCSIS downstream service-flow. It is important to highlight that the downstream Classification can be done by the CMTS without needing any CM config file dependency.

`Figure 12` illustrates using BSoD L2VPN using 802.1Q encapsulation variant. The figures below illustrate using BSoD L2VPN using MPLS encapsulation variants.

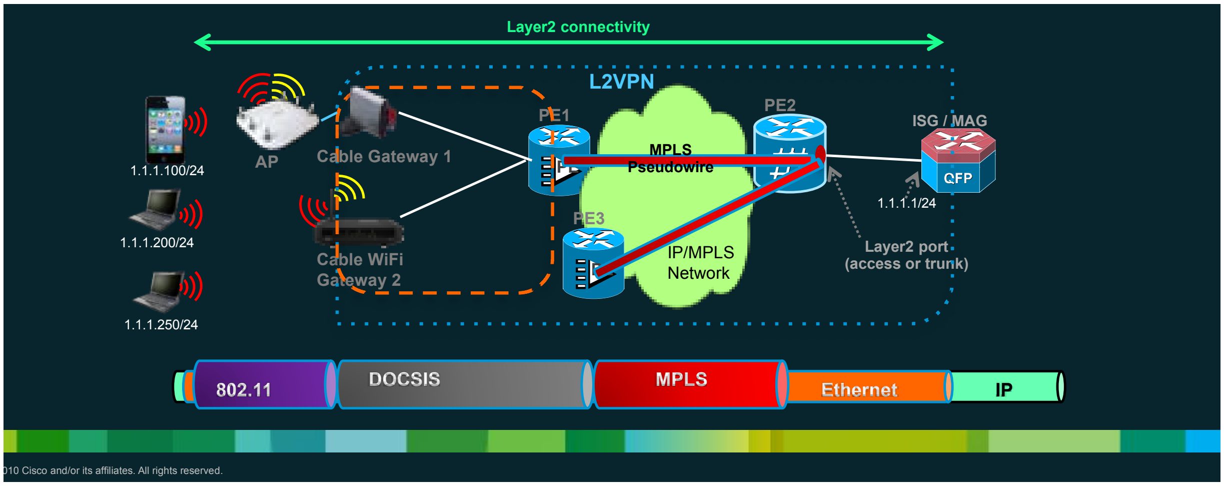
BSoD L2VPN does not require any tunneling from AP or CM, resulting in zero overhead on DOCSIS RFI, hence, avoiding any fragmentation/reassembly possibility, and also resulting in leveraging what's already supported in deployed MSO networks.



The CM config file includes TLVs that describe the mapping of one or more SFs with L2VPN designated for SP Wi-Fi. The config file does not need anything per-modem or AP specific to ensure the DS classification of the SP Wi-Fi traffic.

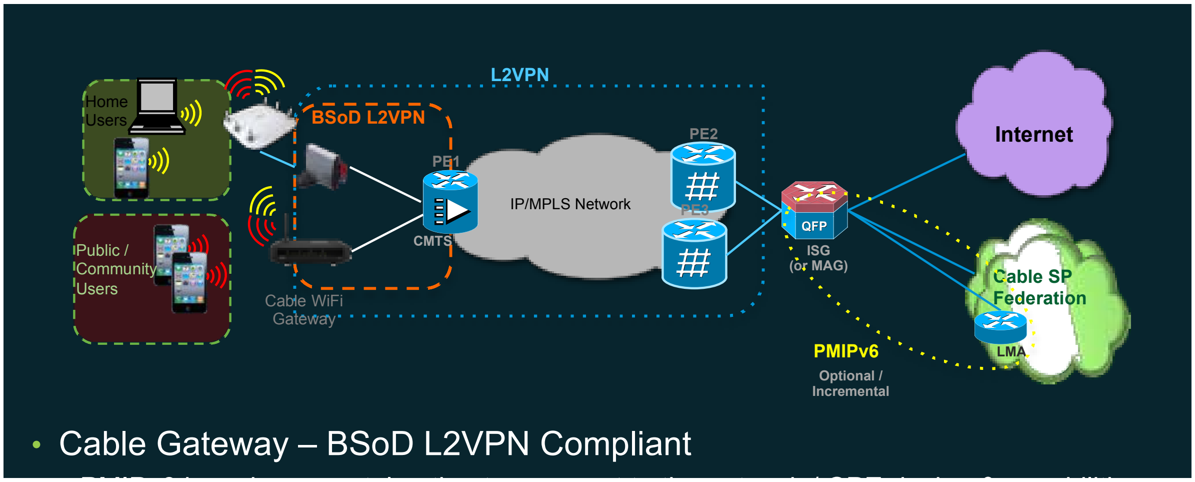
BSoD L2VPN with 802.1Q encap requires one VLAN per CM (if using P2P L2VPN) or one VLAN per network (if using P2MP L2VPN) for SP Wi-Fi.

BSoD L2VPN with MPLS encap



requires one MPLS pseudowire per CM (if using P2P L2VPN) or one MPLS pseudowire per CMTS (if using P2MP L2VPN) for SP Wi-Fi.

What’s really nice about this architectural approach is that it allows for PMIPv6 integration, as/if/when desired by the MSO to infuse mobility during Wi-Fi and Wi-Fi as well as cellular and Wi-Fi handoff. This means that PMIPv6 elements (e.g. MAG and LMA) can incrementally be introduced in the MSO network without changing the existing L2VPN setup, as illustrated in `Figure 14`:



- Cable Gateway – BSoD L2VPN Compliant

The advantages of this approach are: (a) Works in the existing deployments, (b) downstream classification is possible without any config file dependency, (c) Separate traffic management for SP Wi-Fi users and residential users, (d) common config file pertaining to SP Wi-Fi for the CMs with P2MP L2VPN, (e) Seamless mobility in all scenarios is possible with PMIPv6 integration, as/if necessary, (f) requires no fragmentation/reassembly on the last-mile access = better data-plane throughput

The disadvantages of this approach are: (a) unique CM config file per modem if P2P L2VPN is used (note thathe upcoming BSoD L2VPN specification changes (CableLabs work underway) will no longer require unique CM config file, thanks to the dynamic discovery of remote PEs), (b) dynamic SF (e.g. DSx) support may not be available, (c) does not integrate with 3GPP and doesn’t provide mobility in all scenarios.

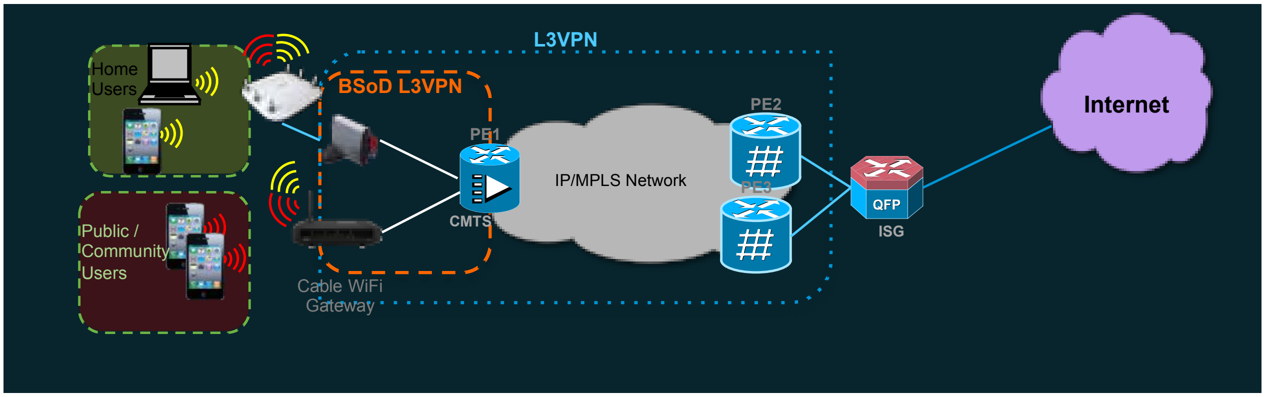
3.1.3 L3VPN

IP/VPN [RFC4364] is one of the most used technologies in SP networks (Wireline or Mobile) for internal purposes (e.g. network virtualization) and/or external purposes (e.g. Business L3VPN service).

This architectural approach allows the CMTS to terminate layer2 and use Layer 3 VPN to deliver the Wi-Fi traffic to remotely located centralized entity at layer3. In this approach, the data plane comprises ‘IP over Ethernet over DOCSIS (or PON)’ in the last-mile access and ‘IP over MPLS’ in rest of the network.

CableLabs standardization of L3VPN is underway (IP/VPN working group).

Figure 15 illustrates L3VPN applicability in sample MSO network topology.



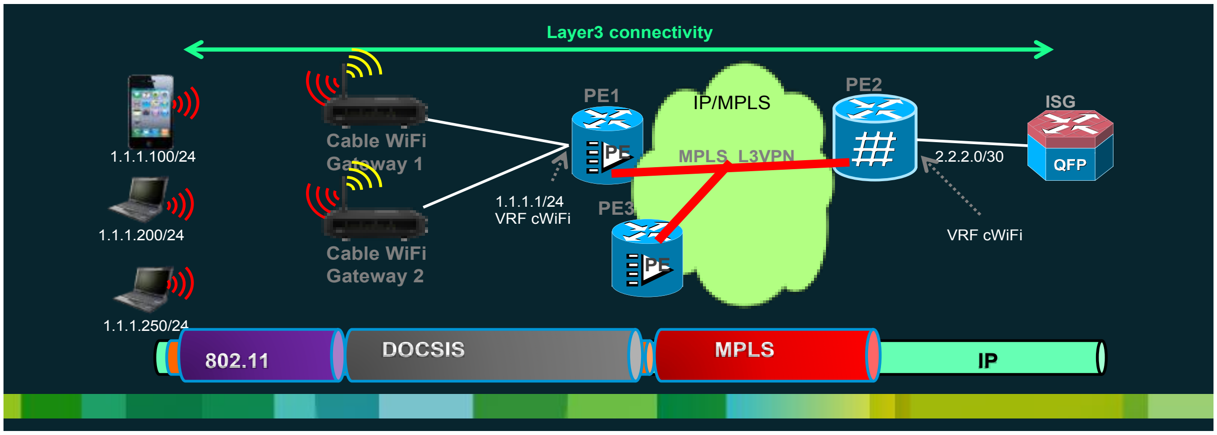
It is important to note that the service-flows used for SP Wi-Fi (e.g. Public/Community users) are different from the ones used by the residential users. This automatically allows for traffic separation and IP prefix/address assignment separation between SP Wi-Fi users and residential users.

A CM is able to classify the upstream traffic (received from the AP) using SSID (in case of embedded CM) or VLAN (in case of standalone CM) present in the Ethernet frames, and forward the traffic over a particular DOCSIS service-flow (e.g. impose DOCSIS Header on the received Ethernet frame) to the CMTS. A CM is also able to forward the downstream traffic (received from the CMTS on a particular DOCSIS service-flow) to the AP (e.g. remove DOCSIS header and retrieve Ethernet frame).

Using IP/VPN, a CMTS is able to forward the upstream traffic (received from the CM) on its uplink e.g. NSI to the network (or towards the centralized entity, if present), after removing the DOCSIS header and imposing an MPLS header. A CMTS is also able to forward the downstream traffic (received from the IP/MPLS network or centralized entity) after removing the MPLS header, to the Cable Modem on a particular DOCSIS downstream service-flow. It is important to note that the downstream Classification can be done by the CMTS without needing any CM config file dependency (e.g. per-CM or per-AP classifier).

The CM config file includes TLVs that describe the mapping of SFs with L3VPN designated for SP Wi-Fi (e.g. cWi-Fi in the figure above).

Figure 16 illustrates the data plane utilized when IP/VPN is used for SP Wi-Fi.



The advantages of this approach are: (a) CMTS could become the per-user policy enforcement point (with or without MAG function), (b) common config file pertaining to SP Wi-Fi for the CMs, (c) downstream classification is possible without any config file dependency, (c) the Wi-Fi traffic could follow the IP routing right from the CMTS, if needed, ((d) Wi-Fi users can be served by any DHCP server, (e) dynamic SF (e.g. DSx) support is available, (f) Seamless mobility is possible if the Wi-Fi user gets handed-off between APs served by the same CMTS

The disadvantages of this approach are: (a) Seamless Mobility is not possible all the time, since IP address preservation can not be guaranteed upon AP hand-off from one CMTS to another (without some additional complexity), (b) does not integrate with 3GPP, (c) cablelabs standardization not completed yet

Like L2VPN, L3VPN does not require any tunneling from AP or CM, resulting in zero overhead on DOCSIS RFI, hence, avoiding any fragmentation/reassembly possibility, and also resulting in leveraging what's already supported in deployed MSO networks.

3.1.4 Future Possibilities

In the previous section, although transport options are discussed as three discrete options, there are various other ways to achieve the requirements set out earlier. For example the benefits of PMIPv6 can be derived without the tradeoffs of tunneling by implementing the MAG in the network. Of course such an architecture brings its own set of tradeoffs. Similarly if subscriber management is implemented at the edge of the network it may eliminate the need for L2VPN/L3VPN architectures that are used to route traffic to a centralized entity. Such advanced architectures and solutions are outside the scope of this paper and are not discussed in any further detail here.

3.1.5 Comparison of Transport Options

The table below compares the three architectural approaches for network transport:

Table 1 Comparison of Various approaches

		IP Tunneling (from AP)	L2VPN	L3VPN
1	CableLabs Standardized	No	Yes	In progress[3]
2	Available	No[4]	Yes	Yes
3	Data Plane (Last-Mile Access)	User Ethernet frame over GRE|L2TP over IP over Ethernet over DOCSIS	User Ethernet frame over DOCSIS	User Ethernet frame over DOCSIS
4	Data Plane (Network)	User Ethernet frame over GRE|L2TP over IP (over MPLS)	User Ethernet frame over .1Q or .1AD or MPLS	User IP packet over MPLS
5	Overhead on Last-Mile Access	Yes	No	No
6	Requires Unique CM config file per Modem	Yes	Yes/No	No
7	User Awareness	ISG, MAG	ISG, MAG	CMTS or ISG
8	CMTS/CCAP Uplink/NSI needs?	IP	802.1Q Trunk, or IP/MPLS	IP/MPLS
9	DOCSIS Upstream Classifier?	IP Address	SSID or VLAN tag	SSID or VLAN tag
10	DOCSIS Downstream Classifier?	IP Address	MPLS label or VLAN tag	MPLS label or VLAN tag
11	DOCSIS Fragmentation & Reassembly (on CMTS, CM)	Yes	No	No
12	5-Tuple[5] based Classification by CMTS	No	Yes	Yes
13	5-Tuple based Classification by other routers	No	No	Yes
14	Mobility (WiFi-WiFi)	Yes	Yes[6]	Yes/No[7]
15	Mobility (WiFi-Cellular)	Yes	Yes	No
16	Accounting/DPI/LI possible at CMTS?	No	Yes	Yes

[3] CableLabs Standardization progressing in IPVPN Working Group

[4] Except L2TP, none of the IP tunneling variants seem to be available at the moment on the Modem / Gateway

[5] 5-Tuple = Src IP, Dest IP, Proto, Src Port, Dest Port

[6] May Require PMIPv6 Integration

[7] Seamless mobility as long as AP handoff doesn't change the CMTS.

4.0 CONCLUSION

A number of network transport options for SP Wi-Fi are discussed in this paper. Some of them are already deployed, whereas some of them are being considered for deployment.

The architectural options that help simplify the SP Wi-Fi architecture and harvest network intelligence would provide not only the cost-effectiveness, but also enable monetization opportunities. Monetization is where the next

REFERENCES

[Wi-Fi-Roam]CableLabs Wi-Fi Roaming Architecture and Interfaces Specification
[Wi-Fi-GW] CableLabsWi-Fi Requirements for Cable Modem Gateways
[eDOCSIS] CableLabseDOCSIS Specification
[DPOE] CableLabsDOCSIS Provisioning of EPON Specification 1.0
[BSODIPVPN] CableLabs IP VPN
[BSODL2VPN] CableLabs Business Services over DOCSIS Layer 2 VPN Specification
[Hotspot2.0]WFA
http://www.cisco.com/en/US/solutions/collateral/ns341/ns524/ns673/white_paper_c11-649337.html.
[RFC4364] IETF BGP/MPLS IP Virtual Private Networks (VPNs)

ABBREVIATIONS

AP	Access Point
BSOD	Business Services over DOCSIS
CAPWAP	Control and Provisioning of Wireless Access Points
CM	Cable Modem
CMTS	Cable Modem Termination System
DPI	Deep Packet Inspection
GRE	Generic Routing Encapsulation
LI	Legal Intercept
LMA	Local Mobility Anchor
MAG	Mobile Access Gateway
PMIPv6	Proxy Mobile IPv6
WLC	Wireless LAN Controller

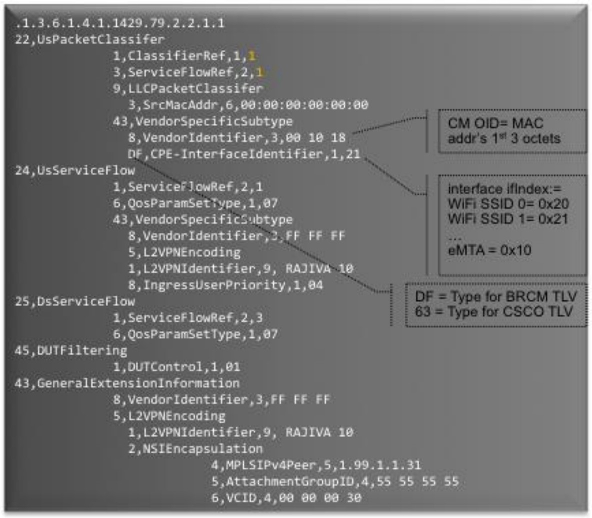
ACKNOWLEDGEMENTS

The author wishes to thank colleagues at Cisco for help and support in writing this paper.

APPENDIX

SP Wi-Fi using BSoD L2VPN – Sample Config file

A sample eCM config file for BSoD L2VPN having SSID-SF mapping is shown below



```
.1.3.6.1.4.1.1429.79.2.2.1.1
22,UsPacketClassifer
              1,ClassifierRef,1,1
              3,ServiceFlowRef,2,1
              9,LLCPacketClassifer
                3,SrcMacAddr,6,00:00:00:00:00:00
              43,VendorSpecificSubtype
                8,VendorIdentifier,3,00 10 18
                DF,CPE-InterfaceIdentifier,1,21
24,UsServiceFlow
              1,ServiceFlowRef,2,1
              6,QosParamSetType,1,07
              43,VendorSpecificSubtype
                8,VendorIdentifier,3,FF FF FF
                5,L2VPNEncoding
                1,L2VPNIdentifier,9, RAJIVA 10
                8,IngressUserPriority,1,04
25,DsServiceFlow
              1,ServiceFlowRef,2,3
              6,QosParamSetType,1,07
45,DUTFiltering
              1,DUTControl,1,01
43,GeneralExtensionInformation
              8,VendorIdentifier,3,FF FF FF
              5,L2VPNEncoding
                1,L2VPNIdentifier,9, RAJIVA 10
                2,NSIEncapsulation
                            4,MPLSIPv4Peer,5,1.99.1.1.31
                            5,AttachmentGroupID,4,55 55 55 55
                            6,VCID,4,00 00 00 30
```

CM OID= MAC addr's 1st 3 octets

interface ifIndex:=
WiFi SSID 0= 0x20
WiFi SSID 1= 0x21
...
eMTA = 0x10

DF = Type for BRCM TLV
63 = Type for CSCO TLV

2012 Spring Technical Forum
CABLELABS • NCTA • SCTE

TECHNICAL PAPER SELECTION COMMITTEE

Chairman, Dan Pike
GCI Cable and Entertainment, Inc.

John Chapman
Cisco Systems

Craig Cuttner
Home Box Office, Inc.

Daniel Howard
Society of Cable Telecommunications Engineers

Mike LaJoie
Time Warner Cable

Kevin Leddy
Time Warner Cable

Dan Moloney
Motorola Mobility

David Reed, Ph.D.
CableLabs

Tony Werner
Comcast Cable

Ken Wright
ARRIS, Inc.

NCTA Science & Technology Department

William Check, Ph.D.
Senior Vice President & CTO

Rex Bullinger
Senior Director, Broadband Technology

Steve Mace
Senior Director, Systems Technology

Andy Scott
Vice President, Engineering

Emily Murtaugh
Department Coordinator

THE COMPLETE TECHNICAL PAPER PROCEEDINGS

FROM:

2012 Spring Technical Forum

CABLELABS • NCTA • SCTE

BOSTON, MA • MAY 21 – 23, 2012

WWW.THECABLESHOW.COM

Published by: cable
National Cable & Telecommunications Association

Compiled by:
Mark Bell, VP, Industry Affairs
Rex Bullinger, Senior Director, Broadband Technology, Science & Technology
Lauren Riordan, Manager, Programs & Events, Industry Affairs

Audio recordings for each session are available for purchase – for information, please check your Cable Show program guide or visit www.thecableshow.com. For availability and price information on previous volumes in the Technical Paper series, call NCTA's Industry Affairs department at (202)222-2430 or visit www.springtechforum.com.

ISBN Number: 0-940272-51-2

CONTENTS

2012 Spring Technical Forum Proceeding

Your Wish is its Command: New Possibilities for Technology Interfaces

Redefining High-Definition: Implications of Ultra HD for Cable Providers

HFC 2.0: The Evolution of Architecture

Where the Wire Ends: Widening Cable's Wi-Fi World

The Gigabit Network: New Possibilities in HFC

Sticking to the Protocol: Implementing Best-in-Class IP Video Delivery Techniques

We Accept Cache: Intelligent Design for Media Storage and Delivery

Flying Colors: Proven Approaches for Network Investment and Improvement

V-REX – VOICE RELEVANCE ENGINE FOR XFINITY

Stefan Deichmann, Oliver Jojic, Akash Nagle, Scot Zola, Tom Des Jardins, Robert Rubinoff,
Amit Bagga
Comcast Labs

Abstract

V-REX is a new platform Comcast is building to provide speech-based applications for television control and other areas. V-REX applies automated speech recognition, natural language processing, and action resolution modules to interpret the user's request and identify the appropriate response. We describe here how we use V-REX to support an iPhone/Android app that allows users to control their cable set-top boxes by speaking into their phones. The primary focus of the work involves building grammar rules and dictionary entries for the range of requests the app can handle. We use the grammar and dictionary both to guide ASR and to allow NLP to extract the actions and entities in the request. We then convert these results into appropriate database queries that extract the information the user needs.

INTRODUCTION

Using a voice interface provides two advantages over traditional set-top-box remote or web interfaces. First, it eliminates the need for typing or other keyboard-based or remote-based text entry methods. (In the case of TV or cable remotes, this can be extremely tedious.) Furthermore, by allowing the use to directly specify what they want, it eliminates the need to wade through a series of menus or pages to find the desired option.

In order to provide a voice interface, we need to answer three questions: what words did the user speak, what action or information are they requesting, and how do we carry out the action or get the information? Each of these questions is answered by a specific module in V-REX. Automated speech recognition (ASR) identifies the words the user has spoken. Natural language processing (NLP) figures out what action or information has been requested. Action resolution (AR) responds to the request.

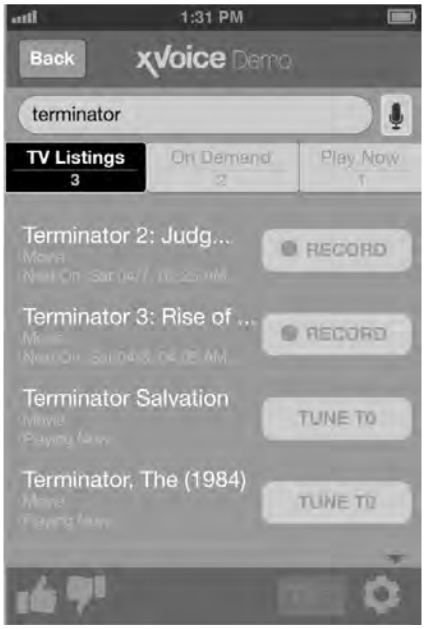


Figure 1 – Sample Results Display

Our initial V-REX application is an iPhone/Android app for Comcast customers that allows them to look up programs and control their set-top box. The app currently can handle three kinds of requests:

1) "What's on" – list what programs are available on a particular channel and/or at a particular time (including right now). The user can also specifically ask for sports games, or for a particular sport such as baseball or basketball.
2) "Tune" – switch the cable box to a specific channel
3) "Search/Find" – find when and on what channel a particular program is playing; this will also show if it is available in the On Demand library

For “what’s on” and “search” requests, the results are displayed on the screen, and the user can select individual programs or channels to get more detail. For example, Figure 1 shows the response when the user asks “What’s on CNN tonight”? In subsequent sections, we will describe each step of the process that produces this response.

AUTOMATED SPEECH RECOGNITION

The ASR module is built with CMU’s Java-based toolkit, Sphinx4. We used Sphinx because it is a well-developed open-source system that we already had experience with.[1] We replaced Sphinx4’s default acoustic model with one from VoxForge, a web site that collects transcribed speech for use with open source speech recognition engines. We built our own application-specific language model, which has two major parts, a pronouncing dictionary and a grammar, incorporating as well a general language model built on the English Gigaword Corpus (www.keithv.com/software/giga).

The dictionary maps words to their possible pronunciations at a phonemic level. The phonemes in our pronouncing dictionary are based on the ARPABet developed for speech understanding systems in the '70s. There are 39 phonemes, not counting variants due to lexical stress. Anything a person can say, including actions like "tune to" and channel names like ESPN, must be stored as a phonetic representation.

The grammar is a textual description of the combinations of words and phrases the system will accept. It is written in Java Speech Grammar Format, an augmented BNF-style format [1]. The grammar contains rules describing how users can ask to change a channel, what exactly counts as a title, and how people can ask about a time of day. Within a limited domain like television, the language model provides a higher level of accuracy in detection than it would normally achieve in an unconstrained system.

The dictionary and grammar are designed around three specific requirements of the application. The first is to handle channel names, many of which are not in the general vocabulary of the basic Sphinx system. For example, “SyFy” and “Tru TV” need to be added. In addition, some channel names may have multiple pronunciations, e.g. “Univision” can be spoken with either English or Spanish pronunciation; and some channel names may contain words that are in the general vocabulary but not commonly used together outside of the domain, e.g. “Fox Business” or “Showtime Family”. To handle these cases, we added the names of all of the channels Comcast provides to the dictionary. Including the channel names in the dictionary does more than just improve recognition of these terms; it also lets us use a more precisely tailored language model, improving the overall ASR performance.

The second requirement is to handle a wide range of time and date specifications. While most of these are part of the general language, there are some that are specific to the television domain (e.g. “prime time”). More importantly, we want to directly handle complex phrases such as “next Tuesday evening after 10” and recognize them as indicating the time of a program as early in the processing as possible. To that end, we include in the grammar a large range of ways of referring to dates and times. A portion of this grammar is shown in Figure 2.

Finally, we need to recognize titles of movies and TV programs. This is a particular challenge, as titles can contain deliberate misspellings or ungrammatical phrases that would be rejected by a general language

[1] We are continuing to evaluate other ASR systems, but so far have found Sphinx4’s performance and accuracy to meet our needs.

```
<temporalAdverb> = (    <weekday>
                      | on <weekday>
                      | this <weekday>
                      | prime time
                      | [right] now
                      | tomorrow <daytime>
                      | today <dayTime>
                    );

<dayTime> = ( morning | afternoon | evening | night );
<weekday> = (sunday | monday | tuesday | wednesday | thursday | friday |
saturday);
<clockTime> =
      (  at  <hour>  o'clock
      |  at  <hour> [ <minute>  ] [ <amOpm>  ]
      );
```

Figure 2 - A portion of the date/time grammar

model, e.g. “eXistenZ” or “De-Lovely”, or subtitles that don’t fit into normal sentence structure, e.g. “Dodgeball: A True Underdog Story”. Even without these kinds of problems within a title, there is the danger of processing the title as a normal part of the whole sentence. For example “What time is Seven on?” is most likely asking about the movie “Seven”, not channel 7 or seven o’clock. In order to handle these problems, we have added movie and TV show titles to our dictionary. This allows us to recognize titles when spoken (in places where titles make sense).

In order to add titles to the dictionary, though, we need to know which titles to add. We can’t simply add **all** of the titles that have ever been produced, because that would involve several million titles. This would drastically increase the size of the dictionary, seriously diminishing both speed and accuracy of the ASR system. Furthermore, the vast majority of the titles would be for movies or TV shows that aren’t currently available and that the user has almost certainly never heard of. Instead we limit the titles to all shows that are currently available (either on a broadcast or cable channel or on demand). We also include the most popular movies and TV shows, even if they are not available, since there is a good chance the user will ask for them.

The top level of the grammar specifies the range of possible requests the system can handle (and therefore needs to recognize). A simplified version of the grammar is shown in Figure 3. The three possible request types each have their own top-level rule, which is further specified in subsequent rules. The various parts of the grammar are combined to provide a language model that constrains and guides the recognition process.[2]

[2] As the application expands to handle a larger range of requests, we anticipate allowing some of the ASR work to use a more general statistical language model, so that the system can recognize unrestricted language. This will be particularly important when we start handling extended dialog. The grammar will stay play an important role in interpreting the request, though, as described in the section on the NLP module.

```
<whatsOn> = <actionPhrase> [ <modifierPhrase> ];
<tuneTo> = <tuneToPrefix> <channel>;
<search> = (<searchPrefix> <title> [now] | <title>);

<searchPrefix> = ( can i watch | play | search | find );

<actionPhrase> = <whatsOnPrefix> | <whatsOnPrefix> <channel>;

<tuneToPrefix> = ( (tune to | change the channel [to] | change [to] );
```

Figure 3 – a portion of the top level grammar

NATURAL LANGUAGE PROCESSING

Once the ASR module processes the user's request, the output (i.e. the request in text form) is sent on to the NLP module. NLP starts by parsing the text, using the same grammar used by ASR.[3] The text is parsed using the JSGF parsing facility, part of the package used to write the grammar (as described above). The NLP module uses the resulting parse tree to interpret the utterance, inferring the semantics from the rules used and the tags assigned in the parsing process.

For example, consider the request "What's on Disney on Saturday?"; the parse structure for this is shown in Figure 4. Here we can determine the request type from the <whats-on-phrase> node, the requested channel from the <channel> node, and the time constraint from the <temporal-adverb> node. These three different pieces of information are actually obtained in three different ways. The request type is determined to be "what's on" simply from the presence of a <whats-on-phrase> node, which reflects that the request has the structure and content of a "what's on" request. The channel is determined to be "Disney" based on the value of the <channel-name> node, which the grammar indicates is the appropriate value for the text "Disney". (In this case the channel name and the text are the same, but that is not the case for all channels.) The time constraint needs more complex processing, which is provided by special-purpose code in the NLP module that knows the range of possible parse structures and how to extract time and date values from them. Once it has identified all the pieces of the request, the NLP module assembles them into a request structure that is passed on to the AR module.

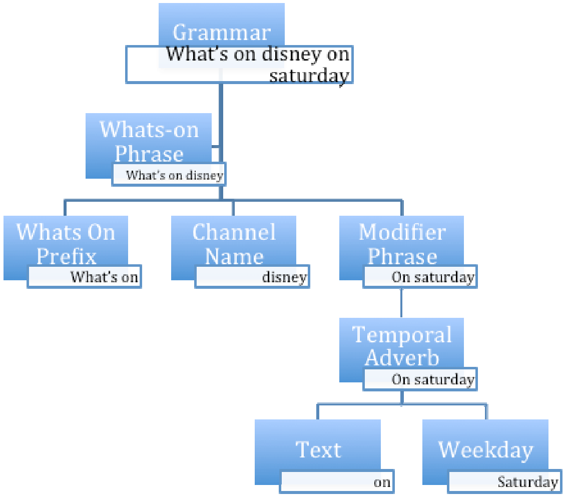


Figure 4 - Parse Structure for "What's on Disney on Saturday?"

[3] The two modules don't have to use the same grammar, although that is the case in the current system. It might be appropriate to use different grammars if, for example, we want to allow incidental comments that are not relevant to the request (e.g. "tune to HBO, please"). In particular, if we switch to using a statistical language model for part or all of ASR rather than a grammar-based one, NLP and ASR will need to use different grammars.

For the current application, we assume that any information not indicated in the request is deliberately left unspecified. For example, if no channel is mentioned, we assume the user wants to know the most popular shows available on any channel in the requested time span; if no time is specified, we assume the request is for programs on during prime time today. Missing information is therefore either left unspecified in the request or filled in with a default value.

In more complex applications, we might need to explicitly mark that the information is missing so that later processing can take necessary action to deal with the situation.

ACTION RECOGNITION

The AR module receives the request structure built by the NLP module and attempts to carry out the request. This involves constructing and sending an appropriate query to Comcast's REX search system. REX is the system we use to index and search through the complete set of programs available on broadcast and cable channels and on demand. The precise form of the query to REX depends on the type of action requested. For "what's on" requests, the query indicates the requested channel (if any) and time span. For "search" requests, the query indicates the title that the user specified. For "tune" requests, the query indicates the name of the requested channel. We need to query REX for tune requests to find the channel number corresponding to the channel name; if the request specified a channel number directly then we can skip the query.

The results returned by REX are packaged up along with an indication of the request type and the output of the ASR module and sent back to the client app. In case of an error, an appropriate error code is returned, along with any relevant information about the error. As mentioned above, a more complex application might require further interaction with the user, either to resolve an error or to get more information needed to carry out the request. The AR module contains a simple text-to-speech component, based on the FreeTTS system [2], to handle such interaction. This capability is not needed currently, though.

THE IPHONE/ANDROID CLIENT APP

The server-side components described above support a client iPhone/Android app that allows the user to speak requests and see and respond to the results. The initial screen for this app is shown in Figure 5; the user can press the microphone and speak a request. A typical response is shown in Figure 6; here the user has asked "what's on CNN tonight?" and the app displays the list of programs returned after processing through the ASR, NLP, and AR modules. The user can select a specific show to get more detail, as shown in Figure 7. Search requests display similar responses, except that the results are organized by relevance to the request rather than by time.

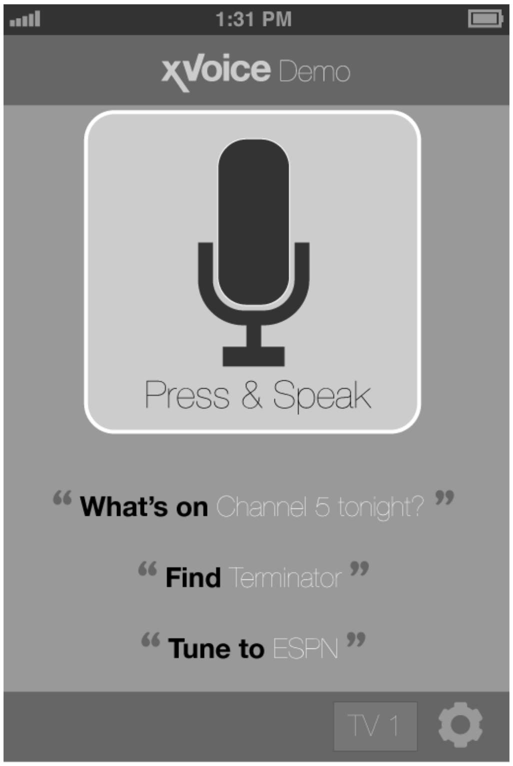
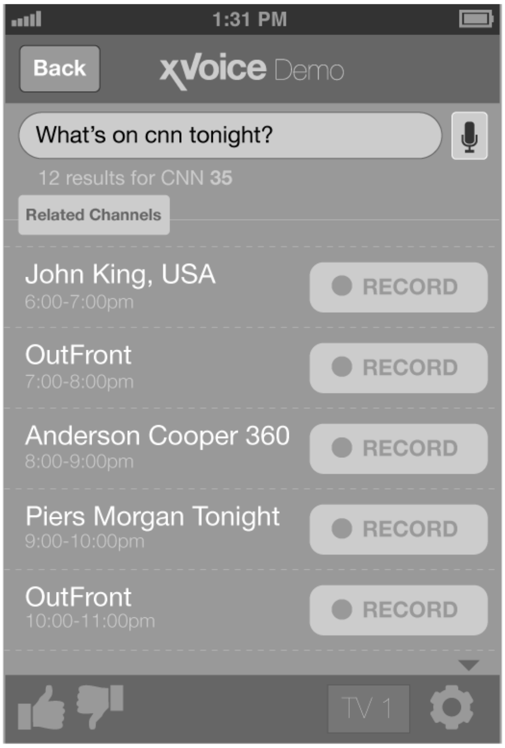


Figure 5 – Initial Client App Screen



What's on cnn tonight?

12 results for CNN 35

Related Channels

John King, USA
6:00-7:00pm

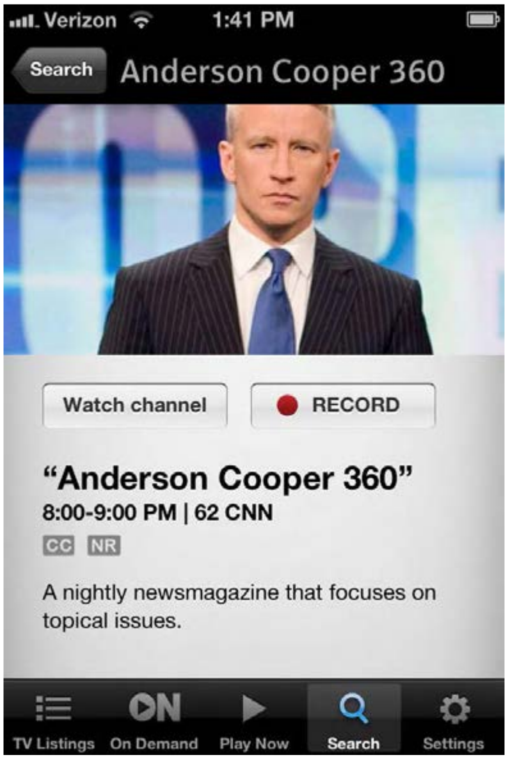
OutFront
7:00-8:00pm

Anderson Cooper 360
8:00-9:00pm

Piers Morgan Tonight
9:00-10:00pm

OutFront
10:00-11:00pm

Figure 6 – Response to "What's on" request



"Anderson Cooper 360"
8:00-9:00 PM | 62 CNN

A nightly newsmagazine that focuses on topical issues.

Figure 7 – Details of a specific program

CONCLUSION AND FURTHER WORK

Speech-based interfaces provide a uniquely simple and direct interface. Users can simply say what they want, without having to type in complex queries or navigate through layers of menus. The V-REX platform combines automated speech recognition, natural language processing, and action resolution to power speech interfaces. Our initial app brings this ability to searching and controlling cable set-top boxes. We are exploring ways to extend V-REX to other applications built on Comcast's cable/internet infrastructure. One possibility is to extend it to our Play Now system for watching programs over the web. A more interesting extension is to apply it to Comcast's home security service, so that people can easily check on the status of their homes while they are away.

REFERENCES

[1] http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/index.html

[2] http://freetts.sourceforge.net/docs/index.php

ARCHITECTURAL CONSIDERATIONS FOR ENABLING HTML5 APPLICATIONS IN AN OCAP ENVIRONMENT

Navneeth Kannan
Motorola Mobility Inc.

Abstract

OCAP has become an established middleware platform for set-top boxes, and while it has enabled application portability, it has not signfiicantly improved application delivery cycles. Meanwhile, operators are developing applications using browser-based tools for tablets and smartphones. The advantages of web-based application development can be realized on a set-top platform by leveraging the robust set of platform services provided by OCAP while replacing the presentation technology with HTML5.

This paper examines the architectural considerations in enabling a browser-based application environment in the context of OCAP middleware for set-top boxes.

EVOLUTION OF THE SET-TOP APPLICATION

The evolution of set-top applications closely follows the evolution of the underlying Application Development Environment, and can be broadly classified into three phases. The Legacy phase, which was defined by the early digital set-top box platforms with proprietary middleware; The OCAP Phase, marked by the Host platforms and OCAP middleware; and lastly, the HTML5-Phase, marked by the general movement towards IP-based technologies in the set-top box, and browser-based applications.

Legacy Phase

Prior to the introduction of OCAP, set-top middleware was predominantly written in C/C++, and based on an underlying real-time operating system. Application interfaces were provided by the set-top vendors, but were proprietary in nature, causing independent software vendors and application developers to design the software to adapt to disparate application interfaces. Set-tops in this era were also limited in terms of CPU power, memory availability and available interfaces.

OCAP Phase

Cablelabs started the OpenCable initiative in 1997, with the goal of helping the cable industry deploy interactive services and create competition in the host device marketplace through a process of standardized specifications for various components of the cable system.

As part of the family of specifications, the OCAP specifications address the middleware component, enabling a common API interface for application developers.

The Java-based application development environment allowed for portable applications, but did not appreciably decrease the application development and deployment cycles. Nevertheless, there is a significant deployment of OCAP based set-top boxes in the field, and there is consensus that at least the video network-facing functionality of the OCAP platform is here to stay.

During this period, there has been a corresponding improvement in the underlying set-top platforms, as they have more CPU power, more memory and much more functionality.

HTML5 Phase

We are now entering this phase, where support for HTML5 based applications will become the norm for the set-top platform. This has been influenced by three independent, but related trends.

1) Emergence of CE devices such as tablets and smartphones as alternate or companion screens along with the TV as the primary screen. Operators are routinely developing companion applications targeting these devices. As the nature and types of the devices proliferate, HTML5 based applications become more appealing given the promise of portability. There is also a need for a consistent look-and-feel on all screens, and thus a desire to not leave the primary screen behind.

2) Migration of traditional data services to being web-based services, largely driven by the previous trend. As an example, metadata services for linear and VOD assets are now routinely available via web-service APIs, and such web servers are in production networks of several operators already.

3) The evolution of the set-top boxes, in terms of CPU power, memory and general capability. Set-tops no longer present a barrier for imaginative and creative experiences heretofore thought of only in the context of richer development environments such as personal computers or tablets.

ENABLING HTML5 APPLICATIONS IN A CABLE SET-TOP ENVIRONMENT

Application support, even in the case of a legacy environment, requires a comprehensive set of back-end server ecosystem. Metadata delivery for EPG and VOD assets, Provisioning interface, Conditional Access and Billing System interfaces, Application download and delivery services are just some examples of the support required as part of an end to end solution.

Many of these requirements still exist even in the case of browser-based applications. The back-end systems will have to be enhanced to be able to support the browser-based applications. Architectural considerations for the server side merit a separate discussion, outside the scope of this paper.

In this paper, we restrict the discussion to the elements of the solution that involve changes to the set-top box software. Furthermore, we are specifically targeting the set-top boxes based on OCAP middleware.

Why retain OCAP?

In the European marketplace, set-top solutions with integrated browser support are already in place. These provide support for HTML-based applications already. These middleware solutions generally do not address the requirements that are unique for the North American cable system.

The network-facing functionality of OCAP already addresses such requirements as support for Cablecard, Provisioning, Regulatory aspects and such, that it makes OCAP a viable platform for enabling browser-based applications.

CO-EXISTENCE OF WEB-BASED APPLICATIONS WITH REGULAR OCAP APPLICATIONS

Browser standards [1] and implementations have evolved to a stage where the Browser is not just a graphics rendering engine, but a full-fledged development environment. Web pages have evolved from simple text and graphics pages, to dynamic pages with animations and now to complex applications, exemplified by websites such as http://www.facebook.com

It is conceivable that an entire application being written in Javascript and run as a browser-based application. However, in the context of a set-top box, there are advantages to allocating functionality to run in the traditional mode. We identify a couple of examples below.

Application Services Layer:

An application service layer is the underlying engine for a navigation application, without the User-Interface rendering portion. EPG metadata ingestion, SDV (Switched Digital Video) protocol handling, and VOD session handling are some examples of functionality that may be retained in a legacy application mode.

Self-contained simple features

One example in this category would be the Diagnostics application that gives a variety of information about the set-top box, and is accessed by the field-support personnel. Another example is the display of MMI screens as part of the Cablecard interactions.

Treatment of Closed-Captions / EAS

Regulatory standards such as Closed-Caption rendering and EAS (Emergency Alert System) require that stringent timing specifications be met. While it is feasible for the platform to extract the Closed Captioned data from the incoming QAM video and provide the text to the web-application for rendering, it is far more efficient to retain platform based rendering.

Implications for the Architecture

The examples discussed above have an impact on the architecture for the solution.

1) Enough memory needs to be allocated to the JVM so that legacy-style OCAP applications may be supported.

2) Graphics planes need to be shared between the browser-based application and the Java-based application. For instance, EAS text needs to be superimposed on top of any other graphics that may be displayed by the current browser-based application.

3) User input (such as keys from remote control) need to be sent properly to the application in control, so some form of application focus management must be implemented, perhaps by the Application proxy.

ARCHITECTURAL ELEMENTS OF WEB-APPLICATION SUPPORT IN OCAP

The following diagram provides a generic view of an OCAP-based solution for supporting browser-based applications. Note that this is a high-level view that allows for multiple implementation choices

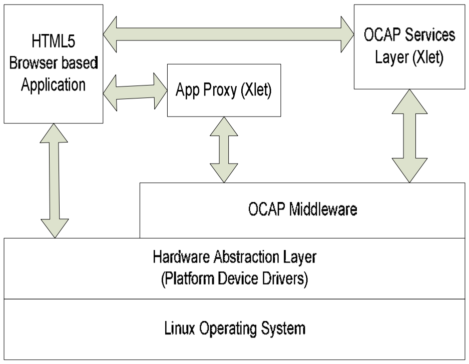


Figure 1: Generic architecture diagram

In order to enable HTML5 based applications in a set-top environment, three things will have to be accomplished.

1) A suitable HTML5 browser needs to be ported to the OCAP-based platform. By a browser, we are not referring to the traditional browser application but the underlying HTML5 rendering and processing engine.

2) The HTML5 standards continue to evolve, and there is a growing consensus to address TV-centric capabilities as part of the

standards body. However, there is a gap today, and support for set-top specific features will have to be exposed to the browser-based applications via standard mechanisms such as Javascript APIs implemented as an NPAPI plugin.

3) The OCAP middleware will have to be sufficiently modified or extended to support meaningful coexistence of browser-based and Java-based applications. Resource management, Life-cycle management and Focus management are the typical aspects that require careful consideration.

Targeting the solution to a broader spectrum of set-top boxes, including those that are already deployed will introduce an additional level of complexity and associated constraints that will limit our degrees of freedom. Nonetheless, it is critical to address these, because the true benefits of any solution are borne only when applied to a large deployed base.

All the above mentioned aspects are explored in greater detail in the following sections.

BROWSER PORTING IN DETAIL

Porting a browser to an embedded platform, and specifically to a set-top, is a non-trivial task. Functionality and performance are key requirements. The general expectation is that any browser-based application must exceed the functionality and performance levels that are available in legacy applications.

In the diagram below, we depict a general purpose browser-engine, with appropriate interfaces that would have to be integrated on an underlying platform. As the HTML5 standards evolve and browsers include additional functionality, additional interfaces will have to be considered. Also, while we indicate OCAP middleware in the diagram below, the discussion should be applicable to other equivalent middleware.

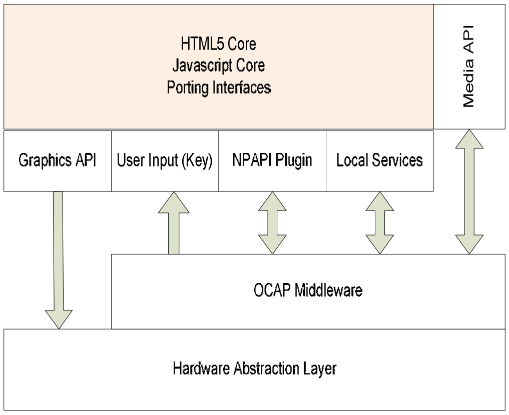


Figure 2: Browser Porting Interfaces

Graphics API Integration:

Given that graphics performance is subscriber-facing and most noticeable, the integration of the browser to the underlying graphics engine is of critical importance. The graphics adaptation layer must allow sharing of the planes across multiple applications. Most new set-top boxes provide support for DirectFB, which addresses these requirements. DirectFB is a thin library that provides hardware graphics acceleration, input device handling and abstraction, integrated windowing system with support for translucent windows and multiple display layers. It is a complete hardware abstraction layer with software fallbacks for operations that are not supported by underlying hardware. The library supports Graphics operations such as rectangle drawing and filling, blending with an alphachannel, colorizing and color keying.

In a typical implementation for a set-top box, the platform utilizes separate windows for system-level functions such as Closed-captioning, EAS and Diagnostics, and a separate additional window is created for Browser-rendered graphics.

While DirectFB is supported ubiquitously in all the new set-top platforms, some of the older platforms may not have the capability. In such cases, assuming that the underlying hardware abstraction layer supports the graphics functions, it is as well easy enough to create a shim layer to provide the subset of the DirectFB API interface as required by the browser. In the same way, not all browser implementations may have a DirectFB interface requirement. In such cases, the south-bound Graphics APIs from the browser have to be married to the underlying hardware abstraction APIs.

Media API Integration:

The audio and video elements were introduced in the HTML5 working draft specifications. The audio element is used for playing audio streams, and the video element is used for playing videos or movies.

Typical ports of the Browser on standard platforms such as a personal computer allow for native support of the video and audio tags to the underlying platform. Support on a set-top will have to be achieved in a similar manner.

The Audio and Video tags will have to be resolved to invoke the associated media player supported by the platform. In general, any browser will allow for hooking the HTML media elements to custom media playback engines. The choice of a specific media playback engine will depend on the functionality, open-source considerations and other architecture needs for the set-top box.

As an example, the open source browser project, Webkit offers a “MediaPlayerPrivate” interface for a port-specific implementation of a media-playback engine. Other browsers will have similar API interfaces for connection to the underlying media pipeline.

User Input:

User input in the case of a set-top box is predominantly with the IR remote control, although it is not uncommon to have additional devices such as a wireless keyboard attached to the set-top device.

When the underlying platform supports a DirectFB interface, it is possible to direct the user input using the DirectFB interface directly to the browser. However, as stated earlier, if we want to retain some functionality in a Java-based application, it might be advantageous to retain the user input to flow via the platform. One example where this feature will come in handy is the case of the diagnostics application.

Keycode mapping is an important aspect in the user-input integration. In general, most browser implementations originated for a personal computing platform with keyboard and mouse inputs. In the case of the set-top box, the remote control keys will have to be mapped appropriately, so that most applications that are hosted on the browser will function reasonably. Browser-based applications that have been written specifically for the set-top device can obviously take advantage of additional remote control keys such as the “Record”, “Menu”, or “Info” keys found on a typical set-top IR remote control.

NPAPI Plugin Interface:

A HTML5 browser will include an execution engine for Javascript. This core engine supports the interpretation of all the standard Javascript constructs. Browsers allow for additional libraries to be supported using the concept of the NPAPI plugins. NPAPI (Netscape Plugin Application Programming Interface) [2] is a cross-browser plugin architecture supported by many browsers. The initial intent for the framework was to provide plugin libraries for specific constructs not supported natively by a browser.

Support for set-top specific functionality may be exposed by a private set of Javascript libraries using the NPAPI plugin interface. This is described in detail in a later section.

Local Services Interface:

Any browser will also need a set of general purpose services from the underlying platform. Typically almost all implementations support running on standard Linux Operating System. Linux is now supported on set-top boxes, and hence most of the local services required by any browser should be available directly from the operating system. An exercise in porting a browser to a real-time operating system will involve providing support for all the local services required by the browser.

A good browser will also support logging functionality that needs to be hooked into the corresponding support provided by the underlying platform.

Other Porting Considerations:

The requirements for an application on a set-top box may require additional changes to the browser, and will need to be addressed as part of the porting process. We give a few examples below, noting that the exact list will depend not only on the choice of the browsing engine, but also the requirements of the application and the selected architecture for the overall solution.

1) Extensions to the browser will have to be made in order to support private Javascript functions such as a function to display Open source information, or for a function to log warnings and errors.

2) The application may need that the *XmlHttpRequest()* function support both synchronous and asynchronous requests.

3) Changes to the access control rules to meet the system requirements: As an example, the solution may require dynamic and secure configuration of whitelist of URLs that the browser is allowed to access.

4) Extensions to the browser to enable secure support for W3C widgets may be required.

5) Configuration of the “First portal page” needs to be established.

APPLICATION ACCESS TO PLATFORM AND OTHER SERVICES

The second big aspect of enabling browser-applications is to provide a mechanism for the applications to make use of platform services. The HTML5 standards continue to evolve, and the establishment of the Web & TV Interest Group is a move in the direction of ensuring that the standards will start addressing the needs of the TV community. The standards body is working on areas related to media pipeline, content protection and even regulatory aspects. Notwithstanding the strides being made on the standards front, there is a gap between what a robust, full-featured set-top application needs and what can be supported by just the HTML5 browser.

Information required by Applications:

There are two categories of functions that require a level of support from the underlying platform, and a third category for higher-level services provided by service layer that runs as an Xlet, a regular OCAP application.

1) Browser-based applications need access to device specific functions. Examples include the front-panel LED, the local DVR, and such items.

2) Browser based applications also need access to service information from the platform. The specific list will depend on the application architecture, and could include

items such as host information including IP address, MAC address, Cablecard Status, User settings for Audio/Video configurations, Channel map information, DVR metadata and so on.

For the functionality described there are corresponding OCAP APIs that are available at the platform layer. The following table provides an example list of some of the application needs, expressed as generic Javascript functions, and the OCAP APIs that would perform the associated action.

App function	OCAP API
getSerialNum()	Host.getID()
isPowerOn()	Host.getPowerMode()
isCAmodule()	CablecardControl.isPresent()
getVolume()	AudioOutputPort.getLevel()
setVolume()	AudioOutputPort.setLevel()
getFreeSpace()	StorageVolume.getFreeSpace()
getTotalSpace()	StorageVolume.getTotalSpace()
clearFP()	TextDisplay.eraseDisplay()
setFPclock()	TextDisplay.setClockDisplay()
setLEDon()	setBrightSpec(on)
setLEDoff()	setBrightSpec(off)

Table 1: Sample list of OCAP APIs

In addition, events that originate from the platform side will need to be handled by the browser-based application. A sample list of OCAP events is given below:

OCAP generated Events
Set-top entered / exited Standby State
IR Remote Control Key Up/Down
Hard disk is full
Error conditions related to a recording request

Table 2: Sample list of Events

In addition, there is a third category of higher level service information that may have to be provided from a service layer that runs as a Java application. These include services such as SDV, EPG metadata information and so on.

Application Proxy and Service Layer Xlets:

In the last section, we have established the need for the browser-based applications to obtain information from the platform and the service layer. In this section, we examine the mechanics of the communication.

The Application proxy is an OCAP application that runs as an Xlet, and serves to act as a proxy for the browser-based application. It acts as a go-between, and does not have any extended complexity. The Service layer, on the other hand, is the part of the application business logic that has been retained to run as a regular Java application. The allocation of functionality between the browser-based application portion and the service layer is a design choice to be made by the Application architect.

We have chosen to indicate the concept of an Application proxy as separate from the services layer, but an implementation of the service layer could very well incorporate the proxy functionality.

Access mechanisms:

There are at least a couple of different ways to accomplish the communication between the browser-based application and the application proxy / service layer Xlets. These are discussed in greater detail in the following sections.

1) JNI based access: Given that browser-based applications are written in Javascript, it will be a natural choice to have a set of private Javascript functions that represent platform functions. Application frameworks typically have an abstraction layer in Javascript to adapt to different types of platforms. Using the NPAPI-plugin interface, a native shared library can be built to bridge the Javascript API bindings to the Application proxy (and the service layers) via JNI (Java Native Interface). This is depicted in the diagram below:

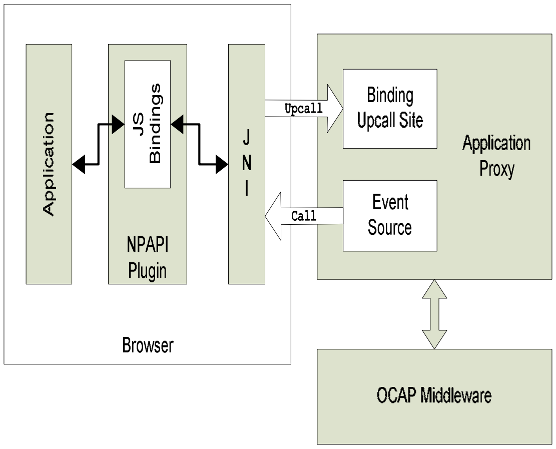


Figure 3: Platform Access using JNI

This method is fairly straightforward, easy to implement, and provides low latency access from the browser-based application to platform support functions. Support for eventing and asynchronous notification is also achieved without any difficulty.

In this case, the browser will be within the same process context as the JVM that also hosts the application proxy (and the service layer). Although we have not shown the service layer in the above example, the same mechanism will apply there as well.

The JNI based mechanism works very well for platform-style functions all of the resources are only locally applicable. For applicability in a home-network, we need additional methods.

2) Web-Service Acces: Access to functionality via RESTful APIs provided by a web-service interface is another way to accomplish our goals. This is typically useful for functionality that is applicable not just to the local device but any other client in the home network. In this case, we will need an additional web-server component, and this could leverage an open-source library such as *lighthttpd* or be a purpose-built library. This is depicted in the diagram below:

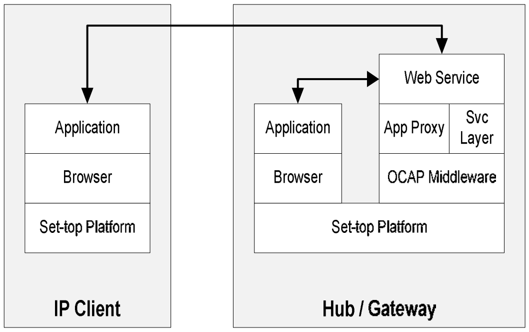


Figure 4: Web Service Access

As can be seen in the diagram above, the web-service based approach allows for hub (Video Gateway) and IP client based applications to access functionality in a symmetric manner.

It should be noted that from the perspective of the IP Client, any local platform resources (such as its own Front-panel LED, as an example) will have to be integrated locally, and in a manner consistent with the underlying middleware and platform architecture. Since web-service requests may be made with standard javascript function calls, this method does not require an NPAPI plugin for this purpose.

The use of web-service interface for local device functionality typically introduces more latency than a simple function call approach (like the JNI mechanism). In a prototype implementation, measurements for identical functionality indicate that the additional latency is within acceptable limits.

3) Hybrid Access Mechanism: A third method will be to use a combination of the above two approaches. Access to the local resources, typically provided by the platform proxy may be accessed via the JNI mechanism, and

access to service layer level information provided via web-services.

There may be other ways to address the requirements as well. The specific choice for an implementation will have to be determined in the context of the overall system architecture and by examining the trade-offs.

OCAP PLATFORM EXTENSIONS

In addition to porting the browser and providing a mechanism for the browser-based application to the platform layer, we have to ensure that the underlying OCAP platform is prepared for the solution.

Considerations based on target platforms:

A carefully thought out architecture will allow the targeting of the browser-based applications to not just new set-top boxes that are rich in resources such as CPU power and memory, but also to set-top platforms that are already deployed in the field. In order to build a viable application platform that developers can take advantage of, it is imperative that we drive the solution to as high a target population of deployed set-top boxes as we can. We discuss some of the considerations below.

1) DRAM Memory Allocation: Memory in the set-top box, previously set aside for applications via allocation in the JVM will now have to be shared across applications that will continue to run on the OCAP side (as Xlets) and appplications that will run as part of the browser. Depending on the specific browser implementation, there may either be a single heap for the browser, or two heaps, one for the basic-core and the other for the Javascript core. The configuration of memory for the browser (including Javascript core) and the JVM objects is typically static (through a configuration file that can be changed before start-up), and hence has to be adjusted correctly. In real-life implementations, we have measured and observed that adequate allocations are achievable in set-top boxes with memory configurations of 512MB for DRAM.

2) Flash Memory Considerations: The OCAP code image will now include the browser as well, and Flash memory requirements will increase to accommodate the increased size. In code download implementations that require ping-pong buffering (requiring the flash to contain two copies of the code image); the requirements will double.

Considerations due to application nature:

Running a browser-based application introduces additional challenges for consideration that will result in changes to the way the OCAP platform. Key changes are described below.

1) Graphics Plane Sharing: In the native layers of the set-top implementation, graphics area must be shared across the three distinct users (browser-based applications, Java-based applications, and native platform elements that utilize the graphics plane (for CC, EAS etc.). This is usually accomplished with the concept of graphics planes. Different planes are allocated to the different entities, and the final composition is made at the lowest levels of the firmware / hardware. If the underlying platform already supports the concept of graphics planes, then this may not require anything extra other than the concept of Application focus management, which can be achieved through the use of the Application Proxy Xlet.

2) Widget Authentication and Storage: If applications are packaged as W3C widgets, there will be additional requirements [3] that will be placed on the platform (in addition to the requirements noted earlier for the browser itself). The platform will have to authenticate the widget (the widget will have to be signed) before it can be stored in Flash.

3) WatchTV Application: If the browser-based application itself is not packaged as a widget, then it is critical to provide at least a minimal WatchTV application that can be run as a regular OCAP application and be stored in Flash.

4) Application Proxy & Focus Management: We already talked about the need for Application Proxy functionality. Depending on the architecture, this may be either an independent Xlet or combined with any services layer Xlet that may be in place. At any rate, application life cycle will have to still follow the OCAP Xlet life cycle concept, and the Application Proxy will have to manage the Focus management aspects, ensuring that the keys are being directed to the current application in focus.

CONCLUSION AND FUTURE TRENDS

Browser-based applications are here, and are here to stay, with a strong developer community support. Set-tops that have supported OCAP middleware are prime candidates for hosting a browser-based application environment, given that the set-top boxes targeted for OCAP had sufficient CPU power, and enough memory. In this paper, we explored the architectural considerations for such a solution, and we believe that it is feasible to design a solution taking specific customer requirements into account. Multiple video service providers are considering moving to browser-based applications.

We believe that with the correct architectural solution, we will be able to utilize the solution not just for new set-top boxes, but also a majority of the deployed OCAP set-top box population.

Rich application frameworks will be required to enable rapid development of applications and ensure portability across different browser environments, multiple screens and such.

We see future work in the following areas:

1) Continued evolution of browser functionality will in turn drive even more functionality towards the browser-based application side.

2) Alternate ways of supporting EBIF or equivalent interactive TV applications in the context of a browser-based applciation will have to be researched.

3) Desire to establish and grow a strong developer community will require additional research on supporting third-party W3C widgets and opening up platform access in a standardized manner.

REFERENCES

[1] HTML5 Specifications from W3C (latest draft): http://www.w3.org/TR/html5/

[2] NPAPI Plugin Interface Specifications https://wiki.mozilla.org/NPAPI

[3] W3C Widget Requirements: http://www.w3.org/TR/widgets-reqs/

HTML5 Framework and Gateway Caching Scheme for Cloud Based UIs

Mike McMahon, VP of Web Experience and Application Strategy
Charter Communications

Abstract

Recent advances in our industry such as TV Everywhere and second screen, companion apps merge video delivery and consumption with web technologies. Similarly, much progress has been made in introducing service-oriented architectures, exposing common web services and enabling a high degree of consistency and re-usability in backend systems.

Video is now clearly being consumed on a wide range of devices and these devices can vary wildly in terms of screen size, capabilities, development platforms, etc. Tablets, game consoles, smart TVs, mobile phones and a number of other devices are all viable video terminals. Processing and delivering video into a variety of flavors, bit rates and such is non-trivial but is generally well understood and now fairly commonplace. In order for the Cable industry to fully embrace an already highly fragmented client platform landscape and position itself to exploit new devices as they become available it is necessary to achieve a similar level of abstraction and re-use in the way user interfaces are built, delivered and maintained. This paper presents an HTML5 based UI framework, built on open standards but optimized and configured specifically for the needs of TV centric applications.

THE HTML5 OPPORTUNITY

Although HTML5 remains a maturing technology, the web development community has actively embraced it and most modern web browsers already support it. In our industry, there has been some speculation surrounding the video tag and the current lack of DRM. The premise here is not "HTML5 video," rather the use of cutting edge web techniques associated with establishing a user interface, built from a singular and re-usable code base which is common and consistent across a wide range of devices. For the Cable Industry, moving the user interface code into the Cloud in this way not only represents an opportunity to address a variety of devices, but additionally empowers us to embrace retail devices as well as add features and extend functionality at web-like velocity, removing the burden of code downloads and complex provisioning scenarios.

Open Source Frameworks

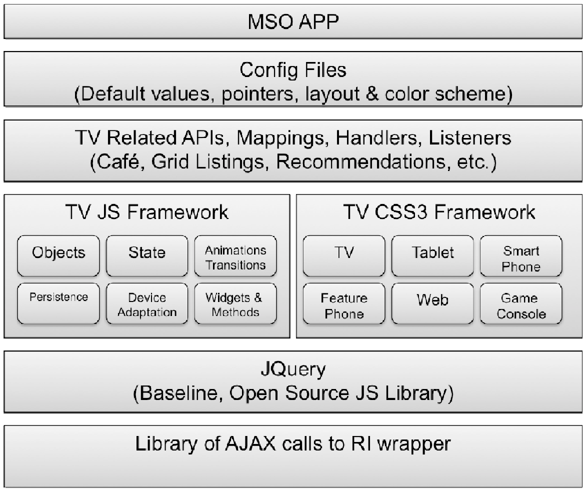
There are countless examples of extremely powerful applications, written entirely as web applications that are as rich in functionality, animation effects and behavior as desktop applications. In practice, these are written as a combination of HTML5 along with an aggressive use of JavaScript and CSS3. It is important to recognize that it is this collection of technologies, rather than HTML5 itself that enable these types of user interfaces. A variety of JavaScript and CSS3 frameworks such as jQuery or Sencha exist for HTML5 development. These typically abstract away platform idiosyncrasies, establish object and state models, provide a variety of animation libraries and generally simplify the development of a single web application to run across a variety of devices.

THE GAP

Without doubt, individual providers will seek to differentiate their brand through unique designs, features and interactivity. While each individual service provider could certainly select a given framework and develop its own, unique cross platform user interface there would be little commonality across the industry and much duplication of effort. We will all have linear listings and VOD search. We will all have cover art and DVR scheduling. Grids and a baseline set of animations are inevitable. We will share the need to support the same range of devices. Furthermore, each provider would be burdened with updating to versions of the framework and addressing new devices, screen resolutions, etc.

AN INDUSTRY FRAMEWORK

Envisioned here is a Cable Industry UI Framework. The ambition is to select among the various open source HTML5 frameworks the core aspects most beneficial to the generic needs of TV centric user interfaces. There would likely be several components involved, the particular assembly of which would constitute an MVC type construct with particular focus on the device and object abstractions required to represent "TV." This baseline component assembly would constitute the core foundation but would require an additional layer of CSS3 and JavaScript abstraction for the Cable specific UI components and underlying object model. The high level stack is represented below:


MSO APP
Config Files
(Default values, pointers, layout & color scheme)
TV Related APIs, Mappings, Handlers, Listeners
(Café, Grid Listings, Recommendations, etc.)
TV JS Framework
Objects
State
Animations Transitions
Persistence
Device Adaptation
Widgets & Methods
TV CSS3 Framework
TV
Tablet
Smart Phone
Feature Phone
Web
Game Console
JQuery
(Baseline, Open Source JS Library)
Library of AJAX calls to RI wrapper

The overriding purpose of this stack is to leverage the generalized foundations of an underlying open source framework such as jQuery and build on top of it the necessary specifics relating the Cable industry. These specifics would include such things as objects for TV listings, recommendations, actors, movies as well as standardized methods and callback routines for fetching recommendations, content searches, etc. Likewise, a variety of UI components representing things like an actual TV listings grid and animation effects such as a cover art carousel would be optimized. CSS3 style sheets for each device family or particular model would cater to the specifics needed in each rendered component. Each layer of this stack is intended to be extensible.

MSO Customization and Extensibility

Each MSO would benefit for the shared plumbing in the underlying framework. Configuration files, unique to each MSO would map to their web service endpoints, define the specific assembly of the various UI components into their presentation and provide a skinning capability via CSS3 overrides.

As new objects, event handlers or animations were envisioned and required; an MSO or third party would develop them within the overall framework. Ideally, these would be contributed back into the community such that other MSOs would benefit. There would likely, for example, be several variations of a TV listings component to choose from as well as useful extensions by way of animation effects.

Inclusion in App Stores

There is often confusion between "Apps" and "HTML5." The two are indeed different things as "Apps" are compiled, installable binaries and "HTML5" represents web pages. App stores and HTML5 are, however, perfectly compatible. There is, of course, very good reason to place applications in app stores. Most users of iOS and Android devices in particular are now familiar with app stores and this is the dominant avenue by which they are likely to search for and discover an MSO application. Applications available in these stores are compiled natively for the specific platform. In order to achieve the benefits of app store inclusion as well as the ability to re-purpose HTML5 across platforms a “native wrapper application” is written which essentially compiles a rudimentary shell for the specific platform and uses the device’s underlying web browser to render all actual user interface screens. This is, for example, the way in which Netflix develops its applications.

ADDRESSING THE BIG SCREEN

With regards to the common retail devices of today such as iOS and Android powered smartphones and tablets, laptops and PCs this framework would provide a robust mechanism to deliver a common and consistent user interface as well as minimize the associated code. The UI is, effectively, a giant web site delivered from the Cloud. Changes made to a single file would propagate to all devices and users would not need to download any updates, they would simply benefit from the new experience during their next session. This is all well and good, but to what extent could the framework be used to deliver the same experience to a STB connected to a 60-inch plasma?

Relevance to the RDK

The RDK recently introduced by Comcast includes a Webkit implementation. Webkit is an open source HTML5 compliant web browser, used in both Apple’s Safari and Google’s Chrome browsers. This provides an alternative to Java as the presentation environment on the CableLabs <tru2way> reference implementation.

This stack can be used in a master-slave in-home architecture whereby a gateway device running the full RDK stack serves as the service termination point within the home, commanding control of tuners, handling conditional access, etc. Additional devices such as laptops, tablets and smartphones can connect to the gateway and both consume tuners and receive the user interface, which is delivered as HTML5 via a web server running inside the gateway. The gateway can be “headed” meaning a television display is actually connected to it or “headless” meaning it serves as the termination point but exclusively provides the UI and services to other devices within the home. Additional outlets need only be very dumb, thin IP STBs, which run Webkit.

To be sure, this description of the RDK does not do it full justice as it is, truly, a very compelling development and much more significant than the brief description above. The point here, however, is that there is an HTML5 presentation layer available and that it can render to the big screen.

Thus, a modern HTML5 compliant web browser is available through the RDK. The RDK, however, does not provide any specific UI or further framework, just that open book upon which things could be written. As is the case with other HTML5 devices, each MSO would need to develop and maintain its own specific UI.

The proposal is to extend the RDK to include the same UI framework discussed in this paper.

GATEWAY CACHING SCHEME

The UI framework would necessarily be hosted in the Cloud. This would allow it to be used independent of RDK gateway architectures as well as ensure that the UI could be rendered outside of the home over any network. By including the framework within the RDK, however, there are additional benefits relating to caching and performance to explore.

Insofar as an RDK based gateway acts as a web server (both to itself and other devices within the home) it is, in effect, a proxy to the actual remotely hosted Cloud. Like all proxies, it acts as the source of truth from the perspective of the client. This presents potential challenges by way of ensuring the gateway is, in fact, up to date but also represents a significant opportunity to be used as a caching node within a distributed architecture. Open source caches such as Varnish could be additionally included within the RDK and would provide a tremendous performance benefit. Specifically, the gateway could be configured to proactively cache guide and VOD listings, cover art, network images, user profiles and targeted ads. This could be done relatively easily via a lazy cache whereby the gateway deferred to the Cloud and simply stored content and data as it passed it through to the client, making it locally available for subsequent requests. It could also take on a more elaborate form whereby server side algorithms proactively pushed information to the gateway, likely during dark hours and with certain targeting parameters designed towards personalization of the UI.

ADDITIONAL CONSIDERATIONS

While I believe the industry would benefit significantly from a common, shared core UI framework it still assumes HTML5 and relatively modern web browsers. What about older PCs or even fairly modern devices with limited rendering capabilities like Smart TVs or game consoles?

HTML5 does not render everywhere. Older browsers like those in many PCs or early incarnations of Smart TVs are capable of rendering simpler versions of HTML. It is necessary that the framework can degrade gracefully by recognizing these devices and rendering a simpler, less animated form of the UI. This would require a somewhat more complex abstraction than would otherwise be necessary but is perfectly feasible. Other devices, like a current Xbox, will require platform specific applications. These devices will benefit from at least a general commonality of the UI in terms of data and objects as served from backend web services, but would still require platform specific, native applications to be written. HTML5 will not currently provide a UI on every device; although it will address a wide range of current devices and it is likely Webkit will continue to proliferate to things like Smart TVs and game consoles.

The Vulgarity

More challenging to the notion of a common UI framework is the fact that there is currently very little consistency in the backend of MSOs. While most of us are embracing web services and these web services are notionally representing very similar things they are far from standardized. The grid for example, is assumed to be in most operators' UI in some form or another. The data used to populate the grid would be fetched through a web service along the lines of:

http://operator.com/apis/getGrid();

The host and specific method call, of course, would be perfectly configurable and each MSO would have their own unique endpoint. This is not a problem. The syntax and structure of the response, however, is a challenge. There are differences in semantic naming conventions as well as overall object models. The semantics of one MSO labeling HBO a "network" and another "provider" or "programmer" are somewhat easier to deal with. Structural differences in the objects or varied sets of interfaces are far more troublesome. One MSO might include actors and detailed descriptions in the getGrid() response. Another might have a secondary web service for getAssetDetails() and a third that does not include actors at all.

These backend variations are not insurmountable but they do require some additional thought. Standardization of core web services is, of course, the ideal solution. It is also possible to establish a JavaScript mapping layer within the framework, although that will likely lead to poor performance. A possible middle ground scenario would involve each MSO establishing a server side transformation layer to its existing web services.

CONCLUSION

As MSOs, we face a similar challenge in providing consistent user interfaces to a growing set of devices. Cloud based UIs allow us to more uniformly deliver and present a user interface as well as extend new features and services in a coherent and efficient manner unlike with traditional STBs. This is something we can immediately explore online and through smartphones and tablet devices and will increasingly become viable on leased CPE through initiatives like the RDK. A common, shared industry UI framework would allow us to further exploit the opportunity and reduce the individual burden of redundant web development. Such a framework could be developed based on best in breed, open source efforts from the web community but configured specifically to suit the needs of TV centric user interfaces.

SURFACES: A NEW WAY OF LOOKING AT TV

Simon Parnall, Kevin Murray and James Walker
NDS

Abstract

The rapid evolution of home display technology offers the potential for an ever-more realistic and immersive experience of media and, within a few years, we will see large and yet also unobtrusive 'lifestyle' surfaces that could cover a whole wall. In the face of such capability the obvious question is "How might the television experience evolve?" and our vision is of a better, more integrated system that provides viewers with both a collective and personal experience and which adapts to a range of sources, including metadata, from both outside and inside the home.

INTRODUCTION

The choice of type and size of television screen for the home is so often a compromise between the extremes of an exciting viewing experience when the device is switched on and the wall or corner space occupied by a dark and dull object when the device is switched off. And, when the screen is on, the size of the picture may well be inappropriate for the type of content and engagement of the occupants of the room.

Science Fiction overcomes such concerns by assuming an invisible and scalable screen – often taking the place of the wall itself, or a window or indeed in mid air. Science Fiction has also assumed an intelligent management of presented material, following the individual and assimilating and prioritizing a range of sources.

Today's mobile phones make the Star Trek communicator look somewhat bulky as advances over the years have successively removed the novelty of such a concept. In the same way today's screen, projection and graphics technologies are slowly and steadily bringing us closer to a reality of the vision of Science Fiction. In fact, we are now very nearly at the point where key aspects of this vision can be realized and could be adopted by consumers in the not-so-distant future.

Walk into a consumer electronics exhibition today and you will find many example components of this vision. There are thin-bezel screens that can be treated as tiles to create larger and larger surfaces, or glass screens that transparently reveal the wall behind when off. We already have sophisticated companion devices offering touch control and each year we are seeing ever more sophisticated gesture and voice recognition.

Our role in this opportunity space will be to create the technologies that integrate such components to produce a sophisticated and intuitive user experience that matches content and mood, and which produces pictures of an appropriate size and position for each circumstance. Furthermore the presented audiovisual content will be supplemented with additional content and so-called domotic feeds (that is material concerning the home).

In this paper, believing in the inevitability of this trend in display technologies and the opportunities this creates, we set out our vision for how the television experience will evolve, some lessons learned from our first prototype implementation of this vision, and touch on our plans for the second-phase prototype which we are currently constructing.

VISION

Our vision of the future is of a viewing environment with one or more large display surfaces. Surfaces that are a) frameless, b) unobtrusive, c) ultra high-definition and d) ambient. These surfaces can be adapted to fill or partially fill one or more walls of a room, and will co-operate to provide an integrated experience. The opportunity is to open up possibilities way beyond the limits of today's devices though:

- content comprising multiple visual elements that can be adapted spatially and temporally, freeing the user from choosing a single element, or the system from having to impose overlays;
- shared, co-operative usage of the surfaces, with connected companion devices becoming personal extensions;
- supporting connected applications and services operating in a more streamlined, integrated manner, reflecting and effecting changes in viewer engagement in TV content;
- dynamic adaption to, and control over, the environment the surface finds itself in – such as physical size, resolution and the room in which it sits (e.g. adapting to the wallpaper or controlling the lighting); and
- introducing domotic content into the TV display in a sympathetic manner.

Based on this vision, a prototype was constructed and demonstrated at both IBC 2011 and CES 2012. This prototype has a single surface occupying most of one wall and a photograph of this is shown in figure 1. This shows a single surface constructed from six screens and one of several companion devices that may be used simultaneously to control and interact with the system.

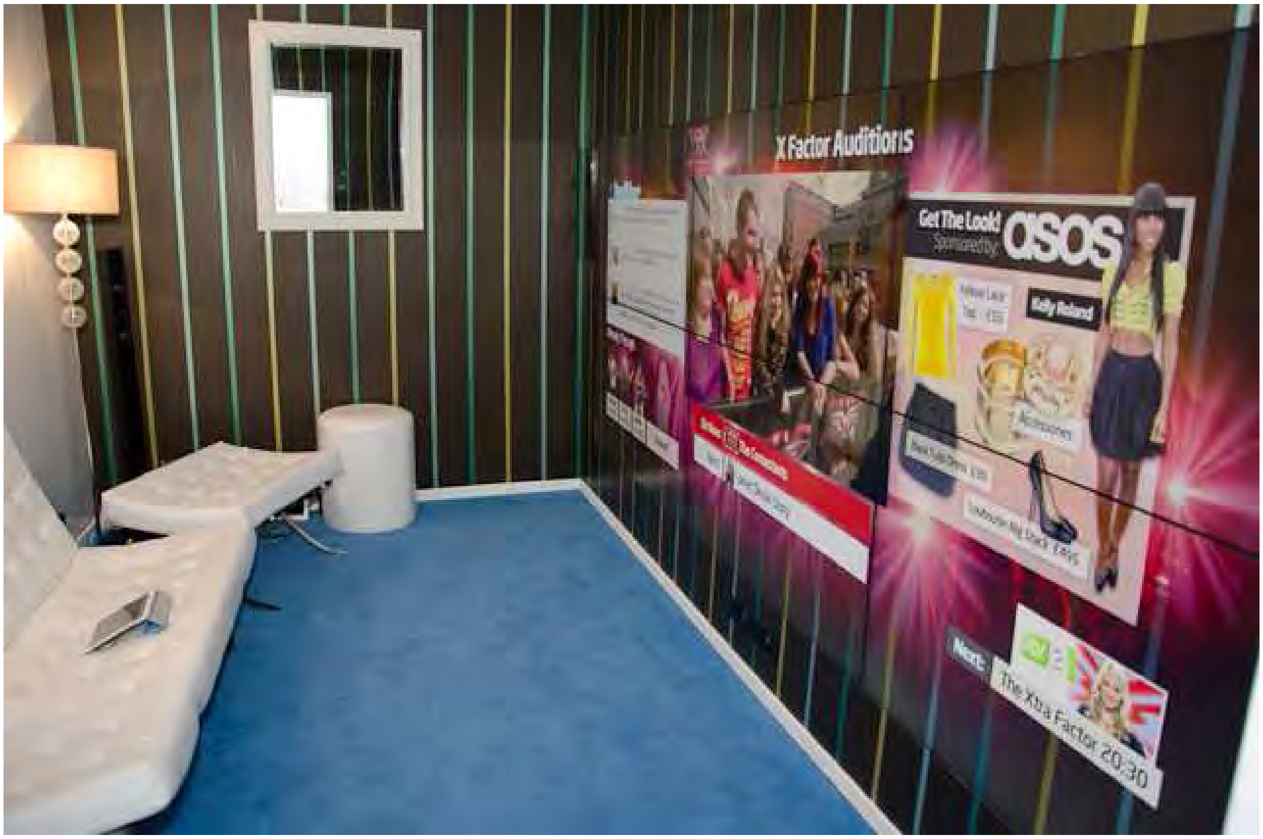


Figure 1: Prototype System

IMMERSION

Many programs have a natural flow and pace – points at which the viewer or viewers are extremely immersed and engaged in the content. Examples of this may be a critical part of play in a sports game, a news story of direct relevance or a very dramatic scene in a soap. Likewise there may be times of lesser immersion or engagement. Examples of this may be waiting for players to take their positions, an uninteresting news item or a section of the soap that is recapping past happenings. In these areas of lesser immersion, the viewer's interest may naturally be taken by other related items, such as the current scores in related games, the next news story or what is being said about the soap by their social contacts.

In our system we have introduced the concept of 'immersion'. Immersion is key to the way that the surfaces are used and the way that the content is presented on them. Put simply, the more immersed in the content the viewer is, the greater emphasis that is placed on the core video, and the less immersed they are the more emphasis comes to be placed on related content which may then be introduced. This related material could be social media, advertising, program graphics, additional material, or virtually anything.

Examples of high and low immersion are shown in figures 2 and 3 respectively, which are screen captures taken from our prototype system. In figure 2, we see how the video roughly shares the surface with other social, voting and advertising graphics and content sources during the scene setting and build-up to the main performance. By comparison, figure 3 shows the high immersion example where the program in figure 2 has moved on to the main performance, and the related items have been removed, and the video increased in size and prominence.

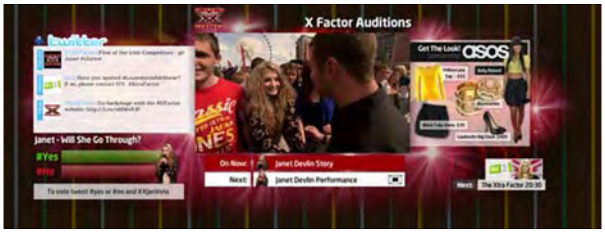

X Factor Auditions
twitter
Janet - Will She Go Through?
#Yes
#No
Get The Look! asos
On Now: Janet Devlin Story
Next: Janet Devlin Performance
Next: The Xtra Factor 20:30

Figure 2: a low immersion example

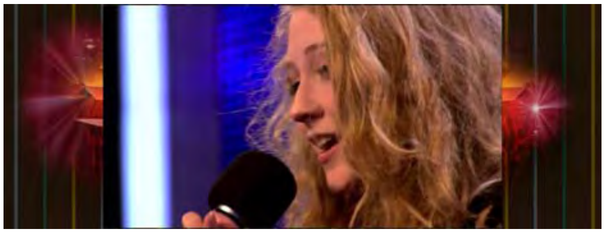


Figure 3: a high immersion example

In our prototype system, immersion is controlled in two ways – via "broadcast metadata" (as was used for the examples above) which indicated the broadcasters expected level of immersion, and also via a slider in the companion device which allows the user to modify the immersion (both up and down) as they wish. Clearly other mechanisms could also be employed, such as audio or video analysis of the room and the viewers, but the prototype shows that these two simple mechanisms work very effectively.

TECHNOLOGICAL MOTIVATORS

Displays

Display technology is continually improving. We have seen that relentlessly the average screen size is increasing year by year, as evidenced by (3). But there are two key technological changes which directly relate to our vision.

Firstly, screen bezel sizes are getting smaller. Our prototype system uses professional monitors with 5mm bezels, but LED backlit consumer displays are approaching similar, or better, bezel sizes and

OLED offers the prospect of a bezel width of near zero. Even with today's widths there is the real option of creating large ultra high definition surfaces out of tiled arrays of inexpensive displays.

Secondly, whilst still in the research laboratories, transparent displays which naturally allow the underlying environment to show through are starting to be developed. These would trivially allow the blending of displays into the room environment.

Video Content

We are also starting to see the first indications of the next jump in resolution beyond HD with the advent of 4K – both in displays and in content. At the same time as this higher resolution content is arriving, the importance of lower resolution content is not diminishing, whether from archives, citizen journalists or from challenging remote locations. Thus it is becoming hard to just assume that any content will look acceptable on any display size.

Non video Content

Outside the display arena, we are seeing ever more related data sources, from social media through games to dedicated websites. In the interconnected world, these are a crucial part of the entertainment experience, but today we are faced with the dilemma of either destroying the television experience by placing graphics over the video, or taking the viewer away from the lean back world of television into the very different and highly-interactive world of the internet.

BREAKING THE SCREEN BOUNDARIES

Today's television makes the basic assumption that "the display is always filled". Thus, video will fill the display, regardless of the size of display, quality of the video, or the resulting impact of an oversized face or object; and it also effectively does only one (main) thing at a time.

With larger, higher resolution display surfaces this implicit behavior and more can be challenged. Content need no longer necessarily fill the display surface, and the display surface can simultaneously be used for many different components.

In turn, these new capabilities mean that the traditional means of laying out video and graphics can be challenged. For instance we might:

- share the display between the content of more than one viewer, helping to make the TV a shared focal point rather than a point of contention;
- 'unpack' the constituent elements that are composited by a broadcaster in post-production, presenting these alongside the 'clean' audio-visual (AV) content, leaving it un-obscured. Obvious examples include digital on-screen graphics such as tickers, banners and sports statistics. To enable this, the composited elements would need to be delivered separately alongside the clean AV and then rendered in the client;
- 'unpack' all of the contextual assets that are composited in the Set-Top Box (STB), such as interactive applications and multi-screen content (e.g. multi-camera sports events);
- present contextually relevant online content alongside the video, for example, relevant web content, social comments (such as twitter hash-tags for the show), relevant online video etc;
- enable navigation and discovery user interfaces to be presented alongside video, going beyond today's 'picture-in-guide' presentation;
- present personal content, which whilst not directly related to the main television content, may still be desirable to end users to be seen on screen. Examples would

include personal social feeds, news feeds, images, discussion forums etc;

- present domotic content, such as user interfaces for in-home devices and systems, which can include video feeds from devices such as security cameras, door entry systems and baby monitors; and
- integrate visual communications, such as personal video calls, noting these may sometimes be used in a contextual way e.g. virtual shared viewing experiences between homes.

Thus, the way the TV experience takes advantage of the surface is by continuously managing a wide range of content sources and types that are combined appropriately for presentation.

Real Object Size

The tradition of a television picture scaling up to fill the display means that an object is effectively displayed at an unknown size. With this assumption broken, it now seems realistic to allow an object to be displayed at its real size, regardless of the display (as displays report their size though the standard connectors). For instance, in advertising it could be interesting to show just how thin the latest phone really is, just as is possible in print media today.

Content Opportunities

In the same way that the composition has always assumed a need to fill the rectangle, so has the creation of video content – which has followed the model of filling the proscenium arch of classical theatre. The proposed systems can offer new opportunities to the content creator.

One simple example of this is shown in figure 4. Here, the movie trailer is blended into the background to give the appearance that it tears its way through the wallpaper, dramatically conveying the unsettling nature of the promoted movie.

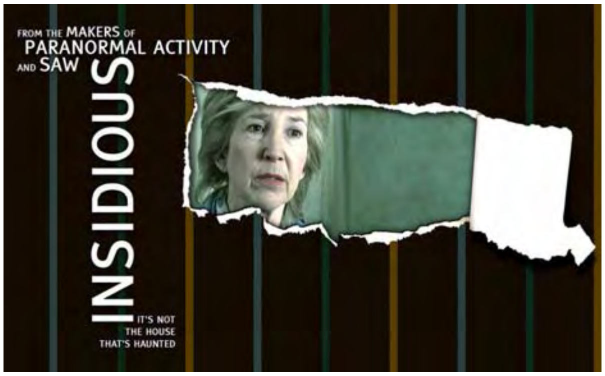


Figure 4: Non-rectangular content

There are numerous other areas where this technique opens up new opportunities. For example:

- editing could become more subtle with gentle fades, and several scenes can co-exist for longer and with less interference;
- content need no longer be fixed into a given size – if portrait content is provided from citizen journalists, then it can be displayed naturally in that form; and
- multiple synchronized videos could be used, in a fashion made popular in TV series such as 24, but without any requirement for their relative placement.

Implicit in this capability is the requirement to support an “alpha plane” style functionality that can be used both to describe arbitrary shapes and to allow for blending of the content into the background. This is, of course, not new and techniques such as luma and chroma keying are well known both in the professional head-end market place as well as supporting functionality in DVD and BluRay media. However, bringing this functionality into a traditional broadcast chain would represent a new usage.

A COMPANIONABLE EXPERIENCE

The growing importance of companion devices (tablets, phones, laptops etc) to the modern TV experience cannot be understated. Such devices permit us to construct an experience which is, at the same time, both collective (involving everyone in the room) and yet personal (allowing each person to interact with the various elements as they wish).

The companion device is key and integral to our prototype experience – and interactions with the companion device are directly connected with what is seen on the main surface(s). This is achieved through several means:

- The companion devices are able, within constraints, to adapt the content on display, including adding or removing components or re-arranging the layout. An example of this is interface is shown in the iPad screen capture of the web-browser in figure 5, where, for instance, the display can be re-arranged by dragging around the icons representing the parts of the content displayed on the surface.

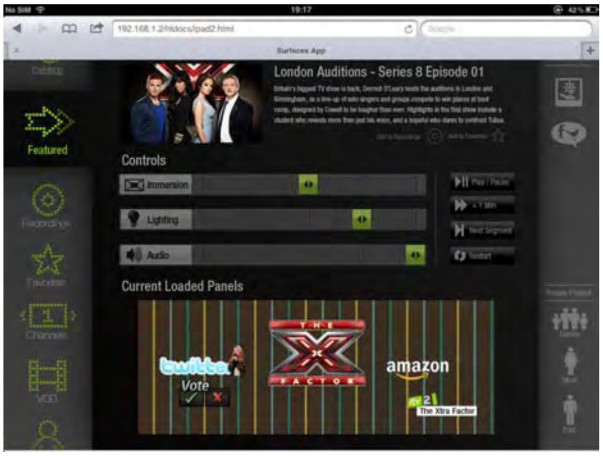


Figure 5: A Companion Application Interface

- Interactions, such as voting or feedback is done on the companion device, but this directly feeds back into the graphics displayed (in addition to the normal feedback one would expect).
- Control over the level of immersion. Although, as discussed earlier, a change in the level of immersion can be triggered through broadcast data and sensors in the room, the companion device is fundamentally able to control the final immersion experienced. In the prototype, as shown in figure 5, this is managed through a slider.

This approach results in interactions with the companion device that end back at the main display surface(s), rather than just with the companion device itself. For example scores from a game played by the whole family during a show could be displayed on the main surface.

A SURFACES SYSTEM ARCHITECTURE

The prototype system constructed to explore our vision was built using a single, six-output computer (an AMD Eyefinity graphics card in a powerful PC) with software that was itself built on standard HTML5 technologies (e.g. javascript and CSS transitions) in functionality largely contained within a standard browser. This approach enabled a fast and flexible development and exploration of the principles. Whilst the HTML-5 toolset proved to be an excellent platform, the use of a single six output graphics card places fundamental limits on scalability, the number of display surfaces that can be supported and, of course, on cost.

In our current work we are moving from the architecture of our first prototype. We are doing so because we will be using multiple display surfaces within a single room, and exploring how these can be combined for the presentation of a single entertainment experience, and co-operate to support multiple simultaneous entertainment experiences (e.g. the big game and the soap).

To achieve the required flexibility in the number of surfaces, scalability, cost and content presentation dynamism, we are developing a more advanced architecture, based around several concepts, including:

- rendering the graphics and video on more than one independent device;

- utilizing synchronization between the rendering devices, such as used in SAGE (1), but tailored for the specific use cases we are tackling;
- a separation of layout policy issues and rendering issues; and
- a single layout with a “world view” of the entire set of surfaces in use.

A high-level overview of the current architecture is shown in figure 6. This shows two separate surfaces, each driven by its own client. These clients then interact with the layout and synchronisation server(s) to ensure a consistent experience across the surfaces. In addition, the diagram shows that the audio is driven from only one surface, a deliberate choice to simplify the architecture.

<u>Synchronization Architecture</u>

It is important to be able to synchronize content spread between different clients. In a more traditional broadcast architecture, this would theoretically be possible using mechanisms such as the PCR values contained within a transport stream, but our approach does not assume either a direct transport stream feed to each client, or even that the content is made available in transport streams (e.g. it could be streamed over HTTP using any one of a number of mechanisms such as HLS or Smooth Streaming).

Instead, we have chosen to synchronize to a master audio playback clock on the main audio output. Where broadcast content is being consumed, there are many techniques that can be used to match this clock to that of the live broadcast content. This master audio clock is then replicated and synchronized via the synchronization server to other clients that are involved in playing back synchronized media.

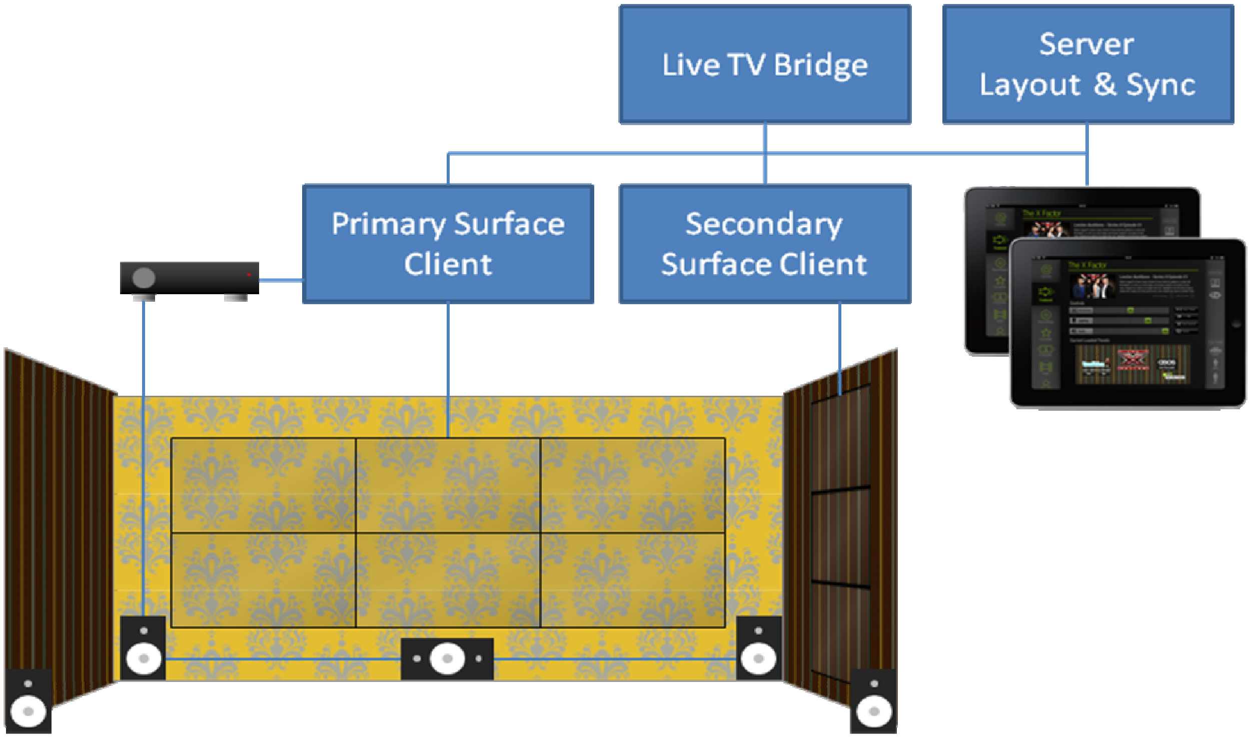


Figure 6: A New Surfaces Architecture

Our initial experiments with this architecture have shown that it appears to provide a reliable synchronization between different clients to a level that is acceptable for lip synchronization. Further experiments are underway to characterize and measure the accuracy that can be achieved.

Audio Architecture

Normally, audio is decoded and presented with simply a level control. However, in our proposed system the audio architecture becomes more complex than in a traditional approach, with various audio processing operations becoming an essential part of the overall architecture.

The most obvious audio processing requirement is positioning. From the proposed layout of surfaces in figure 6, it is clear that the secondary surface is not between the main speakers, and so any video that is presented on this surface with synchronized audio needs to have this audio repositioned. This repositioning needs to be dynamic, for instance as a video is moved from the primary to the secondary surface, the audio should be moved in synchronization. And, given the potential size of a surface, repositioning of the audio is desirable even when the content is moved within a surface. For example a video that occupies only the left third of the surface should have its sound stage correctly placed.

Earlier we discussed the concept of immersion, and how the video element of the experience can be balanced against other components to reflect the levels of interest both through a program's length and of a given viewer or viewers. This has a direct mapping to processing of the audio. Whilst the volume levels are one key part of this, this is best when combined with controlled compression – a reduction of the dynamic range of the content so that quieter parts become louder and the louder parts become quieter. Such processing allows the volume to be reduced in a fashion that retains access to the quiet sections of the content.

Much of the required functionality described above appears to be relatively easy to implement in the proposed Web Audio APIs that have recently become available on various platforms (2). This should make implementing the required audio architecture within an HTML5 environment relatively straightforward, and this work is currently underway.

Layout

One final component of the architecture deals with the layout of the media items to be displayed. Earlier in this paper we discussed how content typically packed together can be transmitted in an unpacked form, with the chosen and relevant components then laid out by the Surfaces system when the content is finally presented to the viewer. This process is not the highly constrained process we are used to where precise locations can be given for each item and, as the surfaces to be used might well be substantially different in each viewing environment, the process must be very flexible, and it is this flexibility that is an interesting challenge.

One aspect of the required flexibility comes from the number of inputs to the layout process to control what is displayed. These come from the local environment such as the range, sizes, locations and properties of the surfaces available and the immersion level of the viewer, and from the broadcaster, such as the list of potential components, their relevant priorities and a potential preferred immersion level. It is the layout engine that balances these inputs and selects a suitable set of components to display and locations for them.

In addition to the "what" of the layout is the "how', the appearance. More specifically, certain components may need to be adapted to the environment into which they are to be placed. For instance, if the room has white walls and the content item is white text, some means of making the text legible must be provided automatically. More generically, the design of an item should be able to adapt to the predominant background colors of the environment.

This introduces challenges at several levels that go beyond that of most current content presentation designs, such as may be found in many websites. Firstly we need an adaptive description of the requirements a broadcaster desires beyond those commonly in use today, and beyond even those of responsive web design (4). Next, we need a mechanism that can quickly and efficiently resolve these requirements in the face of a collection of local inputs. Finally, and perhaps most challengingly, we need the content producers and designers to understand that their content can and will be presented in many different ways, and a complete control over this presentation is potentially very counter-productive to the viewer's engagement.

CLOSING THOUGHTS

Our thinking started when considering the possibilities that the display industry will be offering in just a few years' time when the black boxes in the corners of out rooms disappear and unobtrusive, frameless, ultra high definition ambient surfaces take their place. In exploring the opportunities this technology will offer we have come to consider how content is presented, and the way in which its various components (current and future) will be assembled for the viewer. We have come to an appreciation of the way in which control and interaction with such an experience can work both in a personal and collective manner. And, in contrast to the 'lean forward' experience of today's connected TV we have seen how the 'lean back' experience of Surfaces requires a sophisticated automatic layout control engine.

As we have explored function, so we have explored form, and the PC based solution for a first stage demonstration now begins to give way to a believably scalable and cost effective hardware and software architecture.

It is often commented that the role of television in our lives has changed dramatically as other devices have fought for our time and won our attention. And yet, families and groups still wish to spend time together, sharing space and switching between personal and collective experiences. A developed television experience which embraces this truth, and which invites immersion and interaction at appropriate levels, must surely be for our industry a goal worth aiming for. Surfaces is, for us, a vehicle to explore this space and we are excited by the future we see before us, and the reaction we have received. The future is not one where the medium is marginalized, but a future in which people will truly find a new way of looking at TV.

REFERENCES

(1) High-Performance Dynamic Graphics Streaming for Scalable Adaptive Graphics Environment, *SuperComputing 2006, November 11-17 2006.* http://www.evl.uic.edu/files/pdf/SAGE-SC06-final.pdf

(2) Web Audio API, Chris Rogers, W3C, https://dvcs.w3.org/hg/audio/raw-file/tip/webaudio/specification.html

(3) Of Large LCDs, Unused Fabs, and Projector Killers, Pete Putman, Display Daily, April 9th, 2012. http://displaydaily.com/2012/04/09/of-large-lcds-unused-fabs-and-projector-killers/

(4) Responsive Web Design, Ethan Marcotte, 2011, ISBN 978-0-9844425-7-7, http://www.abookapart.com/products/responsive-web-design

Leveraging Time-Based Metadata to Enhance Content Discovery and Viewing Experiences

Ben Weinberger
Digitalsmiths

Abstract

Today's consumers have more choices than ever before for video entertainment and viewing devices. But with this explosion of choice has come complexity. Finding engaging entertainment has become a time-consuming and frustrating endeavor, resulting in decreased engagement and satisfaction.

The key to overcoming this discovery challenge lies in rich, time-based metadata. By creating time-based metadata at the production level and leveraging metadata-driven solutions to build best-in-class search and recommendation applications, stakeholders can create additional value at every stage of the video content lifecycle.

This paper discusses the superior metadata technologies and how they can be applied to solve today's toughest discovery challenges.

MAKING DATA RELEVANT

To enable state-of-the-art video search and recommendation tools, you need state-of-the-art data. And you need the ability to access, integrate and normalize data from disparate sources.

CREATING THE DATA SET

From dialog to set design, anything about a scene can be tagged. Efficiently and accurately creating this rich time-based data requires advanced algorithms for facial recognition, scene classification, speech recognition, natural language processing, closed-caption time alignment and ad break detection.

To understand the granularity of the resulting metadata, it is worth drilling down to the details. Each video (an asset) may have metadata associated with it. This asset-level metadata can be human authored or directly imported from 3rd party sources. Each asset in turn consists of a series of contiguous scenes. Each scene is named and is time-bound. This is a human authored process.

Metadata is tracked throughout an asset. Individual metadata elements, such as an actor, location, rights, score, etc., are associated with a scene (container) and specific frames of video (location within an asset). A metadata track may be subdivided into subtracks. For example, an objects track could be defined for tracking specific branded elements in an asset, such as cars. The process to create metadata tracks and subtracks is both automated and human authored. For example, once an actor is tagged by name, facial recognition software can tag the appearance of the actor throughout the asset.

A segment refers to a time-bound portion of the asset containing a metadata element. The segment can also have metadata associated with it (referred to as segment attributes). All metadata is automatically tracked to a frame-level timestamp, providing the ability to display the exact frame in which the metadata track or element occurs. The depth of metadata that can be associated within a single frame is shown in Figure 1, a frame from *Spiderman 3*.

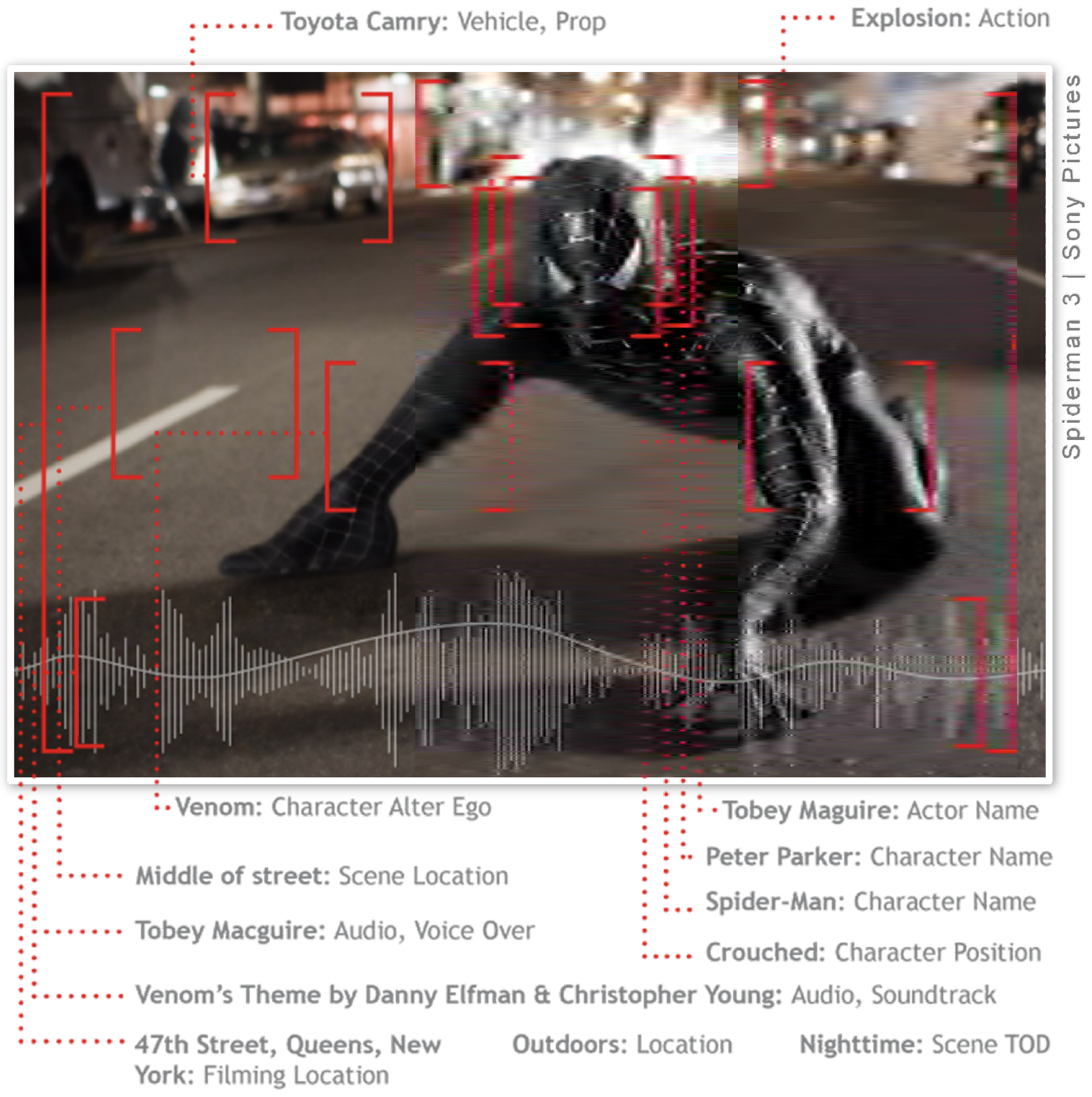

Toyota Camry: Vehicle, Prop
Explosion: Action
Spiderman 3 | Sony Pictures
Venom: Character Alter Ego
Middle of street: Scene Location
Tobey Macguire: Audio, Voice Over
Venom's Theme by Danny Elfman & Christopher Young: Audio, Soundtrack
47th Street, Queens, New York: Filming Location
Outdoors: Location
Nighttime: Scene TOD
Tobey Maguire: Actor Name
Peter Parker: Character Name
Spider-Man: Character Name
Crouched: Character Position

Figure 1: All metadata created around individual frame

Given the depth of information that can be created, it is preferable to tag the data at the time of production when much of the information is known by those closest to the creation of the asset. For example, in Figure 3, the filming location is tagged as 47th Street in Queens with a Toyota Camry in the shot. If the video were tagged by someone other than the production unit, this level of information might be lost.

Information depth and granularity provides for much stronger ability to search and find the specific video segment you are searching for. While tagging at the production level is ideal, post-production tagging also yields deep, rich information that goes beyond the typical descriptions of title, major actors, and plot.

The ideal metadata solution should also support real-time tagging for live content. With the 92nd PGA Championship as an example, time-based metadata was married to the scoring feed to video-enable the Leader Board and the Scorecard. This enhanced viewing experience allowed fans to click directly on the golfer or hole to replay specific shots, increasing viewer engagement and creating new sponsorship and advertising opportunities.

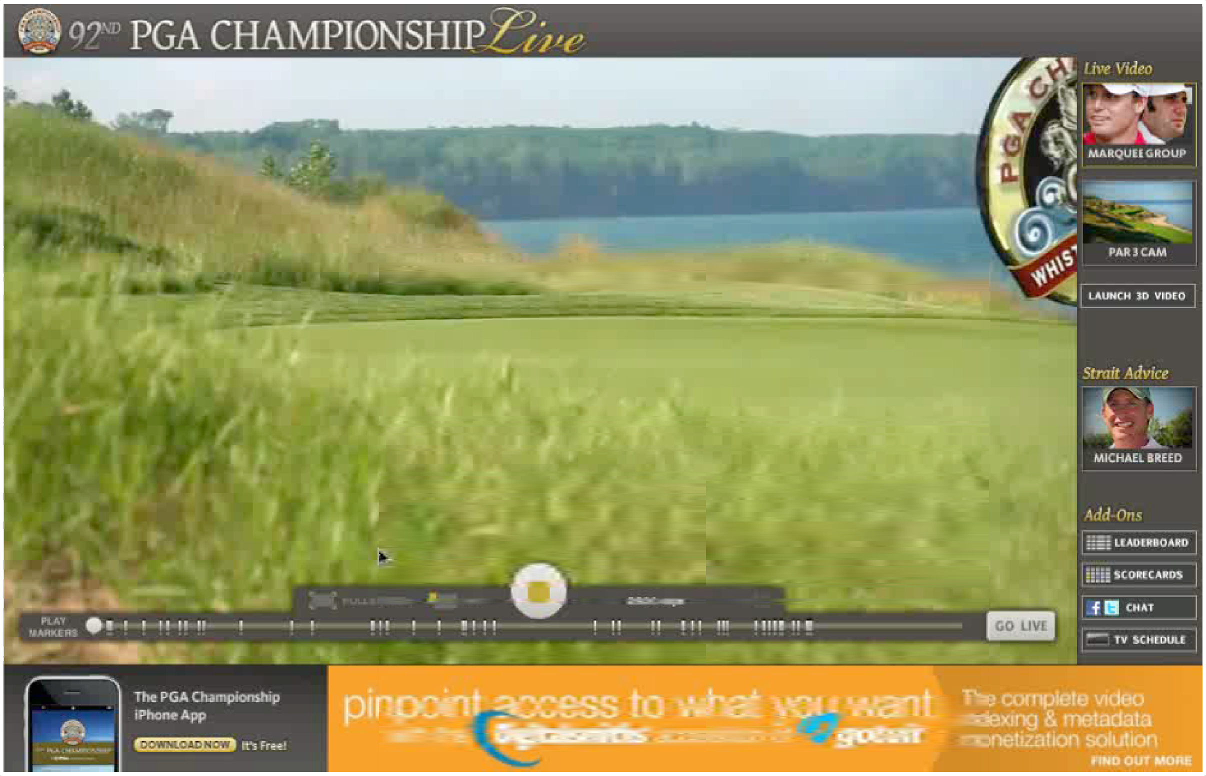


Figure 2: Metadata-driven live viewing experience

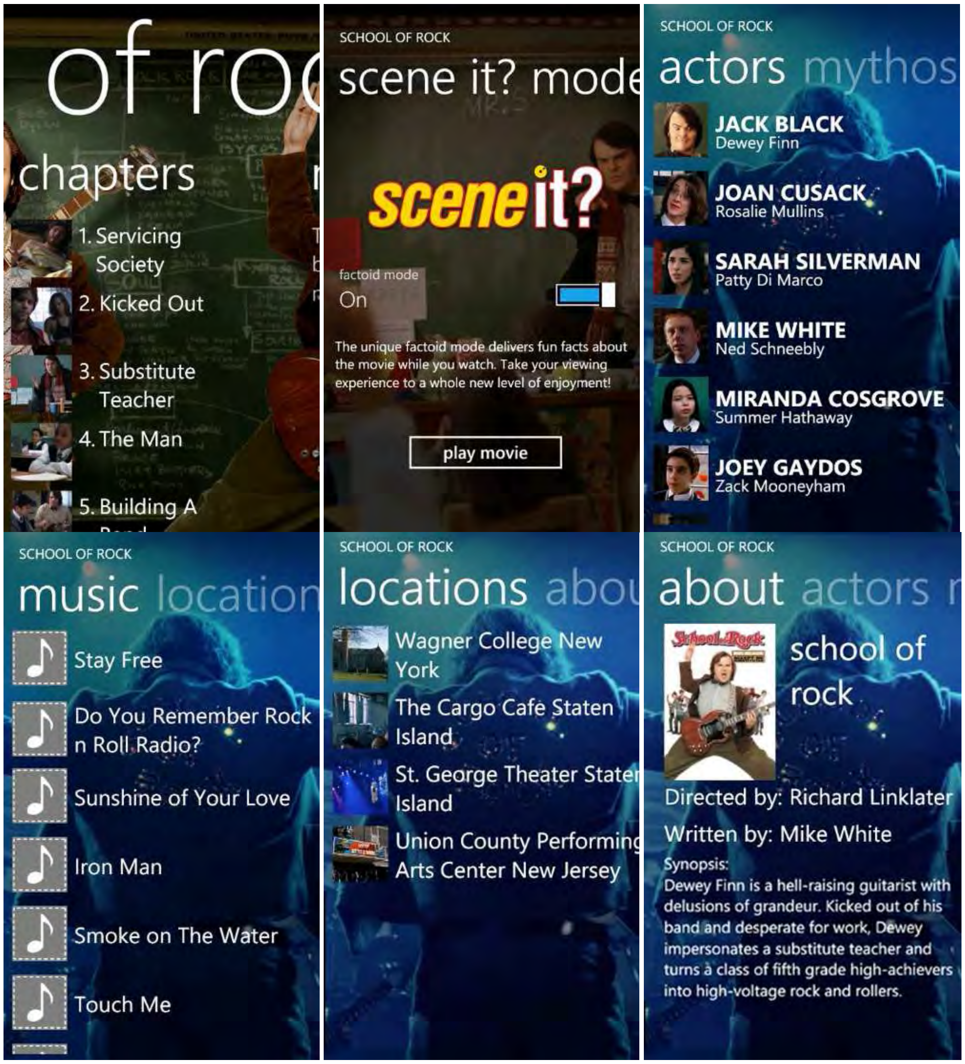
The full potential of creating rich metadata sets is achieved by creating large libraries of tagged assets. This deeper level of intelligence opens the door to unparalleled search and recommendation functionality and accuracy.

For example, while it may be common knowledge that Tom Cruise danced in *Risky Business*, a metadata-driven search for “Tom Cruise dancing” will also deliver *Tropic Thunder*, a movie in which Tom Cruise briefly danced but is not even listed in the credits.

ACCESSING THE DATA

Creating the dataset is the primary task; however, equally important is providing access to the dataset. As stated, the full power of a search and recommendation engine is found in the size of the libraries. But there will be multiple libraries available as they are created by production teams, post-production studios, and third parties (well after post-production). The search engine must be able to interface with multiple libraries to maximize the value of rich metadata and to provide the ability to recommend videos across production houses, studios, and movie libraries.

ENHANCING THE USER EXPERIENCE



The unique factoid mode delivers fun facts about the movie while you watch. Take your viewing experience to a whole new level of enjoyment!

Directed by: Richard Linklater

Written by: Mike White

Synopsis:
Dewey Finn is a hell-raising guitarist with delusions of grandeur. Kicked out of his band and desperate for work, Dewey impersonates a substitute teacher and turns a class of fifth grade high-achievers into high-voltage rock and rollers.

Figure 3: Second screens

With the growing popularity of connected devices, actionable, accurate metadata is needed to deliver the interactive applications that users have come to expect.

Being able to create a unique look and feel for the user that meets the specific need of the licensee is equally important to the success of a video search engine. For example, the needs of the PGA PC application differs widely from the phone application (Figure 3) that leverages time-based metadata from *School of Rock.*

The same dataset could drive a connected television application, a set-top box application or a tablet application.

SUMMARY

Efficiently creating rich time-based metadata around each frame of a TV show, movie or live event requires advanced algorithms for facial recognition, scene classification, speech recognition, natural language processing, closed-caption time alignment and ad break detection.

The ideal discovery platform then integrates and normalizes scene-level metadata with rich 3rd party data from disparate sources to create a deeper level of intelligence around video content, enabling unparalleled accuracy and personalization in search results and recommendations.

With these time-based metadata solutions, stakeholders can develop best-in-class enhanced discovery and viewing experiences that drive engagement and better monetization of video assets.

All images are copyrighted by their respective copyright owners.

BITS, BIG SCREENS, AND BIOLOGY

Dr. Robert L Howald and Dr. Sean McCarthy
Motorola Mobility

Abstract

High definition television (HDTV) has dramatically improved the consumer viewing experience. As such, despite its hunger for precious bandwidth, increased HD programming continues to be a key industry objective. However, as evidenced by the exhibits and technology on display at the Consumer Electronics Show (CES) this past January, today's HD, is just a step in the progression of consumer video. In addition, today's video processing and delivery is also undergoing significant change. In this paper, we will explore new generations of HD video and the innovative enabling technologies that will support them. We then roll-up the components and project their impact to network architecture.

Specifically, we will consider advanced formats, beyond just emerging 1080p60 (blu-ray) HD. Recognizing the expectation of a very long HFC lifespan, we will quantify how QFHD (aka 4k) and even proposed "Super Hi-Vision," or UHDTV, stack up for consumer services. We will assess practical and human factors, including those associated with HD-capable second screens, such as tablets. We will quantify physiological variables to the optimization of the video experience, such as personal through immersive screen sizes, viewing environment, and high frame-rate television.

On the encoding side, we discuss H.265 High Efficiency Video Coding (HEVC) against its own "50%" objective. And, just as we considered human variables associated with the user experience, we can take advantage of human biology to deliver the highest perceived quality using the smallest number of bytes. Using new signal processing models of the human visual system (HVS), the ultimate arbiter of video quality, a unique combination of bandwidth efficiency and high perceived video quality can be achieved. This technique, called Perceptual Video Processing (PVP) will be detailed, and its impact on video quality and bandwidth quantified.

In summary, we will evaluate long-term network prospects, capturing the potential trajectory of video services, innovative encoding techniques, emerging use cases and delivery, and shifting traffic aggregates. In so doing, an enduring network migration plan supporting multiple generations of video and service evolution can be projected.

INTRODUCTION

Decades of broadband growth and an ever-increasing range of video services has given operators a sound historical basis upon which to base future growth trends, which is critical for business planning. Service growth and subscriber satisfaction with the portfolio of media delivered to them provides new revenue opportunities. To meet these demands, key decisions must be made for upgrading hubs, homes and the access networks. The prevailing MSO approach has been a very successful pay-as-you grow approach, capitalizing on technologies as they mature and as consumer demands require. This has worked extremely well because of the latent HFC capacity, which incrementally was mined as necessary by extending fiber, adding RF spectrum, incorporating WDM optical technologies, and delivering digital and switched services.

As IP traffic has grown aggressively, video quality has also moved ahead, albeit at a more gradual pace. The appetite for HD is being fed at this stage of the evolution, but the HD lifecycle itself has only just begun. As cable systems deliver 720p and 1080i formats, the ability to support and deliver 1080p quality already exists in the CE and gaming worlds. Flat panel televisions continue to become larger, more capable, and lower cost. Their size already is breaching the boundary of where a "normal" viewing distance would benefit from a yet higher quality video signal. 2k and 4k (Quad Full HD or QFHD) formats have entered the conversation and the demonstration rooms. These formats are being explored and seemingly will inevitably lead to a new service offering. Beyond QFHD is the Ultra-High Def (UHDTV, 4x QFHD)) format, or Super Hi-Vision, invented by NHK in Japan in the mid-1990's. At that time, it was foreseen by NHK to be a consumer format in the 2030 time frame.

EVOLUTION OF VIDEO SERVICES

Spatial Resolution

With the advent in particular of HDTV, development of QUAD HD (2x in each dimension) and UHDTV, the video and CE industries have a strong understanding of the relationship among resolution required, screen size, and viewing distance.

Just as visual acuity is measured and referenced to object sizes at defined distances, the display size and placement relative to the viewer is a key piece of the resolution requirement equation. Figure 1 is a straightforward way to see how these factors interact [40] based on recommendations provided by multiple professional organizations, home theatres experts, and retail manufacturers. Generally, for a fixed resolution (linear trajectories on the plot), a larger screen size is best viewed further away. For a fixed screen size, higher resolutions are best viewed by sitting closer to allow for the full benefit of the increased detail on the display. Finally, for a fixed distance from the display and the higher the format resolution, the larger the screen size should be.

As a simple example, a 50 inch screen, if viewed more than 20 ft away or greater, will begin to lose the benefit of HD at 720p, and provide an experience more akin to Standard Definition 480p. Sitting too close, such as 5 ft away on a 100" 1080p screen, threatens quality due to distinguishing of pixels. This chart thus also explains the increased pixel count of UHDTV based on a 100" display recommendation and wider viewing angle (closer).

The guidelines come from different organizations and retailers, and while they tend to cluster around similar recommendations, they are not in complete agreement. This is generally due to the varied perspectives of the organizations, such as, for example, what sells more TVs. The range of recommendations varies from about 1.5x-2.5x of display size for viewing HD content, with the lower end corresponding to 1080 resolution.

The recommendations are also correlated to an assumption about visual acuity as it relates to the ability to resolve the image detail. They are also associated with viewing angle considerations. For example, the recommended optimum fields of view are given as about 30° (SMPTE) or 40° (THX) in the horizontal plane. In the vertical plane, simpler guidelines are designed around avoiding neck strain, and so describe maintaining at least a 15° vertical field of view. The maximum recommended, beyond which neck strain is a risk, is a 35° viewing angle.

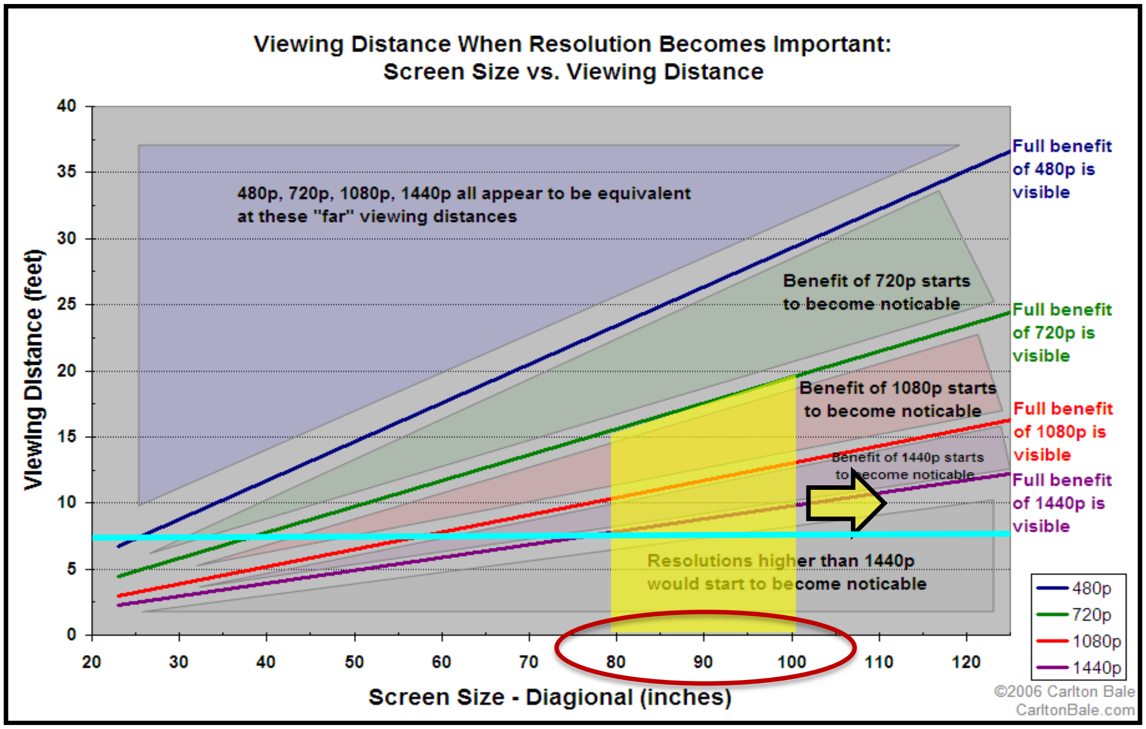


Figure 1 – Screen Size, Viewing Distance, and Spatial Resolution

Let's ponder modern display capabilities. Consider the bottom right corner of Figure 1, shaded yellow. A typical viewing distance in the home today is about 7.5-9 feet, which certainly has been driven in part by historical screen sizes. It is not surprising for anyone who has visited a big box retailer recently that flat panel screens are available now at ever-increasing sizes, such as those shown in the shaded yellow range of Figure 1. At 7.5 feet distance (light blue line), "only" a 55" screen could show perceptible benefits for resolutions better than 1080p (light blue line crosses red line). At 80", flat panels have fully breached the 2560x1440 resolution threshold, sometimes referred to as Extreme HD (4x 720p HD resolution) in the gaming world. The next stop beyond this is QFHD at 3840x2160p. Based on this figure, there is potential viewing value for this format screen size and larger.

Note that UHDTV, was viewed as a 100 inch screen, but also viewed at only about 1 meter (3.3 feet). The intent was to generate the feeling of immersion. Studies by NHK concluded that feelings of discomfort often associated with immersive viewing such as IMAX level off with screen size at a certain point. In the case of UHDTV, the angle at which this occurs is for 80 inch screens. Therefore, a screen fully 100 inches is not expected to present an increased probability of discomfort, but yet yields the level of immersion and video quality desired in the experience.

Now consider Figure 2. Not only do larger primary screens translate into the need for better spatial resolution, our secondary screens also have gotten larger, simultaneously more portable, *and* capable of high quality video such as HD. The explosion of tablets has put an entire new generation of high-quality video capable screens literally at our fingertips for deployment virtually anywhere relative to our viewing perspective.

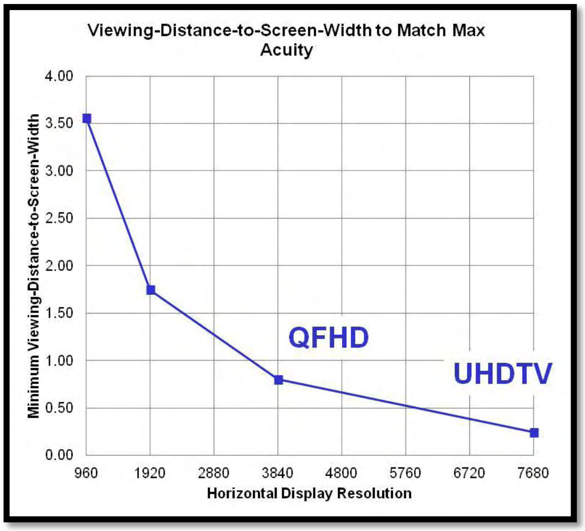


Figure 2 – Screen Size, Distance, and Resolution – Mobile Viewing

In the figure above, it is easy to see how the 1920x1080 resolution can be improved upon for reasonable viewing conditions. For a 10” Motorola Xoom tablet, for example, if the screen is about 17” away, its spatial resolution can be perceptibly improved with a higher resolution format. It is not hard to envision this scenario, for example, on an airplane or with a child in the backseat of a car.

The case for full QFHD or UHDTV on the 10” tablet would be difficult to make based on this figure without some other accompany variables. Nonetheless, clearly screen sizes and portability in this case are combining to change the paradigm of mobile viewing environments far, far away from the legacy of QVGA resolution at 15 frames per second.

Dynamic Resolution

The term “dynamic resolution” refers to the ability to resolve spatial detail of objects in motion. The 30 Hz (interlaced), 50 Hz, and 60 Hz frame rates have origins in AC line rates, and thus are not scientifically tied to video observation and testing. They simply exceeded what was known at the time about 40 Hz rates causing undesirable flicker. Most early analysis on frame rate was to ensure that motion appeared realistic (seamless), as opposed to a sequence of still shots. There was less focus on eliminating other artifacts of motion, such as smearing effects. Yet, as spatial resolution has improved, temporal resolution has not. Interlaced video itself is a nod to the imbalance of motion representation – exchanging spatial resolution for a higher rate of image repetition to better represent motion than a progressive scanning system of the same bandwidth.

The above frame rates have since become embedded in tools and equipment of the production, post-processing, and display industries, and so, with respect to frame rate, we are hostages to the embedded infrastructure and scale of change that would be required to do anything else. As such, HDTV standards today are based on the 60 Hz interlaced or progressive frame rate. It has been suggested [1] that with larger and brighter displays of higher resolution, the frame rates in place based on practical implementation limitations of the 1930s era ought to be reconsidered, of course while recognizing a need to maintain some level of backward compatibility.

As displays become larger and of higher resolution and contrast, the challenges to effectively displaying motion increases, because the edges to which movement is ascribed are now sharper. What is optimal? There is not a firm answer to this question. The human visual system streams video continuously in a physiological sense, so the question is around the processing engine in the brain. Various sources describe tests where frame rates of 100-300 fps show perceived improvements compared to 60 fps [1, 41]. The difficulty of performing this type of testing – content and equipment – limits how much has been learned. There are potentially positive encoding implications to these higher frame rates. Intuitively, more rapidly arriving frames ought to be consistent with better coding efficiency, as it is likely that there is less variation frame-to-frame.

We will not consider any changes to frame rate beyond interlaced/30 to progressive/60. But this is a variable to keep an eye on as larger screens and live sports viewing collide.

Formats and Bandwidth Implications

High Definition has had a major impact on the industry in multiple ways. On the positive side, the Quality of Experience (QoE) delivered to the consumer is tremendously improved. HD has enabled cable operators to strengthen the service offering considerably. And, like the DVR, HD has very much the "once you have it, you never go back" stickiness to it.

Conversely, while HD services certainly act to increase revenue, they also create a significant new bandwidth burden for the operator. Whereas 10-12 standard definition (SD) programs can fit within a single 6 MHz QAM bandwidth, this number drops to 2-3 HD programs in a 6 MHz QAM. This loss in efficiency is compounded by the fact that HD today represents a simulcast situation – programs delivered in HD are usually also transmitted in the SD line-up. For all of the subsequent analysis, we will base MPEG-2 SD and HD program counts per QAM on averages of 10 SD/QAM and 2.5 HD/QAM. Obviously there cannot be a fractional number of programs n a QAM slot. The 2.5 assumes that for MPEG-2 encoded HD, an operator may chose 2 or 3 in a 6 MHz slot based on content type, and the QAMs are equally split with both.

Perhaps most worrisome with respect to bandwidth is that current services are basically HD 1.0. Only the first generation of formats are deployed – 1280x720p and 1920x1080i. The improvement over SD is so vast that it is easy to wonder what could possibly be the benefit of even higher resolution. However, as we showed in Figure 1, it is relatively straightforward to show how the continued advancement of display technology at lower and lower costs, in particular consumer flat panels, leads to reasonable viewing environments where resolution beyond 1080-based systems would be perceptible. In addition to the flat screen scenario, similar analysis in Figure 2 showed similar conclusions for"2nd screen" tablet viewing. All modern tablets support HD quality viewing. Coupled with realistic use cases that are likely to include close viewing distances, higher resolution scenarios may add value here as well.

We will consider the effects of two next generation video formats on the HFC architecture's ability to support them – QFHD and UHDTV. QFHD has had prototype displays being shown since approximately 2006 and has entered the conversation as the big box retailers now routinely display 80" screen sizes. Analyst projections have placed QFHD in the 2020 time frame for deployment timeframes. A comparison of these two formats against standard HD, and other formats, is shown in Figure 3 [19].

Note that QFHD works out to 4x the pixel count as 1080 HD, and UDHTV works out to 16x the pixel count. In each case, there is the possibility of higher bit depth (10-bit vs. 8-bit) as well, which translates into more bits and bandwidth. We will assume this is taken advantage of in the latter case only. As a result, we arrive at the following set of potential scaling factors, without any assumptions about possible latent compression efficiencies on top of conventional gains projections for new display formats.

SD to:

1080i – 4x
1080p – 8x
QFHD – 32x
UHDTV – 160x

It is of course premature to know precisely what compression gain may be available for advanced formats, since these enhanced formats are in their infancy. For now, we will rely on the resolutions to correlate with bandwidth, with the 8-bit to 10-bit pixel depth for UHDTV and the frame rate for p60 (doubling the information rate) as the only other variations quantified.

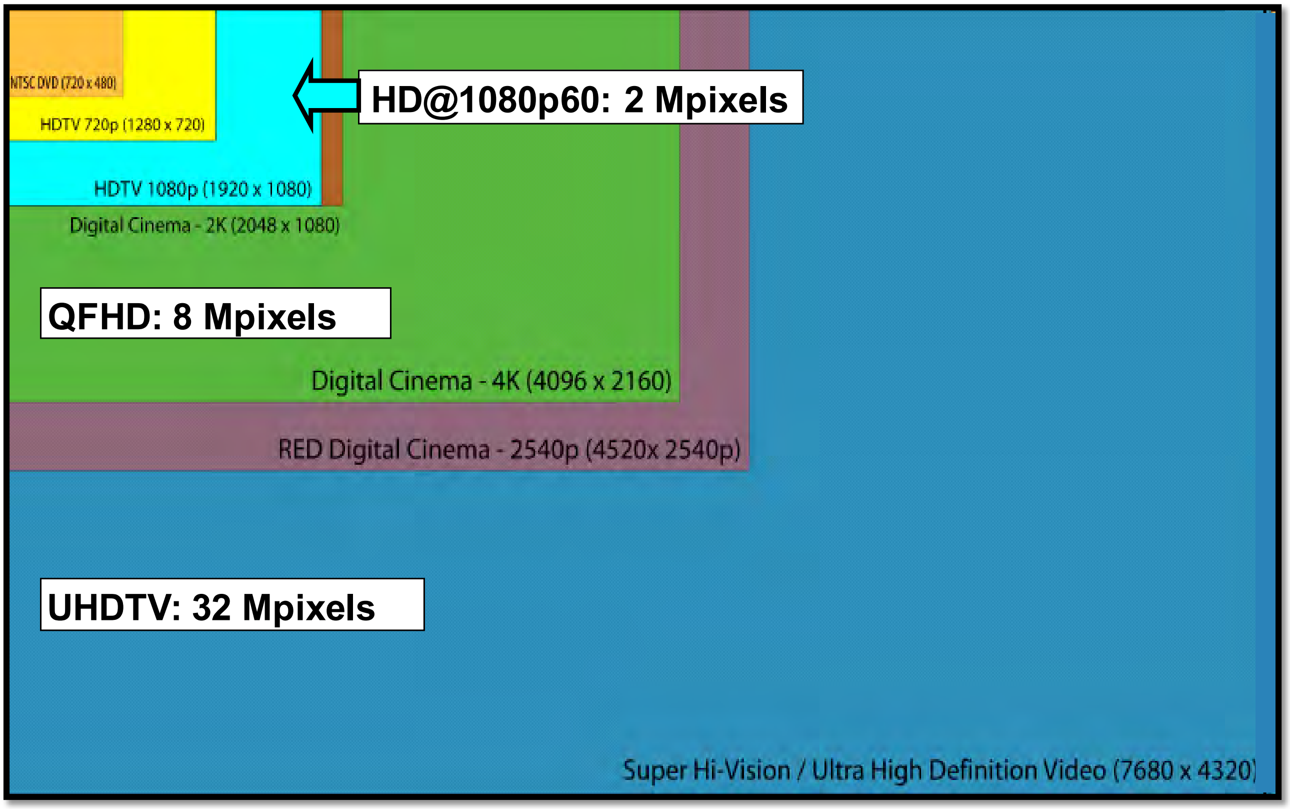

NTSC DVD (720 x 480)
HDTV 720p (1280 x 720)
HD@1080p60: 2 Mpixels
HDTV 1080p (1920 x 1080)
Digital Cinema - 2K (2048 x 1080)
QFHD: 8 Mpixels
Digital Cinema - 4K (4096 x 2160)
RED Digital Cinema - 2540p (4520x 2540p)
UHDTV: 32 Mpixels
Super Hi-Vision / Ultra High Definition Video (7680 x 4320)

Figure 3 – Beyond High Definition Formats Comparison

VIDEO COMPRESSION – STILL ON THE MOVE

It may not seem like long ago, but it is nearly 10 years since the Advanced Video Coding (AVC) [27, 30] international standard was completed in 2003. AVC – also known as H.264 and as MPEG-4 part 10 – has been a remarkable success. It has enabled IPTV and HDTV to take hold and grow commercially. It has enabled Blu-Ray video quality at home. And it has been powering new models of delivering digital video over the internet. AVC and its equally successful predecessor, MPEG-2, are expected to continue to play an important role in the digital video economy for many more years, but they'll be soon joined by a new entrant to the international standard portfolio -- High-Efficiency Video Coding (HEVC) [6, 10, 18, 22, 31, 32].

In many ways, HEVC is a close cousin to AVC. Both are of the same genus of hybrid block-based compression algorithms that incorporate spatial and temporal prediction, frequency-domain transforms, data reduction through quantization, and context-sensitive entropy encoding. Where HEVC stands out is in the wealth and sophistication of its coding tools, and in its superior compression efficiency.

Figure 4 captures the state of the set of core MPEG compression standards in the context of their lifecycle.

Efficiency

First and foremost for any compression standard is the simple question of how much more efficient it will be at compressing video streams. HEVC aims to double the compression efficiency of its AVC predecessor. AVC itself doubled compression efficiency compared to MPEG-2. That means that a consumer quality HDTV program delivered using 16 Mbps today with MPEG-2 (like a cable TV QAM channel supporting 2-3 HD channels) would need only about 4 Mbps using HEVC. As we will see in subsequent analysis, it also means that we might reasonably expect to be able to deliver Super HD (4kx2k) over the bandwidth we use today for regular HDTV, enabling yet another generation of enhanced video services.

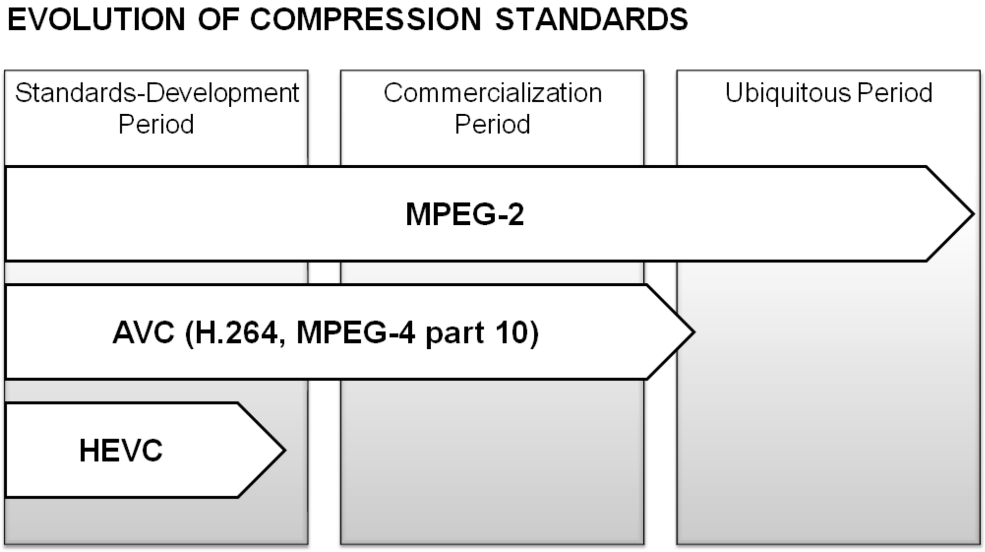

EVOLUTION OF COMPRESSION STANDARDS
Standards-Development Period
Commercialization Period
Ubiquitous Period
MPEG-2
AVC (H.264, MPEG-4 part 10)
HEVC

Figure 4 – State of Current Video Compression Standards

At the onset of the HEVC development process, the ITU-T and MPEG issued a joint call for proposals [33]. Twenty-seven proposals were received and tested in the most extensive subjective testing of its kind to date. Scrutiny of the proposals entailed 134 test sessions involving 850 human test subjects who filled out 6000 scoring sheets resulting in 300,000 quality scores. The conclusion [34] was that the best proposals yielded 50% bit rate savings compared to AVC at the same visual quality. The potential for another 50% gain launched the Joint Collaborative Team for Video Coding (JCT-VC), and HEVC development formally got underway.

In late 2011, JCT-VC reported another series of compression-efficiency tests [35] using objective rather than subjective methods. Those test showed that HEVC had not yet hit the 50% mark with scientific certitude, but was very close and had excellent prospects for additional gain. Table 1 shows the results from objective tests comparing HEVC to AVC High Profile. The tests were conducted using various constraints to examine the efficiency of HEVC for several important potential use cases: broadcast such as over cable, satellite, and IPTV that need random-access features to support fast channel change and trick modes, low-delay applications such as video conferencing, and intra-only compression that uses only spatial prediction within each frame of video to support applications such as contribution-quality video.

The bit rate savings listed in Table 1 represent the point at which HEVC and AVC High Profile produce the same peak-signal-to-noise ratio (PSNR). Though PSNR can be a sometimes inaccurate metric of subjective video quality [25, 39], the data in Table 1 are consistent with the earlier extensive subjective testing [35] and are thus expected to be valid predictors. The data of Table 1 represent overall average performance of the various HEVC use cases for a wide range of resolutions from 416x260 to 2560x1600 [13]. It is clear from Table 1 that HEVC substantially outperforms AVC High Profile.

Other results from the JCT-VC report on objective tests are displayed in Table 2. These results provide insight into how HDTV might differ from mobile devices with regard to HEVC efficiency.

Table 1 - Compression Efficiency of HEVC compared to H.264/MPEG4 part 10 AVC

NOTE: Relative Compression Efficiency is calculated as 1/(1 - Bit Rate Savings)

Example Use Case	Encoding Constraint	Bit Rate Savings	Relative Compression Efficiency
Broadcast	Random Access	39%	164%
Video Conferencing	Low-Delay	44%	179%
Contribution	All-Intra	25%	133%

.

Table 2 – Current Compression Efficiency of HEVC for HDTV and Smartphone

Display	Width	Height	Bit Rate Savings Compared to AVC High Profile	Relative Compression Efficiency
HDTV	1920	1080	44%	179%
Smartphone	832	480	34%	152%

Table 2 points out that HEVC's gains for HDTV resolutions are greater than for smartphone resolutions. They are also greater than the average over all random-access results shown in Table 1. These results hint that HEVC may become relatively *more* efficient for emerging resolutions beyond HDTV, such as 4K (4096 x 2048) and Ultra HD (7680x4320). If such proves to be the case, market forces might help accelerate deployment of HEVC as a way for operators and display manufacturer to offer new beyond-HD options to consumers.

It is important to note that both MPEG-2 and AVC improved significantly as they moved from committee to market. Even today, MPEG-2 and AVC continue to become more efficient as competition pushes suppliers to find new ways of improving quality and squeezing bits. The same dynamic is expected with HEVC. It should experience additional improvements, rapidly, when it emerges from the standardization process, followed by long-term, continuous honing through commercial competition. It is common in industry circles to project that HEVC will achieve its targeted doubling in compression efficiency – it is simply a matter of time.

For purposes of our subsequent analysis of HFC capacity and services, we will assume that HEVC will indeed ably achieve its 50% goal when commercially available.

Under the Hood

Some of the AVC efficiency gains were the result of new coding techniques such as context-adaptive binary arithmetic entropy coding (CABAC). Yet a large part of the gains came from making existing tools more flexible. Compared to MPEG-2, for example, AVC provided more block sizes for motion compensation, finer-grained motion prediction, more reference pictures, and other such refinements.

HEVC also gains its performance edge by using newer versions of existing tools. One of the most significant enhancements is that the concept of a macroblock has morphed into the more powerful concepts of Coding Units (CU), Prediction Units (PU), and Transform Units (TU).

Coding Units are square regions that can be nested within other Coding Units in a hierarchical quad-tree like manner to form an irregular checkerboard. The advantage is that smaller Coding Units can capture small localized detail while larger Coding Units cover broader more uniform regions like sky. The result is that each region in a picture needs to be neither over-divided nor under-divided. Avoidance of excessive segmentation saves bits by reducing the overhead of signaling partitioning details. Judicious subdivision saves bits because the details within each terminal Coding Unit can be predicted more accurately.

Prediction Units extend the "just-the-right-size" coding philosophy. Prediction Units are rectangular subdivisions of Coding units that are used to increase homogeneity – and thus predictability – within Coding Units. If a particular Coding Unit encompasses a region of grass and a region of tree bark, for example, an encoder might attempt to arrange the boundary between Prediction Units so it matches the grass-bark boundary as closely as possible. Together, Coding Units and Prediction Units create a quilt of more homogeneous patches that are easier to compress than regions of heterogeneous textures.

Transform Units are also subdivisions of Coding Units. The objective is to position and size Transform Units such that a picture is subdivided into mosaic of self-similar patches when viewed from the frequency domain. One of the dominant visual artifacts in MPEG-2 and AVC is the distortion that sometimes occurs near sharp edges and around text. This artifact is a result of performing a transform and quantization across radically different textures on either side of the edge. In HEVC, Coding, Prediction, and Transform Units work together to more precisely decouple the textures flanking the edge thereby reducing spillover and avoiding the visible defect.

Other coding tools also get a makeover in HEVC. Intra prediction supports many more directional modes to discriminate the angular orientation of lines, edges, and textures more exactly. Inter prediction has improved interpolation filters to yield higher quality motion vectors. And there are less costly ways of sending motion vector information to the decoder. HEVC also gains at least one new kind of loop filter targeted at improving both objective and subjective visual quality.

Not all the enhancements in HEVC are incremental. HEVC will be capable of delivering high-quality video to every conceivable device from the size of a thumbnail to a wall-filling 8k x 4k display in wide-gamut color palette that rivals the natural world. That is an opportunity for unparalleled consumer experiences.

Commercialization -- Profiles & Levels

Compression standards of the caliber of HEVC are complex amalgams of sophisticated algorithms and protocols. In the past, specific subsets of capabilities and features of MPEG-2 and AVC were organized into Profiles with Levels to aid commercial adoption and facilitate interoperability between vendors. It would be unsurprising if HEVC also adopted a family of Profiles, but at the moment the HEVC Committee Draft [37] specifies only Main Profile, which roughly corresponds to AVC High Profile.

Within the HEVC Main Profile, the Committee Draft does specify a number of Levels. Each Level corresponds to a maximum picture size (in terms of number of samples) and maximum pixel rate for the luma component. From these constraints, it is possible to indicate the minimum Level that would correspond to various consumer devices, as we do in Table 3 for smartphones; HDTV on tablets and flat panels at home; and next-generation beyond-HDTV displays.

Table 3 - How HEVC Main Profile Levels Might Correspond to Various Displays

	Example Format	Width	Height	Frame Rate	Minimum Level
Smartphones	QCIF	176	144	15	1
	CIF	352	288	30	2
	480p	854	480	30	3
	QHD	960	544	60	3.1
HDTV	720p	1280	720	60	4.1
	1080p	1920	1088	30	4.1
				60	4.2
Beyond HDTV	4K	4096	2160	30	4.2
				60	5.1
	Ultra HD	7680	4320	30	6
				60	6.1

Note that most smartphones and sub-HD resolutions would be supported starting at Levels 1 through 3, depending on the picture size and frame rate. Note that any Level above the minimum Level could be used. HD resolutions would be supported starting at Level 4. Beyond-HD resolutions would require at least Level 5 & 6 with one interesting exception. Super HD 4k x 2k resolution at 30 frames per second shares Level 4.2 with 1080p 60 frames per second. It may turn out that operators will be able to leverage Level 4.2 in the future to provide consumers with both 1080p60 sports content and Super HD 4k film content (24 frames per second).

Next Steps

The process of earnest creation of HEVC began in 2010 with a Call for Proposals (CfP). There have now been nine JCT-VC meetings in which approximately 200 attendees per meeting created and debated over 2000 input documents. In **February 2012**, JCT-VC issued a complete draft of the HEVC standard called the Committee Draft [37] which will be refined over the coming months. The Committee Draft also serves as a starting point from which to explore development of commercial HEVC products. The Final Draft International Standard is scheduled to be made available in January 2013 for formal ratification.

HEVC is well on its way. And, as we shall see in the next section, it will be an essential component of future advanced video services for cable operators, based on what we are able to project today for service mix, spectral constraints, and likely migration strategies.

TRAFFIC AND SPECTRUM

Dynamics of the Shift to IP Video

While video resolution affecting bandwidth requirements presents an enormous capacity challenge, it is not the only variable driving spectrum use. In addition to bandwidth growth of the video itself, the nature of the traffic aggregate being delivered is changing as well. There are many variables in play, virtually all of which are driving towards increasing unicast delivery of video content:

- More content choice
- Time-shifting
- Trick play expectations

- Network DVR (nDVR),
- Video capable IP device proliferation (tablets and smartphones)
- Shrinking service groups

And, of course, over-the-top (OTT) viewing from web-based content providers is already unicast delivery.

As a result of these shifts, the gains typically afforded by multicast capability, or bandwidth reclamation gains associated commonly with SDV architectures, begin to evaporate. Consider Figure 5 [23]. On the right edge of the curve, we can see by comparing the DOCSIS channel count required for delivery of unicast compared to multicast that for a large group of active users and predominantly linear content, there is significant, exploitable gain. This converts to important bandwidth savings. This has been the lesson of SDV widely deployed in HFC networks today. However, these deployment advantages are based upon the content choice and the size of service groups of the time. Today, as node splits occur, the growing use of a variety of IP clients consuming video, increased choice etc., the operating point on the curve shifts to the left.

The crosshairs in the figure (60% penetration x 60% peak busy hour viewing on a 500 hhp node) represents a reasonable operating point in a system outfitted with 200 HD and 200 SD programs available as switched services. There is clearly much less gain at this point, suggesting only a modest savings in exchange for the complexity of multicast. Some optimization steps may be taken to most efficiently allocate spectrum, but with an eye toward simplicity of architecture as well. This approach is shown in Figure 6 [23]. This diagram illustrates the concept of broadcasting the very popular content to take advantage of programming where simultaneous viewing is likely to occur regularly, optimizing use of bandwidth while maintaining simplicity in the architecture. Analysis in [23] suggests that the vast majority of gain, around 80-90%, occurs in approximately the first 20 programs.

Thus, a combination of broadcast and unicast may be the end result of an IP Video system weighing the tradeoffs of efficiency and complexity. The modest loss of efficiency of "all unicast" in the figure is recovered through the use of a small tier of broadcast services. And, as service groups continue to shrink, there will be virtually no bandwidth efficiencies lost at all. This is illustrated in Figure 5, for example, for the 80 active IPTV viewers.

Next Generation Video Formats: Parallel Characteristics to IP Video

The dynamics commonly associated with 2nd screen viewing may also come to pass in the next generation primary screen video world. There is a large permutation of video formats for mobile viewing, being usurped today by high quality formats. The likely similar dynamic to emerge for primary screens is simply that new formats will get introduced well before other formats are retired. Historically, this would suggest a need to simulcast formats to ensure all customers have their video needs served based on what formats they can support on the TV sets in their home. With more formats arriving, and an overall accelerated pace of change, this could create a bandwidth Armageddon given the nature of the advanced formats relative to bandwidth consumption. However, as we shift into the IP Video world today built around 2nd screen compatibilities, we are developing and deploying tools for discovery and delivery of a large permutation matrix of formats and protocols based on the different capabilities and interfaces of IP client devices.

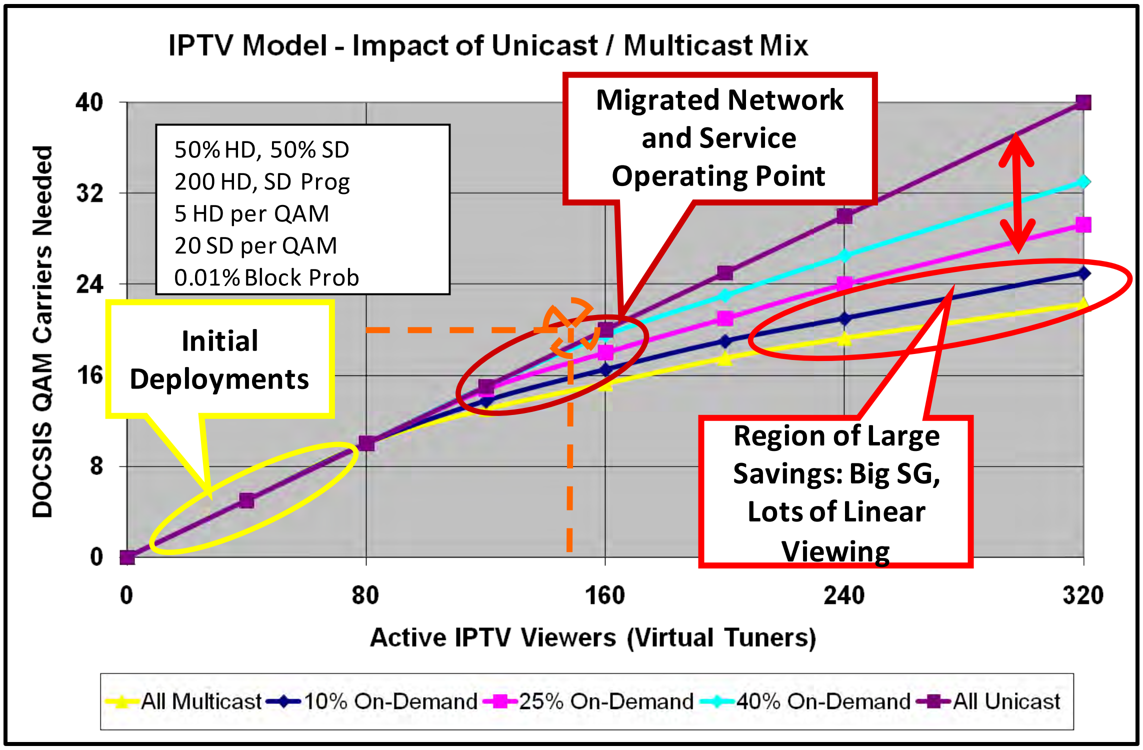


Figure 5 – IP Video Shifts the Spectrum Allocation Methodology

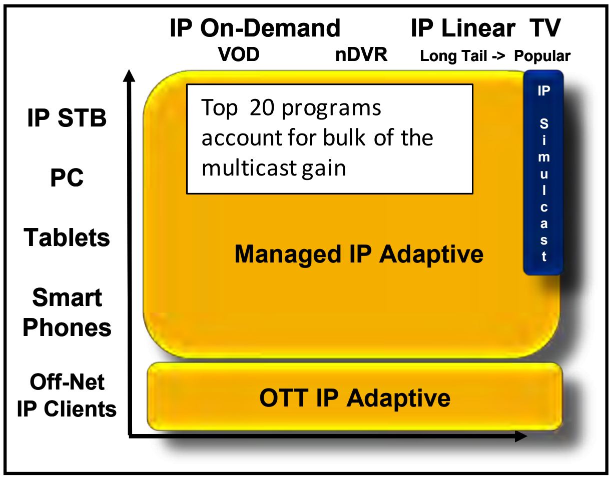


Figure 6 – Optimizing IP Video Delivery

This same dynamic could occur in the future with new high-resolution formats and smart TVs, with the only difference being that the process will take place with respect to discovering and adaptation to *primary* screen capabilities. The intelligence required is being built today to serve those 2nd and 3rd IP screens. By the time, for example, QFHD is a video format scaling in volume, the migration characteristics driving traffic to nearly all unicast will have taken place. As such, primary screen format discovery will be timely for keeping simulcast requirements at bay.

The model that we will assume as we assess the network implications is one of a small set of broadcast (conservatively quantifying with 40 total broadcast programs), with all other traffic as unicast. We will assume that the remaining traffic for video – the video unicast – is inherently captured in the traffic projections as part of 50% CAGR on the downstream. It may, in fact, be precisely what the CAGR engine of growth *is* for IP traffic over the next decade. A contrasting view would be to project HSD growth at 50% CAGR, but add to this video traffic aggregates representative of video service rates of an aggregate [13].

NETWORK IMPLICATIONS

It is quite simple to illustrate a network capacity problem in the face of increasing video quality and resolution, which directly translates into more bandwidth required. In Figure 7 we find the intersection of traffic growth, video services, and time in order to help guide MSO decision timelines. The trajectories moving upward from left to right show a commonly assumed Compound Annual Growth Rate (CAGR) of 50% over a period of ten years offset with two breakpoints over the course of the decade where a (perfect) node split takes place.

While the HFC available capacity in the downstream is over 5 Gbps when considering the highest order modulation profile currently utilized (256-QAM – the yellow horizontal threshold) it is of course not all available to support IP traffic today. The vast majority of today's spectrum is set aside for video services. Figure 7 charts the growth of IP services, but also quantifies the setting aside of spectrum used for video services. These video services that are the moving target that we are looking to quantify here. The bandwidth set aside for video services is subtracted from the 870 MHz capacity to identify the threshold for when the IP traffic would exceed the available spectrum to support it. These thresholds are the horizontal lines on Figure 7.

Four thresholds are shown bounding the available capacity over the course of 10 years. The first, baseline case (red) identifies the available spectrum for data services growth if the video service offering is made up of 60 Analog carriers, 300 SD programs (30 QAM slots), 50 HD programs (20 QAM slots), and 8 VOD slots. The math for this distribution of broadcast and VOD is quite simple: 60+30+20+8 = 118 slots consumed for video services, leaving 18 slots for DOCSIS.

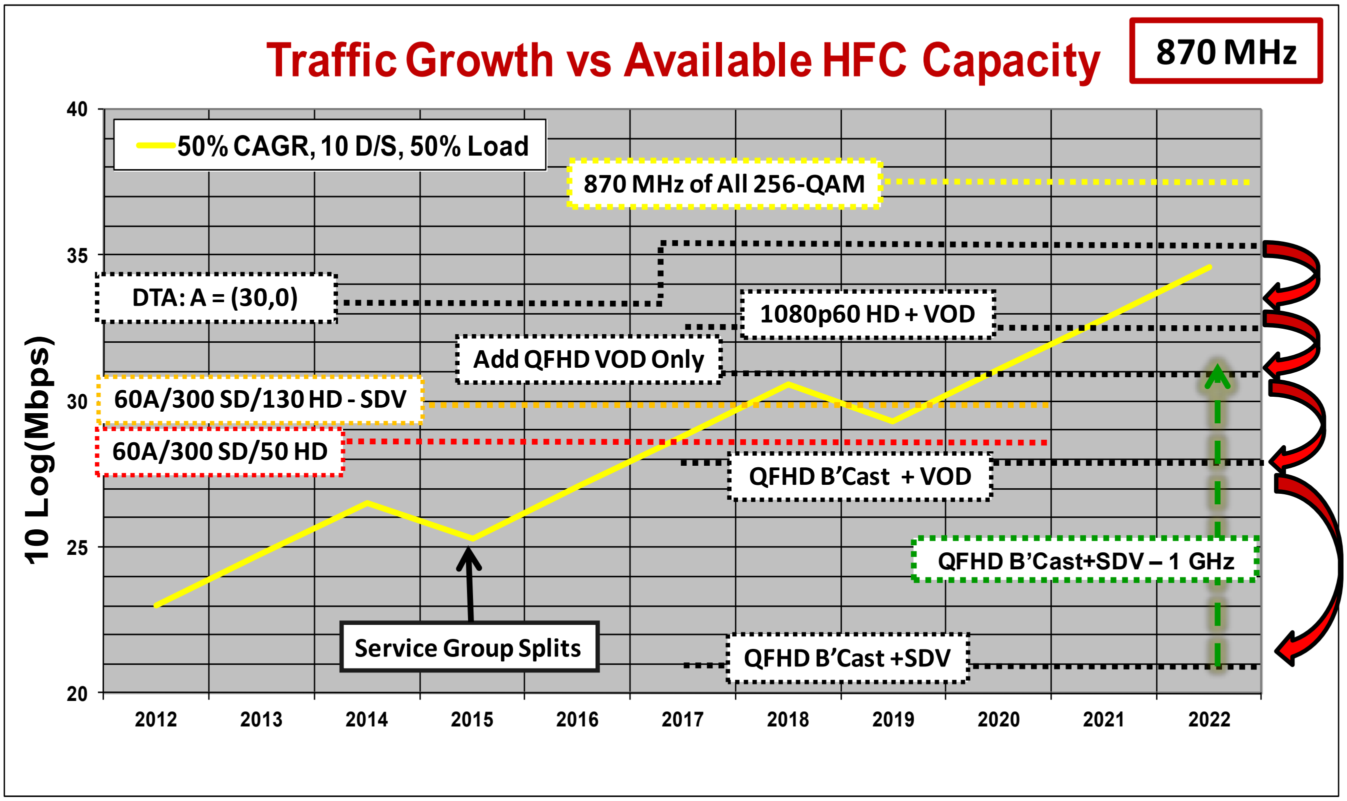


Figure 7 – New Resolutions Project to Massive Spectrum Management Concerns

Under an assumption that today's downstream DOCSIS carriers consume 200 Mbps of capacity (50% peak busy hour usage of 10 deployed downstream slots), then this video service architecture supports IP traffic growth through the year 2016, assuming there is one service group split along the way.

The orange threshold identifies the available headroom for IP growth if Switched Digital Video (SDV) is deployed, and the SDV achieves 3:1 gains for both SD and HD. Also, the HD program count is increased to 130 (modeled after a specific operator example objective). The broadcast tier in this case is limited to 60 Analog carriers, and the top 40 most popular channels offered, which are broadcast in both HD and SD. All other programs are on the SDV tier (about 20 SDV slots). The benefits of SDV are clear in Figure 7. Despite more than doubling of the bandwidth-intensive HD programming, we nonetheless gain new capacity for IP growth.

As powerful as SDV is for reclaiming spectrum, it is only reclaiming QAM spectrum, which is already inherently efficient in delivering digital video. There are further, large spectrum gains available by instead reclaiming spectrum from the Analog carriers through the use of digital terminal adaptors (DTAs). In Figure 7, the implementation is a phased approach. Phase 1 is a reduction of Analog slots from 60 to 30 – the black threshold that extends through 2017. In 2017, it is suggested that Phase 2 kicks in, whereby all Analog carriers are removed. This is the second black threshold, where now well over 3 Gbps has been freed up as capacity for IP growth. This chart and analysis process also identifies the flexibility available in downstream spectrum management. There are many knobs and levers associated with decisions on service mix and use of tools available for bandwidth growth.

Of course, the core issue as new video services evolve is that a 10-year plan demands consideration of these bandwidth-hungry next generation video possibilities. Ten years of tools and projections are encouraging, but the projection is based on video services and technology as we know them today. The plan can quickly implode by considering the capacity when including the integration of new generations of HD.

Four phases of next generation video evolution are identified by the red arrows on the right side of Figure 7. First, consider simply that all of the HD becomes 1080p60 HD – broadcast, SDV, and VOD. It is assumed this format does not require a simulcast phase (existing STBs and HDTVs support the format if it is available to them). The drop in available capacity (the first red arrow on the right hand side of Figure 7) reflects about a lost year of lifespan, all other assumptions the same.

Next consider that a Quad Full HD (QFHD) format is made available on VOD as an introduction to this format in its early days, as capable televisions become available to early adopters. The current VOD allocation remains (1080p60) in this case, so this advanced VOD service is completely additive in terms of spectrum. It is assumed that this format is deployed only using MPEG-4 compression. Nonetheless, as revealed with the second red arrow, we see a larger step downward in available capacity, which now is just over 1 Gbps. Roughly another year of lifespan is lost, all other assumptions the same. Furthermore, this would drive the timing of the second node split for downstream in 2018 if a QFHD VOD tier were to become viable in that time frame.

Now consider the third step, whereby QFHD was used for Broadcast HD and VOD, but not SDV. Note we have not included a simulcast of standard HD, even though it is TBD at this point whether a QFHD format can be "down-resolutioned" to standard HD. Certainly this is not the case in today's televisions or STBs feeding televisions, but it is likely to be a consideration in future iterations. The 4x scaling of standard HD is of course, in part, to make it more likely that systems can take advantage of current processing in the video chain through the simple integer scaling factor of pixels. Not delving into the details of how this might play out, we quantify the impact of a change in the broadcast and VOD to QFHD. The effect identified by the third red arrow is to drop network capacity down to about 600 Mbps, and clearly this is eating into any hope for supporting long-term IP traffic growth.

Lastly, now consider that the SDV tier is converted, but the VOD is not. An example of why this might be practical is that as the IP migration takes place, it might be determined that the legacy VOD infrastructure is not permitted to grow with new MPEG-2 TS based investment. These investments would be made instead in the IP domain, with VOD being one of the first phases of the video services migration to IP. In this case of Broadcast and SDV supporting QFHD as opposed to standard HD, we clearly see, in the form of the lowest black threshold on the chart at about 21 dB (just over 100 Mbps of capacity available for IP traffic, or three DOCSIS channels), the hopeless situation for next generation video without some new ideas and evolutionary approaches to be supported over the HFC network.

To point out a measure of hope that hints at some of the consideration we will account for later in the paper, the upward pointed green arrow shows where this situation would instead fall if there was 1 GHz worth of spectrum to work with. The spectrum freed up by 1 GHz of HFC compared to 870 MHz is about 22 slots, which works out to almost 900 Mbps using 256-QAM. New spectrum is but one tool we will evaluate as a means to enabling the migration of next generation video services

Note also that we have as yet not even attempted to factor in any capacity effects associated with Ultra-High Definition Television (UHDTV) as a potential format.

LONG-TERM VARIABLES: GOOD NEWS – BAD NEWS

We observed in Figure 7 that there was an obvious problem brewing under the assumptions made based on considering HFC architectures, services, and technology, as we know each today.

In Table 4, we begin to make the case for why the situation may not be as dire as these projections. On the left hand side of Table 4, "Losses," we quantify in the decibel language of the projection analysis the potential bandwidth penalty of the new formats, quantified in the row identified based entirely on the resolution difference. Again, it may be determined in practice that the encoding process more favorably compresses the formats than is portrayed in Table 4, but for now we will simply rely on encoding efficiency gains consistent with the average savings attributed to H.264 and H.265 using today's HD format. In each case, this amounts to 50% savings, based on early evaluations of H.265 and our prior discussion on HEVC.

Table 4 – Video & Network Variables: Losses and Gains

Losses	dB	**Gains**	dB
1080p60	3.00	H.264	3.00
QFHD	6.00	H.265	3.00
UHDTV	6.00	Split	3.00
10-bit	0.97	N+0	9.21
Frame Rate	0.00	Mod Profile	0.97
Total	15.97	VBR/D3	1.55
		Total	20.73
Difference	**4.76**	(All)	
	7.76	(HD to QFHD)	

The conservative assumption for 1080p60 is that it is 2x the bandwidth required of 1080i30. For the purposes of this study, as is generally done in practice as well, we will not distinguish between bit rates of 1080i and 720p although the former is roughly 12% more bits of transport rate.

We consider a 10-bit depth of field for UHDTV, but no additional overhead associated with changes to subsampling. We also do not consider any additional frame rate impacts on transport bandwidth. While scan rates of television rates have increased, and, as discussed, studies [1] reveal that frame rates higher than 60 Hz are perceptible by humans, there appears to be no move afoot to standardize in the market place on anything higher. Additionally, UHDTV is standardized around a 60 Hz frame rate. While research noted above has shown perceptibility by human of up to 300 Hz, it would not be fair to impart new bandwidth associated with new frame rates at this stage, even if they are to take shape. It is intuitively likely that higher frame rates lend themselves to more similarities between adjacent frames. We leave the variable in the chart because we believe that in time, formats will begin to experiment with higher frame rate delivery. We should keep this variable in the back of our minds as a possible wildcard.

The "Losses" when added together in the worst case of UHDTV as the final phase is 15.97 dB, which we will round to 16 dB for discussion purposes.

Now, let's take a look at the "Gains" in Table 4. Some of these we have already observed as "HFC as we know it" gains in our projection chart. We identified service group splits by the traffic growth breakpoints in the chart, which recognized the virtual doubling of bandwidth (ideally) in a typical node split. The average bandwidth allocated per subscriber in the split service group is now twice as much.

We also capture the service group split function here identified as N+0. In this case, we are recognizing that rather than perform further business-as-usual node splits after another round of this expensive activity, an "ultimate" split is executed instead, where the fiber is driven deepest – to the last active. The impact to the average bandwidth made available per subscriber is much greater in this case, with the homes passed per N+0 node assumed to be 40. Note that we identify one split prior to N+0 in Table 4. In the actual timeline-based model we will capture the move to N+0 as an extra split (two total) prior to the migration to N+0. We captured the decibel effect (3 dB) within the N+0 adjustment in the table to match what we will show on the subsequent projection analysis.

Lastly, we applied the benefits of MPEG-4 encoding in introducing QFHD – obviously better than MPEG-2, but also clearly not enough itself to compensate the bandwidth growth. This is intuitively obvious enough, seeing as the MPEG-4 gains do not offset the resolution increase in pixel count. However, it should be pointed out that the 1080p60 case shown in the trajectory of Figure 7 may indeed be offset by the introduction of MPEG-4 to deliver that service. In fact, it is reasonable to consider that 1080p60 as a service does not become a video service offering *until* MPEG-4 is available.

We showed in the Figure 7 a hopeful sign in the form of a different total available spectrum – 1 GHz vs. 870 MHz. However, because our starting assumption of 870 MHz may be optimistic or pessimistic, and because the spectrum expansion discussion is a wide-ranging one, we will address the physical bandwidth component in a subsequent discussion dedicated to spectrum.

We identify three other "Gain" variables – the subsequent generation of encoding, H.265, the use of IP Video delivery using bonded DOCSIS channels, and the opportunity to be more bandwidth efficient in an evolved (i.e. N+0) HFC architecture

High Efficiency Video Coding (HEVC, H.265)

As described, HEVC is in the heavy lifting phase of development and standardization, has as an objective a 50% better bandwidth efficiency of video transport, all while also yielding a higher quality. It appears to be on the track to achieve these targets.

The time-to-market for encoding standards and time-to-scale of advanced video formats follow roughly similar temporal cycles in terms of years. They are not necessarily in phase, but in both cases long evolution cycles have been the norm. As shown in Figure 8, it has been to the case in the past that the encoding gains served to continually drive down the rate of video (all SD for a time), even as slow as the pace of encoding development was. This singular fact explains the rise of over-the-top video. Data speeds raced ahead while video rates continuously dropped, crossing paths about seven years ago.

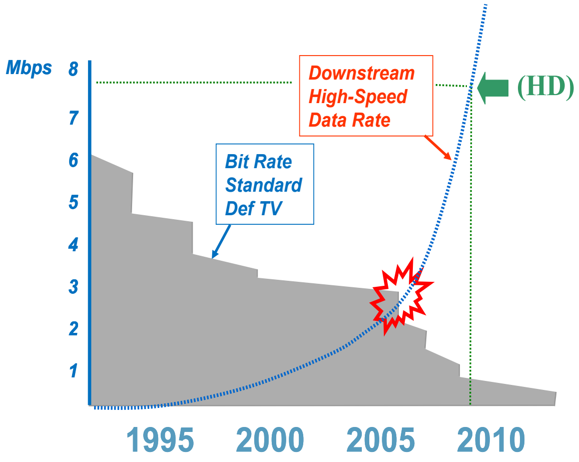


Figure 8 – Compression Meant Video Rates Only Decreased for Many Years

Now, however, demand for more HD has exploded, and display technology advanced significantly as well. It appears that the continuously accelerating pace of technology development will mean that higher quality, better resolution video will proceed faster than the process of standardizing encoding techniques. There is no accelerant to such a process, and arguably the increasingly competitive technology environment could lead to a slower standardization process, with service providers caught in between.

As indicated, early evaluation of H.265 and the conclusions drawn around this work described previously suggests that it will indeed achieve its target objective of 50% savings in average video bandwidth.

IP Video

Legacy architectures are based on simple traffic management techniques that allot an average of 3.75 Mbps per standard definition video stream to fit 10 such streams within a 40 Mbps single-carrier downstream QAM pipe. The heavy lifting of bit rate allocation is done at the MPEG level, whereby video complexities are estimated, and a fixed number of bits in the pipe are allocated to the ten streams under the constraint not to exceed 37.5 Mbps total. The same process plays out over High Definition slots, but in this case only two or three HD streams are part of the multiplex.

The introduction of DOCSIS 3.0 adds channel bonding to the toolkit, which, with the addition of MPEG-4 encoding, increases the stream count by over and order of magnitude relative to the transport pipe size. The net effect is the ability to use law of large number statistics for both SD and HD to the favorable advantage of less average bandwidth. So many independent streams competing for so much more pipe capacity results in a self-averaging effect that yields more efficient use of an N-bonded channel set when compared to MPEG-2 based video over N single channel QAM slots.

Self-averaging suggests that variable bit rate (VBR) streams can be used, recognizing the peaks and valleys will be handled inherently by the statistics (actually a capped VBR). Several prior analyses [16] of DOCSIS-based delivery, taking advantage of favorable statistics of wide channels to better handle the peaks and valleys of video traffic, shows that capped variable bit rate transmission yields a bandwidth savings that can be exploited. We use a 70% scaling as the bandwidth required for VBR-based channel bonded DOCSIS video in comparison to CBR-based single carrier QAM transport.

Fiber Deep Migration

"Business as Usual" HFC migration has been shown to be well-suited to about a decade of video and data traffic growth, without any new or special tools or techniques to accomplish this lifespan [13]. As discussed, use of node splitting in the HFC architecture reaches its ultimate phase when the last active becomes a fiber optic node. This architecture goes by various names – Passive Coax, Fiber-to-the-Last-Active (FTLA), or N+0. Regardless of the name, the architectural implications have two core components: small serving groups - on the order of 20-40 – and the opportunity to exploit new coaxial bandwidth becomes much more straightforward (30 assumed). The lifespan provided by BAU splits will not only make N+0 more cost effective due to RF efficiencies, but it will also leave operators within a stone's throw of FTTP should the need arise as an end state.

An important "side" benefit of an N+0 architecture is that the quality of the RF channel improves dramatically without the noise and distortion contributions of the RF cascade. The result is a higher SNR HFC link in the forward path. Because of this, we then consider more bandwidth efficient modulation formats. In Table 4, we have assumed that 1024-QAM will be readily accessible in such architectures, and in particular if new FEC is also implemented.

Finally, the removal of all RF amplifiers in the plant leaves only taps, passives, and cabling between node and subscriber, a much simpler scenario for flexible and expanded use of new coaxial bandwidth. Prior analysis [12] has shown how 10 Gbps (GEPON) and higher downstream capacities become conceivable in this architecture.

As fiber penetrates deeper into the HFC architecture, ultimately perhaps landing at N+0, the possibility of exploiting more bandwidth efficient modulation profiles exists, especially if the forward error correction (FEC) is updated from J.83 to modern techniques with substantially more coding gain. Here, we assume 1024-QAM supplants 256-QAM, for 25% added efficiency [15].

The dB Balance Sheet

Adding up the "Gain" side of Table 4, we find a total of about 20.7 dB, vs. 16 dB of "Loss." The encouraging information here is that this implies that, in principle, we can convert our current HD lineup fully to UHDTV and this would still be supported over the HFC network. All else equal, HFC lifespan would not be compromised in the face of IP traffic growth – the trajectory thresholds would not drop. This is so because the net of the gains and losses is a positive 4.7 dB. Thus, the thresholds would actually rise. Better yet, if the only format we bother concerning ourselves for business planning purposes is QFHD, then we have another 3 dB or headroom in our net gain.

The flaw in this good news story, of course, is that by the time we are considering QFHD, the IP CAGR is already threatening video service thresholds. We are at or near the end of the ten year window of migration. We are looking to extend HFC lifespan *beyond* this decade to the next while introducing these advanced video services. The excess gain can be viewed as available overhead for a simulcast transition. Based on Table 4, there is 4.8-7.8 dB to work with as part of enabling the possibility. While the services our transitioning, the IPV evolution is taking place, and the network is undergoing BAU migration, there are some "Not Business As Usual (NBAU)" evolutions expected to be taking place as well related to spectrum and architecture.

We will use Table 4 and these NBAU evolution factors to extend the projection through another decade, and draw conclusions on the intersection of video evolution, traffic growth, capacity, and the role of CAGR.

NEW SPECTRUM CONSIDERATIONS

Now let's consider the spectral aspects that were discussed in the last section, but not quantified in Table 4.

Figure 9 illustrates the anticipated spectrum migration of the HFC architecture long-term. A key driver discussed in great depth in [13, 14] is the necessity of operators to do something to address the limited upstream for the future. There are no easy answers to new upstream spectrum, and this figure describes the most effective approach and best performing from a modulation efficiency and flexibility standpoint, and which also yields the most efficient use of spectrum long term. The later is perhaps *the* key long-term primary objective for HFC spectrum evolution.

Because of the reasons outlined in [13] and [14], we foresee a phased approach to spectrum migration, consistent with the way operators incrementally deal with infrastructure changes in the context of dealing with legacy services and subscribers. The end state of the spectrum migration is shown in the bottom illustration of Figure 9, where some level of asymmetry consistent with what supports the downstream/upstream traffic ratio, will remain. No matter where the Frequency Domain Diplex (FDD) architecture lands in terms of diplex split, it is most assuredly going to yield a downstream capable of over 10 Gbps, and an upstream capable of over 1 Gbps.

While Figure 9 represents the most likely evolution scenario, other versions may come to pass. However, for any implementation, it is virtually guaranteed that the 10 Gbps/1 Gbps targets, at least, will be achieved. We will use this certainty in our projections in determining the ability of the evolved HFC architecture to deliver next generation video service in the face of continued growth in high-speed data services.

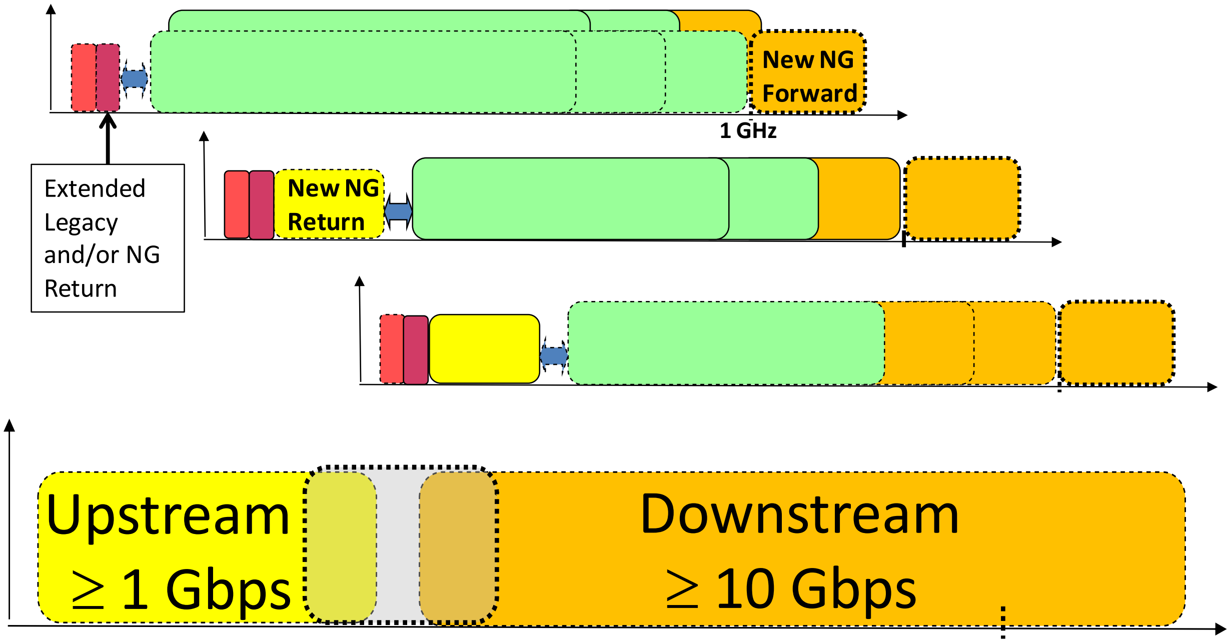


Figure 9 – Probable Evolution of the Cable Spectrum

PUTTING IT ALL TOGETHER

We now revert back to our original problem of capacity growth, and extended timeline of Figure 7 to account for the introduction of new generation of video formats. Beginning with Figure 10, we take into account all of these factors of video bandwidth growth and capacity preservation, placed in the context of HFC lifespan.

Video Service Delivery Assumptions

As we discuss video formats such as QFHD and UHDTV, it is reasonable to assume that HEVC has a key role, that fiber deep migration has continued to take place and is quite far down the path, and that the IP Video transition is in full swing, and possibly even complete. It is also reasonable to suggest that *unless* these evolutions take place, it is not practical to consider new tiers of advanced video services. Under this assumption, QFHD and UHDTV only become service in the cable network over IP, and only when HEVC is available in products for deployment.

The transition model is, of course, critical, as every new format introduces a period of simulcast if a service represents a broadcast. Conversely, in a full IP transition and a fully unicast architecture, the resolution and format become part of control plane and discovery. There is no wasted simulcast bandwidth, just any bandwidth penalty paid if the migration of video service delivery from the "legacy" efficiencies of broadcast to a dominantly unicast architecture is not properly managed (see Figure 5 and 6).

The Intersection of Video Services and Traffic Growth

Now let's consider Figure 10. Figure 10 is a modified Figure 7, extended through the end of the next decade, managed with an N+0 migration, and accounting for various capacity enhancing techniques discussed above.

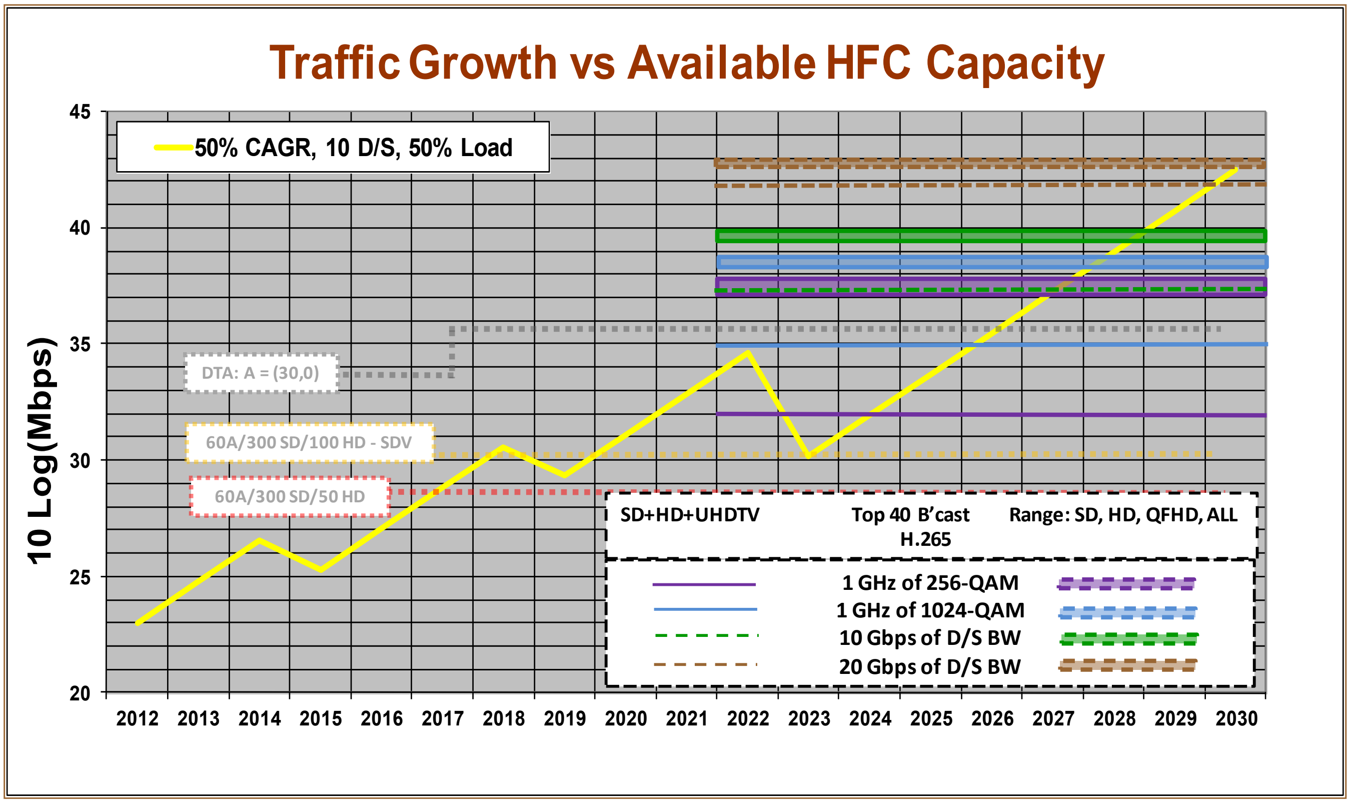


Figure 10 – Next Gen Video, Traffic Growth and HFC Capacity Limitations

The CAGR description is no different than Figure 7, it only goes on for longer, and sees a steeper breakpoint in 2022, representing the final phase migration to N+0. The Figure 7 thresholds are shown, in faded form, for reference against the cases to follow.

The legend at the bottom right is described as follows.

In all cases, we are talking about thresholds set by having a *static IP broadcast of the Top 40* channels. We are therefore taking advantage of IP video efficiencies, only as we know them today and previously identified in Table 4. Recall, we indicated that for a 200/200 lineup of SD/HD, then the top 20 programs would amount to 80-90% of the multicast gain in a switched IP system capable of multicast. The conclusion from that analysis was that a simplified, near-optimal architecture might instead be a mix of full broadcast and unicast, recognizing that all dimensions of network and service evolution are towards more unicast. From that, we have conservatively used a Top 40 program broadcast, which essentially would account for all of the multicast gain. At 40, it will likely come at the expense of some inefficiency of spectrum use versus multicast, but we prefer to err on the side of setting aside more spectrum for the purpose of a conservative analysis.

Also, because we are ultimately after the second-decade phase of HD evolution, we implement the next phase of compression evolution, HEVC, in calculating the long-term thresholds.

Four cases of available capacity are identified:

1) 1 GHz of spectrum carrying all 256-QAM, or 6.32 Gbps (purple)
2) 1 GHz of spectrum carrying all 1024-QAM, or 7.9 Gbps (blue)

3) A 10 Gbps downstream, in light of our prior conversation about the evolution of cable spectrum and key objectives (green)
4) A 20 Gbps downstream – enabled only through an N+0 architecture with a further extended use of coaxial bandwidth, requiring additional plant evolution of the passive architecture, including tap changes (brown)

Four cases of video formats are also analyzed. However, three of them fall close to one another in net capacity impact, and are lumped together in a "range" identified by a *rectangle* of the associated color on Figure 10. The fourth, most burdensome case is, not surprisingly, that which includes the introduction of UHDTV under the bandwidth assumptions we have identified previously – 160x the bandwidth requirement of the SD resolution and format. These UHDTV cases are identified by *lines* of the associated color – note that the green, 10 Gbps line is dashed, only because it overlaps the rectangular threshold range of the 256-QAM case.

The three cases in proximity whereby a rectangle is used to identify the threshold range are (in each case a simulcast of the Top 40):

1) SD + 1080p60
2) SD + 1080p60 + QFHD
3) SD + QFHD only

The latter, for example, makes sense if we consider that the integer relationship of formats (4x scaling of pixels) makes for the potential that next generation QFHD screens are also capable of displaying a "down-res" to 1080p60, or the STB/CPE function is capable of performing this function for the television. The range of remaining capacity in these three cases seems intuitively very close, and in fact is always within about 1 dB. This is a product of three things:

- Large capacity made available by all-QAM to 1 GHz, at least
- HEVC whittling down SD and 1080p60 rates by a factor of one-quarter
- The nature of the chart, based on nonlinear CAGR, is decibel units which tend to compress large numbers, which is illustrated by recognizing we are quantifying the impact of 18 years of aggressive compounding of traffic.

Let's examine what Figure 10 reveals.

First, consider UHDTV as a format that is mid-to-late next decade in scale at the earliest. It is not realistically able to be supported by HFC, at least under the assumptions we have used here. Even the most favorable of evolution deployments shown here – 20 Gbps of downstream capacity – suggests that persistent CAGR coupled with this broadcast video service runs out of room before the end of the decade. The vast majority of the bandwidth is the UHDTV itself, so eliminating the simulcast component is negligible to this conclusion.

By contrast, if we look at the QFHD scenarios, and view this as a format eligible at the end of this decade, then even the least capable case of 1 GHz of 256-QAM bandwidth offers nearly a decade of support for this scenario, with a range reaching exactly to the end of the next decade (2030) before a threshold breach of HFC capacity. This bodes well for the ability of tools available – just as we understand them today – to manage through an aggressive combination of video service evolution and persistent CAGR of IP traffic. It remains to be seen if this form of the evolved HFC network is the most cost-effective approach to enabling this service mix. But, it is surely comforting to know that the possibility exists to support such services with a 2012 understanding of technology, recognizing in addition the long time window of observation in which to adapt strategy and technology accordingly.

Note, of course, that since we have used 10 Gbps and 20 Gbps and not QAM calculations, these threshold apply equally to any access network that would set aside IP bandwidth for 40 channels as described herein. However, since other architectures may be full multicast, a broadcast adjustment (removing this lost capacity) might be in order for an accurate comparison. This is quite easy to accommodate by noting that 10 Gbps is simply 40 dB on Figure 10, while 20 Gbps is 3 dB higher at 43 dB. It is clear that there is very little difference in lifespan implied between these thresholds and those with broadcast allocations in this stratosphere of bit rates and continuance of CAGRs.

Settling of CAGR

In Figure 11, we show a modified case, whereby the assumption is made beyond this first decade that CAGR *decreases* to 32%. We chose this settling of CAGR at 32%, such that the net CAGR for the period through the end of the next decade is an 18-year average CAGR of 40%.

The logic behind this assumption is that we have seen this aggressive march forward of CAGR driven primarily by over-the-top (OTT) video services. In the model developed here, we are already allocating spectrum for most popular video services, and thus using video also as a driver for CAGR could be considered double counting, at least in part (the most-watched part) of this phenomenon. In addition, the vast history of CAGR growth has been around *catching up* with our ability to download and/or consume media – audio, then video. Once these media consumption appetites are satisfied, then it is possible that a CAGR settling will take place, with limits set by behaviors and eyeball counts [11]. Of course, it may simply be replaced by as-yet-to-be-determined non-media consumption applications, or altogether different kinds of media consumption that is bandwidth-busting, such as volume displays. That, however, seems beyond even the extended time frames we are evaluating here.

The above reasoning was completely qualitative, and it may in fact turn out that aggressive 50% CAGR persists indefinitely, or possibly increases. Nonetheless, because of the ramifications of long-term CAGR variation, we thought it useful to show this perspective, and that an 18-year average of 40% CAGR was a reasonable amount of settling to consider. Note that only at the year 2030 exactly would the 40% average and the 50%-32% model meet. The trajectories along the way getting to those points will, of course vary.

Now let's evaluate what Figure 11 below says about video services evolution, capacity, and time.

First, observe now that *every* QFHD case indicates a lifespan of the network *through* the end of the next decade, even the 1 GHz, 256-QAM only case. This is a very powerful statement about the impact a settled CAGR may have on the support of advanced video services. It is also a reminder about the dramatic mathematical and planning implications of 18 years of compounding.

For UHDTV, there still does not appear to me much hope for a lasting solution to broadcast support, under what seems like the reasonable assumption that it does not make a large-scale service appearance until 2025 or beyond. The best case scenario in Figure 11 only suggests that 20 Gbps of network capacity covers the UHDTV scenario plus traffic growth into 2032-2033, which is then very shortly after it would have been introduced.

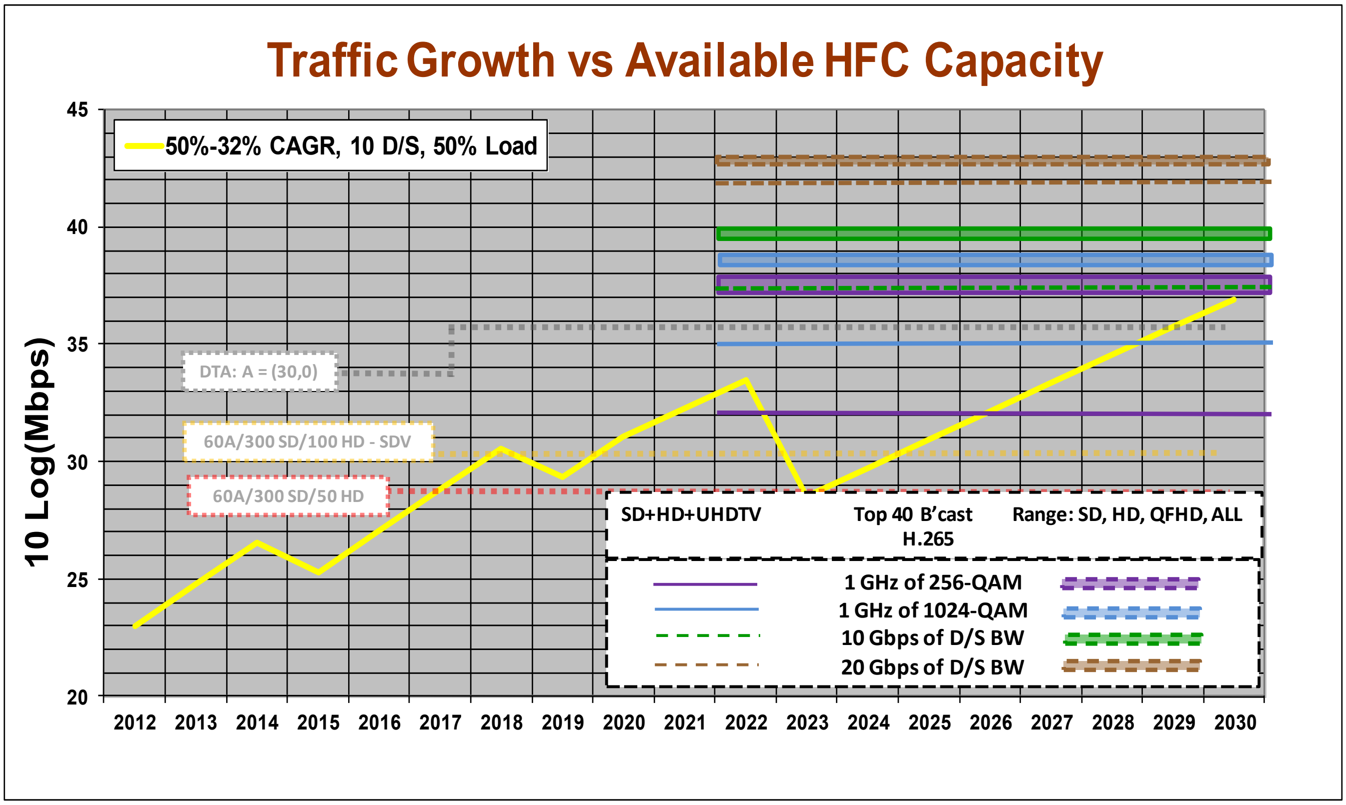


Figure 11 – Next Gen Video, Traffic Growth + CAGR Settling, and HFC Capacity

Conversely, this conclusion might more optimistically be stated by noting that HFC that manages a capacity of 10-20 Gbps *can* clearly support an early phase of UHDTV experimentation and deployment, and provide some cushion of years over which its significance as a scalable service can be evaluated. Does it become a niche scenario, where a very select number of channels become part of a programming lineup, much like 3D is today? For a mid-2020 time frame of experimentation, there are enough years of support in an early, modest phase of deployment where these kinds of questions can be asked and answers provided. These answers can then be used to guide a phase of network evolution, such as Fiber-to-the-Home, if scalability of the service is required. Or, it may lead to the conclusion that UDHTV is not an every-household type of consumer service, but associated with, for example, the penetration of home theatre-type owners. If so, it likely remains largely on the IP unicast service tier, and never become a broadcast scenario to worry about. Though, if this latter situation comes to pass, then this could exactly be the kind of thing that keeps CAGR chugging at 50%, while this model reflects the 18-year, 40% average case.

There are clearly many interrelated variables to consider and scenarios to quantify. Our assessment of the results leave inevitably to the conclusion that these projections are best viewed as living documents, and must be periodically re-assessed for the validity of the assumptions as trends and service mixes evolve over time. Advantageously, though, the projections indicate there are valuable windows of time near term, and again in the long term as efficiency improves. These windows offer the opportunity to observe and make methodical decisions to manage the evolution, without the pressure of an urgent congestion problem on the horizon.

THE EYES HAVE IT

While developing HEVC, compression science was not standing still elsewhere. Recently, a new technology – Perceptual Video Processing (PVP) -- was incorporated into broadcast encoders to improve the efficiency of both MPEG2 and AVC significantly. PVP technology [20] leverages the biology of *human vision* itself to enhance the encoding process. It can be thought of as a compression co-processor. Performance improvements typically range from 20% for moderately-easy-to-encode content to up to 50% for hard-to-encode content. Given the close familial resemblance of HEVC to its predecessors, it's quite possible that PVP could grant similar bonus improvements on top of HEVC's innate high compression efficiency as it has for MPEG-2 and AVC [6, 34, 35].

Signal Processing and Human Vision

Perceptual Video Processing (PVP) technology is an encapsulation of design principles that are thought to be at work in the visual system based on decades of research into the biology of human vision [2, 3, 4, 7, 9, 20, 24, 26]. Though biological in origin, these design principles are rooted in concepts that are familiar to signal processing engineers, namely, the ideas of noise reduction, signal estimation, and error signals. What is unique is that PVP is based on a model [21] of early visual signal processing, which has the following key components:

- Vision is tuned to the scale-invariant statistics of natural images [8]
- First stages of visual processing act as optimal filters designed to minimize the impact of noise
- A second stage of processing makes an estimate of the error associated with the first stage and uses that error signal to self-adapt to changing lighting conditions
- The output stage of processing is a coded form of the error signal, which can be thought of as a visual map of statistical uncertainty associated with the estimation process.

A key insight is that statistical uncertainty equals perceptual significance. The output error signal – the "uncertainty" signal -- highlights two kinds of information:

1. Image features that are uncertain because local correlations in the image are as likely to be attributable to noise as to actual variations in the signal. These are the features that are likely to be ambiguous from a signal estimation point of view and thus may require more attention.

2. Image features that contain local correlations that deviate from statistical expectations associated with natural scenes. In some sense, these are the "unexpected" correlations that might be worthy of closer inspection.

The notion that the output of early vision correlates with local statistical uncertainly provides a potential clue about higher-level perception and visual behavior. Eye-tracking and saccades, for example, might be considered behaviors intended to spend more time inspecting areas of high uncertainty to minimize overall uncertainty. Similarly, areas of high activity in retinal output might correlate with areas of high perceptual significance because they are the most suspicious in terms of statistical expectations – this is a clue that it may be worthy of special attention.

This model of the early visual system might also provide a context for understanding why edges are perceptually significant. According to the model's key components, edges are not perceptual important because they are edges, rather because they are localized correlations that deviate from the global expectation of scale invariance and thus require longer inspection to reduce uncertainty. It is not in fact the edge that has maximum uncertainty, rather it is the area around the edge, which itself might provide insights into the fundamental nature of perceptual masking and Mach bands – the illusion of heightened contrast near edges.

The Engineering View of Retinal Processing

The signal processing described occurs through the biological processing of the retina. The retina is made up of specialized cell layers, and each has a specific task. These can be classified as follows [20]:

Photoreceptors – The rods and cones we learned about in primary school health class. Photoreceptors are the first line of processing, are very densely aligned, and convert light (photons) into neuroelectrical signals.

Horizontal Cells – This second stage of processing cells collect the output of the photoreceptors and share them with adjacent horizontal cells as kind of a spatial low-pass filter operation on the discrete photoreceptor inputs.

Bipolar Cells – In the third stage of processing, bipolar cells collect both photoreceptor and horizontal cell inputs, and essentially acts to subtract the photoreceptor cell inputs, performing a differentiator type of mathematical operation.

Amacrine Cells – Bipolar cell inputs are received by amacrine cells, which come in different types. One important type acts as an electrical rectifier and gives a measure of the mean activity in the bipolar layer. A second type provides feedback to the first two layers to adjust their response properties according to this mean activity observed.

Ganglion Cells – The final stage of retinal processing, these cells take input from both bipolar cells and amacrine cells, and process and package them for delivery over the optical nerve to the brain.

Figure 12 illustrates the visual processing stages as a signal processing operation, described using tools analogous to functions common in signal estimation applications [20].

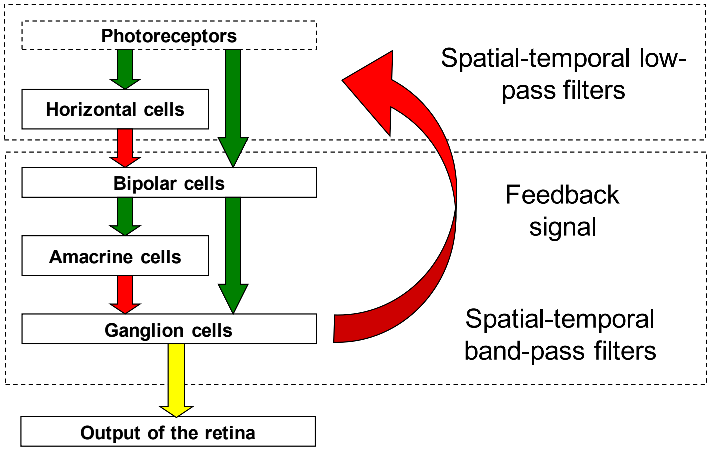


Figure 12 – Visual Cells as Signal Processing Functions

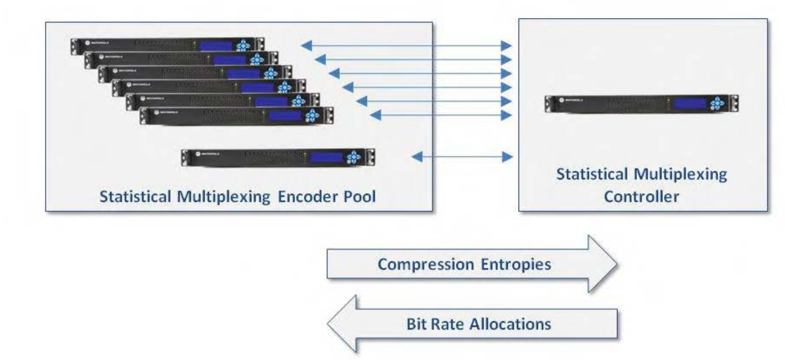
PVP Technology

Considerations for the biology of vision has proven to be very effective in improving compression efficiency in professional broadcast encoders. The key design principles have been extended to encompass space, time, and color and collected into a set of tools and software and hardware implementations collectively referred to as the Integrated Perceptual Engineering Guide (IPeG™). PVP is a particular commercial implementation of IPeG designed to operate in real time to reduce compression entropy and improve predictability in coding.

Internally, PVP identifies features in video that are likely to have high perceptual significance and modifies those features to reduce the number of bits required while preserving video quality. In its first commercial incarnation [20, 38], PVP performs two noteworthy complimentary operations: 3-Dimensional Noise Reduction (3DNR) and Adaptive Detail Preservation (ADP). The 3DNR operation is a combination spatial/temporal nonlinear adaptive filter that is very effective at reducing random noise in areas the eye may not notice. The ADP element preserves visually important detail and attenuates quantization noise, impulse noise, stochastic high-contrast features, and other hard-to-compress detail difficult for the eye to track.

An example of PVP used to improve compression efficiency for statistical multiplexing is illustrated in Figures 13 and Figure 14. The central concept in statistical multiplexing (aka “statmux”) is that more and better channels can be delivered over a limited bandwidth by allocating bits intelligently between the various channels that comprise a statistical multiplexing pool. Channels that are easy to encode at a given point in time are given fewer bits than channels that are hard to encode. This traditional “statmux” operation is illustrated in Figure 13.

Using PVP, this operation is modified with this additional intelligent processing as shown in Figure 14. The statistical multiplexer still does its bit rate allocations, as always, but it now does so based on an enhanced set of inputs from the IPeG processor. PVP improves statistical multiplexing by selectively reducing the greediness of hard-to-encode channels in real time. High compression entropy means more bits would be needed to achieve a target video quality. Low compression entropy would require fewer bits to achieve the same video quality. PVP preferentially reduces the entropy of hard-to-encode features thereby making tough content kinder and more generous neighbors in the pool.


Statistical Multiplexing Encoder Pool
Statistical Multiplexing Controller
Compression Entropies
Bit Rate Allocations

.

Figure 13 – Traditional Statistical Multiplexing

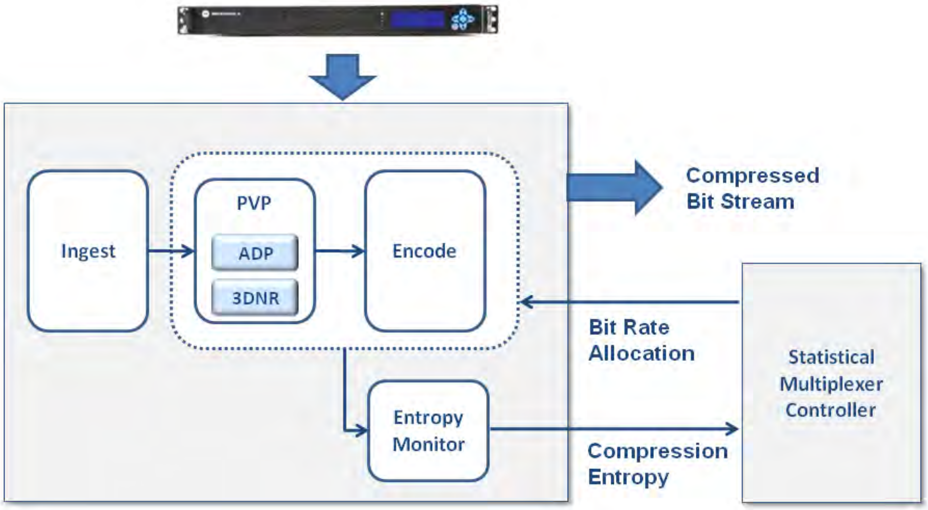


Figure 14 – PVP: Perception-Guided Adaptive Modification of Compression Entropy

An example of the graded impact of PVP on compression entropy is shown in Figure 15. Note that the relative impact of PVP is largely independent of the operational bit rate, which could prove to be a useful feature in statistical multiplexing pools that contain premium channels with higher targeted operating bit rates than other channels in the same pool. The data shown in Figure 15 are typical of moderate-to encode and difficult-to-encode broadcast content. The actual reduction in compression entropy may be optimized for particular use cases by adjusting the strength of PVP from weakest to strongest.

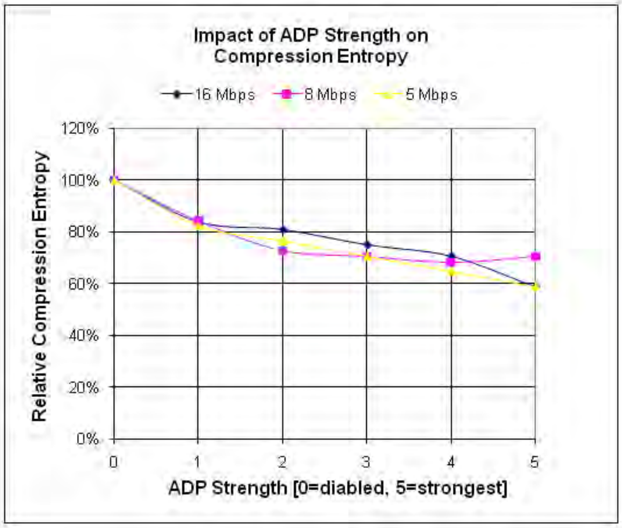


Figure 15 – PVP Reduces Compression and Can be Tuned to Requirements

Complementing HEVC

One of the key advances of HEVC is "just-the-right-size "processing in which each Coding, Prediction, and Transform Unit is sized precisely to capture the self-similarity within the picture detail they encode. It is without question a highly efficient way to squeeze bits -- but it's *not* the way the eye sees.

There are two key questions to examine to predict the impact of PVP on HEVC efficiency:

1) Would PVP enhance predictability and thus promote regions of self-similarity that can be captured efficiently by HEVC Units?

2) Are HEVC's "just-the-right-size" Coding, Prediction, and Transform Units also "just-the-right-size" for the natural scale of vision? If they are, then we would expect PVP to have less of an impact for HEVC than it does for AVC and MPEG-2.

The first question is straightforward, and the answer is *yes*. PVP nudges video towards statistics that would be expected of clean natural scenes when those nudges would not be very noticeable. In other words, the PVP promotes predictability and regional self-similarity. It does this by reducing unpredictable random noise and slightly modifying stochastic high-contrast features that are "unexpected" as described previously. On this basis, we would expect PVP to improve HEVC's innate compression efficiency to approximately the same extent that PVP improves AVC and MPEG-2 efficiency.

The second question is a bit more involved. Getting a handle on the natural scale of vision entails comparing the size of retinal images to the resolving power of the eye.

The visual angles subtended by various kinds of displays are listed in Table 5. The size of the visual field depends on the physical size of the display and its distance from the viewer. For QFHD (4k) and Ultra HDTV, we use the dimensions of recently announced displays [29] and predict that comfortable viewing distances will be only moderately larger than they are for traditional HDTV.

Table 5 -- Expectable Visual Angles for Various Display Types

Display Type	Format	Resolution		Dimensions (inches)			Viewing Distance (inches)	Visual Angle (degrees)
		Horizontal	Vertical	Diagonal	Width	Height		
Smartphone	QHD	960	544	5	4	2	12	19
Tablet	1080p	1920	1080	11	10	6	16	35
HDTV	1080p	1920	1080	55	48	27	76	35
Super HDTV	4K	4096	2160	70	62	33	88	39
Ultra HDTV	8k	7680	4320	85	74	42	90	45

The fovea of the retina sees the central 2 degrees of the visual field with high acuity [17]. It is the part of the retina with the greatest resolving power. We watch television by continually moving our eyes around to bring our fovea in line with particular features one after the other. Our brains integrate this sequence of focal observations into a unified seamless experience.

In Figure 16, we examine the size of the foveal image relative to the size of the visual field subtended by various display types. Our fovea spans only about 1/10th the width of a smartphone display, which means we must still move our eyes about even for the smallest display type. For 1080p and finer resolutions, our fovea sees at any moment in time only a disc of pixels having a diameter about 5% of the width of the whole display. It is worth noting that area of the disc comprises less than 1% of the total pixels in the display. We only see that small 1% of the display in detail at any instant. Research into bit rate reduction of video in other circles has been around trying to figure out how to take advantage of the fact that so little of a screen is actually processed at any given instant [5].

In Table 6, we quantify these relationships across a range of display types. An important insight comes about when we analyze the number of physical pixels seen by the fovea as a function of display size and resolution. A disc about 100 pixels in diameter contributes to foveal vision for smartphones, 1080p tablets, and HDTV. If brightness and contrast were put aside, the equal density of pixels would suggest that we would notice about the same level of visual detail – and same level of compression artifacts – on smartphones and tablets as we would see on HDTV when viewed from normal distances. Visual details and artifacts would likely be less noticeable for 4K and UHDTV because they would be 2-3x less magnified in the foveal image according to the pixel diameters shown in Table 6.

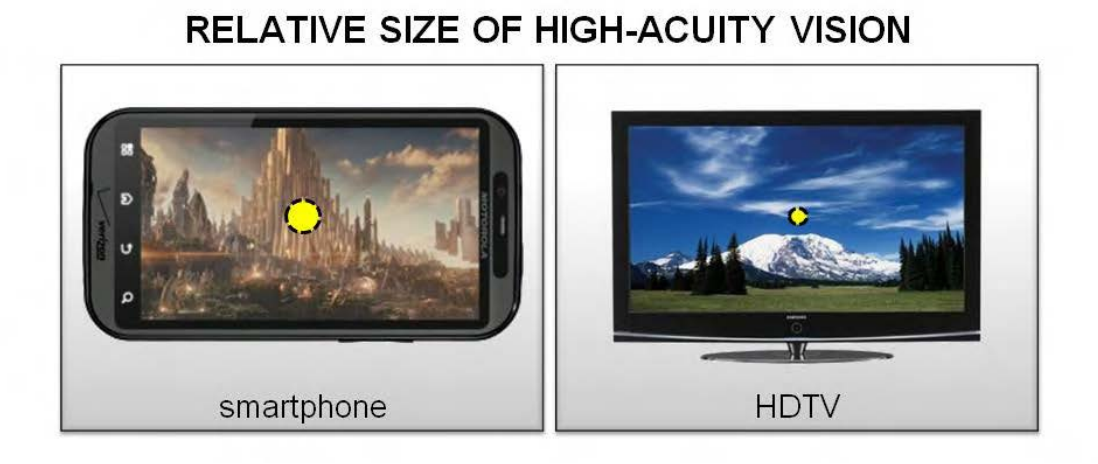


Figure 16 – Size of Projected Foveal Image (yellow) vs. Display Type

Table 6 -- Size of Foveal Field of View Relative to Size of Coding Units

Display Type	Format	Size of 2-degree Foveal Field of View						
		Percent of Screen Width	Pixels (dia.)	Macroblocks or Coding, Prediction, and Transform Units				
				4x4	8x8	16x16	32x32	64x64
Smartphone	QHD	11%	101	25	13	6	3	2
Tablet	1080p	6%	111	28	14	7	3	2
HDTV	1080p	6%	110	27	14	7	3	2
Super HDTV	4K	5%	211	53	26	13	7	3
Ultra HDTV	8k	4%	343	86	43	21	11	5

The scale of MPEG-2 and AVC macroblocks and sub-partitions relative to the size of the foveal image for smartphones, tablets, and HDTV is illustrated in Figure 17a. The homologous HEVC Coding, Prediction, and Transforms Units are depicted in Figure 17b. We noted previously that the fovea covers only a tiny fraction of a display screen at any moment. Smaller yet are macroblocks, sub-partitions, and HEVC Units. Even the Largest Coding Unit (LCU) presently allowed in HEVC (64x64) is significantly smaller than the fovea's field of view.

The homologous HEVC Units for 4k and UHDTV are illustrated in Figure 17c. The difference between HDTV and beyond-HD is a matter of visual scale. The foveal image of a LCU becomes 2-3 times smaller in 4k and UHDTV, respectively, compared to HDTV. Other smaller HEVC Units become visually diminutive, and the smallest 4x4 HEVC Units become tiny.

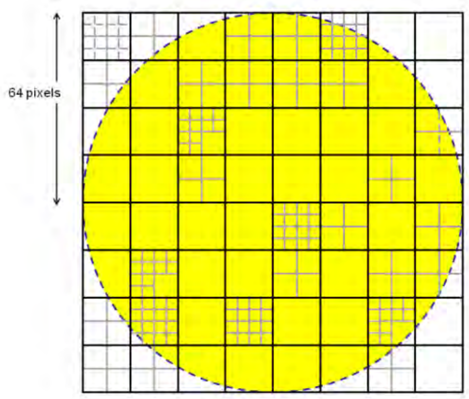


Figure 17a – MPEG-2 and AVC Macroblocks (dark) and Sub-partitions (light) Relative to Foveal Image (yellow) for smartphones, tablets, and HDTV

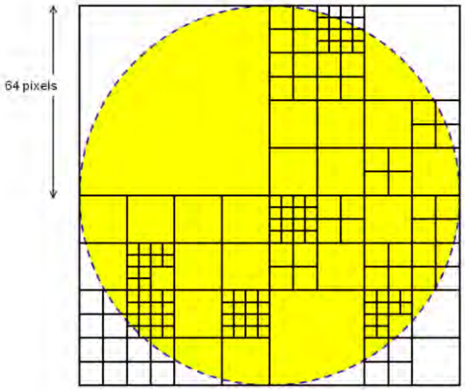


Figure 17b – HEVC Coding, Prediction, and Transform Units Relative to the Foveal Image for smartphones, tablets, and HDTV.

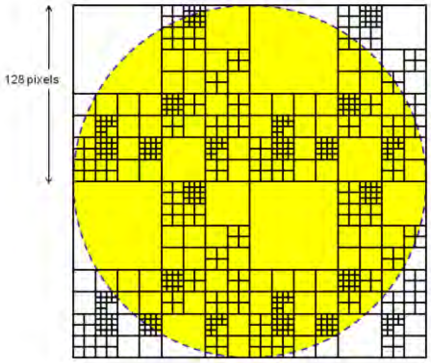


Figure 17c – HEVC Coding, Prediction, and Transform Units Relative to the Foveal Image for 4K and Ultra HD (note the relative size of the Units are smaller than in Figure 17b)

HEVC and AVC use rectilinear segmentation. The specific architecture is different, but the motivating philosophy is the same. More important, the visual scale of the rectangular segments is not dramatically different. HEVC provides a few larger block-size options that are better able to isolate regions of self-similarity without over segmentation, but those block sizes are still smaller that the fovea's field of view.

We can conclude from the above analysis that HEVC Units are, in fact, not always visually "just-the-right-size." Like AVC macroblocks and sub-partitions, HEVC Coding Units will have discrete boundaries within the foveal field of view even when encoding video that is visually smooth across the fovea. Compression artifacts tend to gather around discrete boundaries because those are the places that prediction is weakest. When those boundaries lay within the retina's high-acuity foveal field of view, they will be noticed. HEVC would need larger Largest Coding Units (LCU) to prevent over segmentation of the foveal image and meet the "just-the-right-size" visual ideal. For of smartphones, 1080p tablets, and HDTV the LCU would need to be at least 128x128. For 4K and Ultra HD, LCU would need to be at least 256 x256.

PVP and HEVC Together

Given the overall similarity of HEVC and AVC in terms of coding philosophy and visual scale, we project that PVP will improve HEVC coding efficiency to much the same extent that is improves AVC and MPEG-2 coding efficiency. HEVC's intrinsic compression efficiency is reported in [6, 34, 35]. Relative bit rates expected are listed in Table 7 and plotted in Figure 18. The impact of PVP is very content specific. Nonetheless we have found that PVP provides an overall average bit rate savings of ~20% in national-scale commercial deployments. We use that value in Table 3 to calculate the benefit of PVP to HEVC.

Table 7 -- Expected Bit Rate for Various Coding Modes and Display Types

Coding Method	Expected Bit Rate (Relative to AVC alone)		
	Smartphones	1080p Tablets & HDTV	4K & UHDTV
AVC	100%	100%	100%
AVC + PVP	80%	80%	
HEVC	66%	56%	50%
HEVC + PVP	53%	45%	40%

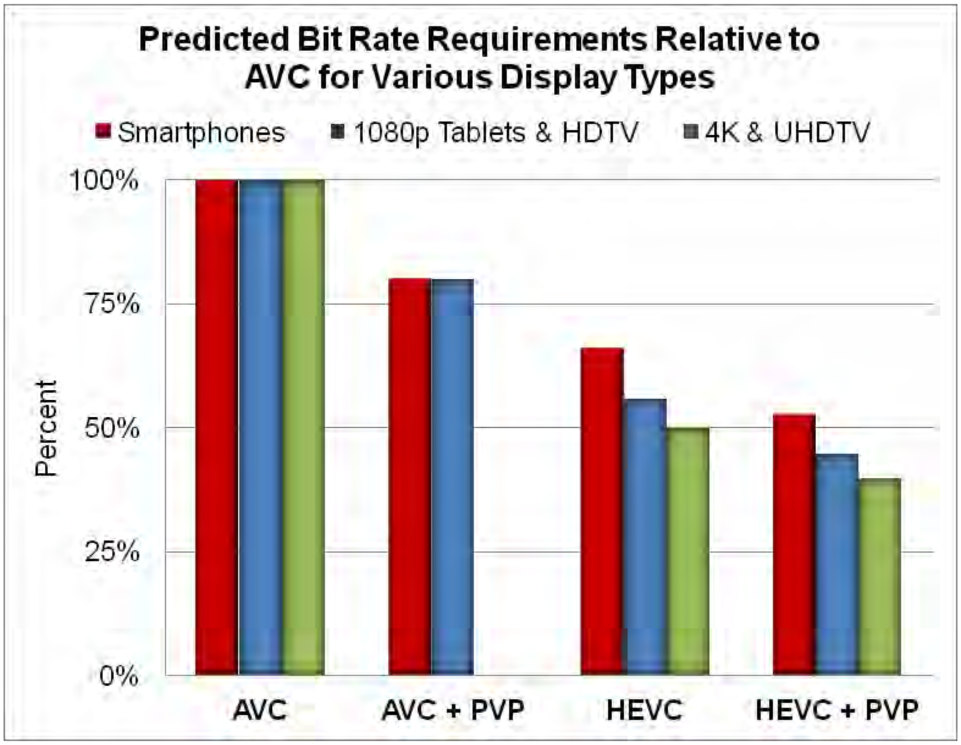


Figure 18 – Projected PVP Efficiencies Bit Rate for AVC and HEVC vs. Display Type

We can take this new knowledge and apply it to the prior figures that quantify traffic growth impacts. Figure 19 does so for the last case evaluated previously (Figure 11, 18-yr average CAGR of 40%). Of course, we do not anticipate tremendous new lifespan effects of PVP with a projected 20% of added efficiency. The expected value, at least early in PVPs evolution, is improved QoE of AVC and eventually HEVC video.

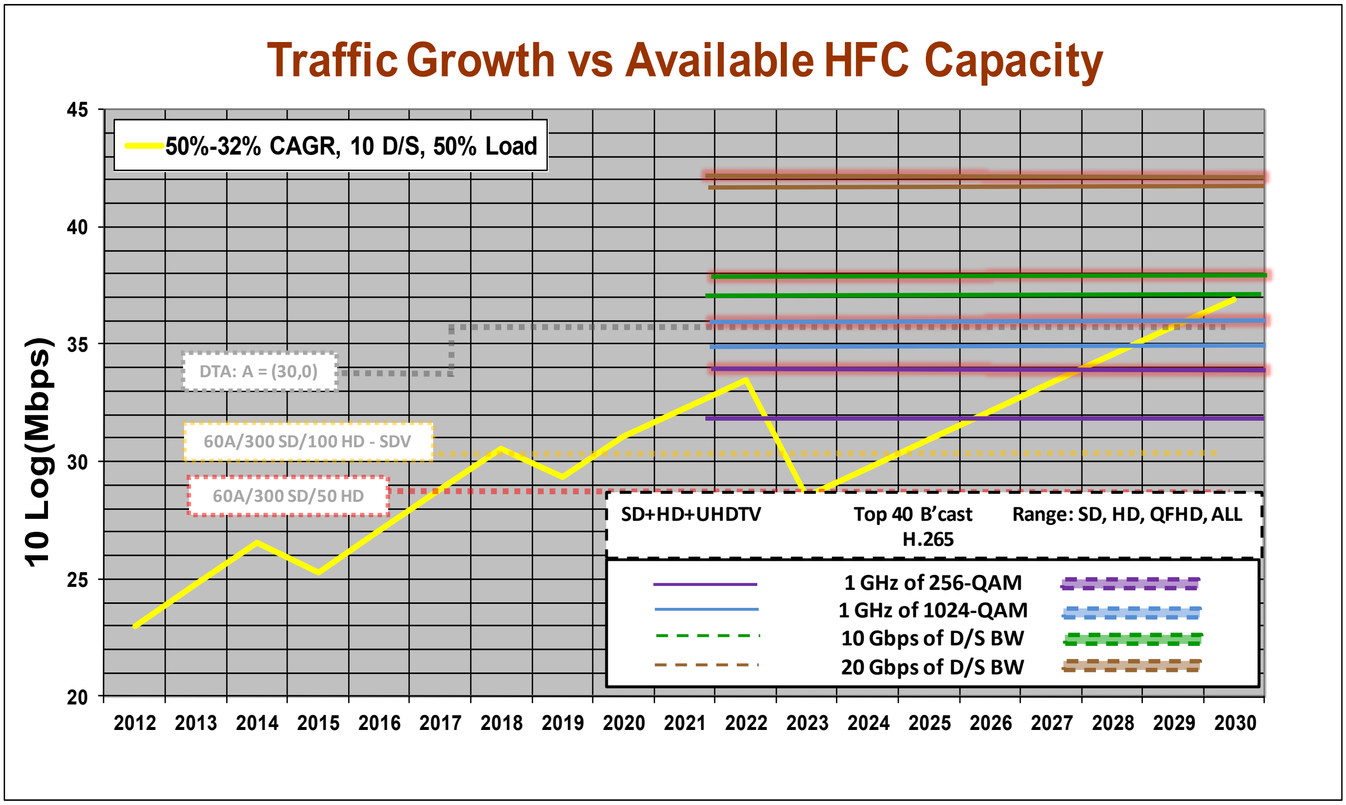


Figure 19 – HEVC + PVP, Traffic Growth and HFC Capacity (Settled CAGR Case)

Figure 19 indicates that the 20% of added efficiency at least has made the least-capable architecture evaluated (1 GHz of 256-QAM, purple) theoretically capable of weathering UHDTV services, or any substitute, similarly bandwidth-hogging applications that might beat it to market, into the middle of the next decade without the threat of breaching the threshold of capacity within a ten-year time frame under the assumptions used here. For that architecture, it also amounts to two extra years of lifespan, with the added burden on the non-PVP case that the final N+0 segmentation must also occur at least two years earlier.

For the higher capacity cases (1024-QAM, 10 Gbps, 20 Gbps), the impacts are less dramatic. Given that the existing network is, in fact, based on 256-QAM and outdoor plant equipment is 1 GHz capable only today, that impact carries more weight regarding preparation for a next generation of video bandwidth utilization.

Now consider Figure 20. Figure 20 is a redo of Figure 7, with the anticipated 20% benefits of PVP rolled up on the case of MPEG-4 AVC used in the Figure 7 analysis. In this case, we can observe a pretty significant impact of the extra 20%, largely because modest increases translate into large dividends when there is so little latent network capacity to begin with. These are shown in the upward pointing black arrows, which show the before/after of PVP being added for each scenario previous calculated. For example, in the worst case scenario in Figure 7 (and shown also in Figure 20) – QFHD in both the broadcast and the SDV tier as next generation HD, the network capacity was essentially completely consumed. Three available slots remained for IP traffic.

Because 20% of that tremendous amount of bandwidth is also a good chunk of bandwidth itself, adding it back to the pool for IP growth is pay substantial dividends as shown in Figure 20. With the savings, QFHD could actually be supported with some data growth runway. And, with a 1 GHz network, the network supports this level of enhanced HD with IP growth through 2020 under the migration assumptions used here of two segmentations. It is very unlikely that enhanced HD resolutions will be this pervasive in the market is such a short period of time. The introduction as VOD may be more practical in the timeframe of Figure 20. However, it is comforting to apply a bandwidth hungry, yet practical, "killer" application example to analyze in the projection analysis, and come out with a conclusion that the system does not only not break, but in fact enabling of such an application to a degree before any new steps or technologies are applied that could increase network capacity.

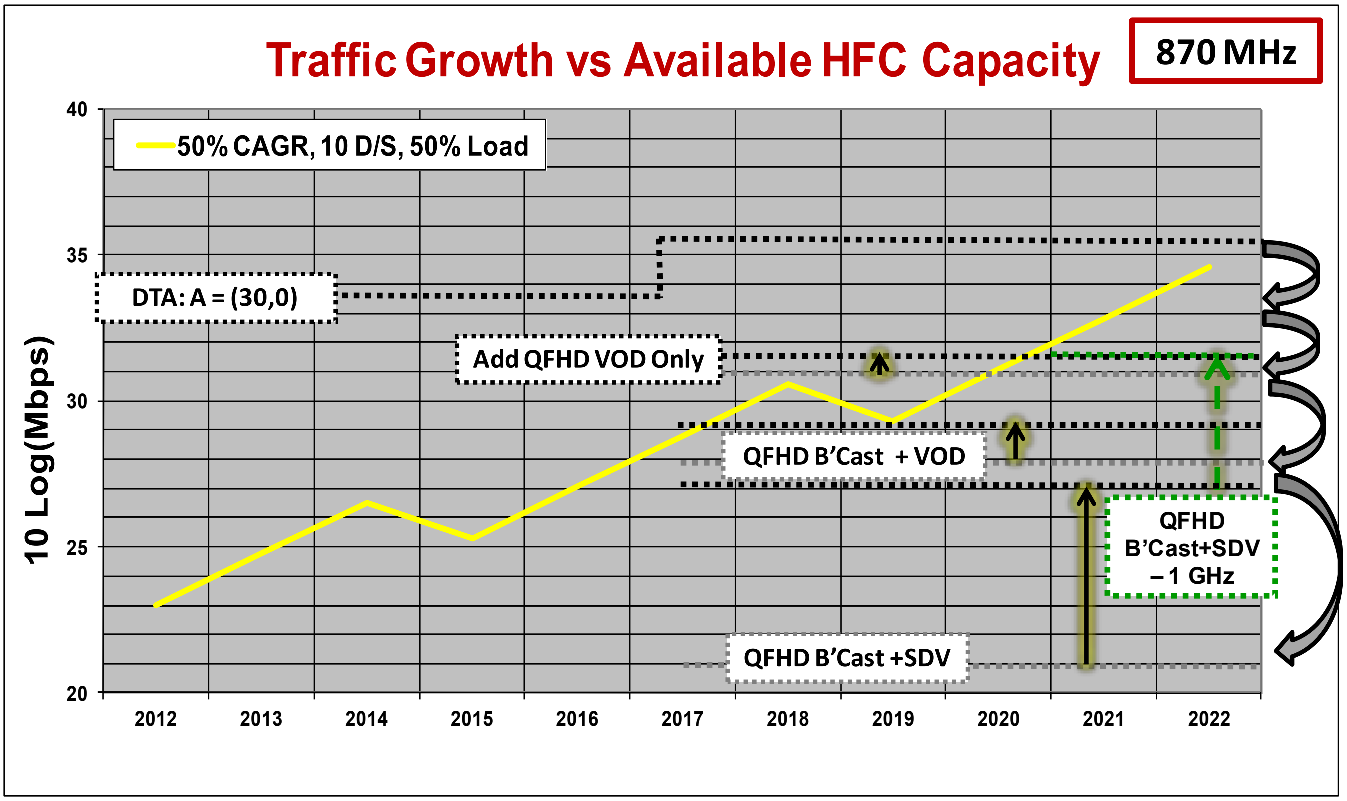


Figure 20 – Added Capacity with 20% PVP Efficiency, QFHD Format Cases (aggressive CAGR Case)

SUMMARY

In this paper, we evaluated network projections for the long-term, including many permutations of scenarios that included current and future services. We included technology and architecture options that are likely to come into play during the time windows observed, and applied these to quantify their effect. These include the shift to IP delivery, "beyond HD" video services, standards-based and innovative new encoding techniques, emerging use cases and delivery, and architecture, spectrum, and RF delivery enhancements. The result is a blueprint for an approach to preparing network service and migration plans – a blueprint that is, however, a "living document" given the accelerating pace of change in technology and services.

It is clear that there are many interrelated variables. However, any solution approach must include a comprehensive understanding that quantifiably describes the effects of network, technology, and service changes, such as shown in this paper. This is critical to properly engage in effective scenario planning, bound the problem, and prepare solution paths suited to an operator's circumstances and expectations.

REFERENCES

[1] Armstrong, M and D Flynn, M Hammond, S Jolly R Salmon, *High Frame Rate Television*, BBC Research Whitepaper WHP 169, September 2008.

[2] Attneave, F., *Information Aspects of Visual Perception,* Psychol. Rev. 61 183-93, 1954.

[3] H.B. Barlow, The Coding of Sensory Messages: Current Problems in Animal Behaviour, Ed. W. H. Thorpe and O.L. Zangwill, Cambridge: Cambridge University Press, 331-360, 1961.

[4] Barlow, H.B., *Redundancy Reduction Revisited*, Network: Comput. Neural Syst. 12:241-253, 001.

[5] Deering, Michael F, "*The Limits of Human Vision*," Sun Microsystems, 2nd International Immersive Projection Technology Workshop, 1998.

[6] De Simone, F et al., *Towards high efficiency video coding: Subjective evaluation of potential coding Technologies*, J. Vis. Commun. (2011), doi:10.1016/j.jvcir.2011.01.008

[7] Dowling, J.E., The Retina: An Approachable Part of the Brain, Harvard Univ. Press. 1987.

[8] Field, D.J., *Relationship Between the Statistics of Natural Images and the Response Properties of Cortical Cells*, JOSA A, 4 (12): 2379-2394, 987.

[9] Hare, W. A. and W.G. Owen, *Spatial Organization of the Biplolar Cell's Receptive Field in the Retina of the Tiger Salamander,* J. Physiol. *421*:223-245, 990.

[10] Ho, Yo-Sung and Jung-Ah Choi, *Advanced Video Coding Techniques for Smart Phones*, 2012 International Conference on Embedded Systems and Intelligent Technology (ICESIT 2012), Jan. 27–29, 2012.

[11] Howald, Dr. Robert L, *Boundaries of Consumption for the Infinite Content World*, SCTE Cable-Tec Expo, New Orleans, LA, October 20-22, 2010.

[12] Howald, Dr. Robert L, *Fueling the Coaxial Last Mile*, SCTE Conference on Emerging Technologies, Washington DC, April 2, 2009.

[13] Howald, Dr. Robert L, *Looking to the Future: Service Growth, HFC Capacity, and Network Migration*, 2011 Cable-Tec Expo Capacity Management Seminar, sponsored by the Society for Cable Telecommunications Engineers (SCTE), Atlanta, Ga, November 14, 2011.

[14] Howald, Dr. Robert L, and Phil Miguelez, *Upstream 3.0: Cable's Response to Web 2.0*, The Cable Show Spring Technical Forum, June 14-16, 2011, Chicago, Il.

[15] Howald, Dr. Robert L**,** Michael Aviles, and Amarildo Vieira, *New Megabits, Same Megahertz: Plant Evolution Dividends*, 2009 Cable Show, Washington, DC, March 30-April 1.

[16] Howald, Dr. Robert L, Dr. Sebnem Zorlu-Ozer, Dr. Nagesh Nandiraju, *Delivering Pixel Perfect*, The Cable Show Spring Technical Forum, May 11-13, Los Angeles, CA.

[17] Helga Kolb, et al, *Webvision: The Organization of the Retina and Visual System. Part XIII: Facts and Figures Concerning the Human Retina*, WorldPress, http://webvision.med.utah.edu

[18] Marpe Detlev , et al., *Video Compression Using Nested Quadtree Structures, Leaf Merging, and Improved Techniques for Motion Representation and Entropy Coding*, IEEE Trans. Circuits Syst. Video Techn., Vol. 20, Nr. 12 (2010) , p. 1676-1687.

[19] McCann, Ken, and Jeff Gledhill, Adriana Mattei, Stuart Savage, *Beyond HDTV: Implications for Digital Delivery*, An Independent Report by ZetaCast Ltd, July 2009.

[20], *A Biological Framework for Perceptual Video Processing and Compression*, SMPTE Motion Imaging Journal, Nov/Dec 2010.

[21] McCarthy, Dr. Sean T., and W.G. Owen, "Apparatus and Methods for Image and Signal Processing,". US Pat. 6014468 (2000). US Pat. 6360021 (2002), US Pat. 7046852 (2006), 1998.

[22] Sullivan, Gary J. and Jens-Rainer Ohm, *Recent Developments in Standardization of High Efficiency Video Coding (HEVC),* SPIE Applications of Digital Image Processing XXXIII, Andrew G. Tescher (editor), Proceedings of SPIE Volume 7798, Paper number 7798-30, August, 2010.

[23] Ulm, John and Gerry White, *Architecture & Migration Strategies for Multi-screen IP Video Delivery***,** 2012 SCTE Canadian Summit, March 27-28, Toronto, CA.

[24] Vu, T.Q., S.T., McCarthy, and W.G Owen, *Linear Transduction of Natural Stimuli by Light-Adapted and Dark-adapted Rods of the Salamander*, *J. Physiol. 505(1):* 193-204, 1997.

[25] Wang, Zhou and Alan C. Bovik, *Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures*, IEEE Signal Processing Magazine, January 2009.

[26] Watanabe, S., *Information-Theoretic Aspects of Inductive and Deductive Inference*, IBM J. Res. Dev. 4. 208-231, 1960.

[27] Wiegand, T, G. Sullivan, G. Bjontegaard, and A. Luthra, *Overview of the H.264/AVC Video Coding Standard*, IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 560-576, July 2003.

[28] Yoshika Hara, *NHK Bets on Super Hi-Vision as Future TV*, EE Times, Sept 17, 2007.

[29] *CES 2012: 4K TV Sets Make Their Debut, Minus the Hoopla*, Los Angeles Times, January 11, 2012.

[30] ITU-T and ISO/IEC, ITU-T Rec. H.264 | ISO/IEC 14496-10 Advanced Video Coding (AVC), May 2003 (with subsequent editions and extensions).

[31] ISO/IEC JCT1/SC29/WG11 (MPEG), “Description of High Efficiency Video Coding (HEVC),” doc. no. N11822, Daegu, KR, January 2011.

[32] ISO/IEC JCT1/SC29/WG11 (MPEG), “Vision, Applications and Requirements for High Efficiency Video Coding (HEVC)”, doc. no. N11872, Daegu, KR, January 2011.

[33] ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/16, "Joint Call For Proposals on Video Compression Technology", WG11 document N11113 and Q6/16 document VCEG-AM91, Kyoto, January 2010.

[34] JCTVC-A204, “Report of Subjective Test Results of Responses to the Joint Call for Proposals (CfP) on Video Coding Technology for High Efficiency Video Coding (HEVC),” Dresden, DE, April, 2010.

[35] JCTVC-G339, “Comparison of Compression Performance of HEVC Working Draft 4 with AVC High Profile,” Geneva, Nov. 2011.

[36] JCTVC-F900, “Common test conditions and software reference configurations,” Torino, IT, 14-22 July, 2011.

[37] JCTVC-H1003, “High efficiency video coding (HEVC) text specification draft 6,” Geneva, CH, November, 2011.

[38] Motorola Mobility SE6601 Encoder, http://www.motorola.com/Video-Solutions/US-EN/Products-and-Services/Video-Infrastructure/Encoders/SE-6300-6500-Series-US-EN.

[39] “Video Quality Experts Group Report on the Validation of Video Quality Models for High Definition Video Content” VQEG HDTV Final Report, vers. 2, June 2011.

[40] www.carbonbale.com

[41] www.100fps.com

WHY 4K: VISION & TELEVISION

Mark Schubin
SchubinCafe.com

Abstract

Throughout the history of functional television, there have been moves towards higher definition, countered by the existence of lower-definition standards. Few of the choices of definition have been related to human visual acuity, however, which varies according to many factors.

The latest push for higher definition, to go beyond HDTV, is being driven in part by considerations unrelated to cable television, such as ease of program production and declining movie attendance and TV-set sales. The current era of bit-rate-reduced digital-video transmission, however, might nevertheless be an ideal time to offer consumers what could be the next level of increased picture definition.

THE ORIGIN OF DEFINITION

Language and Vision

Although citations for other senses of the word *definition* in the *Oxford English Dictionary* date back to the 14th century, the earliest citation for the sense relating to a manufactured system's "capacity to render an object or image distinct to the eye" is from 1878 (with two slightly older citations related to visual distinctness as rendered by an artist or in a natural formation).[1] The date might be associated with a different publication.

In 1862, Hermann Snellen published (in German) a book about something that he called *optotypes*.[2] In English, the book might be called *Sample Letters, for determining visual acuity.* The book introduced two concepts that have lasted to the present day: the idea that "normal" vision is 20/20 and the familiar letters on an eye chart, such as the *T* shown in Figure 1 below.

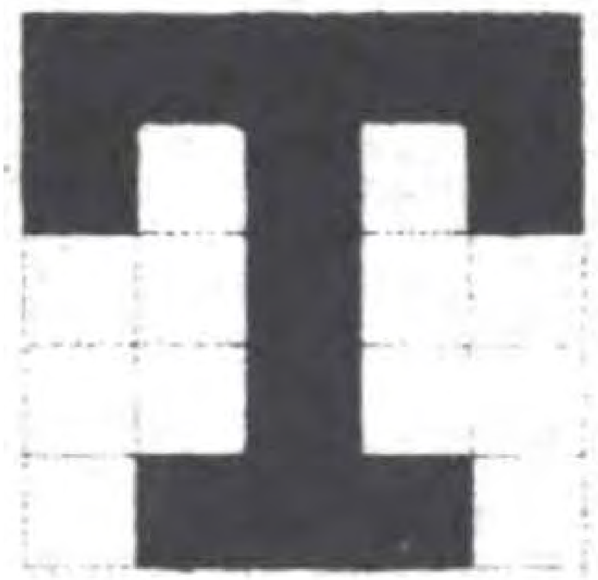


Figure 1: An 1862 Snellen Optotype

As the faint marks behind this optotype taken from his book indicate, the letter occupies a grid five units high by five units wide, and every element of the character, whether black or white, occupies one grid space. Snellen's definition of normal vision involved the ability to resolve features that subtended an angle of one arc minute (one-sixtieth of a degree) on the eye's retina.

If the whole character, therefore, were printed at a particular size and placed at a particular distance so that it would subtend an angle of five arc minutes, it would be able to be read with "normal" vision. The distance chosen was twenty feet so as to avoid visual issues associated with presbyopia (age-related inability to focus at short distances caused by the hardening of the eye's lens), a condition that usually first becomes noticeable around the age of 45.[3] In terms of visual focus, a distance of 20 feet is close to infinite.

Below is the familiar top of an eye chart based on Snellen's optotypes.[4] It was said in 1995 to have had more copies printed and sold in America than any other poster.[5] The top *E* on this chart is labeled "200 ft." on the left and "60 m" on the right.

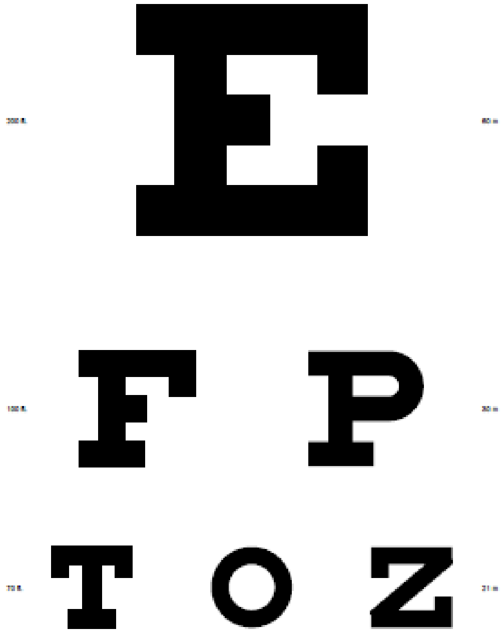


Figure 2: Top of an Eye Chart

The "200 ft." designation means that a person with "normal" visual acuity, as defined by Snellen, can distinguish that *E* at a distance of 200 feet, (or 60 meters). Vision defined as "20/20" (or "6/6") indicates an ability to see at 20 feet (or six meters) what a person with "normal" visual acuity can also see at 20 feet (or six meters).

Someone who could not read any letter smaller than the top *E* would be said to have "20/200" (or "6/60") visual acuity, the ability to read at 20 feet (or six meters) only what a person with "normal" vision can read at 200 feet (or 60 meters).

Issues Associated with the Definition

In nothing described to this point has anything been said about the illumination of the chart, the perceived contrast of the characters, or the definition of the edges of optotypes printed on the chart. Regarding the last, note, for example, that the edges of the *T* of Figure 1 are not as well defined as those of the *E* of the chart of Figure 2.

Snellen, himself, was aware of other issues associated with visual acuity. Below is a portion of another chart from his 1862 book. It shows not only an inversion of the color of the optotypes and the background but also a variation in contrast between the two lines of optotypes shown here. Snellen was clearly aware that contrast could affect visual acuity.

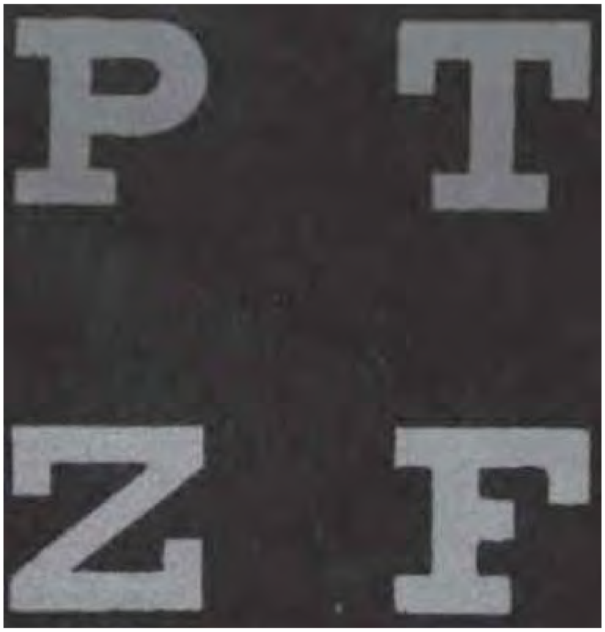


Figure 3: Portion of Snellen 1862 Chart with Color Inversion and Contrast Variation

As for the optotype edges, they contain higher spatial frequencies than the feature sizes would suggest. A pair of lines, one black and one white, suggest a cycle. If each line subtends a retinal angle of one arc minute, there would be 60 such lines in a degree. With half the lines white and half black, the spatial frequency could be said to be 30 cycles per degree (30 cpd).

Unfortunately, at a spatial frequency of 30 cpd, the edges of the optotypes would be soft.

The effect can be observed below. The *E* of Figure 2 was resized several times in image-manipulation software. Is it still readable as an *E*? It should seem clearer farther away.

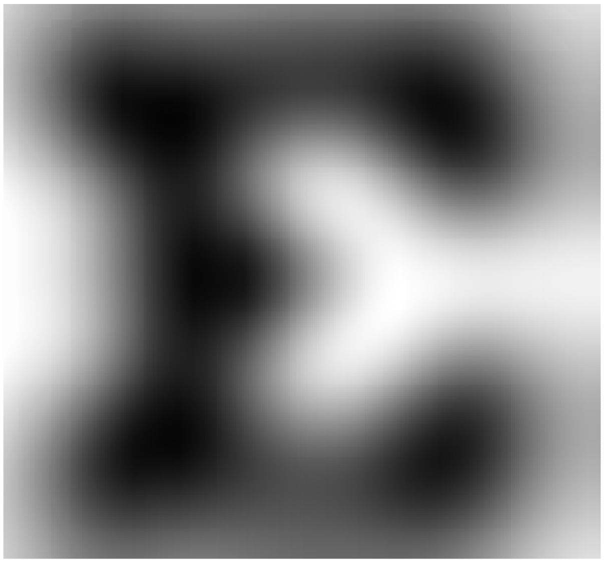


Figure 4: Snellen's *E* Filtered

The images below illustrate how sharp edges require higher spatial frequencies. Snellen's *E* is shown at the upper right. To its left is what its vertical strokes might look like if sinusoidal, and, to the left of that, a graph of the sine function between black and white.

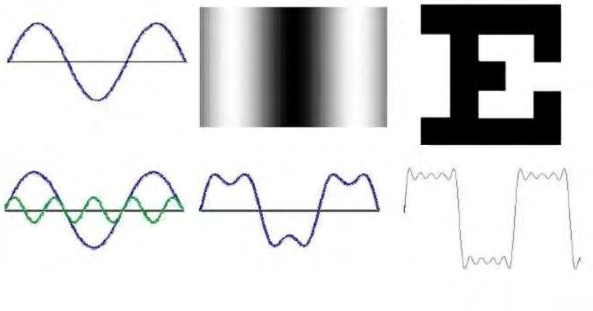


Figure 5: Adding Harmonics for Edges

In the lower row of Figure 5, at left is the same sine wave (the "fundamental") with another sine wave of three times the frequency (the third "harmonic") superimposed on it. To its right is the addition of those two waves. The transitions between dark and light are now shorter and steeper. At the far right is the sum of the fundamental and the third, fifth, seventh, and ninth harmonics. The transitions are shorter and steeper still. A perfect edge would require the sum of the fundamental and all of its odd harmonics, a square wave.

As for contrast, consider the image below. It's called a contrast-sensivity grating. Contrast increases from bottom to top, and picture definition (or spatial resolution) increases from left to right.

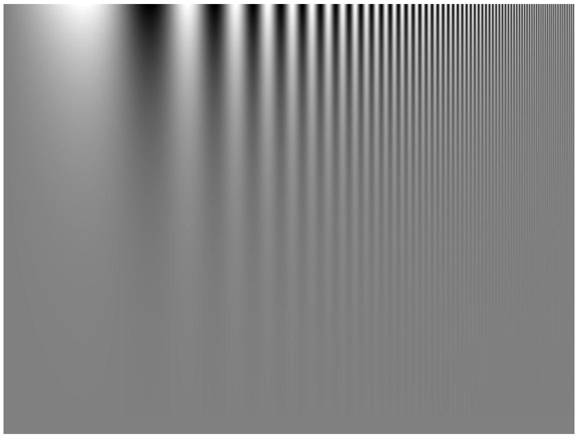


Figure 6: Contrast-Sensitivity Grating

Assuming normal printing or display and relatively normal (or corrected vision), the undifferentiated gray at the bottom of the grating of Figure 6 should appear to have a curve or "V" on top, the left and right edges higher than what is between them. In fact, there is no such curve in the grating. It is being added by the viewer's visual system.

Just as human hearing is most sensitive to middle sound frequencies, so, too, is human vision most sensitive to middle spatial frequencies (varying between about one and eight cycles per degree). Those who are familiar with the Fletcher-Munson curves of loudness sensation might find the visual contrast-sensitivity curve to be similar.[6]

An example of how important the contrast-sensitivity function (CSF) is in human vision may be seen in the pair of composite images of Figure 7 at the top of the next page. They

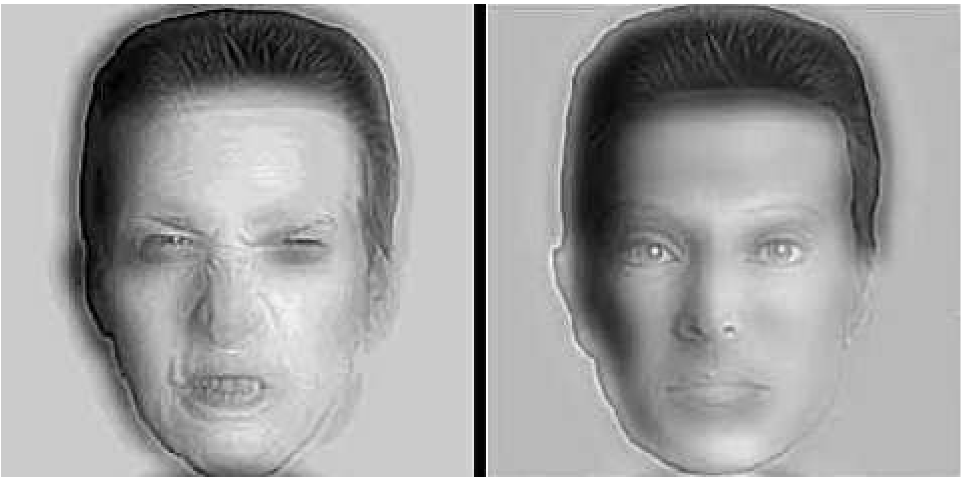


Figure 7: “Angry Man/Neutral Woman,” © 1997 Aude Oliva & Philippe G. Schyns

were created by Aude Oliva of Massachusetts Institute of Technology and Philippe G. Schyns of the University of Glasgow. They are used here with permission.[7]

To a viewer with relatively normal or corrected vision looking at the images from an ordinary reading distance, the image on the left will appear to be that of an angry-looking man, while the one on the right will appear to be that of an emotionally neutral-looking person, perhaps a woman. As the viewer moves farther from the images, however, there will be a distance at which both images appear to be of angry-looking people, followed by a long range of distances at which the angry man appears on the right and the neutral person on the left.

The composite images were created by combining two sets of images. One set, with the angry man on the left, has spatial frequencies intended to be seen near 6 cpd. The other, with the angry many on the right, has spatial frequencies intended to be seen near 2 cpd, in the lower insensitive section of a typical human CSF. As the viewer moves away from the images, both sets of spatial frequencies increase, the lower ones moving into the more-sensitive region of the CSF and the higher ones into the upper insensitive region of the function.

TELEVISION DEFINITION

Early History

The Alfred P. Sloan Foundation’s Technology Series includes a book about the invention of television. Its preface has the following: “But who invented television? Nobody knows.”[8]

Nevertheless, as acknowledged by that book and many other sources, the first person to achieve a video image of a recognizable human face seems to have been John Logie Baird. And the first reception apparatus that he used operated with just eight scanning lines at eight frames per second (fps).[9]

That was a drop from the spatial definition of previous image-transmission systems. Although recognizable-face television wasn’t achieved until 1925, television proposals are older and actual, working facsimile-transmission systems older still. British patent 9745 was issued in 1843 to Alexander Bain for a fax system.[10]

A slightly later fax system, Giovanni Caselli’s Pantelegraph (developed in 1856) saw extensive commercial service. Figure 8, on the next page, shows an actual fax page received via Pantelegraph.[11]

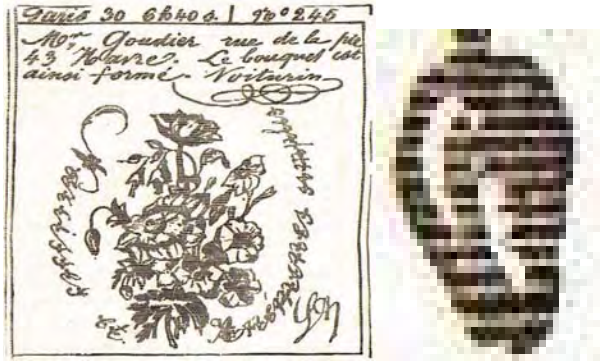

Paris 30 6h40s. | No 245
Mr Goudier rue de la ...
43 Havre. Le bouquet est
ainsi formé - Noitarin

Figure 8: Pantelegraph Fax and Portion

The image at left shows the complete fax page. The image at right shows a magnified portion of just the flower bud on the left side of the arrangement. Fifteen scanning lines can be counted in the bud, alone.

Below is a drawing from German patent 30105, issued in 1885 to Paul Nipkow for an "electric telescope." Television historian Albert Abramson called it "the master television patent" for its video scanning.[10] Each rotation of the scanning disk would produce one video frame. As the scanning disk drawing shows ("D1" through "D24"), Nipkow chose 24 scanning lines per frame.[12]

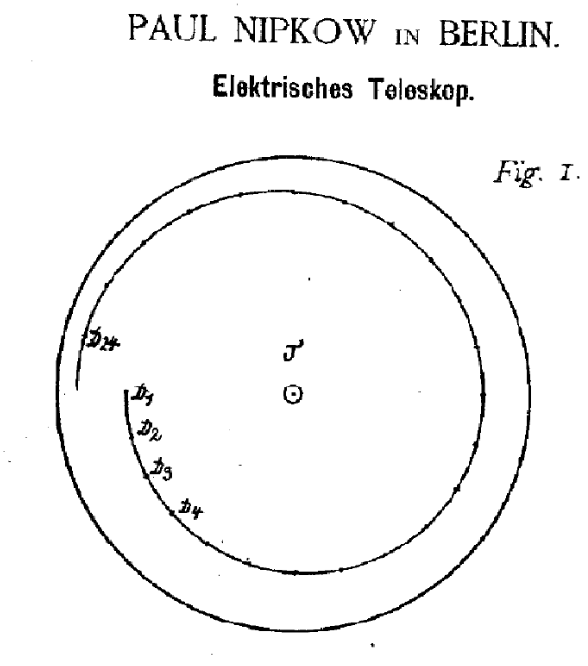


Figure 9: Nipkow's 24-line Scanning Disk

Although Baird and Nipkow might not have conducted studies of optimum image definition, Herbert E. Ives, who headed facsimile and television research at Bell Telephone Laboratories and was also an expert on photography, did. In his introduction to television in *The Bell System Technical Journal* in 1927, he described the definition requirement for what was, at the time, considered primarily an extension of one-to-one telephone service:

"Taking, as a criterion of acceptable quality, reproduction by the halftone engraving process, it is known that the human face can be satisfactorily reproduced by a 50-line screen. Assuming equal definition in both directions, 50 lines means 2500 elementary areas in all."[13]

The 50-line system was soon used, however, to capture larger scenes. In 1928, employees swinging a tennis racquet (as shown below) and a golf club were shown in a video demonstration, and a Bell Laboratories engineer was quoted as saying "We can take this machine to Niagara, to the Polo Grounds, or to the Yale Bowl, and it will pick up the scene for broadcasting."[14]

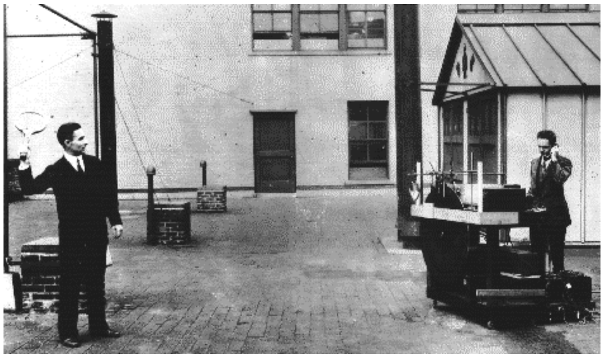


Figure 10: Tennis Swing Televised in 1928 [15]

In fact, the 50-line definition of the Bell Labs system was relatively high compared to that of most of its contemporaries. The second issue of *Television* magazine in the U.S., dated the same month as the Bell Labs demonstration, in its editorial content and advertising listed picture definitions of 24 through 50 lines.[16]

Only August Karolus, in Germany, went to higher definition in 1928. At the 5th German Radio Exhibition that year, he showed images with 96-line definition.[17] They are compared below to 30-line images from Dénes von Mihály at the same event.[18]

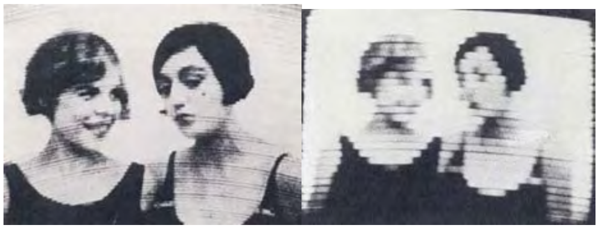


Figure 11: 96- & 30-line TV Pictures in 1928

The First High-Definition Era

Even before television moved from electromechanical scanning to all-electronic systems, there was great interest in higher-definition images. In 1935, a Television Committee, headed by Baron William Lowson Mitchell-Thomson Selsdon, reported to the British Parliament that the government should mandate "high-definition television." It was defined in paragraph 28: "it should be not less than 240 lines per picture...."[19]

Beginning in 1936, British television broadcasts alternated between a 240-line electromechanical system and an all-electronic system with 405 total scanning lines, of which 377 were active (picture carrying). When, in 1952, the Television Society (UK) heard a talk about the events that led to the 405-line broadcast standard, the presentation was called "The Birth of a High Definition Television System."[20]

The use of the term "high-definition television" wasn't restricted to the United Kingdom. Reporting on RCA's 441-line (383 active) television demonstrations at the 1939 New York World's Fair, *Broadcasting* magazine noted,"The exposition's opening on April 30 also marked the advent of this country's first regular schedule of high-definition broadcasts."[21]

When the first National Television System Committee (NTSC) began its work on a U.S. standard in 1940, it surveyed U.S. and non-U.S. proposed and working television systems ranging from 225 to 605 lines. Its last decision (*after* what was supposed to be the committee's final meeting) was a shift from 441 lines to 525 (483 active).[22]

Two other line-number standards saw significant broadcast use after World War II. They were an 819-line standard first broadcast in France in 1949 (with 737 active in the French version and a slightly higher number active in a Belgian version) and a 625-line standard first broadcast in Germany in 1950.[23]

The 625-line (575 active) number was later adopted by most of the world's countries, including France (1963) and the UK (1962).[24] The exceptions were those adopting the U.S.-standardized 525-line system. Although there were many different transmission systems (primarily for the 625-line countries), those two line numbers dominated the standardized, analog, all-electronic television period.[25]

In the previous, largely electromechanical television era, viewers could generally clearly perceive picture improvements with increasing line numbers, as in Figure 11, above left. There were, however, some anomalies even back then.

As early as 1914, Samuel Lavington Hart applied for a patent for interlaced scanning (scanning the image at a lower line number and then re-scanning at a slightly different position to fill-in between the lines), issued the next year.[26] Beginning in 1926, Ulises A. Sanabria applied a three-to-one interlace to electromechanical television, using three, slightly offset scans of 15 lines each to form a complete image of 45 lines.[27] Among advantages reported by an independent observer were a reduction of image flicker and line visibility (called "strip effect),[28] the opposite of complaints about interlaced scanning today.[29]

Studies have shown full, limited, or no effect on perceived definition from interlace over the number of lines in a single scan even in a still image.[30] Increased line visibility and flicker (or "interline twitter") in still images may be seen in video images, which are readily available on the Internet.[31]

Aside from interlace, many other factors could affect image-quality perception. In 1955, a delegation of U.S. engineers went to England to study television there and reported the perceived quality of the 405-line pictures superior to those of America's 525-line pictures. Possible reasons ranged from better operational practices to better allocation of bandwidth to different filtering.[32]

Combining Vision and Television

Ignoring all other television technical characteristics (in scanning system, scene, lens, camera, transmission, reception, display type, and display settings) that could affect image definition, many have reported an optimum viewing distance based only on the number of active scanning lines and Snellen's "normal" visual acuity of one minute of visual arc per scene element. According to that theory, NTSC's approximately 480 active scanning lines, if filling eight degrees of visual arc, would exactly match one arc minute of acuity. That condition occurs when the viewer is slightly farther from the screen than seven times the picture's height.[33]

It is the case that there is an optimum viewing distance. Farther than the optimum distance, the viewer's visual acuity precludes seeing the full resolution being presented. Closer than the optimum distance, elements of the display structure become visible, effectively preventing the viewer from "seeing the forest for the trees."

Consider the image below, an extreme example of this phenomenon. It is possible that you have seen it previously in this paper. It is a lower-case *O* in the Times New Roman typeface used in this text, as it might be depicted on some computer's color LCD screen. As in Figure 7, at a sufficient viewing distance (squinting might help), the image will change, this time from colored blocks to a round, black, lower-case *O*.

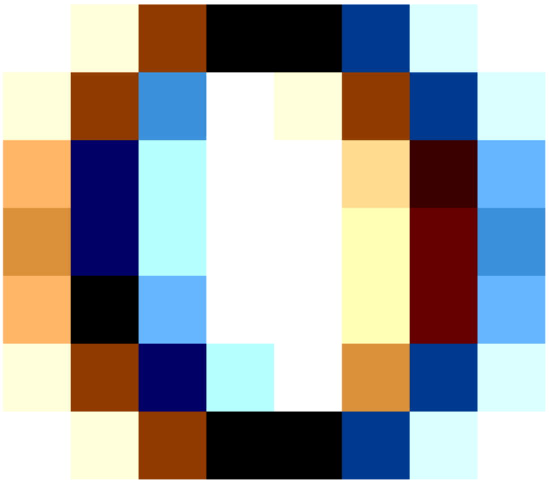


Figure 12: A Fixed-Grid Display *O*

If Snellen were correct about normal visual acuity being 30 cpd, and if active scanning lines (in interlaced television) directly determined perceived resolution, then it *would* be accurate to say that the optimum viewing distance for NTSC video is about seven times the picture height. It would still *not* be accurate, however, to say that viewers typically watched NTSC video at seven times the picture height.

Bernard Lechner, a researcher at RCA Laboratories, conducted a survey of television-viewing distances during the NTSC era. What he found was that those he surveyed watched television from a viewing distance of approximately nine feet, regardless of screen size, a figure that came to be known as the Lechner Distance. Richard Jackson, a researcher a Philips Laboratories in the UK, conducted a similar survey and found a similar three meters, the Jackson Distance.[34]

Assuming, again, that normal visual acuity allows detection of features subtending one minute of visual arc and that active scanning lines determine television resolution, then at the Lechner Distance it would be essentially impossible to see greater than NTSC resolution on a 25-inch four-by-three TV screen. The picture height of a 25-inch screen is 15 inches (1.25 feet); seven times its height is 8.75 feet. The optimum viewing distance for 480 active lines is actually 7.15 times the picture height (1/(2*tan(8 degrees/2))), which means the optimum 25-inch NTSC TV-viewing distance would be 8.94 feet, almost exactly the Lechner Distance.

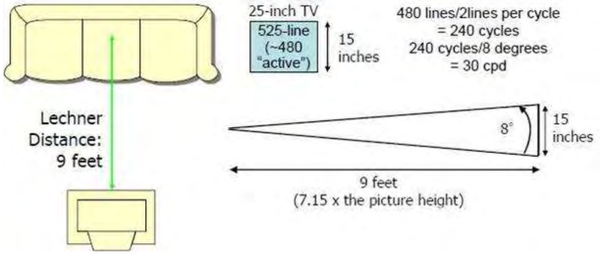

25-inch TV
525-line (~480 "active")
15 inches
480 lines/2lines per cycle = 240 cycles
240 cycles/8 degrees = 30 cpd
Lechner Distance: 9 feet
8°
15 inches
9 feet
(7.15 x the picture height)

Figure 13: U.S. Standard-Definition Viewing

According to figures from the Consumer Electronics Association, the average TV-screen size shipped by factories to U.S. dealers through the year 2000 was under 25 inches.[35] Sales to consumers typically follow a year after factory sales, and TV replacement takes years after that, so, if definition beyond NTSC cannot be perceived on a 25-inch screen, there would seem to have been no incentive for a move to HDTV in the U.S.

In Japan, according to this theory, rooms are smaller, so current HDTV had its origins there. Unfortunately for the theory, TV screen sizes in Japan were also smaller.

Psychophysics

In fact, there *are* reasons why HDTV detail looks better even to American viewers. First, 30 cpd is *not* the limit of human visual perception. In testing conducted in Japan on ultrahigh-definition television (UHDTV), the test subjects were found to have an *average* visual acuity not of 20/20 but of 20/10 (i.e., able to distinguish features twice as small as Snellen's criterion).

That should have meant that their visual acuity was 60 cpd instead of 30. The testing revealed, however, that the subjects were able to distinguish the "realness" of images as high as 156 cpd (the highest spatial frequency that was measured), more than five times supposedly "normal" acuity and more than 2.5 times even 20/10 acuity. "Realness" (degree to which an image was perceived to be comparable to a real object) rose rapidly to beyond 50 cpd and then slowed, but it did increase to 156 cpd (the highest tested), and the data suggest it would continue to increase (slowly) beyond that point.[36]

Realness, like other words ending in the suffix *-ness*, such as *brightness* and *loudness,* is a psychophysical sensation (a psychological response to a physical stimulus). And psychophysical sensations tend to be based on more factors than just the measurable physical phenomenon most closely associated with them. Thus, luminance, alone, does not determine brightness, and sound-pressure level, alone, does not determine loudness.

Similarly, there is a psychophysical sensation associated with picture definition but not determined exclusively by it. That sensation is called *sharpness.*

At the top of the next page is a graph of a typical modulation transfer function (MTF). In the case of spatial definition of the luma (gray-scale) component of images, the modulation is changes between bright & dark.

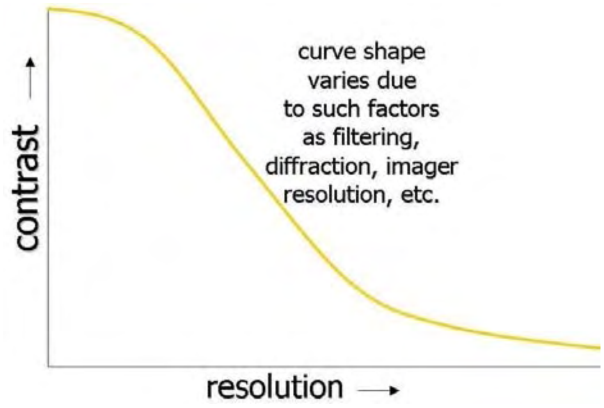


Figure 14: An MTF Curve

The curve of Figure 14 could be that of a lens or a television camera or a complete television system, “from scene to seen.” Of most interest, with regard to the sensation of sharpness, is the area under the curve.

There are two significant schools of thought about the relationship of that area to sharpness. One is based on the work of Otto H. Schade, Sr. at RCA Laboratories.[37] The other is based on the work of Erich Heynacher at Zeiss.[38] The former suggests that the sensation of sharpness is proportional to the square of the area under the curve, the latter that it is proportional to the area.

In either case, as shown in Figure 14, image sharpness is most affected by the area under the “shoulder” of the curve (at left) and least by the area under the “toe” of the curve (at right). Sony took advantage of the low contribution of the highest spatial frequencies to sharpness in the design of the HDCAM recording system. It drops 25% of the luma definition at much less loss of sharpness.[39]

One of the factors affecting the shape of the MTF curve is number of digital samples. Digital sampling and reconstruction require filtering. The graph at the top of the next column is a basic filter shape, the so-called SINC or (sin x)/x function. The horizontal scale is arbitrary; on the vertical scale, *1* represents 100% modulation transfer.

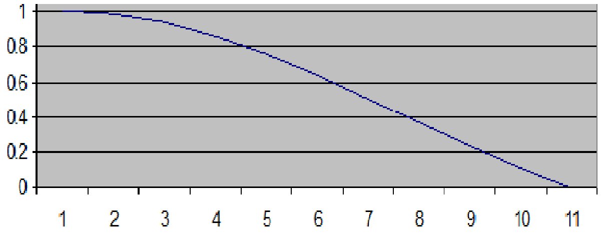


Figure 15: SINC Function Filter

If the vertical axis represents contrast and the horizontal axis represents image definition, then, if number 11 represents, say, 1920 active samples per line, the contrast at 1920 is zero. If, however, number 11 represents 3840 samples, the contrast at 1920 is 64%, a very significant difference.

The next two figures illustrate real-world examples. They are taken from the Bob Atkins Photography web site and are used here with permission.[40]

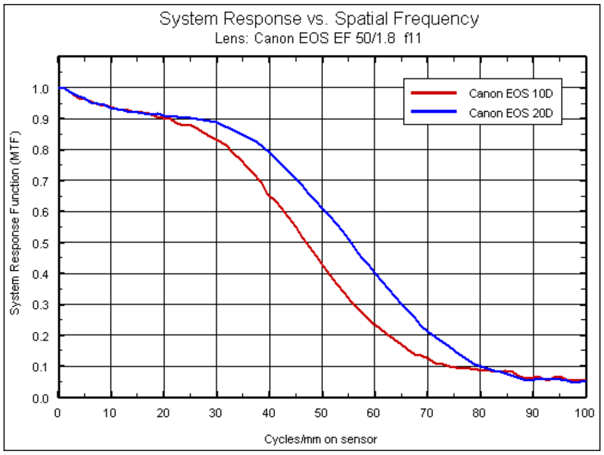


Figure 16: MTF Comparison of Two Cameras
© Bob Atkins

The red (left) curve of Figure 16 is from a Canon EOS 10D still camera, which uses an effective 3072 x 2048 photosite image sensor. The blue (right) curve is from a Canon EOS 20D camera, which uses an effective 3504 x 2336 photosite image sensor.

The horizontal linear increase in definition is just 14%, but significant additional area is created under the MTF curve of Figure 16.

The resulting increase in sharpness can be seen in Figure 17 below.

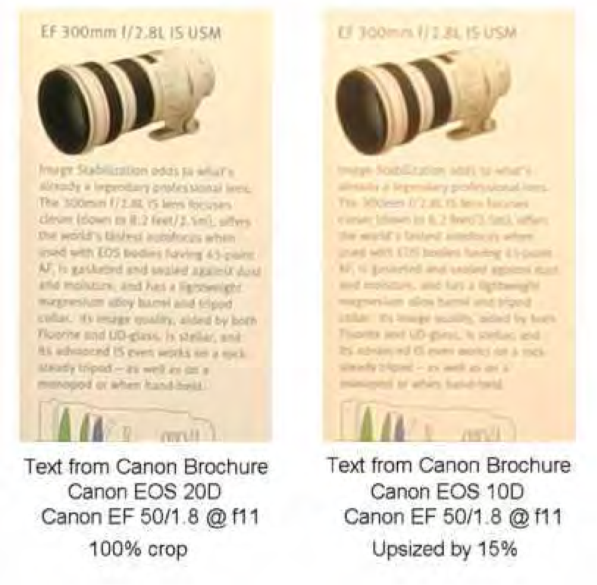

EF 300mm f/2.8L IS USM
Image Stabilization adds to what's already a legendary professional lens. The 300mm f/2.8L IS lens focuses closer (down to 8.2 feet/2.5m), offers the world's fastest autofocus when used with EOS bodies having 45-point AF, is gasketed and sealed against dust and moisture, and has a lightweight magnesium alloy barrel and tripod collar. Its image quality, aided by both fluorite and UD-glass, is stellar, and its advanced IS even works on a rock steady tripod – as well as on a monopod or when hand-held.
Text from Canon Brochure
Canon EOS 20D
Canon EF 50/1.8 @ f11
100% crop
EF 300mm f/2.8L IS USM
Text from Canon Brochure
Canon EOS 10D
Canon EF 50/1.8 @ f11
Upsized by 15%

Figure 17: Sharpness Difference
© Bob Atkins

The text, photo, and drawing details above indicate very little difference in image definition, as might be expected from the small (14%) linear increase. The additional area under the MTF curve, however, makes the increased sharpness of the left image readily apparent. The vertical definition difference between so-called 1080-line HDTV and NTSC is about 225% (it is a similar 213% from 1080-line HD to so-called "4K").

BEYOND HDTV

Differentiating the Theatrical Experience

Average U.S. weekly movie attendance in every year from 1945 through 1948 was 90 million. By 1950, it was down to 60 million; by 1953, it was just 46 million.[41] The cause of the drop was apparently the rise of home television. The movie industry turned to wider screens, larger (higher-definition) film formats, and such offerings as stereoscopic 3D in order to differentiate the movie-going experience from that of watching television.

In 2011, there were just 24.6 million weekly cinema admissions in the U.S.[42] Only 27% of the number of movie tickets of 1948 were sold in the same year that the population grew to 213% of its 1948 level.[43] So 3D and higher-definition formats are still under consideration in a digital-cinema era.

Today, instead of 70-mm film, the movie industry discusses "4K," images having 4096 active picture elements (pixels) per row (with a number of rows appropriate to the image aspect ratio). The earliest digital-cinema projectors were "1.3K" (comparable to 720p HDTV); many current installations are 2K (comparable to 1080-line HDTV).[44]

Traditional cinema seating arrangements created a wide range of viewing distances for audiences, as shown in Figure 18 below, courtesy of Warner Bros. Technical Operations. Figure 19, courtesy of the same source, shows a typical more-recent auditorium with stadium-style seating. The scales are calibrated in picture heights.

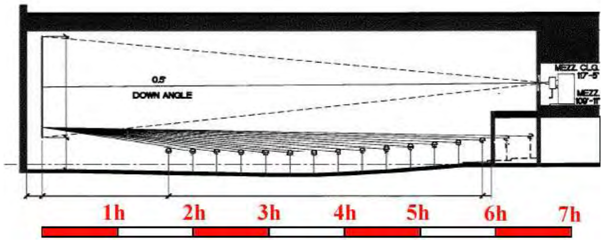


Figure 18: Traditional Cinema Seating

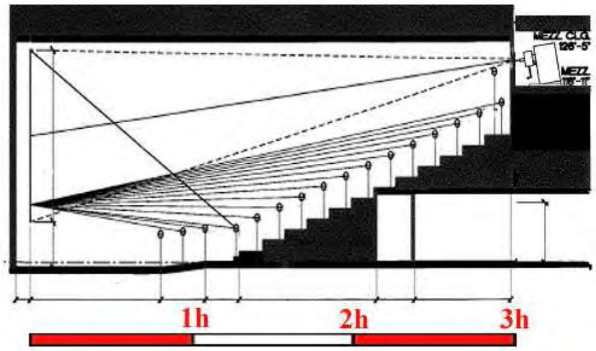


Figure 19: Cinema Stadium-Style Seating

The bulk of the audience is closer to the screen in stadium seating and, therefore, might benefit from additional image definition. Figure 20, below, courtesy of ARRI,[38] shows that definition even beyond 8K (8192 active pixels per row) should be perceptible even at just 20/20 visual acuity in some seats in some cinemas.

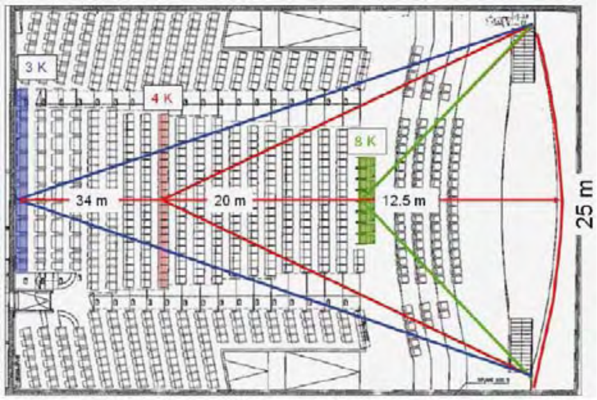


Figure 20: 20/20 Resolvable Definitions

In practice, however, unlike television consumers, who can compare picture definitions side by side in stores (and generally at closer viewing distances than they would experience in homes), cinema attendees cannot easily compare definitions in different auditoriums. Definition of 1.3K was acceptable to audiences when it was used for digital cinema. Perhaps, like "70 mm," "4K" will be a promotional tool.

Production and Post

Even before the modern HDTV era, television-camera manufacturers used oversampling (generally in the horizontal direction) to increase image sharpness. NTSC broadcast video bandwidth restricts horizontal luma definition to approximately 440 pixels; some broadcast-camera image sensors had more than 1100 photosensitive sites per row.

At the beginning of the modern HDTV era, the difficulty of making sensors with even as many as 1920 photosites per line precluded oversampling. The introduction of image sensors into still cameras, mobile telephones, laptop computers, and other devices, however, provided economies of scale allowing multi-"megapixel" sensors to be produced.

Even in the camera-tube era, some cameras used patterned color filters at the faceplate of a single imaging tube to capture color pictures instead of using color-separation prisms with an imaging tube for each of the three primary colors. In the solid-state imaging era, it has become common in still cameras to use a patterned filter over a single image sensor.

In the common Bayer pattern shown below, green-filtered photosites represent half the sites on the sensor. Red- and blue-filtered photosites represent one-quarter of the sites, each. Recreating a full-color image requires a "demosaicking" process to remove the spatial color effects of the filter.

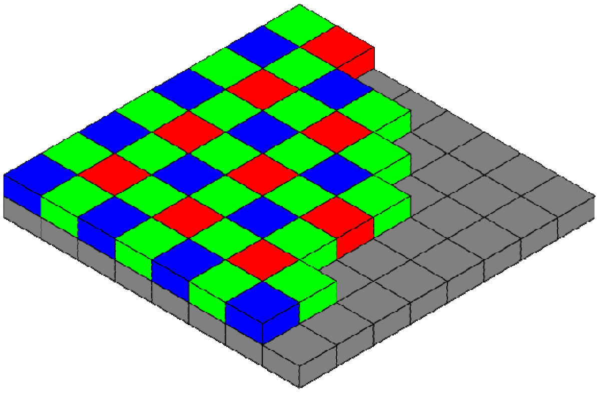


Figure 21: A Bayer-Filter Pattern[45]

There is no consensus about how on-chip color filtering should affect the description of resolution. There is also no consensus about whether 4K requires 4096 samples per line or whether 3840 (twice HDTV's 1920) are sufficient. Thus, one can find labeled 4K, at a single equipment exhibition, cameras with between 8.3 and 20.4 million photosites per image sensor, and with one to three sensors (some older "4K" cameras also used four 2.1-million photosite sensors).[46]

Aside from any advantages in visual definition or sharpness, 4K offers benefits in production and post, as the next three figures illustrate. Figure 22 shows a high-definition video image as a pixel-for-pixel subset of a 4K image, allowing reframing or even zooming after shooting.

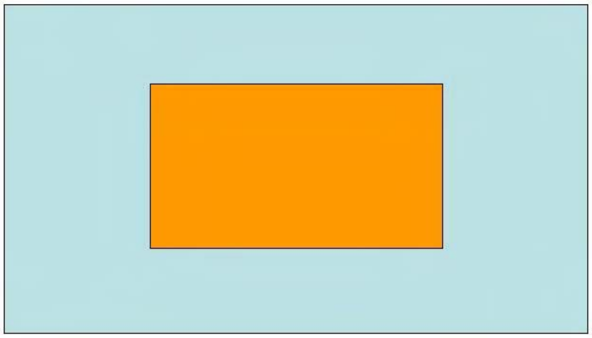


Figure 22: HD as a Subset of 4K

Figure 23, below, shows the effects of image stabilization, which normally causes trimming of the image (illustrated in available short, downloadable video clips[47]). The light inner rectangle (behind the others) is a desired HD image. The skewed rectangle in front of it is the actual image captured by an unstable camera. The smallest rectangle is the trimming that post-production image stabilization would produce. But starting with the outer 4K image allows the full desired HD image to be stabilized.

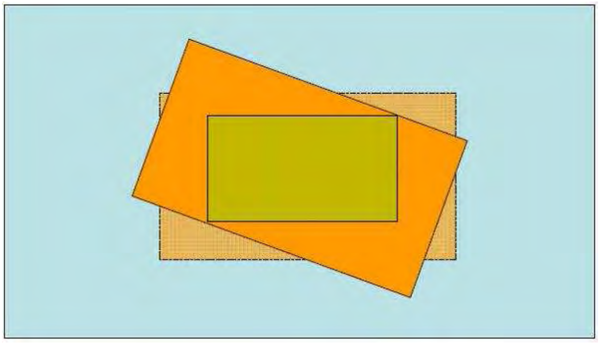


Figure 23: HD Image Stabilization in 4K

Figure 24, at the top of the next column, shows an unusual application of 4K in stereoscopic scene capture. The Zepar stereoscopic lens system attached to a Vision Research Phantom 65 camera, provides side-by-side stereoscopic images on the same image sensor, simplifying processing.[48]

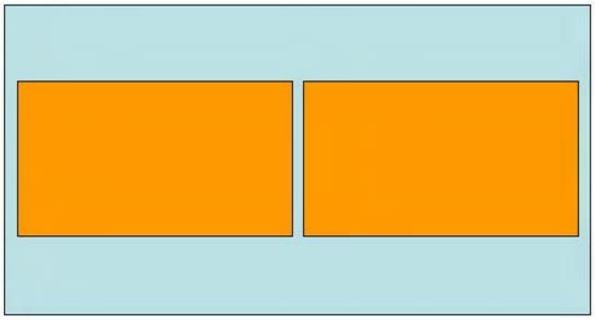


Figure 24: Stereoscopic HD on a 4K Sensor

Very Large TV Screens

When Panasonic introduced its 103-inch plasma TV, at the time the largest consumer flat-panel display, it had the common HDTV definition of 1920 x 1080. As a result, at the Lechner Distance, the image structure could have been perceptible even to viewers with visual acuity somewhat less than 30 cpd.

When the same company later (in 2008) introduced a 150-inch plasma TV, an HDTV image structure would have been even more visible, perceptible even to viewers with impaired vision. The definition of that display, therefore, was 4K (4096 x 2160).

At the top of the next page is an image created by John R. Vollaro for Leon D. Harmon at Bell Labs around 1968, used with permission. Like Figure 12, it doesn't look like what it is when its edges can be seen.

Like Figure 12, the image will appear to be as it should when viewed from a distance. Unlike the blocks of Figure 12, which were created to help make a color, fixed-grid display present black, round edges, the blocks of Figure 25 were created to obscure a natural photographic image for perceptual study.[49]

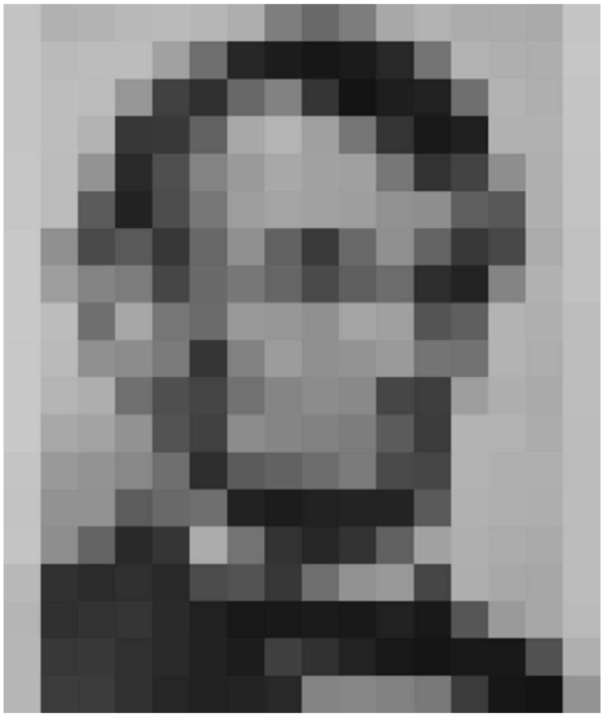


Figure 25: Bell Labs Block Version of Portrait

Surrealist artist Salvador Dali painted a version of the image in 1976, which like Figure 7, changes from one thing to another at a particular retinal angle. The name of the painting is very descriptive: *Gala Contemplating the Mediterranean Sea, Which at Twenty Meters Becomes a Portrait of Abraham Lincoln.*[50]

The effect of Figure 25 occurs because pixels are mathematical points, not little squares, rectangles, or even dots.[51] There are solutions other than increasing display definition, however, such as optical filtering to blur the edges. Holding ground glass or even waxed paper in front of Figure 25, for example, can also reveal its hidden content. Such low-pass optical filtering is commonly used in broadcast television cameras (although it becomes problematic in color-filtered single-sensor cameras: should the filter be appropriate to the luma, the green, or the other colors?).

Aside from their pixel definition and resulting sharpness, very large television displays also stimulate more of the visual field at any given distance. Research into UHDTV (which can be a form of either 4K or 8K) has shown that the increased visual angle not only increases "presence" sensation but also increases dynamic visual acuity (the ability to perceive fine detail moving relative to the eye's retina).[36] Of course, as shown in Figure 22, a very large 4K screen can also be seen as providing an HDTV image in each quadrant.

Distributing 4K

Very large television displays are, and should continue to be, rare in homes. Again, the average screen size of TV sets shipped to the U.S. through 2000 was less than 25 inches. Screen-size increases continued slowly through 2005, followed by a spurt in 2006 as inexpensive widescreen HDTV sets became available.[52]

Size growth then slowed again. According to Display Search, in 2010, the average TV screen size for shipments to North America (which has the world's highest average) was 36 inches. In 2012, it is expected to be 37.8 inches. In 2014, it is expected to be 39.2 inches. Globally, TV screen sizes of as little as 50-inch and above accounted for only 5.3% of shipments in 2010 and are expected to account for only 6.3% in 2014.[53]

Figure 26, below, shows actual and estimated average screen sizes for global TV shipments. Not only is the average below 36 inches, but it also appears to be leveling off.[53]

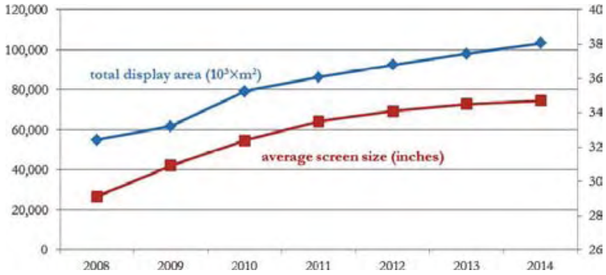


Figure 26: World TV Shipment Average Sizes

Given those screen sizes, it is unlikely that, at the Lechner Distance, average viewers will notice the pixel-grid structure of their television displays. They will also not be close enough to their displays to appreciate the high-definition detail offered by a quarter of a 4K display (or a sixteenth of an 8K display). They *should* be able to appreciate the additional sharpness that 4K image capture offers, but, as Sony's HDCAM filtering indicates, they will see the bulk of that sharpness even on ordinary HDTVs.

It would seem, therefore, that there is little for an average TV viewer to gain from 4K display resolution. The TV-set industry, however, is in a quandary. It was described this way in *The New York Times* in 2011:

"By now, most Americans have taken the leap and tossed out their old boxy televisions in favor of sleek flat-panel displays. Now manufacturers want to convince those people that their once-futuristic sets are already obsolete.

"After a period of strong growth, sales of televisions are slowing. To counter this, TV makers are trying to persuade consumers to buy new sets by promoting new technologies."[54]

That article, which appeared at the time of the 2011 Consumer Electronics Show (CES), indicated that such features as stereoscopic 3D and Internet connections "have not generated much excitement so far." At the 2012 CES, therefore, a new feature being promoted was 4K (and even 8K) definition, with demonstrations from such major manufacturers as AMD, JVC, LG, Panasonic, Sharp, Sony, and Toshiba.[55-58] *Consumer Reports* called 4K "one of the most talked about innovations" at the show.[59]

Aside from promotion by TV-set manufacturers, 4K programming is just starting to become available, primarily in the form of movies, based in part on the production and sharpness advantages of 4K and in part on the use of 4K to differentiate digital cinema from home theater. The late-2011 American version of *The Girl with the Dragon Tattoo,* for example, has been called "the first large-scale end-to-end 4K digital cinema release."[60]

Portions of the 2012 Olympic Games are also expected to be shot and shown in beyond-HDTV resolutions.[61] Thus, cable-television operators might wish to take advantage of all of the promotion by providing a 4K offering. Fortunately, it need not require a large amount of data capacity.

All else being equal, a 4K image sensor has more than four times as many photosites as a 1080-line HDTV image sensor. A 4K display similarly must deal with more than four times as many picture elements.

As might be expected, an uncompressed, 8K (7680 pixels per line, or four times HD's 1920, rather than 8192) "Super Hi-Vision" link would require 16 high-definition serial digital interface (HD-SDI) connections.[62] It is not clear, however, that 4K or 8K require multiples of high-definition data rates in the compressed domain.

There are three main reasons. One may be seen in the previous Figures 14 and 16. As detail gets finer, the energy in the signal is reduced, so there is less to compress.

Another reason relates to motion estimation in bit-rate reduction (BRR) systems that take advantage of temporal redundancy as well as spatial redundancy. The better defined a point is, the more accurately its motion can be estimated and, therefore, the lower the bit rate at which errors will be imperceptible.

Both of those factors suggest that, in an ideal BRR system, the "overhead" to increase from HDTV (or 2K) to 4K will be very much less than an additional three times the original signal value. The third factor is that, absent the very large retinal angle of a cinema screen or very large TV display, viewers are less sensitive to image defects, or, as one BRR-comparison paper put it, "the quality requirements are more stringent when the viewer is in a cinema."[63]

Figure 27, below, is based on a graph in another paper comparing BRR systems for digital cinema. The full graph compares seven BRR systems out to data rates of 260 Mbps for the test sequence called *CrowdRun*. At those high data speeds, the 2K systems all outperform the 4K systems in PSNR.[64]

The small section of the graph shown below, however, is restricted to such low data rates as might be used for delivery of HDTV on a cable-television system. As shown by the five identified data points (all JPEG2000), at those low data rates (starting at 14 Mbps), 4K actually outperformed 2K. It was only beyond roughly 26 Mbps (extrapolated) that 2K outperformed 4K.

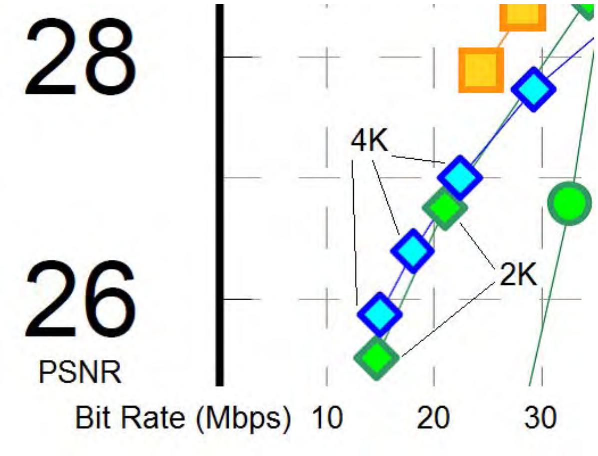


Figure 27: 2K vs. 4K BRR Comparison[64]

BRR quality results can vary according to many factors, but Figure 27 shows that 4K can be transmitted at rates comparable to HDTV with comparable results. It is certainly conceivable that a layered transmission system can also be used, adding only 4K's additional information to that already carried for HDTV. It is not clear, however, what the efficiencies of such layered transmission would be in comparison to the use of a single signal for 4K distribution.

CONCLUSIONS

In program production, 4K is well established and growing. Cameras are available from ARRI, Astro, JVC, Red, Sony, and Vision Research and have also been shown by Canon, Dalsa, Hitachi, Ikegami, Lockheed-Martin, Meduza, NHK, and Olympus.[65]

Though the *4K* designation of some of these cameras can be questioned (largely due to the use of color-filtered single image sensors), all are intended to capture definitions beyond those of HDTV. There is even a technique to extract 4K resolution from masked HDTV image sensors, so as to reduce uncompressed data rates.[66]

In post production, 4K is also well established and growing. *The Girl with the Dragon Tattoo* might be "the first large-scale end-to-end 4K digital cinema release," but all of the individual processes used have been available for some time.

In cinema, 4K is also established and growing. NHK's Super Hi-Vision has been used in cinema-like applications (community viewing of a single, giant screen in a dark room), intended to be viewed at just 0.75 times the picture height.[67]

Super Hi-Vision is also intended to provide a home-viewing experience. Although displays with 4K and even 8K resolutions have been shown, it is not clear at this time either that sufficiently large displays will be purchased by consumers or that consumers will move sufficiently close to smaller displays to give them the “presence” and “realness” intended for Super Hi-Vision.

Figure 28, below, shows a 152-inch plasma TV. Simply getting it into a room in a home is cause for concern. Even a 152-inch size is too small for 0.75-height viewing at the Lechner distance; that would require a 294-inch screen, with a 12-foot-high image (not counting its frame). Clearly, as-intended 8K Super Hi-Vision viewing in the home will require a change in viewing-distance habits.

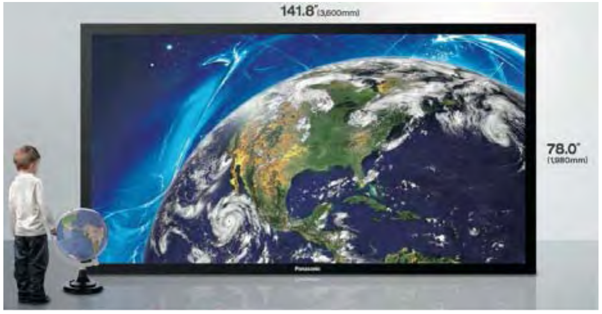


Figure 28: Panasonic 152-inch Plasma TV

The increased sharpness of beyond-HD-resolution imaging is largely available to viewers using existing TV sets. The extraordinary images of 4K and 8K television displays have been reported by observers who could view them at closer than home-viewing distances (e.g., the Lechner Distance).

Cable-television operators can nevertheless take advantage of the promotional aspects of moves to 4K resolution by offering 4K distribution. It is not clear whether *any* increase in bit rate is required, but, due to the low energy of the highest octave of spatial frequencies in a typical, real-world 4K-captured image, improved motion estimation provided by better-defined pixels, and relative insensitivity to compression artifacts at typical TV viewing distances and display sizes, any such increase should be minimal.

REFERENCES

1. *The Compact Edition of the Oxford English Dictionary*, Oxford University Press, 1971

2. Snellen, Dr. H., *Probebuchstaben, zur Bestimmung der Sehscharfe* [Sample Letters, for determining visual acuity], P. W. van de Weijer, 1862 http://archive.org/details/probebuchstaben01snelgoog

3. *A.D.A.M. Medical Encyclopedia*, National Center for Biotechnology Information, U.S. National Library of Medicine, reviewed May 24, 2010 http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002021/

4. Schneider, Joel, “Block Letter Eye Chart,” created May 2002 http://www.i-see.org/block_letter_eye_chart.pdf

5. Bordsen, John, “Eye Chart Still the Standard for Vision,” *The Seattle Times*, August 9, 1995 http://community.seattletimes.nwsource.com/archive/?date=19950809&slug=2135585

6. Fletcher, H., and Munson, W.A., “Loudness, its definition, measurement and calculation,” *Journal of the Acoustic Society of America*, vol. 5, 82-108 (1933) http://www.sfu.ca/media-lab/archive/2011/386/readings/Misc.%20Readings/Loudness,%20Its%20Definition,%20Measurement%20and%20Calculation%20.pdf

7. Oliva, Aude, and Schyns, Philippe G., “Dr. Angry and Mr. Smile: a series,” *Hybrid Images*, Computational Visual Cognition Laboratory, Massachusetts Institute of Technology, 2006 http://cvcl.mit.edu/hybrid_gallery/smile_angry.html

8. Fisher, David E., and Fisher, Marshall Jon, *Tube: the invention of television*, Counterpoint, 1996 http://books.google.com/books?id=eApTAAAAMAAJ

9. Burns, Russell W., *John Logie Baird: Television pioneer*, IET, 2000 http://books.google.com/books?id=5y09hpR0UY0C

10. Abramson, Albert, *The History of Television, 1880-1941*, McFarland, 1987 http://lccn.loc.gov/86043091

11. Prescott, George Bartlett, *Electricity and the Electric Telegraph*, volume 2, D. Appleton, 1892 http://books.google.com/books?id=9_5KAAAAYAAJ

12. Nipkow, Paul, German patent 30101 - 1885-01-15, link to the UK Intellectual Property Office copy: http://bit.ly/H4f5Iu

13. Ives, Herbert E., “Television,” *The Bell System Technical Journal*, volume 6, October 1927 http://www.alcatel-lucent.com/bstj/vol06-1927/articles/bstj6-4-551.pdf

14. “Television Shows Panoramic Scene Carried by Sunlight,” *The New York Times*, July 13, 1928 http://select.nytimes.com/gst/abstract.html?res=F50D13F7395C177A93C1A8178CD85F4C8285F9

15. Schubin, Mark “The First Sports Video,” SchubinCafe.com, July 10, 2009, http://www.schubincafe.com/2009/07/10/the-first-sports-video/

16. *Television* magazine, volume 1, number 2, Experimenter Publishing, New York, July 1928

17. Goebel, Gerhart, “From the history of television - The first fifty years,” *Bosch Technische Berichte*, volume 6 (1979), number 5/6; note: much of the information is available on the web from the Deutsches Fernsehmuseum Wiesbaden (see next reference)

18. “Kapitel 6 (ab 1923)” [Chapter 6 (from 1923)], Deutsches Fernsehmuseum Wiesbaden http://www.fernsehmuseum.info/fernsehgeschichte06.html

19. *Report of the Television Committee*, His Majesty’s Stationery Office, 1935 http://www.thevalvepage.com/tvyears/articals/comrep/comrep.htm

20. Preston, S. J., “The Birth of a High Definition Television System,” *Journal of the Television Society*, volume 7, number 3, 1953

21. *Broadcasting* magazine, May 1, 1939

22. National Television System Committee, *Proceedings*, 1940-1941 http://lccn.loc.gov/45051235

23. Pemberton, Alan, “Line Standards,” *World Analogue Television Standards and Waveforms*, Sheffield, England, 2010 http://www.pembers.freeserve.co.uk/World-TV-Standards/Line-Standards.html

24. Pemberton, Alan, “Timeline” from “Overview,” *World Analogue Television Standards and Waveforms*, Sheffield, England, 2010 http://www.pembers.freeserve.co.uk/World-TV-Standards/index.html#Timeline

25. Schubin, Mark, “Special Report: TV Around the World,” *Videography*, March 1979

26. Hart, Samuel Lavington, “Improvements in Apparatus for Transmitting Pictures of Moving Objects and the like to a distance Electrically,” British patent 15,270, published 25th June, 1915, link to the UK Intellectual Property Office copy: http://bit.ly/H7Gqvx

27. Yanczer, Peter, “Ulises Armand Sanabria,” in “Mechanical Television,” Early Television Museum, http://www.earlytelevision.org/u_a_sanabria.html

28. Dinsdale, A., “Television in America To-day,” *Journal of the Television Society*, volume 1, 1932 http://books.google.com/books?id=O2cPAAAAIAAJ

29. Watkinson, John, *The Art of Digital Video*, Focal Press, 2008 http://books.google.com/books?id=8uLEXlN9ouAC

30. Hsu, Stephen C., “The Kell Factor: Past and Present,” *Journal of the Society of Motion-Picture and Television Engineers*, volume 95, no. 2, February 1986 http://journal.smpte.org/content/95/2/206.abstract

31. Yerrick, Damian, “Demonstration of interlace and so-called ‘interline-twitter,’ based on part of an RCA Indian Head Test Card, ca. 1940,” http://en.wikipedia.org/wiki/File:Indian_Head_interlace.gif

32. *Television Digest*, volume 11, number 36, 1955

33. Taylor, Jim, Johnson, Mark R., and Crawford, Charles G., *DVD Demystified, Third Edition*, McGraw-Hill, 2006 http://books.google.com/books?id=ikxuL2aX9cAC

34. Poynton, Charles, *Digital Video and HDTV: Algorithms and Interfaces*, Morgan Kaufman, 2003 http://books.google.com/books?id=ra1lcAwgvq4C

35. Wargo, Sean, Consumer Electronics Association press presentation, New York, November 2006

36. Sugawara, Masayuki, et al., “Research on Human Factors in Ultrahigh-Definition Television (UHDTV) to Determine Its Specifications,” *SMPTE Motion Imaging Journal*, vol. 117, no. 3, April 2008 http://www2.tech.purdue.edu/Cgt/courses/cgt512/discussion/Chastain_Human%20Factors%20in%20UDTV.pdf

37. Schade, Otto H., *Image Quality : a comparison of photographic and television systems*, RCA Laboratories, 1975, republished in the *SMPTE Journal*, volume 96, number 6, June 1987, http://journal.smpte.org/content/96/6/567, described more recently here, http://www.panavision.com/sites/default/files/24P%20Technical%20Seminar%202.pdf

38. Heynacher, Erich, “Ein Bildgütemaß auf der Grundlage der Übertragungstheorie mit subjektiver Bewertungsskale” [Objective Image Quality Criteria, based on transformation theory with a subjective scale], *Zeiss Mitteilungen*, volume 3, number 1, 1963, described in the *ARRI 4K+ Systems* brochure http://www.scribd.com/doc/52408729/4K-Systems-Arri

39. Thorpe, Laurence J., Nagumo, Fumio, and Ike, Kazuo, “The HDTV Camcorder and the March to Marketplace Reality,” *SMPTE Journal*, volume 107, number 3, March 1998, http://www.smpte-pda.org/resources/The+HDTV+CamorderThorpeMar1998.pdf

40. Atkins, Bob, “Canon EOS 20D DSLR Review,” *Bob Atkins Photography* http://www.bobatkins.com/photography/digital/eos20d.html

41. Vogel, Harold L., *Entertainment Industry Economics: A guide for financial analysis,*, Cambridge University Press, 1986 (with data from *Reel Facts*) http://books.google.com/books?id=3TwrQgAACAAJ

42. “Yearly Box Office,” *Box Office Mojo*, http://boxofficemojo.com/yearly/

43. “Population Estimates,” Historical Data, United States Census Bureau, http://www.census.gov/popest/data/historical/index.html

44. “Help Documents,” *The Big Screen Cinema Guide*, http://www.bigscreen.com/about/help.php?id=36

45. Burnett, Colin M. L., “A bayer pattern on a sensor in isometric perspective/projection,” 28

December 2006, http://en.wikipedia.org/wiki/File:Bayer_pattern_on_sensor.svg

46. Schubin, Mark, “Fun Out of the Sun in Las Vegas - 2011: a different kind of NAB show,” 2011 May 19, http://www.schubincafe.com/2011/06/01/nab-2011-wrapup-washington-dc-smpte-section-may-19-2011/

47. Schubin, Mark, “Things You Can or Can’t Fix in Post: Video Acquisition,” San Francisco Public Television Quality Group, 2010 June 8, http://www.schubincafe.com/2010/06/15/things-you-can-or-can%E2%80%99t-fix-in-post-video-acquisition/

48. “Phantom 65-Z3D System,” Abel Cine, http://about.abelcine.com/wp-content/imported/images/pdf/phantom_65-z3d.pdf

49. Vollaro, John R. “Commentary on the History of ‘Photomosaic’ Images,” March 2006, http://vollaro.com/WebScrapbook/docs/Clinton/NudeStory.html

50. “Gala Contemplating the Mediterranean Sea which at Twenty Meters becomes a Portrait of Abraham Lincoln,” Authentic Society, http://www.authenticsociety.com/about/GalaMediterraneanLincoln_Dali

51. Smith, Alvy Ray, “A Pixel Is *Not* A Little Square, A Pixel Is *Not* A Little Square, A Pixel Is *Not* A Little Square!,” Microsoft Computer Graphics, Technical Memo 6, July 17, 1995, http://alvyray.com/Memos/CG/Microsoft/6_pixel.pdf

52. Wargo, Sean, Consumer Electronics Association, e-mail communication to the author, January 31, 2007

53. Park, Won Young, et al., *TV Energy Consumption Trends and Energy Efficiency Improvement Options*, Environmental Energy Technologies Division, International Energy Studies Group, Ernest Orlando Lawrence Berkeley National Laboratory, July 1, 2011 http://www.superefficient.org/~/media/Files/SEAD%20Televisions%20Technical%20Analysis.pdf

54. Grobart, Sam, “A Bonanza in TV Sales Fades Away,” *The New York Times*, January 5, 2011 http://www.nytimes.com/2011/01/06/technology/06sets.html

55. Pogue, David, “Sampling the Future of Gadgetry,” *The New York Times*, January 11, 2012 http://www.nytimes.com/2012/01/12/technology/personaltech/in-las-vegas-its-the-future-of-high-tech-state-of-the-art.html

56. Putman, Peter, “HDTV Expert - CES 2012: Another Opening, Another Show,” *HDTV Magazine*, January 18, 2012, http://www.hdtvmagazine.com/columns/2012/01/hdtv-expert-ces-2012-another-opening-another-show.php

57. Walton, Jerry, “The Best of CES 2012,” *AnandTech*, January 17, 2012, http://www.anandtech.com/show/5437/the-best-of-ces-2012

58. Healey, Jon, “CES 2012: 4K TV sets make their debut, minus the hoopla,” *The Los Angeles Times,* January 11, 2012, http://latimesblogs.latimes.com/technology/2012/01/ces-4k-tv-sets-make-their-debut-minus-the-hoopla.html

59. “CES 2012 Video: Could 4k TV technology bring better 3D TV?” Consumer News, *ConsumerReports.org*, http://news.consumerreports.org/electronics/2012/01/ces-2012-video-what-is-sonys-4k-tv-technology.html

60. Koo, Ryan, “Fincher Reframes in Post! The 4K Release of ‘The Girl with the Dragon Tattoo,” *NoFilmSchool*, December 28, 2011, http://nofilmschool.com/2011/12/fincher-reframes-post-4k-release-the/

61. Carter, Jamie, “BBC Talks Super Hi-Vision Plans for London 2012,” TechRadar.TVs, *TechRadar*, http://www.techradar.com/news/television/bbc-talks-super-hi-vision-plans-for-london-2012-1068914

62. “World’s First Live Relay Experiment of Super Hi-Vision,” *Broadcast Technology*, number 25, Winter 2006, NHK STRL, http://www.nhk.or.jp/strl/publica/bt/en/to0025.pdf

63. Shi, Boxin, Liu, Lin, and Xu, Chao, “Comparison between JPEG2000 and H.264 for Digital Cinema,” *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2008 http://www.cvl.iis.u-tokyo.ac.jp/~shi/files/Shi_ICME08.pdf

64. Baruffa, Giuseppe, Micanti, Paolo, Frescura, Fabrizio, “Performance Assessment of JPEG2000 Based MCTF and H.264 FRExt for Digital Cinema Compression,” *Proceedings of the 16th International Conference on Digital Signal Processing*, July 2009 http://dsplab.diei.unipg.it/files/baruffa_DSP2009.pdf

65. Schubin, Mark, “Beyond-HD-Resolution Cameras and their Workflows,” 18th HPA Tech Retreat, Hollywood Post Alliance, Indian Wells, California, 2012 February 15, http://www.schubincafe.com/2012/03/11/4k-hpa-tech-retreat-2012/

66. Schöberl, Michael, et al., “Increasing Image Resolution by Covering Your Sensor,” 18th HPA Tech Retreat, Hollywood Post Alliance, Indian Wells, California, 2012 February 17, http://data.memberclicks.com/site/hopa/2012_TR_Pres_SFoessel.pdf

67. “8K Television System ‘Super Hi-vision’ is the TV technology of our dreams,” NHK, http://www.nhk.or.jp/digital/en/superhivision/

CREATING CONTENT WITH EXTENDED COLOR GAMUT FOR FUTURE VIDEO FORMATS

J. Stauder, J. Kervec, P. Morvan, C. Porée, L. Blondé, P. Guillotel
Technicolor R&D France, jurgen.stauder[jonathan.kervec]@technicolor.com

Abstract

New technologies in capturing and displaying images with extended color gamut and new standards for wide gamut color encoding enable a new market of extended-color-gamut content (video, images, games, electronic documents). What is the challenge and what are the issues when feature film production goes for extended color gamut? This paper discusses two topics: digital capture of extended color gamut scenes and color correction of wide color gamut footage. In film production, proof viewing and initial color decisions migrate from the post-production facility to the production site. When capturing digitally scenes with extended color gamut, what can be expected to be seen on the proof monitor? This white paper discusses the issues of sensitivity metamerism, color resolution and color clipping. Once captured, color correction creates the aimed looks for digital cinema viewing, TV home viewing, and other possible means of consumption. This paper discusses the issue of color correction with the constraint of multiple means of color reproduction. A new method is presented that supports the colorist to handle multiple color gamuts using the concept of soft gamut alarm.

INTRODUCTION

When looking into history of motion pictures and technology of argentic film, people always tried to enhance image quality and user experience. In 1932, Technicolor invented the 3-color-dye system starting worldwide the transition from black and white to colored motion picture. More recent efforts aimed to enhance resolution and image size from 35mm to 70mm argentic film [1] or from classical 2D film projection to 3D projection [2]. In all these examples, people tried to enhance image quality while preserving as much as possible from existing infrastructure. The color print of 1932 could be projected using the state of the art film projectors of that time. The film reels were the same. When testing 70mm film stock, the constraint was to keep the Digital Intermediate workflow of 35mm technology. For 3D film projection, the inventors [2] used classical film projectors and same film stocks, they just added an optical system.

In television and video, current standardization efforts include the increase of fidelity of color reproduction and the extension of color gamut. Aiming the fidelity of color reproduction, the EBU specified recently the reference monitors to be used in production and post-production [3]. The IEC specified a metadata format called “Gamut ID” to transmit color gamut information for better color reproduction [4, 5]. In order to increase the color gamut (and the image resolution) from High Definition (HD) to Ultra High Definition Television (UHDTV), the ITU-R (WP6C) looks into extending the color gamut. More precisely, they specify a video signal encoding format [6, 7] that allows conveying colors that are more saturated than specified in current HDTV color encoding format ITU-R BT. 709 [8]. Similar efforts have been done in SMPTE and IEC [9,10,11] but these solutions are not widely used.

If the video industry intends to migrate from HDTV to UHDTV, *production*, *distribution* and *consumption* of video needs to be adapted. For *consumption* of extended color gamut, display makers announce for 2012 first OLED TV screens able to show 40% and more of all visible colors (current displays are limited to 33%). Video *distribution* is addressed by ITU-R.

This paper focuses on the *production* of video with extended color gamut and presents two aspects.

First, extended color gamut will have impact on acquisition using digital cameras. While sets usually are prepared in a way that illuminance of surfaces and colors keep within usual ranges, directors now start to use lights and colors with peaky spectrum, or higher saturation. Three issues of digital acquisition will be discussed: sensitivity metamerism, color resolution and color clipping.

The second topic concerns color correction aiming multiple color displays with different, extended, color gamut and viewing conditions. The concept of soft gamut alarm will be introduced and illustrated.

EXTENDED COLOR GAMUT IN DIGITAL ACQUISITION

New requirements in production using digital cameras include the capture of scenes showing colors with wider color gamut. Directors start to light scenes on production sets with colors that are out of the color gamut of usually used proof viewing devices (such as Rec. 709 monitors). For example in music life events, modern spot lights use programmable color filters able to generate light of high degree of saturation. In traditional production using digital cameras, such colors are avoided. In straight forward signal processing, illegal RGB values may be simply clipped somewhere in the imaging chain. This causes the color output on the reference screen to be widely different from the colors that can be seen in the scene. There is a need of controlled handling of out of gamut colors, in which the errors are minimized.

Color encoding

Before discussing camera specific issues, some basic terms are recalled. The skilled color scientist will skip this section. When a color is expressed by color space coordinates, this is called color representation. When color representation includes aspects such as binary encoding and reduced validity such as device or observer dependence, this is called color encoding.

One type of color encoding is scene-referred color encoding. The principle of color encoding has been structured by the ISO [12] for the field of digital photography and desktop publishing, but the definitions are valid for the video domain, too. Scene referred color encoding identifies color coordinates that are meant to be directly related to radiometric real world color values. The raw RGB output values of a digital camera are usually transformed to scene-referred RGB values, such as defined by ITU-R BT.709 [8]. However, we will see later that this relation is ambiguous due to sensitivity metamerism.

Another type of color encoding is output-referred color encoding. As opposed to scene-referred color encoding, output-referred color encoding is used to represent reproduced colors. Output-referred color encoding identifies color coordinates that are prepared for specific output devices with their defined characteristics and viewing conditions. For example, RGB values of a video can be said to be output-referred color encodings since they are intended for a reference display under reference viewing conditions. Well-known output-referred color encodings are for example sRGB display input values or CIE 1931 XYZ values.

Output-referred color encodings are obtained by color matching experiments. An output-referred color space and the related color matching experiment are characterized by:

- the characteristics of the output device driven by the output-referred color coordinates;
- the characteristics of the observer that perceives the colors reproduced by the output device.

Let us take as example the output-referred RGB coordinates being input to a display. The related trichromatic color matching experiment is classical [13] and involves the CIE 1931 standard (human) observer, corresponding to the average behavior of a small group of test persons. In the experiment, an observer compares the color reproduced by the display with the color of a monochromatic light of a specific wavelength. For each wavelength, he adjusts the RGB values such that both colors match. The result of a color matching experiment are three color matching functions (red, green and blue) indicating, for each wavelength, which RGB coordinates should be input to the display in order to match the monochromatic light.

The classical color matching function results in the output-referred RGB color space of the specific RGB display that was used at the time of the experiment. An RGB space can be defined for any other RGB display.

Better known is the output-referred CIE 1931 XYZ space based on an ideal display with XYZ input signals and mathematically derived XYZ primaries. XYZ coordinates encode a color according to these standardized primaries and according to the CIE 1931 standard observer.

Less known is that we could build an $R^C G^C B^C$ or $X^C Y^C Z^C$ output-referred color space that is based on a digital camera as observer. Let us recall that output-referred color spaces not only depend on the aimed display but also on the referred camera used as observer.

Linear output-referred color spaces can be transformed into each other using a linear coordinate transform as far as the same observer is considered. Hunt [13] shows this for RGB-XYZ transform and the SPMTE [14] for different RGB spaces of different displays. Trichromatic observers (such as the human eye or a digital RGB camera) are characterized by the spectral sensitivities of their photoreceptors. The set of three spectral sensitivities are directly linked to a set of three XYZ color matching functions. One set can be derived from the other but they are of different nature.

Color characteristics of digital cameras

The color performance of a camera is determined by a series of elements:

- Optical system (chromatic aberration, transmission);
- Color filters (shape and coverage of spectrum);
- Primaries separation (beam splitting or CCD RGB pattern);
- Color signal processing (noise, colorimetry transform).

From color science point of view, a classical color image camera is a trichromatic observer. Another well-known trichromatic observer is the human standard observer.

A digital camera is characterized by its spectral locus, defined by the coordinates of all responses to monochromatic light in $R^C G^C B^C$ or $X^C Y^C Z^C$ or even $x^C y^C$ spaces. The spectral locus is the characteristic of a camera that corresponds to the color gamut of a display. The camera spectral locus is less known than the spectral locus of the human observer, but is of the same nature since a classical color image camera is just another trichromatic observer. The spectral locus is represented in an output-referred color space and can be derived directly from the corresponding color matching experiment (see further below). For example, from CIE 1931 XYZ color matching functions, a pair of xy coordinates can be calculated for each wavelength. Plotted in the chromatic xy diagram, these points define all together the curve of the spectral locus. The spectral locus circumscribes all colors that are visible by the observer.

Sensitivity metamerism

Metamerism happens when different spectral power distributions result in the apparent matching of colors for a human eye, or matching of color coordinates for a camera acquisition.

A camera transforms a real-word color stimulus, defined by a spectrum, into three RGB tristimulus values. Similarly to human vision, cameras are subject to metamerism. This raises issues in two directions:

- A given camera may produce identical tristimulus values for two (or more) different spectral stimuli, called a metameric pair (or metameric set, respectively);
- A camera with sensitivity curves different from the human eye differs in their metameric pairs from a human observer.

The link between scene-referred camera RGB values and CIE 1931 XYZ coordinates cannot be trivial since two different spectral sensitivity curves sets are involved, that of the camera and that of the human eye, respectively. Camera and human eye may differ in their metameric pairs leading no non-invertible relations between RGB and XYZ coordinates such as illustrated in Figure 1. Distinct *rg* points can correspond to the same *xy* point and vice versa. *rg* and *xy* chromaticity coordinates are obtained from the RGB scene-referred camera output values and from the output-referred CIE 1931 XYZ values, respectively, by normalization [13].

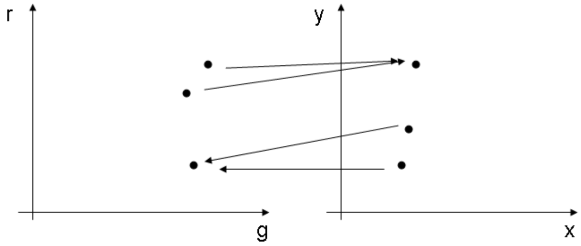


Figure 1: Non-invertible relation between rg and xy due to sensitivity metamerism

This problem is referred to as sensitivity or observer metamerism and can be avoided completely only if the camera satisfies the Luther condition [15] i.e. if its spectral sensitivities are linear combinations of the color matching functions of the CIE 1931 standard observer. Another solution is multispectral cameras [16].

Color clipping in proof viewing

A solution to the problem of sensitivity metamerism would require the estimation of scene-referred and human observer related color values, for example CIE 1931 XYZ values, from camera raw RGB output [15,17]. However, in proof viewing we have a different problem: How to reproduce captured colors on a given proof viewing monitor?

When proof viewing a camera raw RGB output signal on an RGB proof viewing monitor, the raw RGB values should be transformed into output-referred RGB values. We call this a proof viewing color transform. As shown in before, such a proof viewing color transform can exist only up to metamerism difference between the camera and the human eye.

For analysis, let us develop a straight forward proof viewing color transform. For presentation purpose we neglect any non-linearity. For a given camera and a given proof viewing monitor, a straight forward proof viewing color transform can be determined by the following steps:

- Determining the three scene colors that are within the color gamut of the proof viewing monitor;
- Measuring the camera output *RGB* values for these three colors;
- Determining the monitor input $R^m G^m B^m$ values for these three colors;
- Set a linear *RGB* transform *RGB* to $R^m G^m B^m$.

When applying this transform to the camera RGB output values, attention has to be paid to $R^m G^m B^m$ values that are outside of the valid coordinate range, for example [0;1] for normalized RGB values or [64;940] for 10 bit encoded RGB values in TV systems. The values should either be clipped, or soft clipped or compressed into the valid coordinate range.

Figure 2 shows an example for simple color clipping. We set a series of scene colors outside of the proof viewing monitor color gamut and captured them by a digital film stream camera. We applied the straight forward proof viewing color transform and RGB clipping. We displayed the processed RGB values on the proof viewing monitor and measured the CIE xy chromaticies on the monitor and in the scene.

As observed in Figure 2, color clipping modifies hue and saturation. While desaturation may be accepted by a director watching a proof viewing monitor, hue changes are not acceptable. A proof viewing color transform should address and solve this problem.

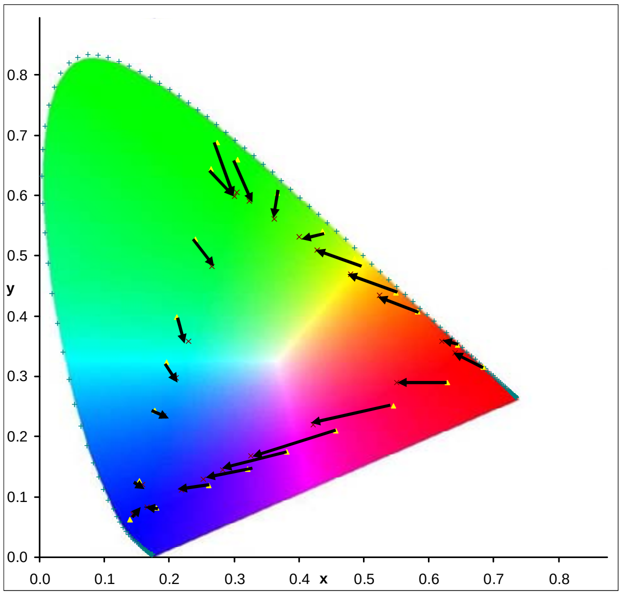


Figure 2: Color clipping (see arrows) of sample real scene colors when displayed on a Rec. 709 proof viewing monitor

Color resolution in digital acquisition

Another issue of digital acquisition when capturing scenes with extended color gamut is the color resolution:

- Difference of filter spectrum from spectral sensitivities of human eye;
- Restricted capacity to distinguish saturated colors;
- Impact on precision of captured hue.

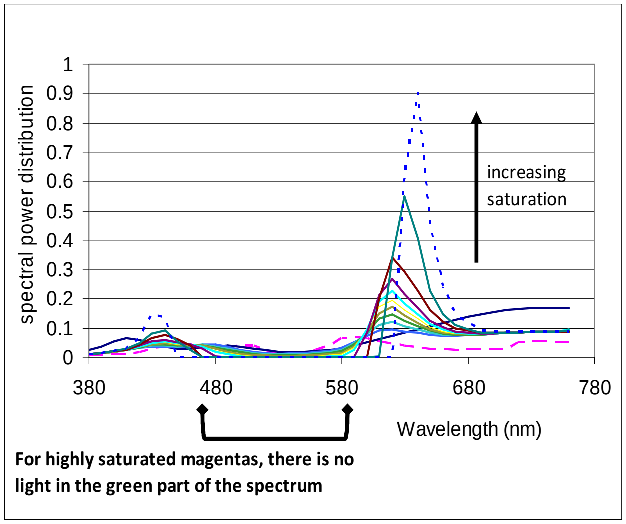
We want to show in the following that these issues result in additional errors on a proof viewing screen:

- Hue shift;
- De-saturation and color clipping.

We will use in the following an ideal proof viewing monitor without color gamut limitations. Color clipping errors such as discussed before are thus excluded.

Let’s take a series of test colors at constant magenta hue and with increasing saturation in perceptually uniform IPT color space [18].

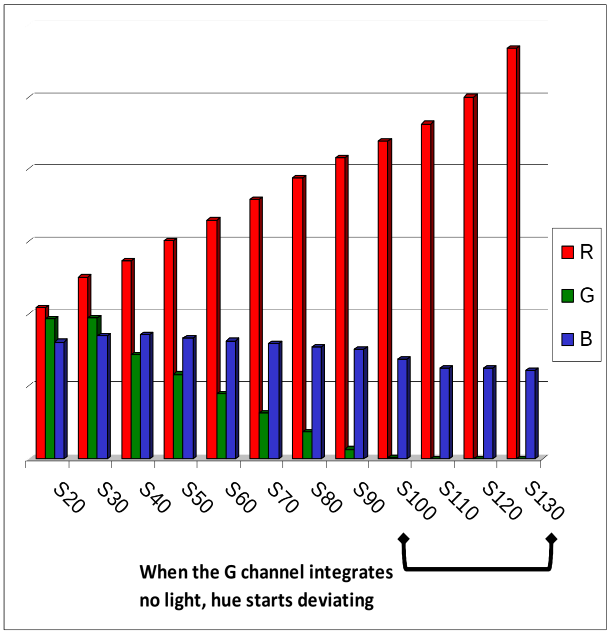
Figure 3 shows one of the possible sets of spectral power distributions that correspond to the chosen test colors. (Note that an infinite number of spectral power distributions may result in the hue and saturation of a given test color.) The spectral power distributions in Figure 3 are representative for spectra becoming sharper with increasing saturation. As observed in Figure 3, the luminous contribution of the spectrum for wavelengths between 480nm and 580nm decreases with increasing saturation. The four most saturated test colors have even zero contribution.



For highly saturated magentas, there is no light in the green part of the spectrum

Figure 3: A set of spectral power distributions corresponding to magenta test colors with increasing saturation from low (magenta dashed) to high (blue dotted)

In such a case, one channel of the camera (here the green G channel) will have no signal and then the camera no more exhibits trichromatic characteristics, but only two channels are active/excited. Figure 4 shows how R, G and B channels evolve with increasing saturation at constant hue according the stimuli from Figure 3. We see the system becoming di-chromatic for stimulus S100 and above, where only the R and B channels integrate light. For these stimuli, hue and saturation deviate as the acquisition system is no more coherent with the usual three channel system behaviour.



When the G channel integrates no light, hue starts deviating

Figure 4: RGB output with increasing R channel (red) decreasing B channel (blue) and decreasing and cropped G channel (green)

A solution to this problem involves the optimization of the spectral sensitivity curves and is beyond the scope of this paper. Such a solution should include an evaluation of color precision such as carried out by Pujol et al. [19] on the number of distinguishable colors inside the McAdam limits.

EXTENDED COLOR GAMUT IN COLOR CORRECTION

One of the artistic steps in production is color correction. Often a first phase is carried out to adjust roughly film footage or raw streams acquired by digital film stream cameras. Large mismatches in color balance and transfer function are compensated by linear matrices and non-linear one-dimensional transfer functions, respectively. Frequently, specific 3D Look-Up-Tables (LUT), also called Cubes, are applied to produce a more pleasant version than the raw version. In a second phase, the director of photography and the colorist apply artistic color changes in order to obtain the desired look of the images. In this artistic phase, the director of photography describes the intent of color correction while the colorist or a skilled operator has to translate the intent into an actual color transform applied to the footage. Such a color transform may include an increase of saturation, a change of color hue, a decrease of any RGB channel or an increase of contrast, for example. Color correction can be applied to an entire frame, to a set of frames, to a specific region in one single frame or even to all image regions in several frames corresponding to a specific color or semantic object (tracking).

Color reproduction during color correction

During this process, the director of photography and the color grading operator have to keep in mind what will be the impact of the applied color correction on the final reproduction medium. For example, if argentic film is first scanned and digitalized and then color corrected using a dedicated, digital proof-viewing projector, the operator verifies the applied color correction on the projection screen while the final reproduction is done by a film printer and then the film is projected.

Differences between the proof viewing display device (for example a digital proof-viewing projector) and the final reproduction device (for example a film printer followed by film projection) should be taken into account during color correction. Differences are due to different media, different equipment but also to different viewing conditions. Viewing conditions include ambient light, surround, background, reference white and adaptation state of the human eye. Differences between the proof viewing display device and the final color reproduction device can include objective, measurable differences of CIE 1976 hue angles, changes of CIE saturation, changes of contrast, differences in CIE 1976 luminance, differences in dynamic range, differences in color gamut as well as differences in color appearance such as changes in lightness, saturation and chroma. The latter three differences can not be photometrically measured.

A known solution to this problem is colorimetric color management (CMM) [14]. For CMM, the characteristics of the proof viewing device and the final reproduction device are measured, mathematically modelled and then compensated using a color transformation. CMM takes into account the color gamut of the devices. When an image contains colors outside of the color gamut of a display device or close to the border of the gamut, the applied color transform may contain color gamut compression, color clipping or other specific operations such that the transformed colors are inside of the device color gamut.

Issues of color correction

The difference of color gamuts of display devices is a problem for color correction. It may happen that the operator applies a color correction that generates the desired image on the proof-viewing device while the final reproduction device is not capable to reproduce some of the colors since the color gamut of the final reproduction device is different from the gamut of the proof-viewing device. It may happen that the operator wants to apply a specific color correction which would generate acceptable results on the final reproduction device but which cannot be visualized on a proof-viewing device with different color gamut.

A known solution is

- to detect out-of-gamut colors for the final reproduction device;

- to detect out-of-gamut colors on the proof view device;
- in the framework of CMM and
- to show a gamut alarm to the operator when an out-of-gamut color has been detected.

Figure 5 shows a typical example how gamut alarm is signaled to the operator. Each pixel that contains a detected out-of-gamut color is shown white.

Classical color correction systems offering gamut alarm functionality however do not address a series of problematic cases.

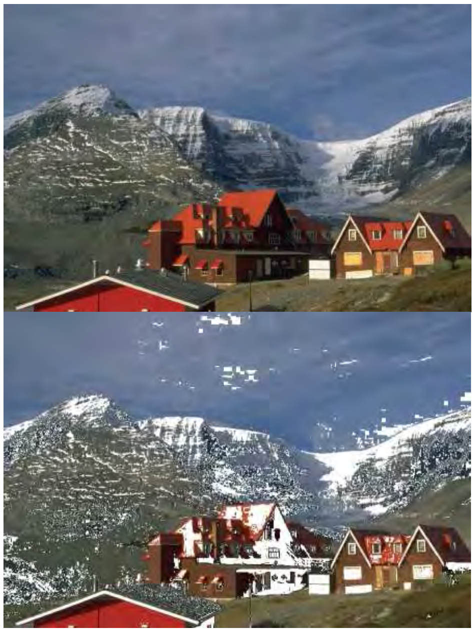


Figure 5: Original image on the screen of the colorist without gamut alarm (top) and with gamut alarm (bottom)

The first case is the difference in viewing conditions. The gamut alarm mechanisms are limited to colors that can not be rendered on a display in the framework of colorimetric color management. In this framework, colors are usually measured by CIE 1931 XYZ coordinates. These coordinates do not consider viewing conditions that influence the human observer while watching the display. In an appearance-based color management framework (appearance-based CMM), such influences are compensated. In such a case it may happen that a color that the operator desires on the proof viewing device can be reproduced on the final reproduction device in colorimetric terms but can not be reproduced when viewing conditions are compensated.

A second case is the consideration of an original reproduction device. When an operator works on footage that is aimed for a final reproduction device and proof viewed on a proof viewing device, it may be important to consider where the content comes from, i.e. for which device the content was originally prepared. This device is called here original reproduction device. It may happen that a color after color correction is well reproduced on the proof viewing and final reproduction devices but not on the original reproduction device. This case needs to be detected and indicated to the operator.

The third case is the uncertain nature of viewing conditions. In an appearance-based CMM framework, influences of viewing conditions are compensated. As soon as colors need to be modified since they are out of the gamut of reproducible colors taking into account viewing conditions, they should be indicated to the operator. This could be an advanced case of classical gamut alarm. Such colors could be marked on the proof viewing screen by specific false colors, for example red. Classical gamut alarm is binary: either on or off. This is well adapted for the case of out-of-gamut alarm considering well-defined color gamuts of display devices. A binary gamut alarm is not adapted to the gamut of reproducible colors considering viewing conditions since characteristics of viewing conditions are less well mastered and known than characteristics of display devices. A binary gamut alarm would be finally not useful for the daily work of the operator.

The fourth case is when the operator wants to modify out-of-gamut colors. There is a difficulty of interpretation of classical gamut alarm. If classical gamut alarm is shown on the proof viewing device, those regions of the image are marked with a false color that represents out-of-gamut colors. An example is shown in Figure 5. When the operator looks at the image with gamut alarm, he aims to identify the colors (their hue, their saturation, their luminance) that are out of gamut. Either he switches on and off the gamut alarm or he analyzes the image as it is.

There are situations where this is easy. In Figure 5, he will identify the blue tones in the sky that – once getting clearer – approach the gamut border and go slightly outside. The blue tones are easy to analyze since the blue sky region contains a variety of tones and transitions. By the position and shape of the out-of-gamut regions the operator can easily analyze the problem.

There are situations where the identification of out-of-gamut colors is difficult. In Figure 5, the red roofs and the brown walls are out out-gamut. Since transitions are lacking, the operator can not be aware which portion of red and brown tones is concerned. This problem is increased in animated and painted images where the color palette is often restricted. It is not visible whether the correction to be applied to these colors needs to be weak or strong. From the image in Figure 5, it is not clear to the operator what may happen to similar colors, those that may occur on the same objects but in following frames where light is slightly different.

This problem is solved today by trial and error as well as by switching on and off the gamut alarm. The operator applies corrections and verifies the gamut alarm. By “trying around” a couple of neighboured tones, he will understand the position of the concerned colors within the color gamut and apply an appropriate correction. This procedure takes time. Furthermore, the operator can not separate out colors being largely outside the gamut that need to be worked first. By watching the image in Figure 5, he can not establish a priority list for his work. This prevents from being quicker by neglecting colors which are only slightly out of gamut.

The fifth case is the growing variety of display technologies in the consumer world, when video productions are to be distributed to consumers with different display technologies, the color correction process using a single final reproduction device will fail to produce content that has controlled quality on displays with other characteristics than those of the targeted final reproduction device. In this case, there may be non-detected colors that are out of the color gamut of the actually used reproduction device.

METHOD OF SOFT GAMUT ALARM FOR COLOR CORRECTION

This section introduces the new concept of soft gamut alarm that assists the colorist in future tasks of color correction with extended color gamut.

Overview

The proposed method aims at proof viewing the visual content introducing the new concept of alarm.

The method has the four following advantages with respect to classical color correction:

- Differences between viewing conditions of different color reproduction devices are considered;
- The uncertain nature of knowledge about viewing conditions is taken into account and content can be created considering this uncertainty.
- The variety of final reproduction devices is considered and content can be created with regard to this variety;

- Reduction of degradations of content with respect to its original/raw version.

The proposed color correction method aims to correct original colors of original images targeting an original color reproduction device with respect to a set of final color reproduction devices. Each of these color reproduction devices is characterized by its color gamut of reproducible colors in device independent, absolute color space and its viewing conditions for color perception by human observers.

The method can be summarized by the following steps:

1. The original colors of the original images are displayed on a subset of the final color reproduction devices, these devices are called proof viewing color reproduction devices;
2. For each of the color reproduction devices, the distance of the original colors to the color gamut of the color reproduction device is determined;
3. For each of the color reproduction devices, the color appearance of the original colors and of the color gamut of the color reproduction device are determined, taking into account the viewing conditions of the color reproduction device;
4. For each of the reproduction devices, the visibility of the original colors is determined, each visibility being the distance of the color appearance of the original color to the color appearance of the color gamut;
5. On one of the proof viewing color reproduction devices, false colors are displayed instead of the original colors, where the false colors reflect the correspondent distance and visibility of the corresponding original color.

The original colors of the original images are color corrected by an operator. Original colors are replaced by modified original colors in a way that the corresponding distance is minimized and the corresponding visibility is maximized.

Figure 6 shows the color processing flow path according to the proposed system. From original colors, false colors are determined that depend on distances to color gamuts. Original and false colors are displayed.

The process can be assisted by automatic gamut mapping [20,21,22]. For all proof viewing color reproduction devices, gamut mapping is applied in such a way that the false colors can be switched off and a reproducible, mapped color is shown. Gamut mapping is preferably carried out in color coordinates representing the color appearance of the colors.

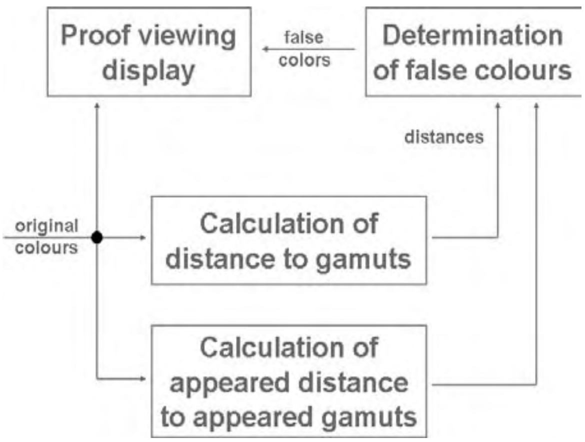

Proof viewing display
false colors
Determination of false colours
distances
original colours
Calculation of distance to gamuts
Calculation of appeared distance to appeared gamuts

Figure 6: Principle of the soft gamut alarm system

The distance to the color gamut is determined as follows. For each of the reproduction devices, the distance of the original colors to the color gamut of the reproduction device is determined using the Euclidean or a weighted Euclidean distance. The distance is forced to zero for original colors being inside the color gamut.

The visibility of an original color for a human observer is determined from the so-called appeared distance that is determined as follows. The original colors aimed for the original reproduction device are transformed into an original device independent color using the device profile of the original color reproduction device. The original device independent colors are transformed into original appeared colors according to the viewing conditions of the original reproduction device, where the appeared colors reflect the color appearance for a human observer. For each of the color reproduction devices, viewing conditions of the reproduction device, the color gamut is transformed into an appeared color gamut. The appeared distance is determined as distance of the original appeared color to the appeared color gamut. For original appeared colors being inside the appeared color gamut, the appeared distance is forced to zero. The visibility is a monotonic function of the appeared distance.

The concept of soft gamut alarm can include more than one false color to be calculated shown instead of one single. For example, two false colors can be calculated as follows. A first false color is calculated from the distance between the original color and the color gamut of a selected color reproduction device. A second false color is calculated from the appeared distance between the original appeared color and the appeared color gamut of the selected reproduction device.

In the following, the proposed method of soft gamut alarm is applied to the case of proof viewing for color correction during post-production of a digitalized film.

Reproduction devices

Three reproduction devices are considered:

- A proof viewing digital projector under dark conditions;
- A digital cinema projector under dark conditions;
- A broadcast reference monitor under dim lighting conditions.

All devices are fed with RGB color values. By device characterization, for each reproduction device, a forward and an inverse device model is established. The forward device model calculates device-independent XYZ color values from device-dependent RGB color values. The inverse device model realizes the inverse operation. The devices model provides also the color gamut of the device.

Consideration of color appearance

The appeared color values and appeared color gamuts are established in the perceptual color space JCh of CIECAM-02. In this color space, J is lightness, C is Chroma and h is hue angle perceptual estimate.

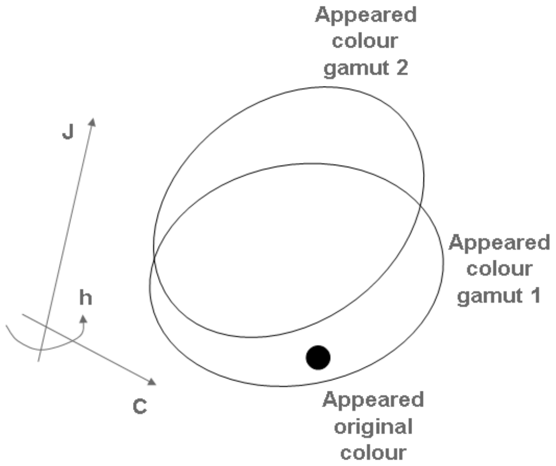


Figure 7: Example of an appeared color that cannot be reproduced on device no. 2

Figure 7 shows a sketch of an appeared original color and the appeared color gamut of two color reproduction devices no. 1 and no. 2 with different viewing conditions. On device no. 1, the appeared original color is close to the appeared gamut and has thus a bad visibility. On device no. 2, the appeared original color is outside of the gamut and is thus not reproducible.

The color appearance model (CAM) CIECAM02 is defined by the following viewing conditions parameters:

- The XwYwZw tristimulus values of the reference white; it can be set to the white point of the display obtained from the forward device model;
- La: this is the adapting luminance to which the observer is adapted; it is expressed as an absolute value in cd/m². It can be set to a value corresponding to 20% of the reference white luminance (mean video value).
- Yb: this is the background luminance which corresponds to the entire screen (or display) average white luminance. This value depends on the video content and may be specified as a percent of the reference white luminance. e.g. 20 for 20%.
- The surround type : there are four possible states:
- Average for day light vision (Yb>10cd/m²);
- Dim for dim viewing conditions (3-5 < Yb < 10 cd/m²);
- Dark for night viewing conditions (Yb<3-5 cd/m²);
- Intermediate this is a linear combination between each of the three other states.

For the use of CIECAM-02, all these parameters need to be known. For the three reproduction devices, the parameters are chosen as follows:

- Proof viewing digital projector
  - XwYwZw: display white measured in the center of the screen
  - Yb: 20% of Yw
  - Dark surround
- Digital cinema projector
  - XwYwZw: display white measured in the center of the screen
  - Yb: 20% of Yw
  - Dark surround
- Professional television monitor
  - XwYwZw: display white measured in the center of the screen
  - Yb: 20% of Yw
  - Dim surround

Generation of soft gamut alarm

The false colors showing the gamut alarm are calculated for the original colors of the images. For each image pixel, and for each of the two other color reproduction devices (the DC projector and the reference monitor), two false colors are calculated for the original color of the image pixel. For each pixel in total, four false colors are calculated. In the following is explained, how two of these false colors are calculated for one of the two reproduction device, selected by the operator.

A first false color is calculated from a function of the color components of the distance vector that is related to the distance between the original color and the color gamut of the selected color reproduction device. More precise, the distance describes the Euclidian distance between the original color and the closest point of the color gamut.

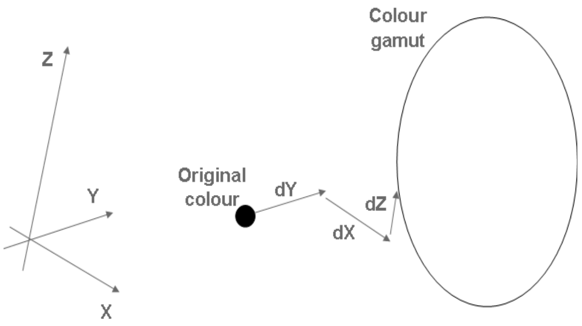


Figure 8: Calculation of a first false color in CIE XYZ space from the distance between the original color and the color gamut

For each color reproduction device, the distance of an original color to the color gamut of the color reproduction device is forced to zero for original colors being inside the color gamut. When the distance is zero, the related first false color is disabled and not calculated.

A second false color is calculated from a function of the color components of the distance vector that is related to the appeared distance between the appeared original color and the appeared color gamut of a color reproduction device. The components of the distance vector are calculated in the perceptual JCh color space of CIECAM-02 representing lightness, hue and saturation. By this choice, the second false color reflects the distance of the appeared original colors from the appeared color gamut of a reproduction device in aspects of lightness, hue and/or saturation, see Figure 9.

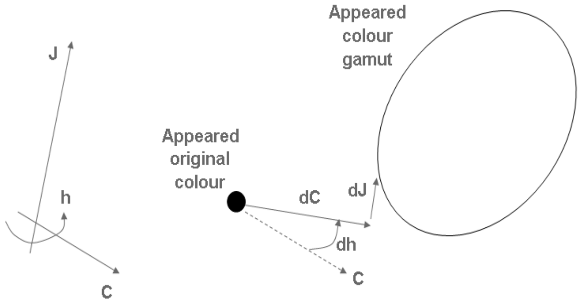


Figure 9: Calculation of second false color from the distance between the appeared original color and the appeared color gamut

The false colors are displayed according to the choice of the operator and will considerably help the management of wide color gamut.

CONCLUSIONS

This paper discusses issues in digital acquisition and color correction of images with extended color gamut such as camera sensitivity metamerism, proof viewing color clipping and gamut alarm in color correction. Production equipment builders should address the increasing demand of directors to capture and proof view scenes with extended color gamut. Optimized color filters and wide color gamut processing modes need to be developed for cameras. Post-production and color correction facilities should adapt color transforms and the related functions of gamut alarm to extended color gamut including evolving viewing conditions, new display technologies and color appearance.

This paper provides some inputs to ease the production of extended color gamut content. However the distribution of this content raises additional issues to be considered, such as the adaptation to the device characteristics or the viewing conditions. However, it is clear that future video formats will integrate extended color gamut so as to better approximate and serve the human visual system capabilities.

REEFERENCES

[1] R.R.A. Morton, M.A. Maurer, G. Fielding, C.L. DuMont, Using 35mm digital intermediate to provide 70mm quality in theaters, SMPTE 143rd Technical Conference and Exhibition, November 4-7, 2001.

[2] Technicolor 3D, www.technicolor.com

[3] EBU-Tech 3320, User requirements for Video Monitors in Television Production, Eurpean Broadcast Union (EBU), Version 2.0, October 2010.

[4] IEC, Multimedia systems and equipment - Color measurement and management - Part 12-1: Metadata for identification of color gamut (Gamut ID), 2011.

[5] A. Roberts, Coloring the future, tech-I, European Broadcast Union (EBU), March 2012.

[6] J.Stauder, C. Porée, P. Morvan, L. Blondé, A gamut boundary metadata format, 6th European Conference on Color in Graphics, Imaging, and Vision (CGIV), Amsterdam, May 2012.

[7] S. Y. Choi, H. Y. Lee, Y. T. Kim, J. Y. Hong, D. S. Park, C. Y. Kim, New Color Encoding Method and RGB Primaries for Ultrahigh-Definition Television (UHDTV), 18th Color Imaging Conference (CIC), San Antonio, USA, November 8-12, 2010.

[8] ITU-R BT.709-5, Parameter values for the HDTV* standards for production and international programme exchange.

[9] ITU-R BT.1361, Worldwide unified colorimetry and related characteristics of future television and imaging systems

[10] IEC, Multimedia systems and equipment – Color measurement and management - Part 2-4: Color management - Extended-gamut YCC color space for video applications – xvYCC, IEC 61966-2-4 Ed. 1.0, November 2006.

[11] Y. Xu, Y. Li, G. LI, Analysis and Comparison of extended color gamut in ITU-R BT.1361 and IEC 61966-2-4, Journal of Video Engineering, Vol. 33, No. 3, 2009.

[12] Photography and graphic technology - Extended color encodings for digital image storage, manipulation and interchange - Part 1: Architecture and requirements, ISO 22028-1.

[13] R.W.G. Hunt, The reproduction of color, Sixth Edition, Wiley, 2004.

[14] SMPTE, Derivation of Basic Television Color Equations, Recommended Practice RP177-1993.

[15] P. Urban, R. S. Berns, R.-R. Grigat, Color Correction by Considering the Distribution of Metamers within the Mismatch Gamut, Proc. 15th IS&T Color Imaging Conference, pages 222-227, 2007.

[16] Yuri Murakami, Keiko Iwase, Masahiro Yamaguchi, Nagaaki Ohyama, Evaluating Wide Gamut Color Capture of Multispectral Cameras, Proc. of 16th IS&T Color Imaging Conference, November 10-15, Portland, 2008.

[17] Jack Holm, Capture Color Analysis Gamuts, Proc. 14th Color Imaging Conference, pages 108-113, Scottsdale, Arizona, November 2006.

[18] N. Moroney, A radial sampling of the OSA uniform color scales, Proc. 11th IS&T Color Imaging Conference, pp. 175-180, 2003.

[19] J. Pujol, F. Martínez-Verdú, M. J. Luque, Cobija, P. Capilla, M. Vilaseca, Comparison between the number of discernible colors in a digital camera and the human eye, Proceedings of CGIV 2004, Second European Conference on Color in Graphics, Imaging, and Vision and Sixth International Symposium on Multispectral Color Science, April 5-8, 2004.

[20] J. Morovic and M. R. Luo, The Fundamentals of Gamut Mapping: A Survey, Journal of Imaging Science and Technology, 45/3:283-290, 2001.

[21] Montag E. D., Fairchild M. D, Psychophysical Evaluation of Gamut Mapping Techniques Using Simple Rendered Images and Artificial Gamut Boundaries, IEEE Trans. Image Processing, 6:977-989, 1997.

[22] P. Zolliker, M. Dätwyler, K. Simon, On the Continuity of Gamut Mapping Algorithms, Color Imaging X: Processing, Hardcopy, and Applications, edited by Eschbach, Reiner; Marcu, Gabriel G. Proceedings of the SPIE, Volume 5667, pp. 220-233, 2004.

Strategies for Deploying High Resolution and High Framerate Cable Content Leveraging Visual Systems Optimizations

Yasser F. Syed PhD, Dist. Eng/Applied Research
& Dan Holden, Fellow, Comcast Labs

Abstract

This paper examines how and why to deliver higher resolution and framerate content in an HFC system, especially focusing on 4K Video delivery with an advanced audio experience. It examines how to deploy this content in a bandwidth constrained environment and concentrates on improvements to the viewer's quality of experience through video compression technologies and leveraging potential video compression gains through sensitivities in the human visual system.

INTRODUCTION

The launch of higher resolution video with greater frame rates will allow MSOs to develop new business opportunities, and provide a competitive advantage against new entrants in the video marketplace. In this paper we will examine the road to better delivered video quality, especially how to leverage the existing HFC infrastructure to deliver 4k video with an advanced audio experience. The paper will concentrate on video compression technologies and the potential for leveraging the human visual system model to provide 4K video in a bandwidth constrained environment. For deployment, we will look at required upgrades to the HFC infrastructure, and what engineering requirements are needed for 4K delivery. New technologies and approaches to reduce costs will also be examined, as well as how the complexity of high-resolution video changes delivery methodology.

4k television technology was introduced at the Consumer Electronics Show in 2012. It is based on a display that has approximately 4000 pixels in the horizontal resolution. 4k differs from previous television standards (480i, 480p, 720p, and 1080P/I) in which the vertical pixel count was annotated. In a 4k display the horizontal resolution is maintained around 4000 pixels, and the vertical resolution is allowed to vary as a function of source content. This technique was adopted to allow support for various aspect ratios and letterboxing. Figure 1 shows the scale of 4K content compared to the resolutions that are supported today.

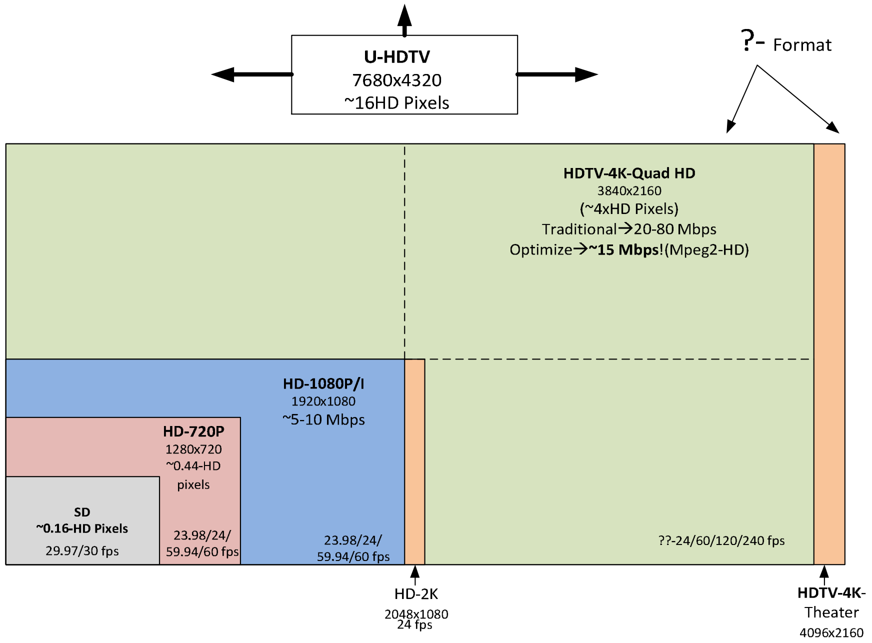


Figure 1 Comparison of 4K to Different Video Resolutions

BUSINESS OPPORTUNITY

One of the most compelling cases for higher quality video is to gain a competitive advantage in the video marketplace. 4k will require “big pipes” at a time when there is clear movement on the part of industry competitors to adopt a mobile strategy utilizing technology that will be limited by available spectrum. Newer video compression techniques will certainly reduce the size requirement for the pipes, while the demands of newer display technologies, (8k and 256 fps) will tax any future video distribution system.

Cable has a reputation for being the leader in delivering an exceptional video experience. First generation 4k delivery platforms will need a vast amount of bandwidth, which is most likely to require a full QAM in order to deliver a quality experience. If we look at historical data, early generation H.264/AVC video was around 9 Mbps for High Definition (HD) 1080i video. A few short years later, we have been able to reduce this bandwidth to 4.3 Mbps.

Until display technology retail prices drop to a reasonable level, it is expected early adopters for 4K televisions will be bars, restaurants, and high-end home theaters. Here are the key assumptions:

- Mass deployment of 4k televisions will not take place until the cost per unit is less than $3,000 per unit
- The introduction of 4k will follow the same general path as the

introduction of High Definition video, which has currently penetrated more than 70% of US households

- Adoption of 4k video will be slower than HD video
- Volume of 4k encoded VOD assets will grow exponentially over the next three years
- Studio post-production already supports a 4K workflow which can be extended to downstream VOD content delivery
- Additional revenue will be generated when customers select to watch assets in a 4k format
- 8k video will not be introduced until at least 2016
- MSOs will not simulcast 1080p60, but may select to offer this format in VOD
- Bars, restaurants, and elite home theaters offer a significant up-sale opportunity

MSOs should take the lead on the introduction of high resolution video delivery. Rather than focus solely on video, it is suggested by the paper authors that the entire sensory experience be enhanced, which includes the addition of 3D audio channels. Background noise in a bar can be very distracting, and providing a high quality audio experience will set our video offering apart from the competition. The adoption of 4k video with 3D audio will most likely not progress at the same pace as HD. HD had the added benefit of changing the format to 16:9 from 4:3, and the elimination of large cathode ray tubes, which drastically reduced the size of the television footprint in the living room. The adoption of HD televisions has been relatively quick historically, whereas, the migration to the distribution of higher resolution video has yet to be established.

ADOPTION WILL BE DIFFERENT FROM 3DTV

There have been many attempts to categorize 4k video to the 3D television experience. This type of comparison is probably not the correct model, as 4k will not suffer from the infamous 3D glasses gaffe. Additionally, massive libraries currently exist at studios that can be easily scanned or transcoded into higher resolution video for distribution. We believe comparing 4k adoption to 3D would be a mistake, since 4K will most likely follow the adoption and general operational patterns developed for HD.

The first linear 4k channel will most likely be an occasional feed that is brought up when a live 4k event is aired. Under this model, 4 HD channels would need to be taken down in order to broadcast a single 4k event. With a few enhancements to the backoffice systems, it should be possible to sell access to a 4k stream on a pay-per-view fashion. The broadcast of huge events, like the Olympics or Super Bowl, could lead to enormous up-sale opportunities.

4k VOD will most likely be the first place where we see significant inroads of high resolution video. Encoders have already been developed that can process 4k video, and it is believed VOD pumps will not have issues with the larger file sizes or MPEG-2 transport stream wrappers. Adaptive streaming technologies should also be suitable for 4k VOD distribution. The video encoding process for QAM and adaptive streaming can be identical. Fragmentors should not require modifications, unless they are "just in time," which may suffer from data transfer rates and latency. The largest gap in the distribution system will be the ability to handle 3D audio, and finding a suitable video player.

ALTERNATIVES

There are many alternatives to 4k video, including Quad HD and 8k. While Quad HD has slightly less resolution than 4k, 8k has twice the resolution and twice the bandwidth requirements. Should Quad HD TVs be introduced into the marketplace, it would be preferred if they have the capability to ingest true 4k content, as MSOs would not want to simulcast both Quad HD and 4k streams. For VOD delivery, it would be possible to support both formats, but as the VOD library grows the added storage expense would prove challenging.

Rather than rolling out 4k, another possibility is to move forward with 1080p60. Encoders and STBs were released in 2012 to support this format. Formal analysis of 1080p60 video quality is beyond the scope of this paper.

Current compression technology will most likely prevent the delivery of 8k content over a QAM, but 8k delivery could conceivably be done utilizing the CMTS and IP delivery methodologies. Both products deliver the same benefits as 4k, higher video quality.

BENEFITS OF 4K

The first implication of moving to 4k video is the size of streams and files will be massive. A single mezzanine, linear stream from a live event may reach up to 500 MBps and a stream sent to a set top box could be on the order of 38 MBps. This implies four high definition channels would need to be taken down in order to place one 4k signal on the plant (Figure 2).

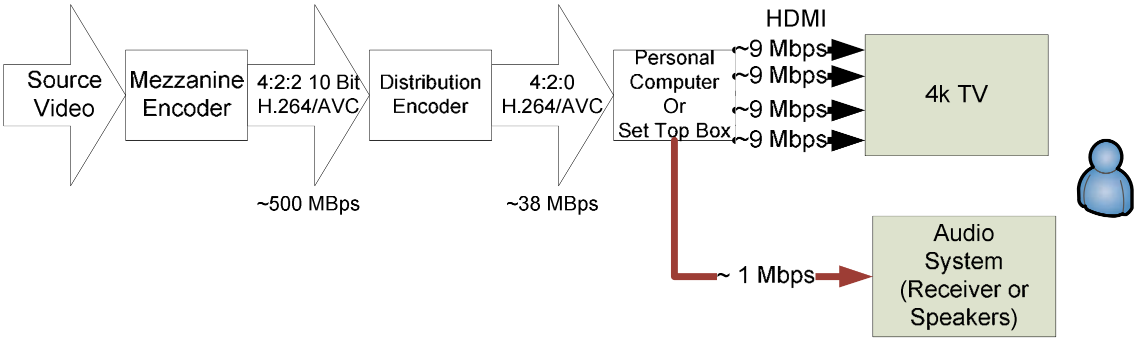


Figure 2 Delivery of 4K content from Ingestion to Consumer

Higher resolution video will allow MSOs to compete with both BluRay and local movie theaters. Many movie theaters currently delivery digital projects in 2k resolutions with a maximum audio experience of 11.1. It is theatrically possible to delivery 4k video with a 22.2 audio experience across an existing QAM to a personal computer (PC) which will replace the current functionality provided by a set top box (Table 1).

Thus, a completely optimized and compressed 4k/HEVC asset should be around 19 Mbps. When we compress this asset utilizing HEVC, we expect to gain around a fifty percent reduction in bandwidth, putting our 4k asset at approximately 10 Mbps. Next, consider that 50% of that potential gain is taken back 50% due to inefficiencies in first generation encoding technologies, frame-rate allocations, and make allowances for content types, then our 4k/HEVC asset can be distributed in the same band width as an HD asset compressed with MPEG-2 (~15Mbps).

Phase	Video Type	CODEC	Bandwidth (MBPS)	Notes
Initial	1080i	MPEG-2	19.3	19.3 was part of a specification. The first generation HD at some MSOs was set to 18 MBPS.
Today	1080i	MPEG-2	9.7	With 4:1 statistical multiplexing, it is possible to send 4 HD streams down a single QAM
Today	1080i	H.264	4.3	Average bit rate for H.264/AVC streams
Initial	4k	H.264	38	Target bit rate for lab trials
Production 4k[1]	4k/60 fps	HEVC	15	Target bit rate for 4k/60 with 22.2 audio

Table 1 Projected and Historic Bandwidth Consumption

[1] Note the projected bandwidth for a production 4k asset. The basis for the projection is calculated as follows:4k video is slightly larger than four HD signals: 4 * 4.3 ~ 18 Mbps in H.264/AVC Add additional audio bandwidth of approximately 1 Mbps for a total of 19 Mbps in HEVC

AUDIO

In addition to an enhanced video experience, the opportunity exists to upgrade the viewer's audio experience. Cable MSOs understand that audio can enhance or detract from the video quality of experience.

Many new audio technologies are under development that will put additional audio channels into the home. Old content can be remixed to support new formats, and additional microphones can be utilized to capture a true "3D" audio experience.

In the short term, consumers will need to add additional speakers to gain the improved audio benefit; and in the near future we will see sound bars that will reduce the complexity and cost of delivering this technology into the home theater and entertainment based businesses.

A typical 22.2 audio experience would require almost 1.5 Mbps when utilizing 24 channels at 48 kbps with constant bit rate (CBR) encoding. By switching this to capped Variable bit rate (cVBR) encoding, a substantial reduction in audio bandwidth utilization will be realized. Additionally, new sound bar technologies will reduce the cost, complexity, and number of speakers required to bring a true 3D audio experience to the customer.

As part of the distribution process, care must be taken to monitor every channel and to ensure multichannel audio is down-converted to basic stereo for playback though the television speakers. While it is assumed 4k content will be viewed with enhanced audio, consumers may select to view the content while utilizing the stereo audio capabilities of the display.

COSTS

The costs to enhance the end-to-end solution for 4k can be broken into their representative components. Here is a partial list of items that may require upgrades.

Encoders – Existing VOD encoders have the ability to deliver 4k video with few modifications, while linear encoders will need to be developed that can handle massive amounts of data in very short periods. Additional modifications will be needed to handle advanced audio technologies such as Dolby Adaptive Audio, SRS Multi-Dimensional Audio, and 22.2 specifications. There will need to be a clear roadmap to get from initial 4k video with H.264/AVC encoding to HEVC. For the initial launch, a single linear 4k encoder should suffice. It will allow a MSO the ability to replace four HD streams with a single 4k stream. For VOD, it is possible to scale the number of encoders to match the size and refresh rate of the library to be converted.
SRM – A next generation SRM will need to be deployed in order to allow the VOD pump to select a 4k asset.
Metadata – New fields will need to be included to indicate the asset is 4k.
Content Encoding Profile – New profiles for 4k encoding will need to be defined.
Storage – 4 times the storage per asset, as compared to HD.
Video Player – Support for new Video and Audio formats.

Adaptive Dynamic Streaming – Support for additional audio CODECs or video CODECs.
Backoffice – Enhancements for billing.
Set Top Box – Faster single or multi-core CPUs and bigger pipes.
Direct Fiber – Larger pipes for mezzanine sources.
Mixing new audio – New mixing technologies for audio.
Trucks, Cameras, Post – Enhancements to editing systems, graphics, and source acquisition equipment for live capture content.

SERVICE AND INFRA-STRUCTURE VIEWS

It has already been demonstrated in the laboratory that 4K video encoding for VOD can be accomplished on existing encoders. A single 4k transport stream is generated and sent across the plant for decoding on a Personal Computer (PC). This stream is then split into four separate streams for delivery to the display across four separate HDMI cables. Once the HDMI interface is upgraded, it is expected a single stream and HDMI cable will be attached to the television.

Linear encoding could be done by handling the encode as 4 separate HD processes that need to be synchronized (hence QuadHD) and distributed as a single stream on the wire. It is important to note this implies that within the video encoding process, the input stream could be split for processing and then combined into a single stream for transport.

While this approach may be viable, newer, multi-core CPUs will most likely be able to handle the entire encode as a single transport stream. A single stream approach across the entire plant will increase operational efficiencies and simplify the operational model. In the case of adaptive streaming, fragmenting a single transport stream would require the identification of a single boundary point in the source video.

The same intuitive logic applies to network DVR. As previously stated, utilizing the same encoding techniques for linear and VOD is optimal due to simplicity and overall operational models for distribution of 4k video (figure 3).

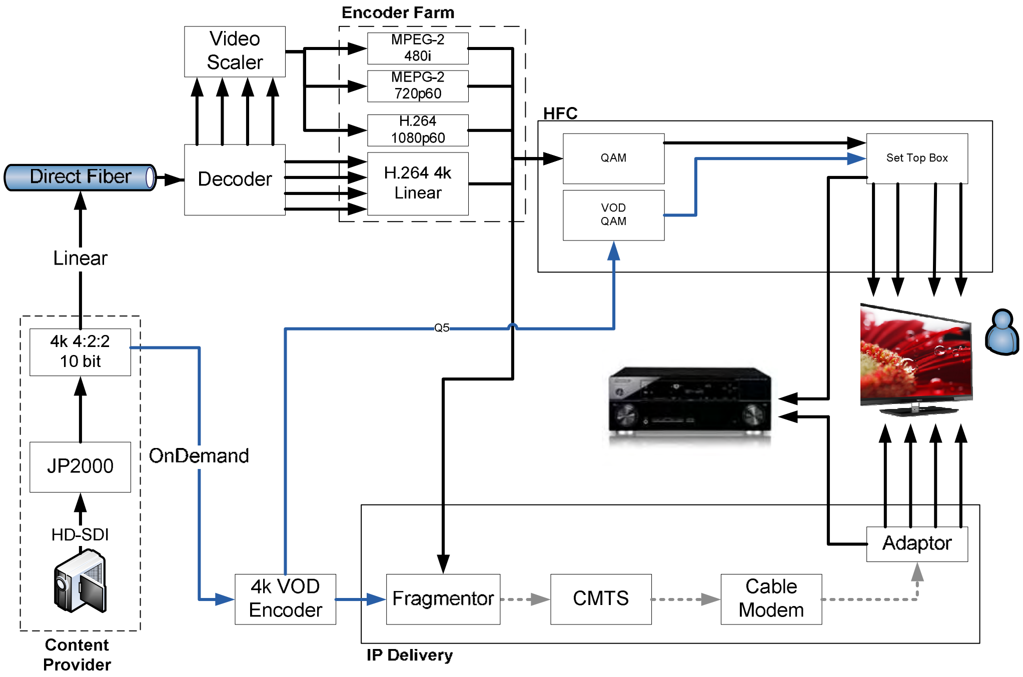


Figure 3 Operational Model for 4K Distribution Video

HD encoding of 1080I30/1080P24 using newer encoding techniques could range from 5-10Mbps when compressed with AVC/H.264. Offline VOD compression will most likely be superior due to multi-pass encoding. If 4K is supported at the same frame-rate, this could imply an encode bit rate from 20-40 Mbps in a cumulative data sense. This does not assume further compression efficiencies due to increased pixel density.

Can the infrastructure support a 40 Mbps 4K stream? A single 40 Mbps 4K channel would:

- Require the same bandwidth as 4-8 HD channels,
- Not fit into a 38.8 Mbps QAM
- And would likely not be carried by an ISP over the public internet

The bandwidth infrastructure modifications for this approach would be cost prohibitive. One bound stream could possibly be fit into a single QAM with bandwidth of 38.8 Mbps, which would replace about 2^+ MPEG-2 HD channels (or 4 HD streams on a 4:1 Mux). To meet a 4K service for HD, each QAM would need two 4K channels. This would mean each 4K channel would need to be bounded under 19 Mbps which would be about 1 HD channels and 1 SD channel.

Is it possible to move from a 1:4 upper bound bit processing ratio to a 1:1.3 ratio? With new coding tools from MPEG such as HEVC, a 50% improvement in compression can be expected. Additionally, having greater pixel density should create some further compression efficiencies to decrease the 1:4 ratio. Even more efficiencies can be gained by the way of improvements to perceptual modeling of our visual system and applying this to coding.

There is room to create more compression efficiencies, especially since encoder design is evolving and new compression tools are becoming granular. And even if a greater frame-rate is needed, pixel processing burden would be less than expected due to increased efficiency in motion vector accuracy and longer GOP length for the same amount of time.

As we examine all of the factors of better compression, filters and modulations, it does become possible to create a 4K stream that should ultimately approach 15 Mbps in the near future.

The next part of this paper will look at potential places to leverage the human visual system model to increase compression efficiencies through perceptual coding.

PERCEPTUAL CODING AND THE HVS MODEL

HVS (Human Visual Systems) attempts to describe how we actually see [from the photoreceptors in our eyes into the visual cortex and other parts of the brain]. Perceptual video coding is used in “lossy” compression at a target bitrate to mask, transform/quantize, or conceal information that is not seen by our visual systems (psycho-visual redundancies) or is optimized to improve what we can see. This is not coding efficiencies due to manipulation of the bit-stream to improve bit/symbol rate of the stream. It attempts to narrow the total information rate to what is just needed for our visual systems.

Our eyes are made up of 127 million photoreceptors in the retina (120 million rods and 7 million cones) that feed a million neurons in the optic nerve that is connected to the brain [Figure 4]. That already represents about 127:1 convergence of information. The rest of the eye is there to focus, shape, and control the amount of light going into the retina. The rods are used for vision at very low light levels (scoptic) and do not contribute very much to color perception. However, the cones deal with vision at higher light levels (photopic) and with resolving fine spatial details and color. These cones are divided into three types (S-short, M-medium, and L-long) that are sensitive to different wavelengths of light and they are the basis for our ability to match any color through a combination of three primary colors (trichromacy) [Figure 5]. The cones are concentrated in a central part of the retina called the fovea which provides the majority of information traveling along the optic nerve. The fovea matches to what we perceive as “the center of focus” for our vision.

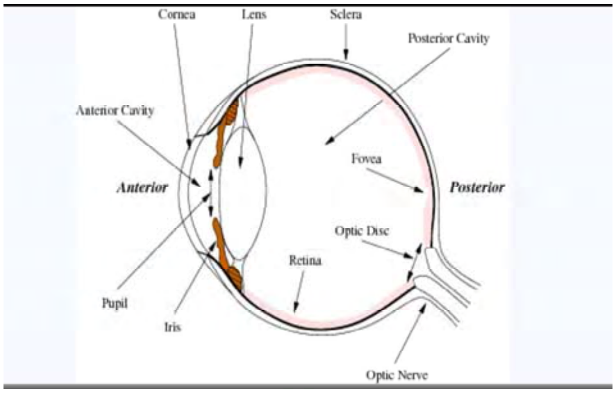


Figure 4 Eye

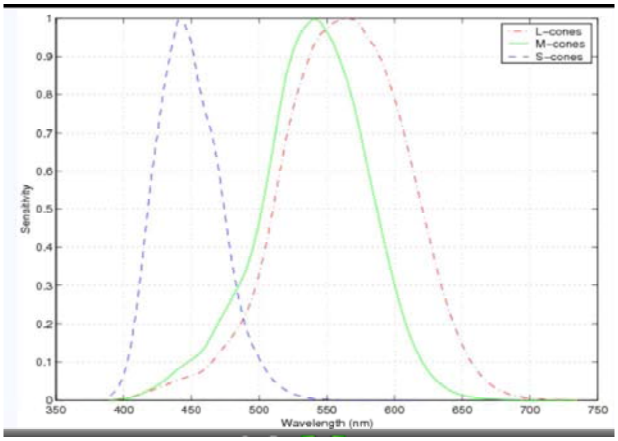


Figure 5 Different Cone Type Wavelength Sensitivity

This information electrically stimulates the optic nerve which feeds into the visual cortex of the brain for semantic and feature processing based upon differences to a windowed-steady state visual model. Eye movement, both right tied with left (saccades), is based on spatiotemporal sensitivities to capture these differences to the brain. From what we see in the human visual system, the visual cortex in the brain does not try to process all information but just what is needed to provide a semantic visual model. Perceptual coding attempts to move past the photoreceptor stage to keeping just the information that will make it into the visual cortex.

So, in trying to model HVS, it can be split up into three areas: 1) a visual attention model, 2) spatiotemporal visual sensitivity model, and 3) a visual masking model. This is basically what is interesting to see, what we can make out of it, and what we could never see at all. Our visual system is sensitive in a number of ways:

- **Contrast**- we aren't sensitive to a level of brightness, we are sensitive to differences in brightness between areas in our vision. This equates to sensitivity to edges in an image and can be affected by the brightness in the background.

- **Spatial Frequency**- as spatial frequency increases, we become less sensitive to variances in spatial details (when does edges become texture?). This can equate to tolerance in coding artifacts in high texture areas as opposed to more constant areas. In color we are even less sensitive to variances in spatial frequencies. Hence one of the reason we can sample color difference less frequently than luminosity (4:2:2 or 4:2:0).

- **Visual Acuity**- This is the ability for the eye to resolve details. One can have reduced visual acuity in fast moving objects (though eye tracking can reduce perceived motion of the object --- reduced retinal velocity). One can also reduce visual acuity by moving further from the object or screen. For ideal viewing, Viewer should be far enough away to not be able to discern pixels on the screen. Increased resolutions can allow for the observer to sit closer to the screen without being able to discern pixels [Figure 6].

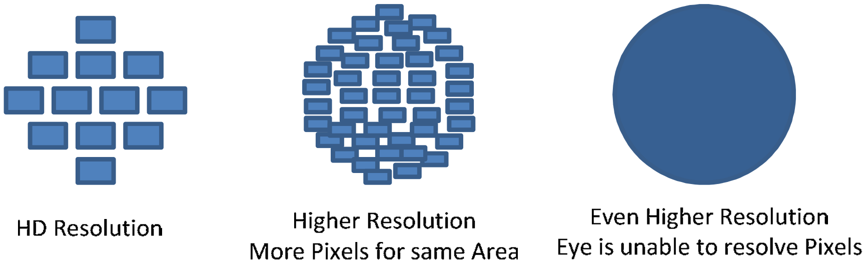


Figure 6 Visual Acuity and Denser Pixels

- **Noise**- These are unnatural changes in contrast due to the image capturing process. This could be due to the scatter on photo sensors in the CCD/ CMOS, heat on electronics carrying the pixel values, or celluloid processing leaving film grain artifacts [Figure 7]. The eye is sensitive to noise at different spatial frequencies which is why low-pass/

band-pass filtering is used as a preprocessing technique to remove these unnatural artifacts.

- **Temporal Frequency**- we are more sensitive to temporal cues rather than lack of spatial details. This is one of the reasons why interlacing can happen because it is a tradeoff of spatial frequency for temporal frequency to address bandwidth issues. It is believed below 50-60 Hz (fps), flicker can be perceived in a series of played out still frames. For this reason, 24fps material sometimes is flashed twice in frame playout on display devices and now material traditionally being shot at 24fps is being shot at 60 fps or even 120 fps for this reason. Additionally, movement that follows natural movement speed and direction is less surprising than erratic movement and speed.

- **Perceptual Uniformity**- This basically means keeping a consistent quality across a video sequence. We are sensitive to quality changes in spatial details of a moving object when viewed in the fovea area of the eye.

To mimic HVS, the attention model needs to identify areas of the image that are tracked by eye movement (saccade) to keep interesting areas in the fovea. Things outside of the fovea do not have to retain as much detail due to change blindness. Object size, and movement (predictable and unpredictable) can be used as cues to identify areas in the video sequence that need more spatial detail. Artifacts can cause a miscue in the eye to areas in the video sequence that are not natural areas of interest and need to be minimized where possible. The spatial temporal model can affect how to maintain a natural sequence with consistent quality over a content scene. Visual masking is a preprocessing function that can hide information in areas that don't need as much spatial detail such that it is coded in a fewer number of bits.

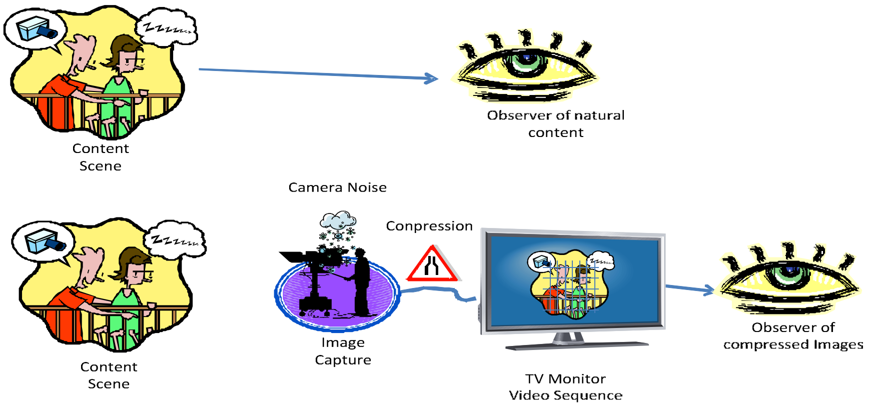

Content Scene
Observer of natural content
Camera Noise
Conpression
Content Scene
Image Capture
TV Monitor Video Sequence
Observer of compressed Images

Figure 7 Capturing Natural Content on Screen

EARLY PERCEPTUAL CODING TECHNIQUES IN COMPRESSION

When we directly see a natural scene, our eyes have a filter (mentioned in the sections above) that reduces the amount of information that reaches the visual cortex. We use our eye muscles, focus and movements to change what the cones in the fovea are seeing such that attention is there for important information in the scene.

To capture the image such that we can recreate what we see (Film/ TV without compression), we represent the scene through a series of still pictures being played at a specific temporal frequency (24 fps (2x)/ 30 fps (60 fields)/60 fps). Consistent quality is maintained between each frame, and interlacing techniques are used for further reducing bandwidth using a tradeoff of spatial resolution and temporal frequency.

However, in the capturing of the image, noise is introduced into the content scene through CCD/CMOS camera devices. To avoid seeing the pixels instead of the content scene, we sit back far enough (2H-4H) such that our visual acuity cannot discern a pixel and blends them together.

With the evolution of an analog medium (6MHz analog program) to a digital medium (10 Channels in 6MHz), we now have the ability to manipulate each pixel value and only send difference information between each frame (i.e. compression). In terms of pre-processing, the noise is being removed through low-pass, band-pass, and temporal filters like MCTF. The encoder then uses block-based transforms to change the coefficient values to be measures of spatial frequency energy.

At this point, the coefficients of higher energy frequencies can be quantized with less precision and use less bits because we have less sensitivity at high spatial frequencies. Additionally, this helps with reducing data redundancies in the bit streams since many of these coefficients are quantized to zero.

In terms of motion, movement of natural objects can only move at certain speeds and are predictable which factors in to some of the coding algorithms that reduce computational complexity. This allows for a reduction of motion search space, and a reduction of number of motion vectors based on size of the object. The “errored” differences between frames can also be quantized in the same manner since errors are mostly in high spatial frequency details. In post processing, the blocking artifacts along transform boundaries can then be removed from the image.

To avoid seeing artifacts from the medium (pixels) rather than the content scene, it is important to be able to view the display screen at the proper viewing distance. If one moves in closer, visual acuity increases to the point where pixels can be discerned (visual acuity is inversely proportional to distance). In terms of monitors, we are getting larger monitors going from 40” to 50” to now 60-70” sets, and the viewing distance from these monitors is remaining mostly constant.

Additionally, we are also getting display devices like tablets and PCs that are being viewed at much closer distances than the 2H-4H recommendations.

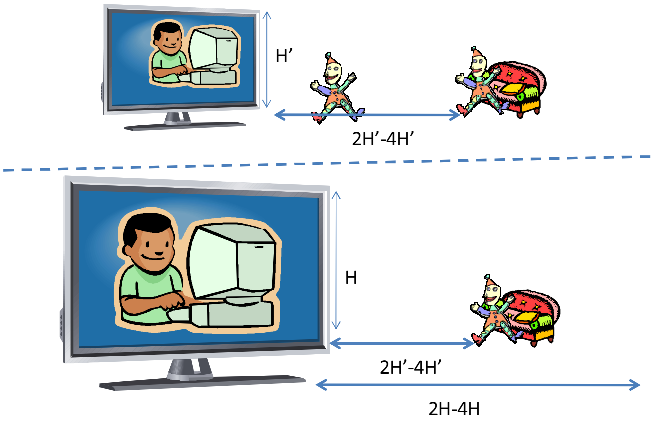


Figure 8 Screen Sizes, Distance, and Visual Acuity in Monitors and PC/Tablets

AFFECTS OF 4K AND HIGHER FRAME-RATES

Going to 4K can create more natural content scenes. Increasing pixel density does not have to create a larger picture; it creates a more densely sampled picture. Each pixel now represents a smaller area which allows for:

- Sharper Edges
  - ✓ Fonts on letters are sharper. The viewer can read documents. [It's "Resolution-ary"].
  - ✓ Less aliasing artifacts and "jaggies" around edges
  - ✓ Textures are more detailed
- Increased pixel density
  - ✓ Approaches visual acuity limits. See less pixel definition and more of the picture at closer viewing distances and angles.
  - ✓ The Viewing distance becomes more flexible. We can get closer to pictures in both large and small displays (This aligns better with the attention model)
- Better contrast
  - ✓ Pictures look brighter/ more natural due to contrast differences and more gradient increases and decreases (This was always an issue for compression)
  - ✓ Neighboring pixels are more correlated since they represent a smaller area

Going to higher frame-rates can create more natural content scenes cues, by sampling motion in content scene to make it more linearly predictive. This is becoming more helpful as CGI (computer-generated imagery) effects in film and video content introduce faster moving objects in sequences. It is also very helpful in sports content where motion is quick and erratic. If the frame rate is too slow for the motion in the content scene, we can get "juddering" artifacts especially if the picture is flashed multiple times to simulate higher frame-rates:

- Smoother Motion
  - ✓ Movement between frames is shorter and can be predicted better
  - ✓ "juddering" can be reduced due to more sampling of motion and less repeated flashing of the picture
- Less Noise from Image Capturing devices
  - ✓ Noise is not temporally correlated and can be filtered through comparisons of sequential frames.

LOOKING AT CODING WITH RESPECT TO HIGHER RESOLUTION AND COMPRESSION

With increases in resolutions, there are going to be more pixels to process. The encoder picks a target bitrate and then tries to make decisions in coding based upon that. Generally, the encoder attempts to conduct:

1) ***Pre-filtering*****:** remove noise and apply a low-pass filter to remove information and details that would never be resolved at that bit rate anyways. Basically, to remove the information that makes the encoder work harder than it needs to be working.
2) ***Transform/Quantization*****:** change the information order of the data stream to make it more compact and quantize high spatial frequency information. Apply entropy coding to the output of this stream
3) ***Predict Subsequent Frames*****:** Use a reference frame(s) to produce a set of motion vectors and "errored" difference frames (P& B Frames). Calculation of motion vectors need to go through a motion vector search which can be a complicated encoding process.
4) ***Post processing*****:** Conceal artifacts created by the encoding process such a blocking and boundary artifacts through post filtering approaches

Places where we can improve this process, due to having higher resolutions and frame-rates, include:

1) ***Pre-filtering:*** Removing noise may be easier because it is approaching the granularity of our visual acuity while natural content scenes would not have this level of granularity. Using the stronger correlation between neighboring pixels, there can be improved techniques for filtering and dithering to handle noise. Additionally with the improvements in CMOS, we may be able to do this earlier at the point of image capture.
2) **Transform/Quantization**: The transform represents a smaller area and more correlation between the pixels which can help in energy

compaction. Some savings can be achieved as well because quantization levels can be changed for a smaller area. However, there are more transform blocks to deal with at higher resolutions.

3) **Predict Subsequent Frames**: With higher resolutions, movement can go beyond the motion search space, which would mean more bits to encode. With higher frame rates, movement is shortened between frames and is much more predictable, which could reduce the amount of bits that are expended. Objects are also bigger (have more pixel density), which would require less motion vectors to support this process. With ½ pel (pixel) motion accuracy across a smaller portion of the picture, the effect of this approach could be fewer errors in the "errored" difference frame. With more accuracy this can save on bits as well. Lastly another effect is longer GOPs over the same time period (just more frames in the same period) which can reduce the expected increase in data through temporal compression.
4) **Post-processing**: There would still be blocking artifacts that would need post processing it would just be smaller in the picture and may only need simpler post-processing techniques.

ENCODERS AND NEW CODING TOOL ABILITIES

With new demands for multiple bitrate encoders and addressing multiple devices, encoders have been evolving to output streams at multiple target bit rates. In many encoders, there is already a calculated quality metric used to make encoding decisions used for the purpose of meeting multiple target bitrates. Additionally encoders are also deploying "look ahead" to analyze the source content to optimize encoding decisions. Both these mechanisms help out in maintaining perceptual uniformity and enabling better visual masking throughout the video sequence through the use of dynamic adaptable filters.

The newer coding standards (i.e. AVC/HEVC) have also been evolving that are developing advancements in coding tools to handle each sub-area of the image and sequence in a different manner. The objective is to use as few bits to convey parts of the image or sequence that don't need as much detail such that more bits can be spent elsewhere. For instance, the background may not need as many bits as a moving object in the foreground. Also, a moving object may not need as much motion vectors since the object travels at the same relative speed against the background. Some tools being developed or refined are:

- Spatial Intra-frame compression Techniques
- Better motion pixel motion search down to ¼ or 1/8
- More granularity in quantization across coding units or transforms
- Changing the transform block size- 8x8, 4x4, 8x4, 4x8

- Changing the size of the macro-block (16x16 to 64x64)
- Changing the number of motion vectors needed for a macro-block
- Reducing the number of motion vectors needed for coding

These different tools contribute to being able to identify and handle separate areas of the image, treat specific bands of spatial frequencies with alternate options, and to code objects as separate temporal frequencies. Combine this with the ability to analyze content and a calculated quality metric in the encoder, and you have the basic tools for creating an attention model along with further refinements in the spatiotemporal sensitivity model and visual masking. From this, the HVS model used in encoding can rapidly improve encoding and reduce the amount of bits needed that can be processed by our HVS system beyond the 50 % reduction already claimed by the latest codecs.

CONCLUSION

The first phase of 4k video delivery should focus on a quality experience for the customer. It is expected that 4k will start with a single, linear occasional channel and a small library of 4k VOD assets encoded with H.264 compression techniques. Should 4k prove to be a success, it will be easy to expand the VOD library by transcoding studio content into higher resolution video.

In order to support new audio formats, assets would need to be remixed. MSOs could have a very basic 4k solution in place in the very near future; and HEVC encoding will allow a production 4k solution using substantially less bandwidth. Based on our calculations, and leveraging coding algorithms sensitive to the human visual system (HVS), 4k assets may in the near future consume the same bandwidth on the local loop as an existing HD asset encoded with MPEG-2.

References

[1] Tang, Chih-Wei, *"Spatiotemporal Visual Considerations for Video Coding"*, IEEE Trans. On Multimedia, Vol. 9, No. 2, Feb. 2007, pp. 231- 238.

[2] Naccari, Matteo and Pereria, Fernando, *"Advanced H.264/AVC-Based Perceptual Video Coding: Architecture, Tools, and Assessment"*, IEEE Trans. On CSVT, Vol. 21, No. 6, June 2011, pp.766-782

[3] Wu, H.R. and Rao, K.R. eds., Digital Video Image Quality and Perceptual Coding, CRC Taylor and Francais Group, New York, 2006.

[4] JCTVC- G1113 WD5: Working Draft 5 of High-Efficiency Video Coding, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 7th Meeting: Geneva, CH, 21–30 November, 2011

[5] ITU-T Rec. H.264 | ISO/IEC 14496-10, (2005), *"Information Technology – Coding of audio visual objects –Part 10: Advanced Video Coding"*

[6] *"Understanding CCD Read Noise"*, www.qsiimaging.com/ccd

[7] Additional Conversations and some eye diagrams Dr. Damian Tan and Dr Henry Wu, School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Melbourne, Victoria Australia.

A Software Friendly DOCSIS Control Plane

Alon Bernstein
Cisco Systems

Abstract

It has been 15 years since the initial set of DOCSIS specs have been authored. In those 15 years software engineering has seen an explosion in productivity at the same time that the DOCSIS control plane has remained fairly unchanged. Can we apply these productivity tools to the DOCSIS control plane to facilitate greater simplicity and feature velocity?

This paper will outline both software trends and protocol design trends that are relevant to the above discussion and how they can be applied to DOCSIS.

OVERVIEW

DOCSIS is primarily an interface protocol between a CM and a CMTS. There is a wide palette of options for a protocol designer and all are relevant to DOCSIS design. Each option has its tradeoffs and the role of the protocol designer is to choose the option that fits the system requirements and constraints the best. The list of options include:

- Generalized interface vs. mission specific interface
- Legacy protocol vs. mission specific protocol
- Stateless vs. stateful
- Client-Server vs. Peer-to-peer

And more…In many cases there are no simple rights and wrongs and a choice that might have made sense at the time of the protocol design turns out to be sub-optimal as systems often end up getting deployed in a manner that is different then what they were designed for. All of these choices have an impact on software. Its not always the case the choice that is optimal for software is the ideal for meeting the system requirements, still its unfortunate that in many cases the software implementation ease is considered as a relatively low priority item. This observation is patricianly in place for DOCSIS since the amount of software resources needed to support the DOCSIS set of protocols is significant.

Before proceeding, a word of caution: in cases where there are requirements and constraints that supersede software requirements then clearly the guidelines explained in this document will not apply. The challenge is to identify these requirements and constrains correctly and not to pre-optimize at the expense of “software friendliness”. To quote the author of the “Art Of Computer Programming”:

> *"We should forget about small efficiencies, say about 97% of the time: premature optimization is the root of all evil”* (Donald Knuth)

SOFTWARE CONSIDERATIONS

Software engineering and protocol design share the same approach to simplifying complex system requirements:

- Modularization: sub-divide a large and complex system into simple and easy to test modules with well-defined inputs and outputs.
- Layering: Define the hierarchy of modules, what services each component provides to another.
- Abstraction: identifying common services that can be shared across modules

Software design methodologies have evolved around the same timelines as the Internet revolution and the creation of networking protocols. But while the timelines are similar the amount of change software went through is much larger the amount of change in the suite of networking protocols that drive the Internet.

Software has evolved from the "C" programming language to object oriented languages such as C++ and Java which allowed for further modularization/layering and abstraction of code to web technologies that brought amazing scale, speed and flexibility. At the same time network protocols stuck with the OSI 7 layer model [6] as possibly the only attempt to apply modularization/layering and abstraction to networking. Case in point: many of the routing protocol RFCs have pages upon pages of interface specifications and message formants. If written from with a "software friendliness" point of view they could have had a well-defined separation of the methods to distribute data across a group of routers (which is similar in many of the routing protocols) and the actual routing algorithms.

Obviously the issue outline above has a wider scope then DOCSIS, so to keep the discussion focused here are a couple of examples of why the current DOCSIS specifications does not follow basic software implementation guidelines along with a high-level proposal on how to fix it (going into fine details is outside the scope of this paper):

Example 1: DOCSIS registration.

The DOCSIS registration process starts with bringing up the physical layer, jumps to authentication and IP bring up, then to service provisioning then back to physical layer bring up.

Initially services are created in the registration process. Additional services are added with a different mechanism then the one used in registration (DSx).

Why is it not software friendly? The fact that the cable modem bring-up has a dependency on the IP layer bring-up makes it difficult to independently develop (aka "feature velocity") and independently test (aka "product quality"), the registration process.

This is a good example where an idea that seemed to offer:

1. Simplification: because the same mechanism used to provision services is also used for the modem bootstrap
2. Optimization: fewer messages since all the various layers are squeezed into the same

Turns out to be not-such-a-good-idea when it comes to software implementation. This becomes painfully obvious when the system is physically distributed. Imagine an implementation where DOCSIS functionality is segregated into processor A and Layer 3 functionality into processor B. Because of the way registration is handled the DOCSIS processor needs to know a little about the IP layer and the IP processor needs to know a bit about DOCSIS. Clearly these are solvable problems and "anything can be done in software" but as mentioned above there is a price to pay in speed of implementation, testability and debug of system issues.

How would a software friendly registration protocol look? A software friendly specification would have clear and independent stages as depicted in the figure below:

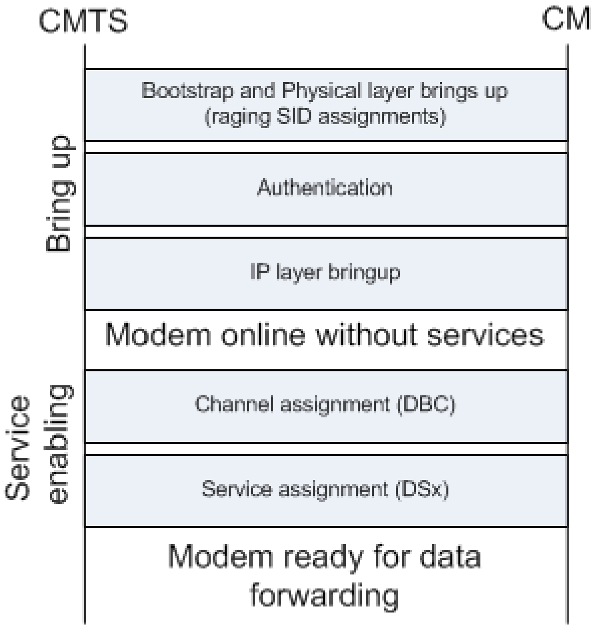

CMTS
CM
Bring up
Bootstrap and Physical layer brings up (raging SID assignments)
Authentication
IP layer bringup
Modem online without services
Service enabling
Channel assignment (DBC)
Service assignment (DSx)
Modem ready for data forwarding

Figure 1 SW friendly registration

1. bootstrap: initial physical layer bring up

2. Authentication: validation of the cable modem for network access
3. L3 bring up: DHCP processing and IP address assignment

Each one of the above steps would be treated as an independent transaction and the three of them would be the workflow needed to bring a cable modem online.

For these 3 steps the major deviation from the current registration is that we don’t rely on TFTP for service provisioning. There are two reasons to skip TFTP; the first being that the IP layer is not even up so we can’t access anything beyond the CMTS and the second being that we want to postpone the service providing part to a later stage anyway. Having said all that, the CM still needs to communicate with the CMTS and it still needs some form of a service flow to do it. If we don’t have any services provisioned how do the CMTS and CM communicate? The “temporary flow” that DOCSIS creates anyway for registration would just leave on until after the IP bring-up phase and only after that would be replaced by the “real” in the service enablement phase

As far as the next steps go we fortunately have clear transactions to handle:

- Service provisioning using DSX
- Changing physical layer parameters with the DBC

These can be used to change services and channel assignments after the modem is online. Note that the CMTS can still use TFTP to retrieve service parameters and those will be parsed into a DSA message.

Naturally there are tradeoffs to this proposal; the number of messages has increased and a new form of service enablement has been added, however the payoff is significant in terms of software modularity.

Example 2: The Mac Domain Descriptor MDD message is a dumping ground for information about plant topology, IPv6, error message report throttling, security, physical layer parameters and more. A software friendly specification would create independent messages for each of the functional areas. Though one can argue that MDD is “just a transport” for data that can be managed by independent modules in the software, however the inclusion of all of them in the same message creates dependencies that are easier to avoid were the MDD to be broken into separate messages.

END-TO-END PRINCIPAL

Some link-level protocols (such as DOCSIS) assume reliability is required and come up with their own set of timers to assure delivery at the link layer. This might be justified if the link layer is highly unreliable, and even in that case the timeouts set for retransmission must be an order of magnitude shorter then the timeouts of the end-to-end application. If retransmission timers are too long then all sorts of odd corner cases might occur. For example: the DSx-RSP timeout is about 1 sec and there can be 3 of them. If an end-to-end signaling protocol, such as SIP [4] has a message re-transmission time of the same order of magnitude then a DOCSIS implementation might release a SIP message when the application level has already timed out and re-transmitted its own copy. This will not cause the system to break since a robust implementation knows how to deal with messages that are duplicated or out of order, but it will clutter the error counters and fault logs with a "duplicate message" event which would have been avoided if the DOCSIS link layer counted (as it should have) on an end-to-end session establishment protocol. The reader might ask, "what if the end-to-end protocol is designed to be unreliable"? Even in that case it’s not the role of the DOCSIS link-layer to assure delivery if the higher lever application does not require assurance in order to operate correctly.

The DOCSIS software may trigger a timeout for a DSx-RSP, however the expiration of this timeout would be only used for recording a failure and releasing system resources allocated for the DSx, not for triggering a retransmission.

If the media is highly unreliable and failures are a common occurrence then they might be room for link-level error repair but in that case the timeouts need to be an order of magnitude shorter then the application timeouts - a suggested range would be 100ms or so.

A further simplification based on the end-to-end principal is to remove the DSx-ACK phase. DOCSIS uses a 3-way handshake for DSx. The rough outline of the conversation at the service activation phase goes like this:

1. CMTS -> CM: please start a service (DSA-REQ)
2. CM -> CMTS: ok, I started the service (DSA-RSP)
3. CMTS -> CM: cool, my CMTS resources are ready you can start sending data (DSA-ACK)

But this third step is not really needed for the same end-to-end argument. For example consider this zoon-in of a PCMM message sequence (figure 9 in ref [2])

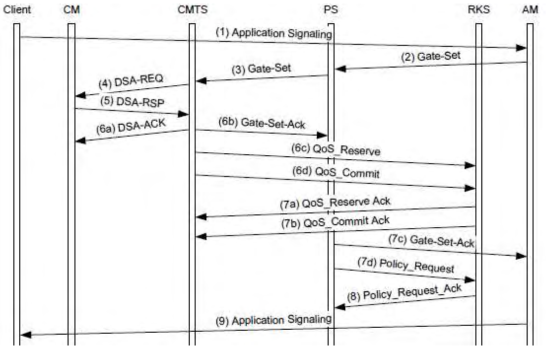


Figure 2 PCMM application signaling

It's obvious from this diagram that the DSA-ACK is not fulfilling any useful function. For one it's a "dead-end" not resulting in any further action, and since it is sent at the same time as the Gate-Set-Ack it is useless in guaranteeing any sequencing of events.

As a side note, some have suggested that the DSA-ACK is needed for extra reliability but this would be an even worst violation of protocol rules since the DSA-RSP is already an acknowledgement and a protocol should not acknowledge and acknowledgment.

ENCODING

DOCSIS uses TLVs to serialize information however TLVs are not common in modern networking stacks and not supported in many of the productivity tools and code generation tools used today. Non-TLV types of encodings include JSON/HTTP/XML/google "protocol buffers" and others. The advantages of the above mentioned tools are:

1. They come with code generation tools that relive the software developer from the burden of parsing messages into native data structures.

2. Most of them encode information in human readable strings that makes debugging easier.

TLVs are a more compact form of serializing data but as bandwidth available on the cable media increases this is becoming a non-issue.

TLVs might also be easier to parse, but CPU power is much less of an issue then it was at the time the DOCSIS specification were written. In fact, the modern cable modems have more powerful CPUs then early CMTS products!

OPEN SOURCE

Another software trend that has been going strong is the movement to open source. As a development methodology it has proven to deliver on wide scale and highly complex software projects. How can open source apply to cable? The CMTS/CM interaction is not likely to be of interest to the open source committee since its so domain specific and for product differentiation reasons it's highly unlikely that CMTS/CM vendors will open their source code.

This document proposes to use source a companion to the CableLabs standard documentation process. For example, if a new registration process was to be pursued then high-level function calls and JSON encodings could be published as open source. This would hopefully promote better interoperability and shorter ATP cycles as it removes a lot of ambiguity in the interface design.

DATABASE TECHNOLGIES

A CMTS implementation needs to manage a database of cable modems, plant topology and more. In many cases this information needs to be shared with a CM and so one view of a DOCSIS system could be that it's a distributed database of CM state and resources. With that observation it's clear that the only type of data sharing that DOCSIS allows for is the transactional type. That used to be the only model for data sharing in the database industry in general but the scale that companies such as google and facebook had to grow to gave rise to a new model, one that priorities performance over accuracy. Clearly for some types of data this model will not work well (financial transactions for example) while for others it makes sense (searching through web pages).

An interesting observation made by the Internet community is captured in what's called the "CAP theorem" [5]. In a nutshell what the CAP theorem states is that when a database designer is requested to support a distributed database that provides [1]:

1. Performance
2. Consistency of data across components of the distributed database
3. Resiliency in cases for system malfunctions, for example, packet drops.

Only two out of these three requirements can be met. The designer still has to choice of which two are fulfilled, but it is not possible to meet all three.

As mentioned above DOCSIS supports a transactional sharing of data that represents a choice of consistency and resiliency over performance, and as long as performance requirements are met (for example, number of voice call created per-second) it is a win-win situation. But as new applications become available and the load on the control plane increases it may make sense to consider other choices. The proposal in the previous section to avoid re-transmissions represents the option of demoting resiliency. Another option is to assume an "optimistic model" where the CMTS can allocate and activate resources on a DSA-REQ, assuming that a positive DSA-RSP will follow and intentionally allowing for short period of times of inconsistency if cases where the DSA-RSP was not successful.

Another useful tool from the database world is the concept of "data normalization". Its outside the scope of this paper to go into the detail of data normalization (see ref [3]), but in a nutshell it's a set of guidelines on how to break complex data into a list of tables with rows and columns where each row is fairly atomic. When inspecting some of the DOCSIS MIBs and MAC messages its obvious that some break at least one of the normal forms. For example, the inclusion of a "service flow reference" in the same table as the "service flow id" violates the normal form that prohibits dependencies between columns of a table. Without getting into too many details this paper only makes the observation that management constructs that are "normal" are easier to implement in software.

SECURITY

An obvious security hole in DOCSIS is letting the cable modem parse the configuration in order to reflect it back to the CMTS. It's worth mentioning it in this document because (ironically) this might have been the single attempt in DOCSIS to help software by offloading the task of parsing the cable modem configuration to the cable modem. However, in order to plug this security hole the CMTS needs to parse the configuration anyway and overall it's a great example of how premature optimization can create more harm then good.

[1] [ab] CAP stands for "consistency, availability, and fault tolerance". I took the liberty of translating the above to terms familiar to the cable community since the original terminology might be confused with existing cable terms.

CONCLUSION

DOCSIS has obviously been a very successful protocol. The DOCSIS provisioning and back-office system is part of this success, especially when comparing it to its DSL counterparts where a strong standard for provisioning and service enablement does not exist. Nevertheless a 15-year critical review, and possible updates would certainly help DOCSIS to become even better in facing the challenges ahead.

This paper suggests that software implementation ease and modern software tools need to play a bigger role in the design of future DOCSIS protocols and while some of the proposals made here are of academic and demonstrative value only, others can be relevant to future versions and enhancements of DOCSIS.

REFERANCES

1. DOCSIS MULPI : http://www.cablelabs.com/specifications/CM-SP-MULPIv3.0-I18-120329.pdf
2. PCMM: http://www.cablelabs.com/specifications/PKT-SP-MM-I06-110629.pdf
3. Codd, E.F. (June 1970). "A Relational Model of Data for Large Shared Data Banks". Communications of the ACM 13 (6):377–387.doi:10.1145/362384.362685.
4. Session Imitation Protocol, RFC 3261
5. CAP Theorm : http://www.cs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf
6. ITU-T, X.200 series recommendations: http://www.itu.int/rec/T-REC-X/en

PUSHING IP CLOSER TO THE EDGE

Rei Brockett, Oleh Sniezko, Michael Field, Dave Baran
Aurora Networks

Abstract

The ongoing evolution of cable services from broadcast video to narrowcast digital content (both data and video) has fuelled corresponding technical innovations to solve and support operators' operational and capital requirements. One area of particular interest is the QAM modulator. Accelerating subscriber demand for data and narrowcast video services will require a surge of new QAM deployments over the next several years, giving rise to a host of operational difficulties.

In this paper, we present the case for distributed headend architecture for HFC networks and discuss architectural and operational benefits of the Node QAM form factor, where the conversion of digital payload into QAM-RF signals is pushed from the headend to the cable TV optical node. In addition, we analyze the Node QAM in the context of the CableLabs® Converged Cable Access Platform (CCAP) architecture.

BACKGROUND

Distributed Architecture Drivers

A key topic when discussing next-generation cable infrastructure is the balance between analog optical transmission, including the transmission of multicarrier QAM-RF signals, and baseband digital transmission of signals such as native Internet Protocol (IP) signals. Cable operators have gone through several transitions already, with the introduction of digital television; the growth of high-speed data; the use of IP-based distribution in the headend; and the use of native baseband IP-based communication between headends and hubs. The driving force has always been efficiency and cost.

The imperative to meet subscriber demands results in certain bottlenecks: physical space and power within the headend, bandwidth capacity in the deployed HFC, distance between headend and subscriber, limitations of hard-wired infrastructure.

For each of these areas, there are solutions, but a distributed headend architecture that extends the boundary point where content enters the RF domain addresses all of these:

- Headend space and power consumption can be mitigated by consolidating functionality and increasing port densities in next-generation CMTSs and Edge QAMs. Alternatively, functionality can be distributed to the hubs and nodes, leaving only the IP network and MPEG2-TS processing in the headend. Direct generation of RF output at the edge of the network eliminates the need for an RF combining network at the headend. This reduces headend space and power requirements and simplifies network operations by avoiding the need to mix signals in the RF domain.

- Distance limitations can be relaxed by pushing deeper the conversion of digital signals to RF. Analog optical transmitters and amplifiers are at the limits of their capabilities, and add expense and design complexity. However, by extending the headend IP domain to the node, not only is optical transmission distance extended, but RF signal loss budgets are mitigated and higher loss budget at higher frequencies can be accommodated, thus increasing bandwidth capacity of the subsequent coaxial section of the HFC network. For example, baseband optical links to the node would eliminate analog link contributors to signal degradation, thus allowing for higher modulation levels and hence better spectral efficiency in the available coaxial bandwidth. This can be especially effective and fruitful in passive coaxial networks (PCN), also known as Fiber Deep, Fiber to the Curb (FTTC), or Node-plus-zero (N+0) HFC networks.

- In addition to the effect of explicit signal impairments due to analog optical transmission, bandwidth capacity in the HFC network is further constrained by the complexity of carrying analog (RF) signals over distance. In the optical links to the nodes, the use of multiwavelength systems, while justified by fiber scarcity and revenue opportunities, introduces severe constraints on the usable number of wavelengths and their link performance. Impairments from analog (RF) modulated optical transmitters and erbium-doped fiber amplifiers (EDFAs) further limit the capacity of individual wavelengths. Converting from RF modulated transmitters to baseband digital optics would eliminate these impairments and increase the number of cost-effective wavelengths to 88 (yielding 880 Gbps of capacity to each node) using current technology, with room for growth in the number of wavelengths and the wavelength capacity of next-generation optics.

- The challenge of managing bandwidth allocation between unicast, multicast, broadcast, and data QAM signals is eliminated by mixing content dynamically in the headend IP network. This allows bandwidth to be allocated as-needed in response to market requirements without requiring “hands on” labor.

Accelerating Demand for Narrowcast Services

Rapidly evolving subscriber behavior surrounding the consumption of multimedia is driving cable operators to confront two challenges. The first is the need to significantly accelerate the deployment of narrowcast services while also accommodating bandwidth-intensive services such as HDTV and 3DTV. These narrowcast services typically include high-speed data and packet voice, video on demand (VoD), and switched digital video (SDV), but also encompass other unicast and multicast services such as cable IPTV, network-based digital video recording (nDVR), and other services that leverage the IP cloud at the headend. The second challenge is the difficulty of planning a graceful and cost-effective migration from inefficient and obsolete service silos to new, dynamic methods of flexibly allocating capacity to different services in the face of constantly shifting customer demands.

The need to deploy an unprecedented volume of new QAM modulators is common to both challenges, and this raises concerns over issues including headend environmental constraints, flexibility of service allocation, RF combining issues, HFC transmission considerations, and the need to accommodate legacy equipment.

In these circumstances, one viable solution that achieves the benefits listed above is to relocate the QAM modulators to the HFC node, pushing the native baseband IP domain even further to the edge (closer to the user — the ultimate edge of the HFC network).

DESIGNING A NODE QAM

A Confluence of Technology and Need

Quadrature Amplitude Modulation (QAM) is a spectrally efficient way of using both amplitude and phase modulation to transmit a digital payload on an analog carrier. Cable QAM modulators[1] operate on packets in the MPEG2-TS format, and modern QAM modulators include integrated upconverters as well.

In the decades since the first baseband QAM modulators were assembled out of discrete components, silicon technology has increased a thousand-fold in processing price performance, and decreased a hundred-fold in size, giving rise to a surprisingly rich selection of special-purpose, general-purpose, and programmable chips, based on which we can re-design our modulators.

These advances can finally be used to their full advantage now that demand for modulators has swelled from tens and twenties per headend to hundreds and even thousands. Part of the advantage is in the availability of brute-force processing power, but a companion advantage is in algorithmic efficiencies derived from being able to perform certain steps in bulk. One result is that existing headend Edge QAMs can be made much denser, with thousands of QAM channels in a chassis. Another result is that it is now operationally feasible to put a full gigahertz' worth of QAM channels (or more) in the node.

Node QAM Requirements

The node is a hostile environment for advanced electronics. Power budget and space are limited; cooling is passive; operating temperatures can be extreme; and accessibility is limited. In order for a Node QAM to be operationally neutral when compared to a headend Edge QAM, it must meet the following criteria:

- Low power. In order to avoid the need for non-standard node powering, a full-spectrum Node QAM must be able to generate at least 158 (6 MHz) QAM channels using the same amount of power as a traditional optical receiver. This eliminates the need for active cooling.

- Compact. The Node QAM should be designed to fit within the existing, field-proven node housings.

- Industrial grade operating temperature range (–40°C to +85°C). Unlike climate-controlled headends, or even cabinet-based hubs, components in the node must be able to withstand large fluctuations in temperature.

- Reliable. Servicing a node is logistically cumbersome and operationally expensive. A Node QAM must be robust and uncomplicated. Additionally, remote monitoring is critical. Ideally, cost, space, and power consumption profiles can be kept low enough to enable the deployment of spare modules, which would allow operators high levels of redundancy, even at the node level.

- Simple to install — "Set it and forget it". Installing a Node QAM must be as simple as plugging in a module and verifying the output with a field meter. Complex procedures such as configuration and management should be done centrally, to simplify operations.

- Low cost. Per-channel equipment costs need to keep pace with the cost of headend Edge QAMs.

- Future-proofed. Given the logistical difficulties of servicing nodes, the distributed Node QAM modules should have a margin for upgradability so future technological changes and additions can be accommodated by re-programming the existing modules. This not only simplifies architectural evolution, but also extends the operational lifetime of each module.

This is also applicable to the interfaces between the node modules and the headend/hub infrastructure; new modules can be introduced in a very scalable manner if they leverage the standard data networking interfaces used in the IP network in the headend.

These requirements, while difficult to achieve, are attainable given modern silicon capabilities and careful design, opening up the option to move to a more distributed architecture, with many of the benefits.

ARCHITECTURAL BENEFITS

Generating some or all QAM signals at the node results in a number of advantages.

Exploiting Digital Optics

A major advantage of moving the QAM modulator to the node is the ability to shift to digital optics between the headend and the node. In traditional usage, electrical RF signals are amplitude-modulated onto an optical signal. These signals are extremely sensitive to various fiber nonlinear distortions like cross phase modulation (XPM), stimulated Raman scattering (SRS) and optical beat interference (OBI) caused by the four-wave mixing (4WM) products that come into play depending on power, distance, wavelength count, and other factors. Together with other nonlinear and linear fiber impairments, they limit the capacity of the links and significantly impair transported signals. Designing and "balancing" optical links to the nodes in an HFC system is a delicate art. Furthermore, the lasers modulated with analog (RF) multicarrier signals have limited Optical Modulation Index (OMI) capacity due to the fact of high sensitivity of these signals to clipping. The limits reach up to 30% for directly modulated lasers and approximately 20% for externally modulated lasers. These limits, with the operational back-off of 2-3 dB, severely limit the capacity of every single wavelength in any multiwavelength system of practical distance.

Using baseband digital transmission is much simpler. Because data is not as sensitive to nonlinearities and other impairments, not only can distances be extended, but more wavelengths within a single fiber can be employed, resulting in higher bandwidth capacity to the node. Simpler and more economical optics and amplifiers can be used, as well. With their OMI approaching 100%, digital optics enable significant increases in the capacity and distance of each fiber optic link. Using existing technologies, they can support cost-effective transmission of 88 wavelengths with 10 Gbps/wavelength over distances in excess of 100 km from the IP headend/hub infrastructure. This opens significant opportunity to provide unparalleled bandwidth to the nodes for residential services as well as significant opportunity for additional revenue.

More cost optimizations for capacity and distance can be achieved by leveraging lower-cost optical amplifiers, simplified optical filters, and symmetric and asymmetric SFP, SFP+ and XFP transceivers. Furthermore, deploying distributed architecture and transmitting native baseband IP signals to the node finally enables HFC to take advantage of the high-volume economies of scale in modern digital (data) networking infrastructure, which outperforms the economies of scale for analog cable TV optics a thousand-fold.

Another benefit of using baseband digital optics between the headend and the node is the elimination of an HFC weakness: the analog link contribution to end-of-line noise budgets. The analog (RF) optical links to the nodes with analog (NTSC or PAL) video signals are designed for 47 to 50 dB carrier to noise ratio (CNR) for occupied bandwidths ranging from 700 to 950 MHz. For QAM signals placed on the same link, it translates to modulation error ratio (MER) between 39 and 42 dB. For links with QAM-only load, this limit is usually lowered by designers to 37 dB MER to take advantage of cost tradeoffs and increase fiber utilization efficiency and reach. This is sufficient to support a modulation order of 256-QAM, but it limits the capacity of the HFC link to between 5 and 6.4 Gbps. Improved noise budgets by using the Node QAM would allow the support of 1024-QAM modulation, over a bandwidth range up to 1800 MHz, resulting in throughput capacity of 15 Gbps, nearly triple the current capacity.

Digital baseband transmission would unlock practically unlimited capacity in the fiber links to the node. With the proximity of the node to the furthest service user, especially in PCN networks, distributed fiber to the home (FTTH) solutions like Next-Gen RFoG and xPON can be extended from the nodes selectively, based on the demand and opportunities.

Simplification of RF Combining Network

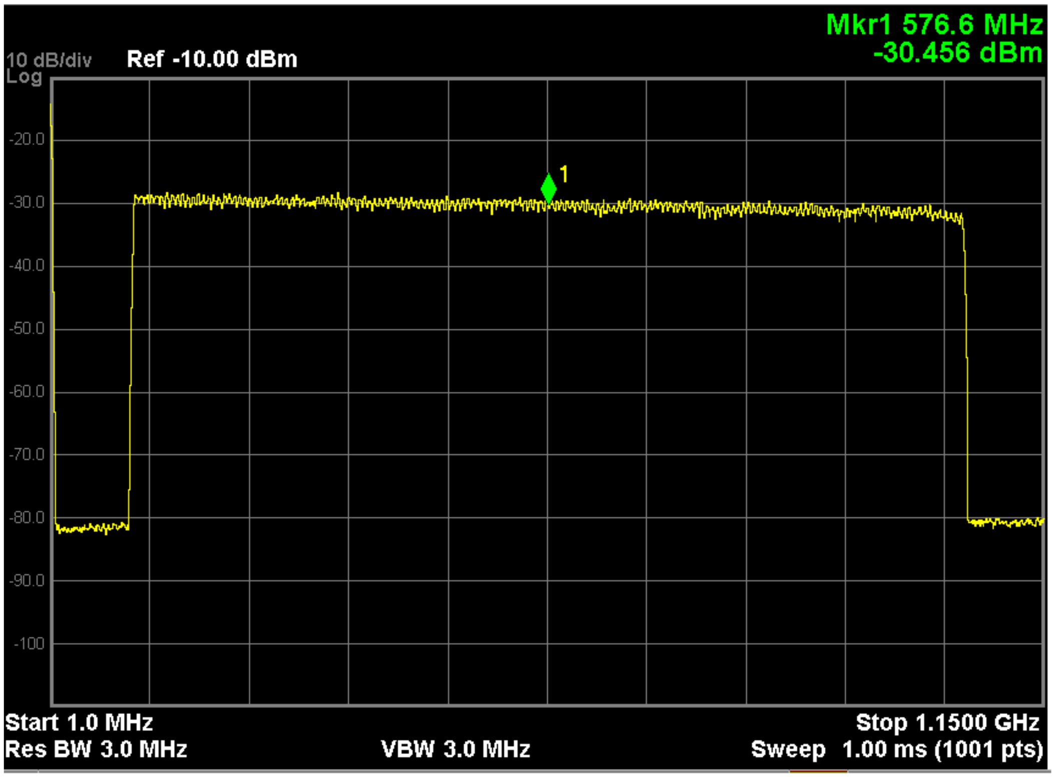
Generating QAM signals in the node allows those QAM signals to bypass the RF combining network. Node QAM output signals can be combined *at the node* with traditionally carried HFC signals in a single stage. New narrowcast QAM signals can be added at the node as needed, with no impact on either the existing RF combining network or the HFC plant alignment.

Besides removing the complexity of recalculating the headend combining plant each time new RF ports are added, it avoids both the signal and power losses associated with combining, splitting, and directional coupling, as well as the power, cooling burden, and significant space inefficiencies. Many of these advantages are delivered by the CableLabs Converged Cable Access Platform (CCAP)[2] architecture, as described later. A distributed architecture goes a step further by allowing legacy signals to be combined in a single passive combining stage at the node.

RF Signal Quality and Node Alignment

When QAM signals are generated in the node (Figure 1), with given output levels and the same or better output signal quality as headend-generated QAM signals, the resulting RF signal in the node is much cleaner. This is because it bypasses the signal losses, noise, attenuation, and distortions that are typically introduced in the RF combining network and the amplitude-modulated optical links to the node.

Operationally, it reduces the amount of RF aligning needed at the node; output power and tilt are generated exactly according to configured specifications, defined by the operator. The signal is not subject to any of the traditional distortions. The impairment contribution of combining network and analog (RF) optical links to nodes is eliminated, with the benefit of unlocking coaxial plant capacity as described above. This allows a 43+ dB MER (see Figures 1 and 2) at the node and gives the operator more options in the coaxial portion in terms of loss budget/coverage and, most importantly, bandwidth. In certain conditions, it makes higher-order modulation rates possible as well, resulting in better spectral efficiency.

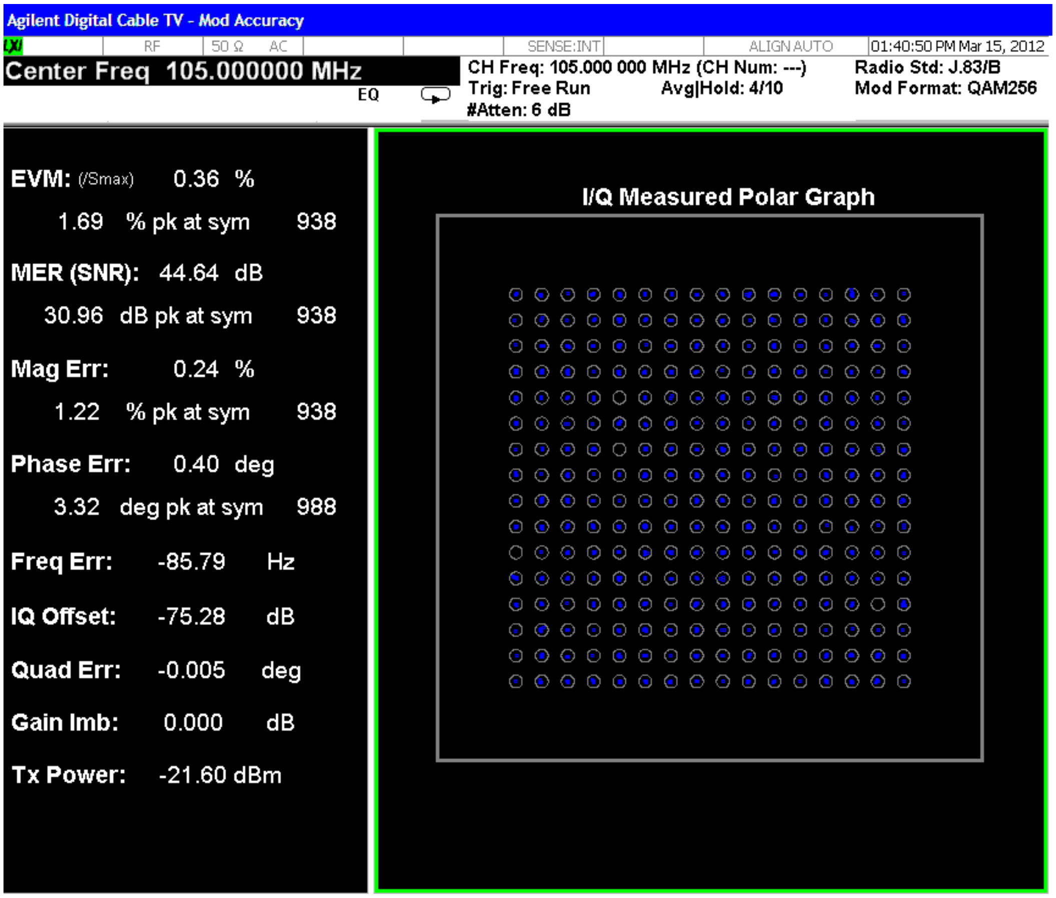


Mkr1 576.6 MHz
-30.456 dBm

10 dB/div
Log
Ref -10.00 dBm

Start 1.0 MHz
Res BW 3.0 MHz
VBW 3.0 MHz
Stop 1.1500 GHz
Sweep 1.00 ms (1001 pts)

Figure 1: 158 Node QAM channels



Agilent Digital Cable TV - Mod Accuracy

01:40:50 PM Mar 15, 2012

Center Freq 105.000000 MHz

CH Freq: 105.000 000 MHz (CH Num: ---)
Trig: Free Run
Avg|Hold: 4/10
#Atten: 6 dB
Radio Std: J.83/B
Mod Format: QAM256

```
EVM: (/Smax)    0.36  %
      1.69  % pk at sym     938
MER (SNR):  44.64  dB
     30.96  dB pk at sym    938
Mag Err:        0.24  %
      1.22  % pk at sym     938
Phase Err:      0.40  deg
      3.32  deg pk at sym   988
Freq Err:     -85.79     Hz
IQ Offset:    -75.28     dB
Quad Err:     -0.005     deg
Gain Imb:      0.000     dB
Tx Power:     -21.60 dBm
```

Figure 2: Node QAM increases RF loss budget or bandwidth capacity.

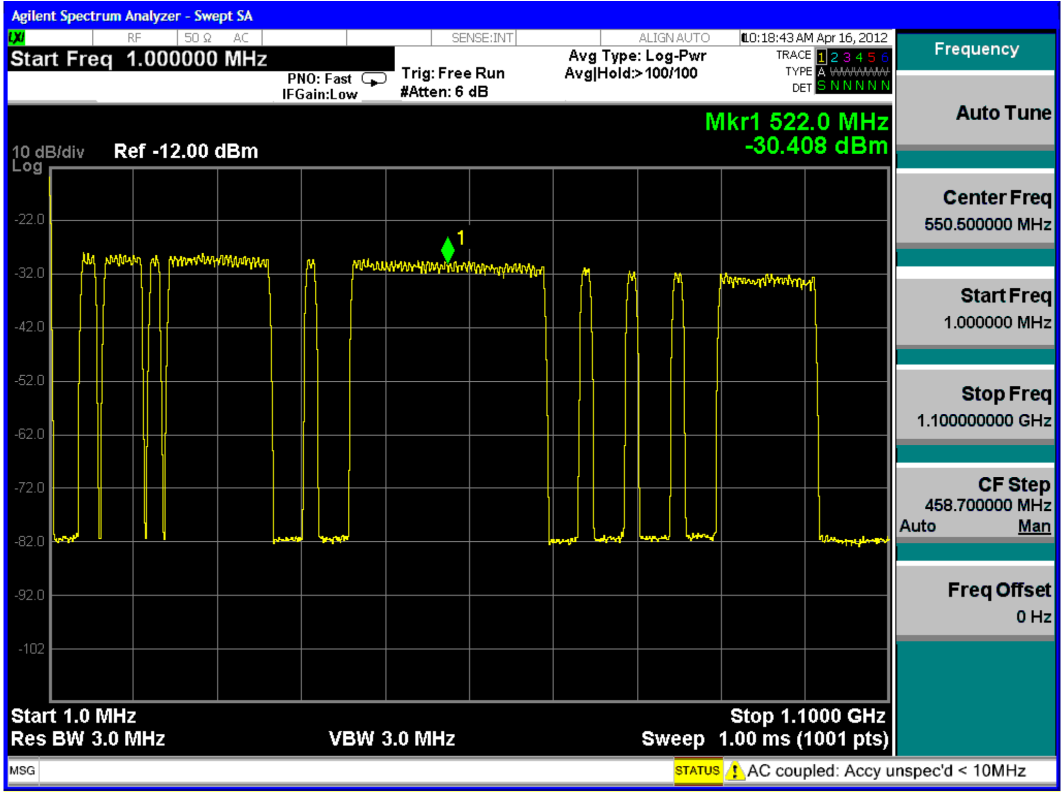
Service Flexibility

An important side effect of the Node QAM is that the optical network feeding it is a de-facto extension of the headend IP network, with access to all of the system’s digital content — broadcast, narrowcast, unicast, and data.

The Node QAM itself is agnostic to the digital payload; it simply modulates the MPEG2 formatted transport streams that are delivered over the optical interface. The payload carried within the transport stream could be a groomed and re-quantized statistical multiplex; it could be an encrypted variable bit rate broadcast multiplex; it could be a simple multiplex of fixed-rate VoD streams; or it could be a DOCSIS M-CMTS-compliant data stream.

The contents of the transport streams are dependent only on the capabilities and sophistication of the headend service manager(s) and resource manager(s), and switched IP connectivity. Artificial service group constraints imposed by the hard-wired RF combining network are removed, leaving only a general-purpose pool of QAM signals to feed the population of subscribers attached to each node or node segment.

An enhancement enabled by the generation of QAM signals in the node from native IP input is the ability to selectively reserve local bands of frequency for other modulation and encoding schemes as well. See Figure 3.


Agilent Spectrum Analyzer - Swept SA

RF 50 Ω AC SENSE:INT ALIGN AUTO 10:18:43 AM Apr 16, 2012

Start Freq 1.000000 MHz

PNO: Fast IFGain:Low

Trig: Free Run #Atten: 6 dB

Avg Type: Log-Pwr Avg|Hold:> 100/100

TRACE 1 2 3 4 5 6 TYPE A WWWWW DET S N N N N N

Mkr1 522.0 MHz -30.408 dBm

10 dB/div Log Ref -12.00 dBm

Start 1.0 MHz Stop 1.1000 GHz

Res BW 3.0 MHz VBW 3.0 MHz Sweep 1.00 ms (1001 pts)

Center Freq 550.500000 MHz

Start Freq 1.000000 MHz

Stop Freq 1.100000000 GHz

CF Step 458.700000 MHz Auto Man

Freq Offset 0 Hz

MSG STATUS AC coupled: Accy unspec'd < 10MHz

Figure 3: Spectrum Allocation Agility. Individual QAM signals can be turned on or off.

Some examples of practical applications include:

- Customized broadcast lineups. Certain niche customers, such as hotels, apartment complexes, hospitals, and campuses can receive their own broadcast lineups, created on the fly, without affecting the existing RF combining network.

- Uneven service usage. Usage of individual types of narrowcast and unicast services may vary unpredictably from node to node. Node QAMs with headend service switching allows each node to have a different service mix, without having to pre-allocate resources.

- Dynamic service allocation. Service usage may also vary within a single node, based on time of day or season. For example, a suburban node might experience heavy VoD usage during the day due to toddler addictions to children's programming, but switch to heavy internet usage late at night when parents use Netflix. With the Node QAM, a single pool of QAM signals can feed all services, without having to provision under-utilized service silos.

- Mixed services within a single channel. With sufficient sophistication from the headend multiplexers and resource managers, the Node QAM can deliver any mix of QAM services — broadcast, narrowcast, CMTS, VBR, CBR in a single channel, giving the operator complete flexibility.

Environmental

While modern headend QAM modulators are an order of magnitude more energy-efficient than earlier incarnations, and two orders of magnitude more compact, the addition of large quantities of new QAM channels via traditional methods creates a significant impact on the headend, in two ways. Headend Edge QAMs create a direct impact by their intrinsic consumption of power, rack space, and cooling mechanisms. They also have an indirect impact, due to the rack space occupied by the combining network; the power loss due to combining, splitting, and directional coupling of service groups, as well as the power consumption of intermediate amplification stages; and the power burden of heating, ventilation, and air conditioning (HVAC).

By moving QAM modulation to the node, not only are power and rack space requirements distributed, but overall per-QAM power and space consumption are reduced due to the fact that lower output levels are needed to drive the existing node RF amplification modules. This helps the Node QAM to live within the design constraints imposed by the node housing, including the use of passive cooling instead of fans. Node QAMs also eliminate the Edge QAMs' impact on the headend HVAC system.

In addition, by bypassing the RF combiner network at the headend, Node QAMs avoid wasting the signal power maintained by the RF combiner network's amplification stages, which end up being discarded when the signal is carried in its baseband digital format. Furthermore, power and space requirements are reduced when optical analog (RF) transmitters are replaced by low-power optical digital baseband transceivers.

These Node QAM benefits mesh well with the fundamental goals of the CCAP architecture, with the added advantages that Node QAM leverages digital optics, and that these benefits accrue on a node-by-node basis, allowing both small and large operators to migrate gracefully to CCAP.

CCAP

CableLabs' CCAP architecture is a bold step in addressing many of the challenges related to the growth of narrowcast services. It leverages heavily the existing body of Data-Over-Cable Service Interface Specifications (DOCSIS) with the goals of increasing the flexibility of QAM usage and configuration; simplifying the RF combiner network; possibly adding content scrambling; creating a transport-agnostic management paradigm to accommodate native support of Ethernet Passive Optical Network (EPON) and other access technologies; improving environmental and operational efficiencies; and unifying headend configuration and management capabilities. CCAP includes a new Operations Support System Interface (OSSI)[3] specification and also takes particular care to ensure compatibility with existing DOCSIS resource management and service management and configuration specifications, in order to facilitate the migration from current CMTS/Edge QAM infrastructure.

CCAP Reference Architectures

CCAP unifies digital video and high-speed internet delivery infrastructures under a common functional umbrella, allowing a CCAP device to be operated as a digital video solution, a data delivery solution (both CMTS and M-CMTS), a Universal Edge QAM, or any combination. Each of the CCAP reference architectures (Video, Data, and Modular Headend) describe physical and functional interfaces to content on the "network" side, operational and support systems within the headend, and the HFC/PON delivery network terminating in various devices at the subscriber premises. Ancillary service and resource managers are allowed to exist both within and externally to a CCAP device.

CCAP OSSI

The lynchpin of the CCAP architecture is the CCAP OSSI, which defines a converged object model for dynamic configuration, management, and monitoring of both video and data/CMTS functions, but also makes provision for vendors to innovate within the framework. By creating a unified standards-based operational front-end to the video and data delivery infrastructure, CCAP OSSI provides a solid foundation for the headend's metamorphosis from a collection of separately managed service silos into an efficient service delivery "cloud".

CCAP and Node QAM

In the CCAP video and data reference architectures, the CCAP interface on the subscriber side is the HFC network. Traditionally, that interface exists within the headend. However, there is nothing inherent about the provisioning and management of QAM signals that *requires* the QAM modulators to be in the headend. Extending the logical boundary of the headend out to the node and minimizing the analog portion of the HFC remains consistent with the goals and specifications of CCAP.

NODE QAM EVOLUTION

Initial Architecture

The initial configuration of the Node QAM topology can be envisioned as one presented in Figure 4. In this configuration, analog and operator selected QAM broadcast channels (*e.g.*, from a different location than the remaining QAM channels) are transported to the node in a traditional fashion but without the burden of combining with the remaining QAM channels in the headend/hub. The number of QAM channels originating in the Node QAM can be adjusted dynamically by the operator.

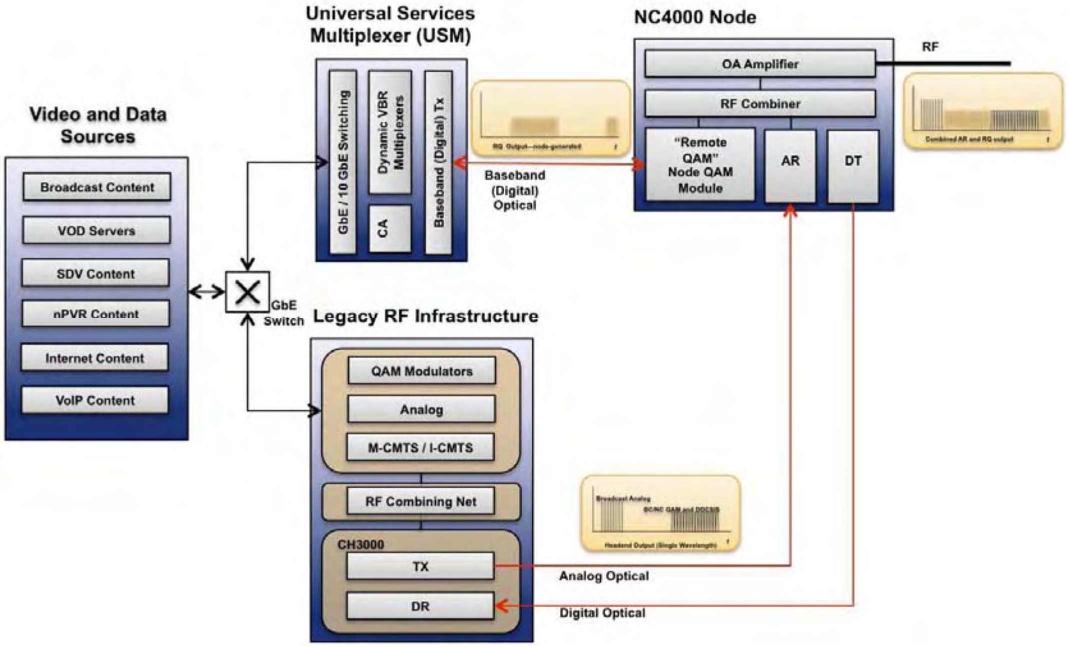


Figure 4 Node QAM Initial Implementation

Conversion to Complete Digital Baseband Node Transport

The next incarnation of the distributed architecture is presented in Figure 5. All analog channels and maintenance carriers are digitized in the headend and transported over the same transport (capacity allowing) to the node where they are frequency-processed and converted back to analog channels at their respective frequencies on coaxial plant. Some additional carriers (*e.g.*, ALC pilot signals) are synthesized in the Node QAM module.

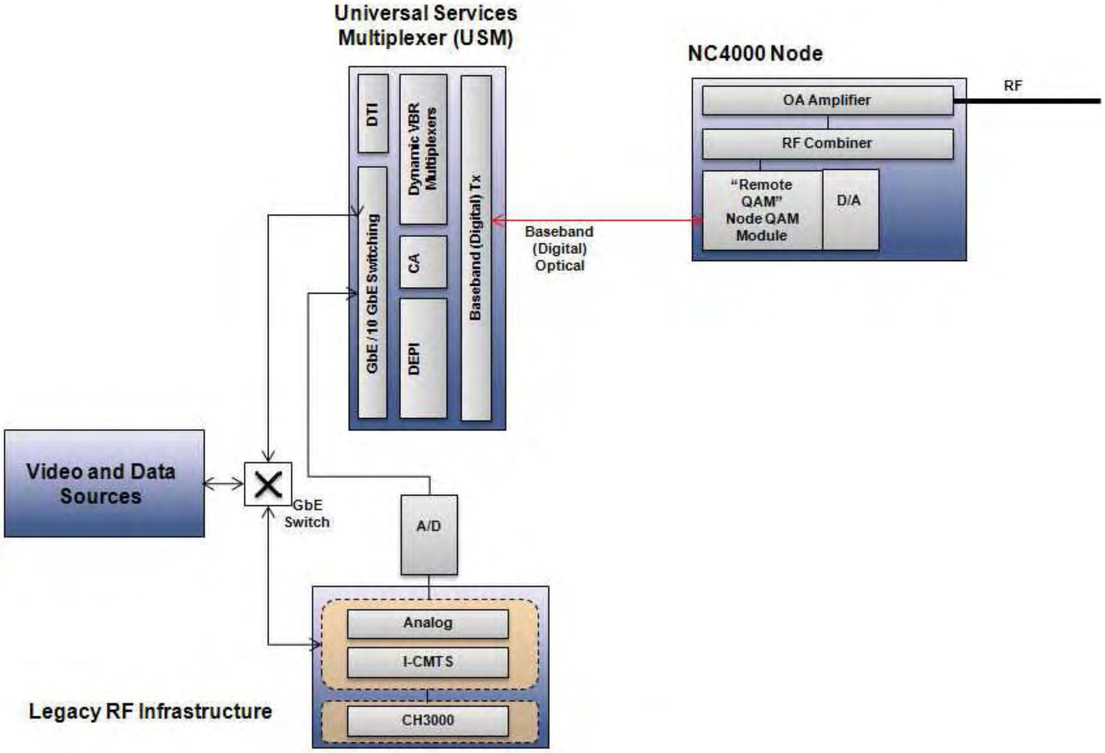


Figure 5: Node QAM Next-Generation

The reverse channel(s) from the node to the headend can also be converted to baseband digital optics, resulting in similar benefits. Options include traditional digital return (digitization of the return spectrum at the node), developing a node-based CMTS (or node-based DOCSIS burst receivers), or even next-generation native IP-over-coax technologies.

A related enhancement arising from the Node QAM's dynamic frequency agility is the ability to support flexible, remotely configurable frequency splits or capacity allocation between downstream and upstream communication, either using frequency division duplex (FDD) or time division duplex (TDD) transmission. This would enable full flexibility and adaptability to downstream and upstream traffic patterns and capacity/service demands.

Future Enhancements

The Node QAM is an ideal platform to be modified to support other modulation schemes for next-generation transport mechanisms, such as EPON Protocol over Coax (EPoC). Implementing EPoC in the node allows significant reach expansion, preserving and facilitating headend and hub consolidation without deploying additional signal conditioners or RF-baseband-RF repeaters with their additional cost, power consumption, added operational complexity of provisioning and additional space/housing requirements in the field or hubs.

OTHER ELEMENTS OF DISTRIBUTED ARCHITECTURE

Node PON

A distributed node-based EPON architecture shares the Node QAM architectural advantages. Node PON modules allow for selective fiber placement from the node for commercial services in node areas where construction costs and effort are limited to fiber extension from the node. In PCN architecture, this is usually below 1 km, and mostly below 300 m if the node is placed strategically. In conjunction with DOCSIS Provisioning of EPON (DPoE) and CCAP, Node PON can address the needs of fast deployment of dedicated fiber links to selected high capacity demand users.

Next-Generation RFoG[4]

In situations where fiber exists all the way to the subscriber, RF over Glass (RFoG) in a distributed architecture has the potential, with minor changes, to exceed the throughput of 10G PON/EPON, without the complexity of adding a PON overlay. This allows for seamless expansion of fiber from RF optical nodes to residences without replacing the distributed architecture node modules. Taking fiber from the node all the way to the subscriber with a FTTH network would allow for additional capacity enhancement beyond 15 Gbps downstream and 1 Gbps upstream facilitated by distributed coaxial architecture, especially with PCN and residential gateways deployed. With RFoG in a distributed architecture, 20+ Gbps downstream and 3 Gbps upstream is achievable today without PON overlay.

SUMMARY

No-one knows precisely what the future will bring but it is clear that subscriber-side demand for IP-delivered multimedia continues to grow as “smart” home and mobile electronic devices proliferate. The cable industry is blessed with the most extensive and highest bandwidth conduit to that last-mile “IP cloud”. At the same time, cable headends have largely already made the transition to IP-based distribution. Moving the native baseband IP-to-RF transition point from the headend to the node brings the convergence of IP headend and IP home one step closer.

As discussed in this paper, there are many advantages to extending the digital headend domain as far into the network as possible, in terms of performance, resource utilization, operational simplicity, and service flexibility. There are many paths for the evolution to digital HFC: the Institute of Electrical and Electronics Engineers (IEEE) is proposing a new physical layer standard called EPON-Protocol-over-Coax (EPoC) to deliver IP traffic natively at 10 Gbps over last-mile HFC; fiber vendors continue to innovate on bringing fiber to the home; new silicon may enable conversion of large bands of RF spectrum at the headend into digital bitstreams that can be converted back to analog at the node. By bringing IP closer to the edge, the Node QAM helps pave the way to a distributed headend and digital HFC.

ABBREVIATIONS AND ACRONYMS

10G-EPON	IEEE 802.3 Ethernet PON standard with 10 Gbps throughput
3DTV	3D Television
4WM	Four Wave Mixing
ALC	Automatic Level Control
BER	Bit Error Rate
CBR	Constant Bit Rate
CCAP	CableLabs® Converged Cable Access Platform
CMTS	Cable Modem Termination System
CNR	Carrier-to-Noise Ratio
DOCSIS®	Data over Cable Service Interface Specification
DPoE™	DOCSIS Provisioning of EPON
EDFA	Erbium-doped Fiber Amplifier
FDD	Frequency Division Duplex
FTTC	Fiber to the Curb
FTTH	Fiber to the Home
EPoC	EPON Protocol over Coax
EPON	IEEE 802.3 Ethernet PON standard with 1 Gbps throughput, a.k.a. 1G-EPON, G-EPON or GEPON
Gbps	Gigabits per second
HDTV	High Definition Television
HFC	Hybrid Fiber Coaxial
HVAC	Heating, Ventilation and Air Conditioning
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPTV	IP Television
M-CMTS	Modular Cable Modem Termination System
Mbps	Megabits per second
MER	Modulation Error Ratio
MPEG2	Motion Picture Experts Group 2 standard
MPEG2-TS	MPEG2-Transport Stream
nDVR	Network-based Digital Video Recording
NTSC	National Television System Committee
OBI	Optical Beat Interference
OMI	Optical Modulation Index
OSSI	Operations Support System Interface
PAL	Phase Alternating Line
PCN	Passive Coaxial Networks
PON	Passive Optical Network
QAM	Quadrature Amplitude Modulation
RF	Radio Frequency
RFoG	Radio Frequency over Glass
SDV	Switched Digital Video
SFP	Small Form-factor Pluggable
SRS	Stimulated Raman Scattering
TDD	Time Division Duplex
VBR	Variable Bit Rate
VoD	Video on Demand
XFP	10 Gigabit Small Form-factor Pluggable
XG-PON	ITU-T's broadband transmission standard with 10 Gbps throughput
XPM	Cross Phase Modulation
*x*PON	any of a family of passive optical network standards (*e.g.*, GPON, GEPON, 10G PON (BPON, GEPON or GPON)

[1] ITU-T J.83 Digital multi-programme systems for television, sound and data services for cable distribution. April 1997.

[2] TR-CCAP-V02-110614. CCAP Architecture Technical Report. June 2011.

[3] CM-SP-CCAP-OSSI-I02-120329 . Converged Cable Access Platform Operations Support System Interface Specification. March 2012.

[4] O. Sniezko. *RFoG: Overcoming the Forward and Reverse Capacity Constraints.* NCTA Spring Technical Forum 2011.

EVOLVING THE HOME ROUTER TO AN APPLICATIONS DELIVERY GATEWAY

Joe Trujillo and Chris Kohler
Motorola Mobility, Inc

Abstract

The home router has become a power house of performance, enabling a dizzying number of devices in the home to communicate with each other and the internet at ever growing bandwidth and capacity. With all this impressive brawn, it is easy to overlook the router's potential for brains.

The home router is an always-on device that is completely intimate to the physical and logical connectivity between devices on the home network and their connections to the internet. That intimacy makes the home router uniquely positioned to host a variety of applications.

In this paper, the authors discuss some of the applications that can supply a brain to accompany the brawn for next generation routers. Some example applications discussed relate to Machine-to-Machine (M2M) communication for home control and security, Personal Content Management, and Advanced Home Network Management. While this list is not exhaustive, it gives a fair idea about the possibilities and opportunities for the Service Provider to move up the value chain, while continuing to delight the customer.

INTRODUCTION

Until now, the nearly complete focus of the home router's evolution has been on improvements in the performance of IP connectivity, while the router's own participation in using that connectivity has been suppressed, maybe even discouraged. One could say that the focus has been on brawn - faster speeds - over brains. The time has come to turn some of that focus towards developing gateway intelligence by way of hosted applications for which the home router is uniquely positioned and qualified.

WHAT KIND OF APPLICATIONS AND WHAT MAKES THE ROUTER QUALIFIED?

A home router is not suitable for every kind of application. It has no keyboard, no joystick, no screen nor speakers of its own. Hosting games, word processors or corporate payroll applications makes no sense at all. The best applications for it to host are those that leverage and extend its innate properties. Simply put, those key properties are 1) It is always on; 2) It is connected to the internet; 3) It is intimately connected to every IP device in the home. Taking the concept one step further, an integrated home router with built in broadband access, such as DOCSIS® 3.0, xPON or bonded DSL, would expand the reach of the hosted applications into the WAN (see Figure 1).

The always-on nature and therefore its ability to continuously access both the internet and devices on the home LAN make the integrated home router the perfect place to host applications that need to provide one or more of the following properties [brain functionalities]:

- Anytime or always-on availability [always thinking]
- On demand or near real time access/control to devices on the LAN [gross motor skills]
- On demand or near real time access to devices on the *once-removed** network [fine motor skills]

* The "once-removed" network is the collection of devices in the home/office that are not necessarily directly IP connected, but can be controlled and/or monitored by other devices that are in turn IP connected. Examples of some technologies that can act in the once-removed network are Bluetooth® (1), ZigBee® (2) and Z-wave® (3)

- A high degree of local abstraction to hide, when necessary, the complexity of the local network or the once-removed network. This allows for more uniform and less complicated communication protocols between the Cloud or other internet devices on the LAN or the once-removed network [can process the environment to abstract and simplify clutter]
- A high degree of local autonomy and in-depth local knowledge to discover WAN, LAN and once-removed topology [is self aware and can communicate it's condition]
- A high degree of local autonomy to help in scaling or offloading from the Cloud or management system [thinks for itself, but is a member of a community]

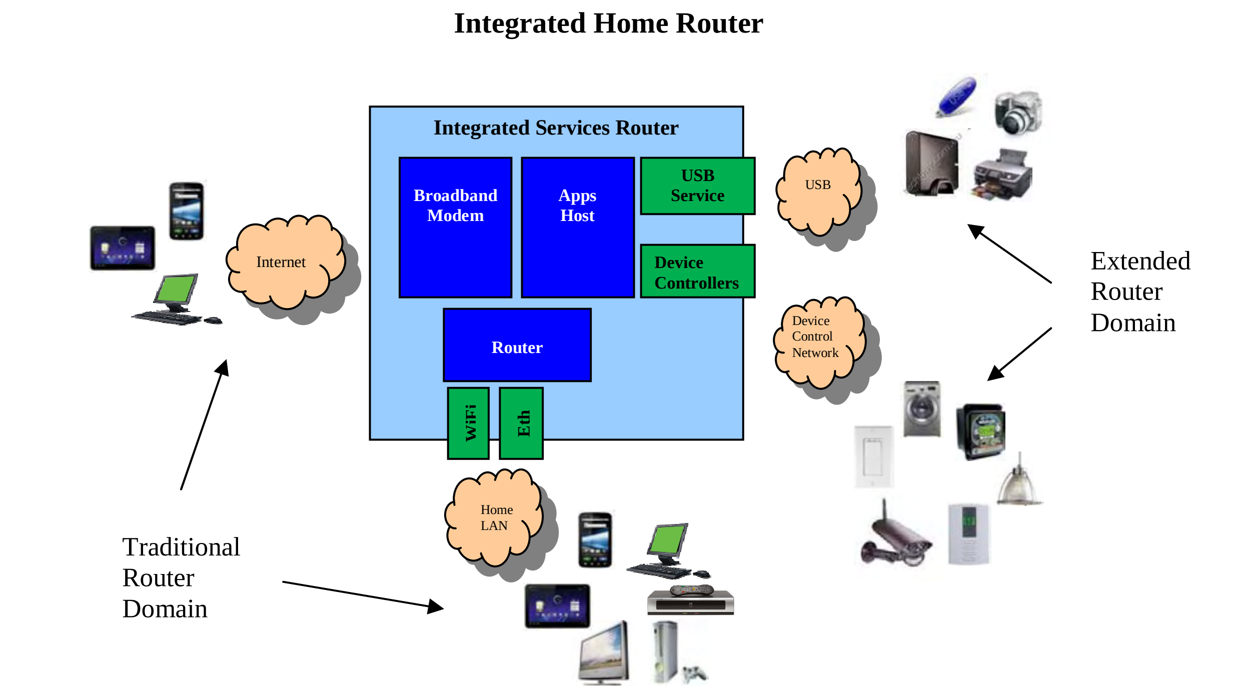


Figure 1

M2M CONTROL POINT APPLICATIONS

Some of the most interesting examples of applications which are ideal for integration into the home router are Machine-to-Machine (M2M) control points. Of course the concept of one device "remote-controlling" another device is not new to the internet. In the most basic sense, M2M is one smart device talking to another smart device via a communication network (4) . In industrial applications, such as on a complex factory floor, M2M has had natural and wide adoption, albeit for a closed environment and a non-consumer market. For the home/consumer market new possibilities are just beginning to open up.

There are several emerging genres of M2M applications for the home. Each of these genres is best serviced from a Cloud portal vantage that can homogenize the presentation to the end users, simplify presence and discovery of devices from across the internet, and be the integration and launching point for service extensions or other services supplied by the service provider. That said, hosting the control point portion of the application in a home router with its integrated WAN or broadband access and direct connectivity to the LAN and once-removed network provides the best solution for the service provider to deliver, control and manage the entire experience.

Home Automation

Examples of features in this genre include the ability to remotely turn on your sprinklers, turn off your air conditioner, turn on or off lights or even unlatch the dog door from a smart phone, computer or hosted scheduler. These are convenience features once only available to high end homes via highly custom installations.

Home Security

Features in this genre would include remote enable/disable of the alarm system, monitor/control of individual sensors (window, door, and motion), and control of camera pointing, scanning and live/recorded viewing of video feeds.

Home Energy Management

Features in this genre would include remote monitoring of total home power usage and/or usage on a per-device basis. Historical analysis of telemetry can be used to detect and correct consumption patterns. Available interfaces to the utility company's portal can be utilized to create useful correlations and validations of power consumption, including actual costs incurred due to specific power consuming devices such as air conditioning, clothes driers and entertainment clusters. Triggers could be used to inform the homeowner of a "violation" in progress, such as the drier being turned on during peak usage or peak billing hours. That alert could come, for instance, as an SMS to a cell phone or an alert ring and pop-up on a custom smart phone app.

Senior Care Monitoring

Features in this genre would include monitoring door sensors, motion sensors and pressure sensors to allow passive monitoring of the elderly or infirmed. Cameras could be added for more complete, but more intrusive monitoring. Triggers such as lack of motion for a prolonged period (have they fallen?) or opening of an off limits door (the front door leads to traffic or dangerous stairs) could alert a care giver and prompt a phone call, visit or emergency action.

Advanced Medical Monitoring

Features in this genre would include gathering telemetry from scales or other medical equipment such as heart rate monitors and glucose meters. For advanced medical monitoring, security, senior care and home automation could be combined. For critical care, perhaps FDA certified/approved devices for M2M applications have a market place.

It is important to note that these home oriented M2M features are not just about one-way remote control services into the home. Their best utilization is when a diversity of machines takes advantage of their local capabilities to build something more useful.

Here's an example of a fully automated M2M scenario that one could envision being easily "programmed" by an end user from a properly equipped smart phone. Using the phone's GPS, the phone can detect when it has moved one mile away from home. Using this event as a trigger, the phone can interact with the M2M network (via the Cloud to the home M2M control point) and cause the home doors to lock, the home alarm system to enable, verify and close the garage door, send an SMS or email from the phone to the elderly care service provider that the person has left the house and even pop open the doggy door to prevent an embarrassing accident.

Some major operators have already entered the home M2M market place and are deploying solutions. These kinds of engagements are expected to grow and help drive technologies and monetized deployments at an accelerated rate. Industry initiatives, such as the TIA's TR-50 (5) and ETSI M2M (6) promise to further standardize the M2M ecosystem and bring a plethora of interoperable service opportunities to the telecommunications industry.

PERSONAL CONTENT APPLICATIONS

Another natural set of applications for an always-on home router have to do with file storage and media access. Network attached storage (NAS) systems for the home are not new, but their presence in the marketplace appears to be growing. Digital photo, music and movie collections grow rapidly, but are almost always spread out over many devices (phone, tablets, cameras, computers). The desire to ease the ability to collect files from these devices to a central location is becoming more urgent.

Collecting the media (copying) to a central location provides a back up to the phone or camera against disaster and provides a place to store when the internal storage of the device becomes full. When a consumer consolidates media, they typically choose to use a home computer's hard drive. This approach is fine for back up and overflow storage, however, it can have some serious limitations.

Setting up an environment where other devices on the home LAN can access that computer's hard drive is complicated and not guaranteed to interoperate across varying devices' operating systems. Remote access from the internet to the computer's hard drive is not possible without special software on the computer. Maybe most limiting is that a computer can be turned off or in the case of a lap top, not even be at home. An always-on integrated home router with attached storage capabilities provides a platform to overcome these limitations.

There are several NAS devices in the market today that can be plugged into the Ethernet port of an existing home router. With enough patience to configure the NAS and the IT properties of the router, many solid features become available to the end user. These features typically include: SAMBA (LAN) access to files available on the NAS; DLNA-Server streaming of media files stored on the NAS to the growing list of compatible devices on the home LAN, including game consoles, MAC and Microsoft OS computers, Wi-Fi™ enabled TVs and Blu-ray Disc™ players; and remote access to files on the NAS from the internet. Remote access capabilities can be extended to social media and file sharing features, with mailing lists and automated posts to social media outlets.

All these features can be supported with a NAS application integrated in a home router. Several additional benefits over a standalone NAS are available if the Router/NAS combination also contains an integrated broadband modem.

Automated Configuration

Since the NAS, router and broadband access are integrated into a single box the configuration is automated. The user doesn't need to know how to configure the router to grant the NAS access, configure DHCP to get it on the network, or assign ports and port forwarding rules to allow internet access.

Advanced Management

It was noted above how an integrated device can automate the configuration tasks. In a service provider deployment there are additional advantages in the ability to manage and monitor the modem, router and integrated NAS as a single entry. A standard retail standalone NAS has no remote management capacities, such as TR-69 or SNMP. A full integration eliminates this problem, enabling the operator to have a much better position to manage a deployment. The combination of automated configuration and advanced management can be a great aid in customer satisfaction and customer loyalty.

Hardware Cost

The cost savings to the operator or as passed on to the end user of a consolidated box could be significant. The cost of buying separate modem, home router and NAS devices can stack up as compared to buying an all-in-one integrated router/NAS device.

Converged Commercial Media Routers

The advantages above will become even more pronounced as the traditional video set-top box continues its evolution towards the IP video gateway. The need to distribute live, on-demand or recoded video to devices on the home LAN will magnify the need for an integrated home router. SOCs which enable IP video distribution capabilities inside an integrated home router will start to appear in the market place in 2012.

ADVANCED HOME NETWORK DISCOVERY AND MANAGEMENT APPLICIATIONS

As stated at the beginning of this paper, until now the focus of the home router's evolution has been on improvements in the performance of IP connectivity. This performance increase is the great enabler of our time. The importance and continued evolution of throughput performance can't be overlooked and must continue for the foreseeable future. However, with all these improvements comes a drawback that must be overcome – high complexity.

Year over year the worry has been stated that lack of bandwidth would cripple quality of service (QOS). There have been numerous strategies to head this crisis off with advanced QOS methodologies, only to find that timely, cost effective technology advances in performance bail us out. It seems that a lack of bandwidth may not be the killer of QOS. The pipes keep getting bigger, symbol rates denser, spectrum more available and diversity transmission techniques ever more standard. However, it may be the complexity and digital clutter associated with this level of improved performance which could be the killer of QOS.

Bonded DOCSIS® 3.0, bonded DSL, 3G, 4G, Gigabit Ethernet with more ports, MoCA®, HomePNA®, multiple SSIDs per multiple Wi-Fi radios, HomePlug®, L2 tunnels, VLANs, VPNs, dual homed WAN - the list seems endless. The technological complexities and home-by-home variations of devices and interfaces have exploded. A typical home is starting to look like an enterprise deployment. But unlike an enterprise, every household cannot afford its own IT department. Compound this with the fact that traditional TR-69, SNMP and other call center techniques are insufficient to scale to the situation without some paradigm shifts. For the most part, current management systems are set up to query the discrete values of pre-known parameters internal to the router's configuration. These techniques are almost blind to the fluid nature of the devices on the home LAN.

A solution to solve this scalability and variability problem is to put much more intelligence and autonomy in the home router. This locally hosted application can analyze the network, detect issues and alert the user or customer care agent of a problem and where to fix it. Better yet, take this local intelligence to the next step for analyzing trends and alert and/or correct an impending problem before it becomes service affecting.

Network Discovery

Keeping track of what devices are on the home LAN can be a challenge. IP enabled computers, tablets, phones, games, set-tops, TVs, Blu-rays, printers, file/media servers and many other devices are popular in the home. How does an operator, customer service agent or even the home user know what devices are connected right now and what the expected properties these devices have so they can help setup or debug the home network? Current TR-69 or SNMP techniques can query some standardized MIBs to get some modem, DCHP, Wi-Fi information and perhaps a few more general router stats and try to interpolate a bigger picture. This can take many queries and still leave the agent without critical information.

Of course the integrated home router is the perfect location for hosting a Network Discovery application. It intrinsically has access to many pieces of information such as DHCP lease table and switch/Wi-Fi learning tables. It can ARP scan for devices that may have statically joined a subnet. Further probes and traffic monitors can discover UPnP devices and their capabilities and probe local IP devices for HTTP Web page capability. This gathered data can be used to create a small database representing the discovered nodes on the home LAN, how they are connected and most important, useful information on each device.

This database is easily exposed for use on the router's local UI to draw a network map that can be drilled down with mouse clicks or as a file which is available to a management system for it to draw the map for a customer service agent. The management system application that uses the topology database can then further augment diagnostics and corrective action by using traditional SNMP or TR-69 management objects.

Network Histogram

The Network Discovery application embedded in the router can automatically refresh the topology database at regular intervals. Changes in targeted parameters from a baseline can be recorded at regular intervals. With this method the database then becomes a histogram that can be useful in capturing variations and instabilities in the network. For instance, it could see that a fixed position Wi-Fi device intermittently drops on and off the network. Imagine the frustration saved by the customer care agent who can actually react with more than just sympathy to a customer saying “Well, it was happening this morning before I called!”

Trend Analysis and Alert Triggering

This application realm dives deeper and takes a running statistical look at the core access technology interfaces. Using various interfaces’ instantaneous measurements and counters available on the integrated router, the application can collect, record, average, filter and analyze trends that can be used to take preventive action before an outage can occur.

Let’s take a DOCSIS® 3.0 bonded down-stream connection as an example. In a typical DOCSIS® 3.0 modem the downstream could consist of data distributed on 8 individual channels (QAM modulated data on 8 frequencies) that are captured and re-sequenced in the modem to create a 300Mbps connection. Each channel is subject to its own analog variations in signal quality due to minor Tx power fluctuations and interferences. In nominal operating conditions, digital receiver techniques are transparent to this “noise”. However, if one or more channels degrade such that transmission errors become significant, then performance and connectivity will quickly degrade and perhaps result in an outage and a truck roll. Having a remote management system poll many measurements 24/7 across 8 channels is neither realistic nor scalable across a large population of devices. Furthermore, any single sample measurement has almost no meaning as far as “good” or “bad”.

A statistical application local to the integrated router could monitor a history of vital signs like raw Frame Error Count (FEC), corrected errors and downstream power. For example, on a per channel basis, a rolling database window could record averaged samples over a statistically significant period of time and show if the frame error count, translated to a frame error rate, is trending up indicating a problem on any channel. It could be useful to graph the table to show this trend visually. Better yet, the application can track the trend itself and on a threshold, send an alert (SNMP trap, TR-69 inform) informing the management system or customer care proactively. This technique could be extended to Wi-Fi, Ethernet, MoCA®, HomePNA® and other interface types in the system.

SUMMARY

We’ve stated that the integrated router is the best choice for hosting applications needing the properties described in the opening paragraphs. The always-on nature guarantees access *when* it’s needed. Its connectivity to the internet guarantees access from *where* it’s needed. And its intimacy to all devices on the extended home network guarantee access to *what* is needed - simply, conveniently and at high quality. The example applications outlined reinforce this point of view.

M2M applications demand all the brain qualities the router can provide. They need to always be on and ready, connected through internet and provide on demand access to devices on the extended home network. These applications need a high degree of local abstraction and autonomy to hide complexities from the user experience and scale to the Cloud.

Personal Content applications are more valuable when the content can be accessed and exchanged from anywhere and anytime. The local autonomy and intimacy of the application with the router make configurations automatic and remote management seamless.

Advanced Home Network Discovery and Management applications take great advantage of the intimacy between the router and broadband modem systems, performing continual measurements and diagnostics not available or scalable from traditional management systems alone. This helps ensure the technological complexity of the networking environment doesn't subtract from the reliability and usability of the connection.

This is also a good time to circle back and thank our friend, performance. Thirst for greater performance has driven the silicon industry to higher densities making more processing power available to router applications in the form of faster CPUs and multiple cores. In older generations of silicon the desire may have been there for hosted applications on the router, but the processing platform was not. It's the brawn of the modern integrated router that has made the brain possible.

References

1. **Bluetooth Special Interest Group (SIG).** Specification: Adopted Documents. *Bluetooth Special Interest Group (SIG).* [Online] Bluetooth SIG. www.bluetooth.org/Technical/Specifications/adopted.htm.
2. **ZigBee Alliance.** ZigBee Standards Overview. *ZigBee Alliance.* [Online] ZigBee Alliance, 2012. [Cited:] http://www.zigbee.org/Specifications.aspx.
3. **Z-Wave Alliance.** Z-Wave Products. *Z-Wave.* [Online] Z-Wave Alliance, 2011 . http://www.z-wave.com/modules/Products/.
4. *The Promise of M2M: How Pervasive Connected Machines are Fueling The Next Wireless Revolution.* **Syed Gilani.** 2009, Embedded Systems Magazine - White Paper
5. **Telecommunications Industry Association (TIA) .** TR-50 - SMART DEVICE COMMUNICATIONS. [Online] http://www.tiaonline.org/all-standards/committees/tr-50.
6. Machine to Machine Communications. *ETSI - World Class Standards.* [Online] 2011. http://portal.etsi.org/m2m.

TRANSFORMING CABLE INFRASTRUCTURE INTO A CLOUD ENVIRONMENT

Gerry White
Motorola Mobility Network Infrastructure Solutions

Abstract

The paper outlines a methodical evolution strategy from today's RF centric, headend-based infrastructure to a digital-centric one, taking advantage of mature Internet, cloud computing and data center technologies. It presents a phased approach, identifying incremental steps leading to the ultimate goal of an efficient delivery infrastructure, and most importantly one that is aligned with Internet and data center technologies, and henceforth is able to leverage their continued development. Each transitional step is evaluated in the context of current and expected changes in technology, products and services. For each step the advantages provided are highlighted and potential risks are noted.

A number of practical options to deploy subsets of the phases are provided, depending on the individual circumstances of an operator, such as service needs and timing, network characteristics, and risk tolerance.

INTRODUCTION

One of the most significant developments in service delivery in the last few years has been the evolution of cloud computing and the massive data centers used to deliver it. The combination of high bandwidth network connections together with low cost computing and storage platforms has enabled companies such as Amazon and Google to deliver sophisticated services at very low cost points. To date, cable operators have used some of this technology for services such as TV Everywhere but in general it has been competitors such as over the top (OTT) video providers who have best leveraged the new technology. This paper proposes a way for the cable industry to take better advantage of data center technology and outlines a number of steps to achieve the transition.

EVOLUTION

In order to speculate on the evolution of the cable network infrastructure we need to consider the evolution of the services which it must support. Figure 1 shows these parallel evolutionary paths.

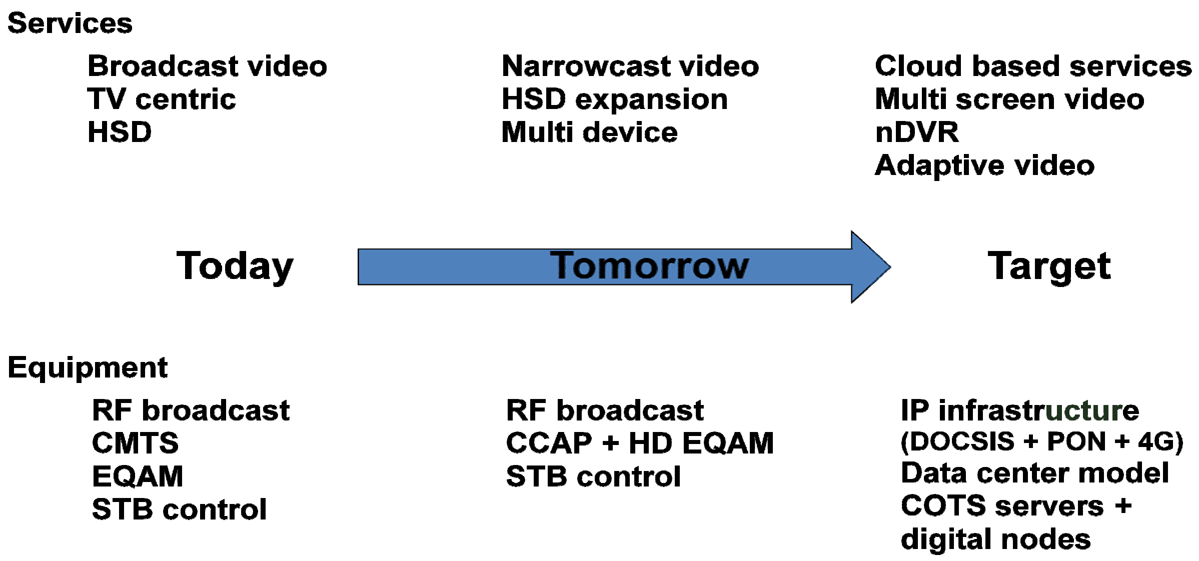


Figure 1: Service and Equipment Evolution

Service Evolution

Current services are heavily focused on delivering linear video programming to a broadcast audience via an STB/TV combination.

In the immediate future we expect to see significant expansion of this service set as narrowcast video services delivering video on demand (VOD) and network based DVR to STB/TV platforms are deployed in parallel with the broadcast service. At the same time high speed data service expansion with compound annual growth rates in the order of 50% is being driven by over the top video services delivered to both TVs and other screens [SAND].

Looking further into the future operators will continue to expand on demand and nDVR services increasingly moving from broadcast to narrowcast services. At the same time competition with OTT video providers will require MSO's to deliver equivalent or better services to multiple types of CPE devices, in the home and on the road. Thus cloud based adaptive video services to multiple devices will become a key component of the service mix.

Infrastructure Evolution

The cable infrastructure must evolve in parallel with the services. Today video services are delivered over an RF broadcast infrastructure, primarily to set top boxes using a proprietary control system. High speed data services are delivered via a parallel CMTS based infrastructure sharing the same physical HFC network as video services but little else.

In the immediate future, as high speed data and narrowcast video expands, existing CMTS and EQAM equipment will be augmented or replaced by higher density platforms to add more narrowcast channels. Systems based on the CCAP specifications will combine CMTS and EQAM functions into a single edge platform but retain the same frequency division multiplexing to share the HFC network and data and video will continue to use independent control systems.

As multi screen video delivery expands the rapid deployment and short lifetimes of the CPE devices will require that service delivery to these devices be based on standard internet protocols with minimal or no changes for the cable infrastructure. The infrastructure must evolve to support IP video delivery to these devices. Thus over time more video will move to an IP delivery mechanism sharing the same resources and technology as high speed data services. This will use existing IP backbone technology for distribution to access networks. The access networks will initially be based on DOCSIS but will include PON and wireless alternatives. This move to a standard IP solution enables the use of standard data center and cloud based services to reduce costs and increase service velocity.

CLOUD SOLUTION

Currently the cloud based environment provides a platform for applications such as OTT video which run over the cable operator's IP broadband service. OTT vendors have taken advantage of cloud services to improve efficiency. Operators have followed suit to some extent with their own OTT like offerings but in terms of leveraging cloud technology are at best on a par with their competition. The remainder of the paper examines how the operator can gain additional advantages from cloud technology by migrating some components of the broadband service itself into this same cloud environment. The technology steps required and the benefits and impacts it may have will be reviewed.

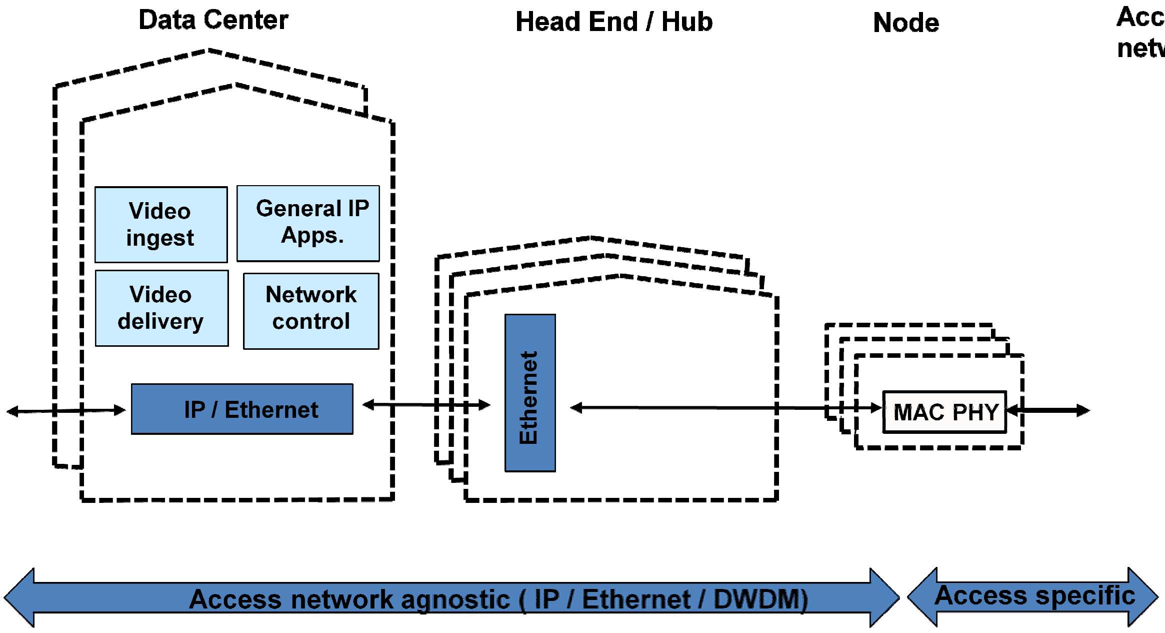


Figure 2: Cloud Centric Solution

Figure 2 shows a possible implementation of this type of architecture illustrating the major components and their location in the network. The philosophy behind this approach is simple; to put as much functionality in the data center as practical, to leverage Ethernet and digital optics where possible and to keep the node simple. The reasons to migrate functions to the data center are to leverage off the shelf hardware and software for reduced cost, to leverage virtualization for redundancy and scaling and to centralize complexity for simpler operation. Ethernet and digital optics are used to provide low cost and long distance options. The node is kept simple for low cost and ease of operation. As a consequence of the migration of functions to the data center and node the head end and hubs become much simpler.

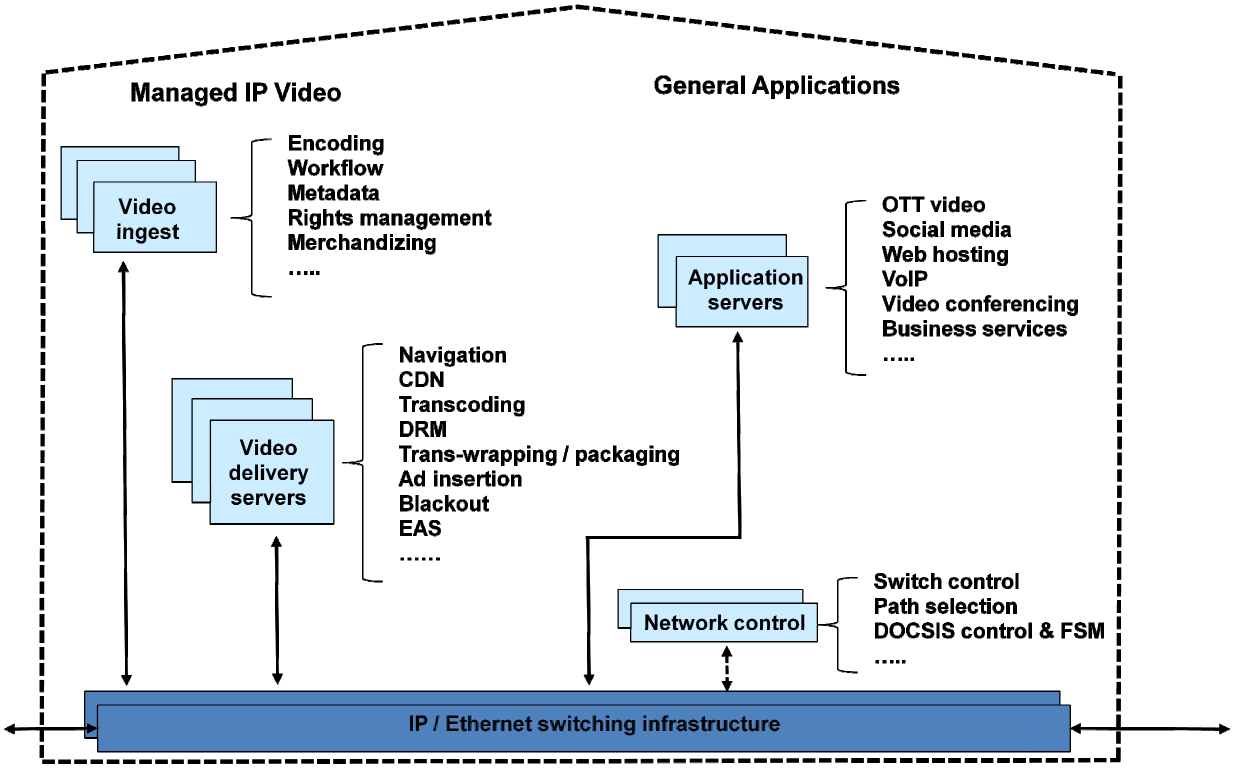


Figure 3: Data Center Functions

Data Center

As with existing cloud based services the data center is based on off the shelf servers running general applications software in a virtual environment as shown in Figure 3. These include general applications software such as OTT video, social networking and Internet access.

Additional servers provide the functions needed to create a multi-screen video service. Video ingest servers receive content from multiple sources and provide encoding, metadata and content management functions.

Video delivery servers provide transcoding and packaging functions to create multiple bit rate program streams along with additional functions such as advertising insertion, and content protection that are required by a full service video provider. A description of a layered architecture for an end to end IP video system can be found in [MSIPD].

This is a simplified description in that these functions may be implemented in a central data center or in a more distributed model based on multiple data centers interconnected with CDN distribution networks. For the purposes of this paper the simple model will suffice as the evolutionary steps proposed for the network are independent of the model selected for data center deployment. The video service architecture described above is in fact deployed currently in both centralized and distributed modes and can be used with both existing HFC delivery networks and the evolved network proposed.

The next step in evolving the network is to use additional servers in the data center to run the access network (e.g., DOCSIS) control plane. In a traditional router or CMTS the data plane typically runs in specialized hardware while the control plane runs in a general purpose CPU embedded in the platform. In this case the general purpose CPU has migrated to a server and standard IP/Ethernet switches provide the data plane forwarding within the data center and from the data center to the head end. A control protocol between the server and the switch such as OpenFlow [OPENF] is used to control the forwarding path.

Head End

The head end in this architecture continues to support traditional analog and MPEG based broadcast video. High speed data and narrowcast video processing has moved to the data center as described above. HSD and IP video traffic passes through the head end via an IP/Ethernet network and is forwarded to the node using the existing fiber links which it shares with the broadcast video using WDM.

Node

In the node the broadcast video is converted from optical to coax media as today. The Ethernet traffic is converted from baseband Ethernet to the protocol to be used for the node to home portion of the network. This may be DOCSIS, other Ethernet over coax (EoC) access technologies, point to point Ethernet or even wireless technologies such as WiFi or LTE. Operation of the node is controlled from the data center (via the head end).

ADVANTAGES

The architecture described above has multiple advantages over a traditional HFC video delivery infrastructure.

Leverage Data Centers

It leverages the work done to provide massively scalable Internet based applications over the last decade to provide cost savings and efficiency:

- The use of COTS technology reduces costs by using general purpose servers as processing engines and standard Ethernet networking equipment for connectivity.
- It leverages standard virtual environments to provide horizontal scaling and redundancy
- It provides a simpler and more robust environment for networking software development.
- It provides a friendly platform for application level software development where familiar toolsets and environments enable software to be created by operators, equipment vendors and third parties to accelerate service velocity.

Head End Simplification

The head end becomes a much simpler environment leading to lower operational costs:

- Complex software functions are centralized in the data center where expertise can be concentrated.
- Power, cooling and rack space needs at the head end are reduced as devices such as CMTS are removed.

- IP services are delivered via a standardized Ethernet network as low level DOCSIS and QAM functions migrate to the node and the need for RF combining in the head end is reduced or ultimately removed.
- Multiple head end / hub locations may be collapsed or run remotely as a "lights out" operation saving operating and real estate costs.

Network Simplification

The use of IP transport to the node simplifies the network in the following ways:

- Standard Ethernet and digital optics can be used between the head end and the node eliminating distance limitations and enabling distribution hubs and small head ends to be consolidated.
- Ethernet switching is used in the data center, head end and hub to reduce costs in the transport network.
- The network from the data center to the hub is independent of the last mile technology from the hub to the home allowing these parts of the network to evolve independently and more cost effectively.

TRANSITION STAGES

From the above it appears that there are significant advantages to moving to the proposed architecture but as always the problem is in how to facilitate the transition at a reasonable cost while continuing to provide service. The following sections of the paper address this transition and break it down into a number of potential stages. The discussion identifies six possible transition stages between today's HFC network and that shown in Figure 2. These are:

1. Introduction of CCAP
2. Split packet processing from physical media dependent and physical layer (PMD/PHY) processing
3. Move PMD/PHY to the node
4. Move MAC processing to the node
5. Move narrowcast processing from head end / hub to data center
6. Retire broadcast processing and remove from head end /hub

Not all stages are required and each operator can select the most appropriate path based on their existing network, operational needs and competitive demands.

Stage 1

This first phase of the transition, shown in

Figure **4**, addresses the change in services from broadcast to narrowcast. Narrowcast channels for MPEG and DOCSIS are added to the head end using high density CCAP based equipment. This may augment or replace existing EQAM and CMTS equipment. The CCAP platform connects to the core network through the Ethernet distribution network in the head end as for existing equipment but will use 10Gbps rather than 1Gbps Ethernet links. Connectivity to the HFC remains RF based and connects to an optical shelf through the existing RF distribution /combining networks in the head end. The optical shelf converts the signals to analog optics and transmits them to the fiber nodes in the outside plant.

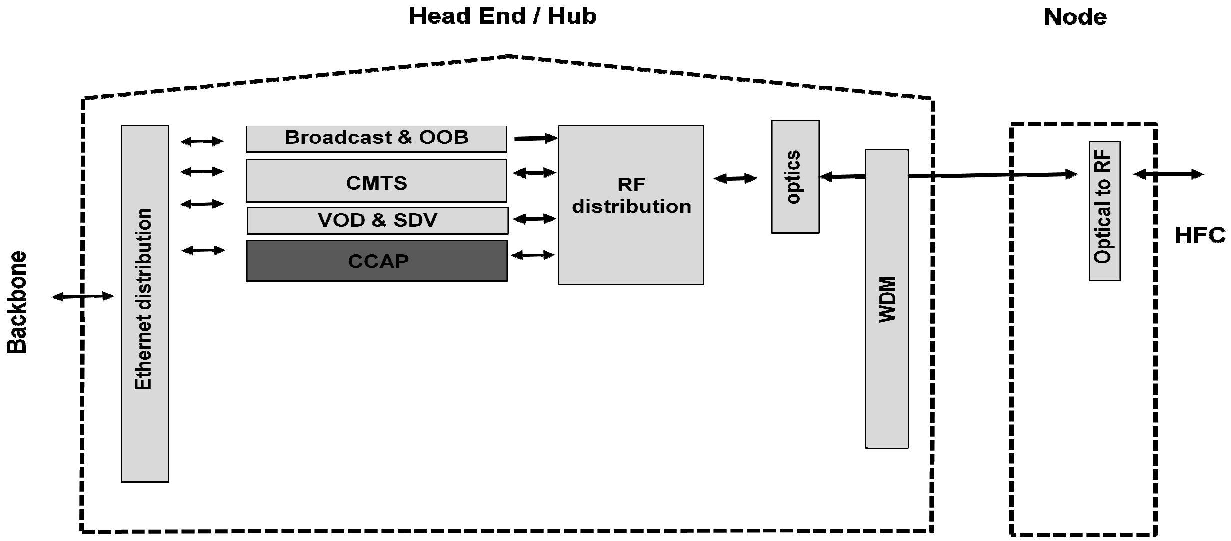


Figure 4: Phase 1, CCAP

The advantages of this transition are the savings in cost, real estate and power consumption provided by the increased density of next generation platforms [CCAP].

The technology changes required and risks associated with this change are those associated with the CCAP program and are well understood at this time.

Stage 2

The next stage of the proposed transition requires a slightly more radical change and is shown in Figure 5. It leverages some of the work done for modular CMTS [M-CMTS] and distributed CCAP architectures [D-CCAP] to move to an Ethernet based distribution system in the head end. The core principal is to decouple the packet processing from the physical media dependent (PMD) portion of the MAC and the PHY. The PHY and PMD dependent functions move out of the core processing engines to be collocated with the optical shelf which may be within the head end or in a distribution hub. The CMTS and EQAM core engines output MPEG streams over Ethernet using standard framing [M-CMTS]. The downstream modulation function is included with the optical transmitters and the upstream demodulation function is included with the optical receivers.

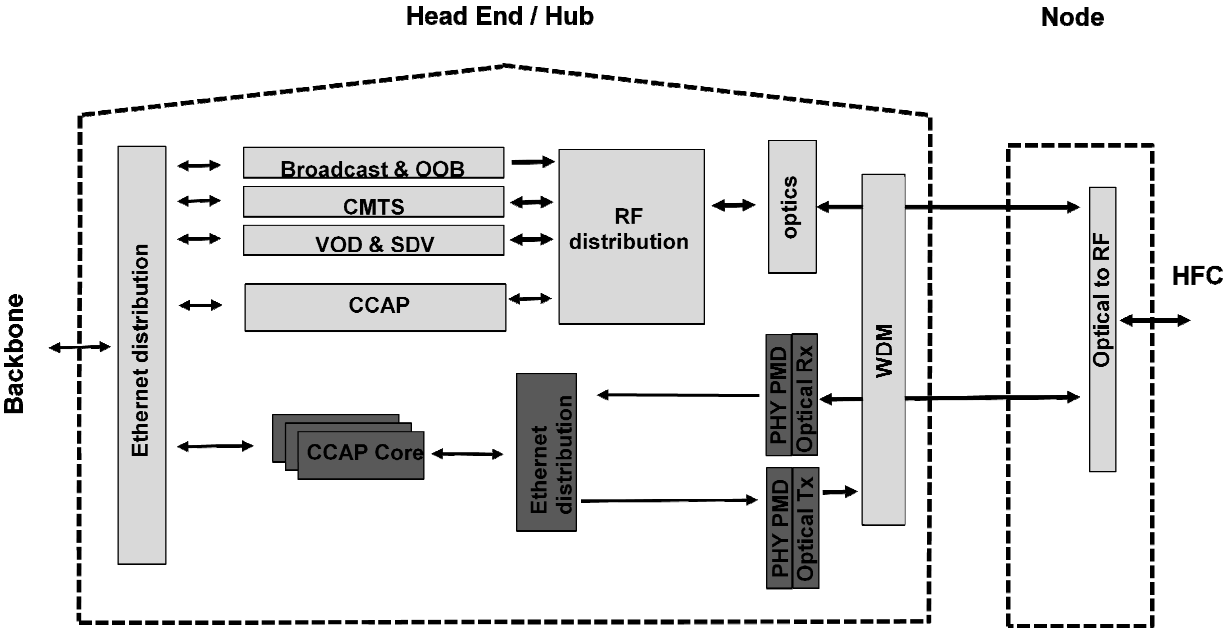


Figure 5: Phase 2 with Local CCAP core

This separation of digital processing from modulation has several advantages. Scaling the system becomes more efficient as CMTS and EQAM core engines scale based on processing needs while the PMD and PHY layer functions scale based on port counts. Thus the core engines scale up as the total number of DOCSIS and EQAM channels to be delivered increases while the PMD/PHY functions scale as the number of serving groups increases. While these factors (channel count vs. serving group count) are certainly related it is typically not a strictly linear relation so there is benefit to the separation.

Redundancy may also be simplified in this case. Ethernet based redundancy can be used for the core processing engines so that there is no need for complex RF switching logic in the core chassis in the data center. RF redundancy can be provided in the optical shelf for large serving groups but for smaller serving group sizes the failure group in the optical shelf may be small enough that RF redundancy is not needed providing further cost reductions.

With the replacement of the RF combining network by Ethernet switching changes such as node splits are made simpler; ports are added to the optical shelf for the new node and processing engines added to the core shelf if required. The operations and management network for the head end can share the Ethernet infrastructure to provide central configuration and monitoring using standard IP tools. Thus path traces and testing across the head end become trivial using tools such as ping and traceroute.

The most interesting feature of the PHY-PMD separation is the increased flexibility it enables for equipment deployment. The standard Ethernet links between the core processing engines and the optical shelves effectively remove any distance restrictions between them. Thus the optical shelves could be deployed in a remote head end or hub while the processing engines reside in a data center. This enables “lights out” operation of the remote facility and significantly reduces power and cooling needs at these locations. Centralizing the complex equipment also reduces the operations skill sets needed at the remote locations.

Figure 5 above shows a potential deployment scenario in which additional services are added via the new model as needed. Existing broadcast, CMTS and EQAM equipment can remain unchanged or be replaced over time.

Figure 6 shows an alternative deployment model with the CCAP core platforms located in a data center remote to the head end.

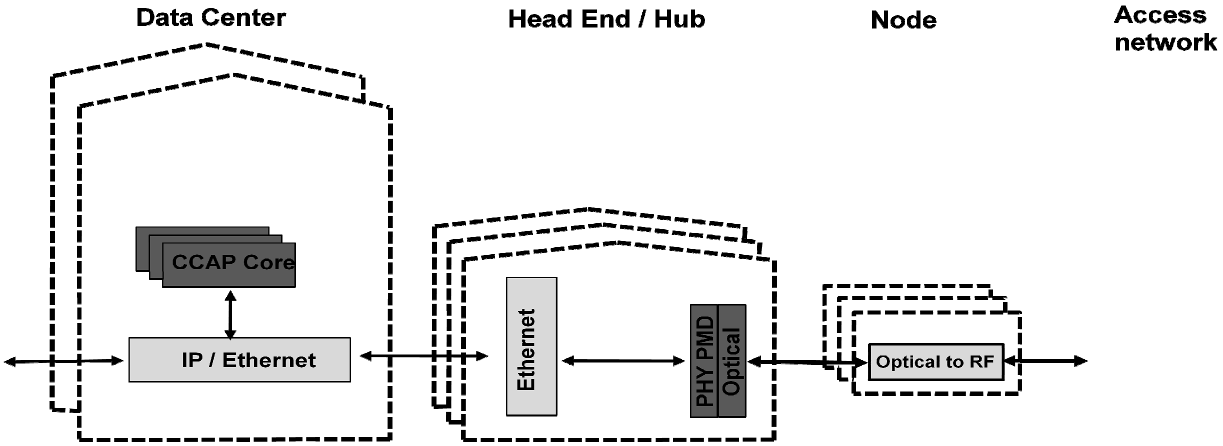


Figure 6: Phase 2 with Remote CCAP core

The technology risks associated with this phase are primarily the timing issues associated with the separation of the DOCSIS MAC and PMD functions.

Stage 3

As mentioned above the use of an Ethernet distribution network between the core processing engines and the optical shelf eliminates distance restrictions between them. The next phase takes advantage of this fact to move the PHY-PMD function out of the head end to the node and extend the Ethernet distribution to the node using standard Ethernet optics as shown in Figure 7 . New services are added using the Ethernet to the node transport while legacy services continue to be supported using an RF overlay. Ethernet and analog wavelengths are multiplexed onto the same fibers using existing WDM equipment.

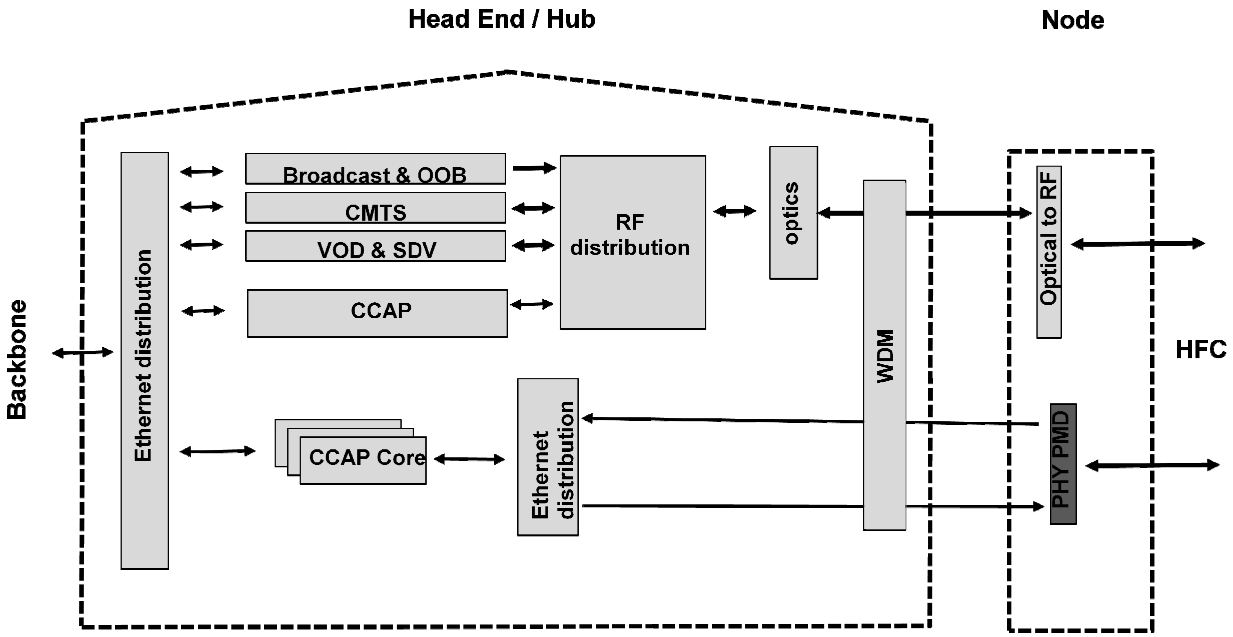


Figure 7: Stage 3, Remote PHY-PMD

Deployment of this phase of the transition plan brings several advantages to the operator. Service expansion is no longer limited by the head end to node transport as the use of digital optics and DWDM provide essentially unlimited bandwidth on this link. The links leverage the costs, speed and density trajectory set by Ethernet systems and standard DWDM platforms.

If the RF overlay can be removed then distance limits between the head end and the node are eliminated. It may then be possible to centralize head end functions and retire some distribution hubs and small head ends providing savings in real estate and operational costs.

The technology risks associated with this phase are the density, powering and cooling of the components in the fiber node and the provision of timing services to the node.

Stage 4

This phase, shown in Figure 8 is a conceptually small change in which the upper layer MAC functions are moved to the node so that they are co-resident with the PMD and PHY functions. As in the previous phase new services are added using the Ethernet to the node transport while legacy services may continue to be supported using an RF overlay

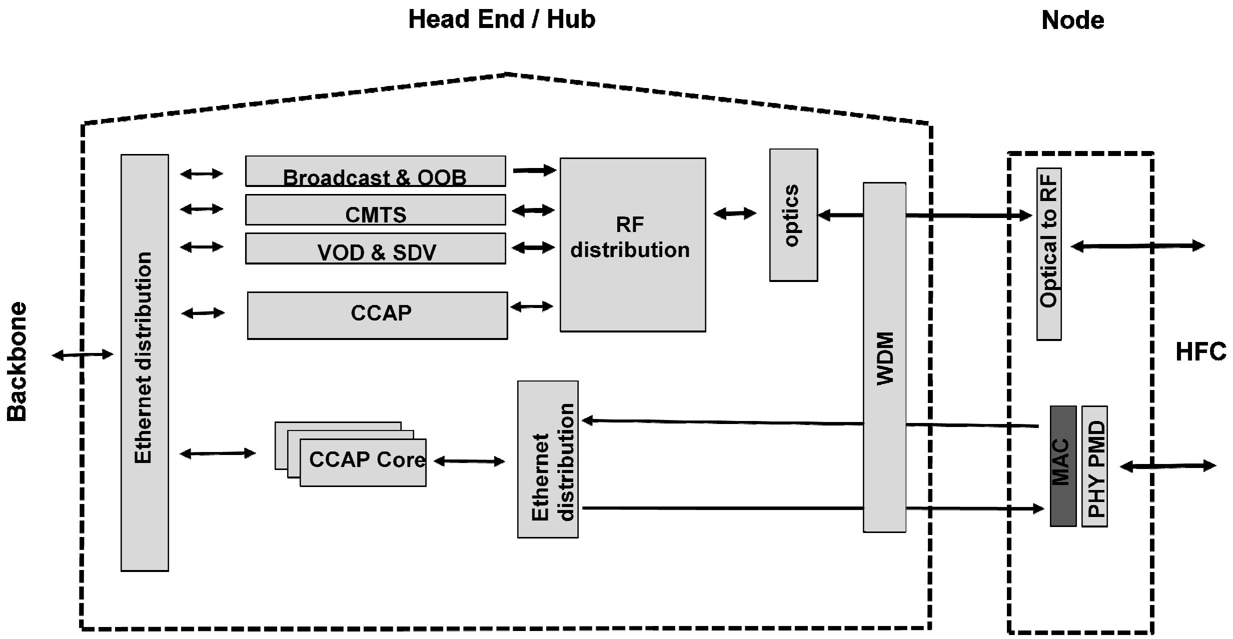


Figure 8: Stage 4 Remote MAC

The advantage of moving the MAC to the node is that it simplifies the timing issues relative to the split implementations described previously. More importantly it decouples the technology for the head end to node transport from that used between the node and the home. This allows the technologies to evolve at their natural and different paces. The head end to node link uses general enterprise networking technology while the node to home link remains specific to the HFC plant. Node splitting and the progression towards an n+0 architecture becomes a process of replacing the DOCSIS MAC/PHY module in the node with an Ethernet switch and moving it further downstream. The impact of transitions to next generation technologies such as PON or EoC is restricted to the node to home portion of the network.

The technology risks are similar to the prior phase as more functions are moved to the node increasing powering and cooling needs.

Stage 5

In this phase, shown in
Figure 9 the legacy narrowcast equipment in the head end is retired and all processing elements migrate to a central data center. The HFC specific MAC elements have migrated to the node and IP/Ethernet transport is used throughout the network from the core to the node. Application and control plane logic are no longer in the head end which is only used for legacy broadcast services. It still acts as a pass through for narrowcast services but this is reduced to an IP/Ethernet switching function.

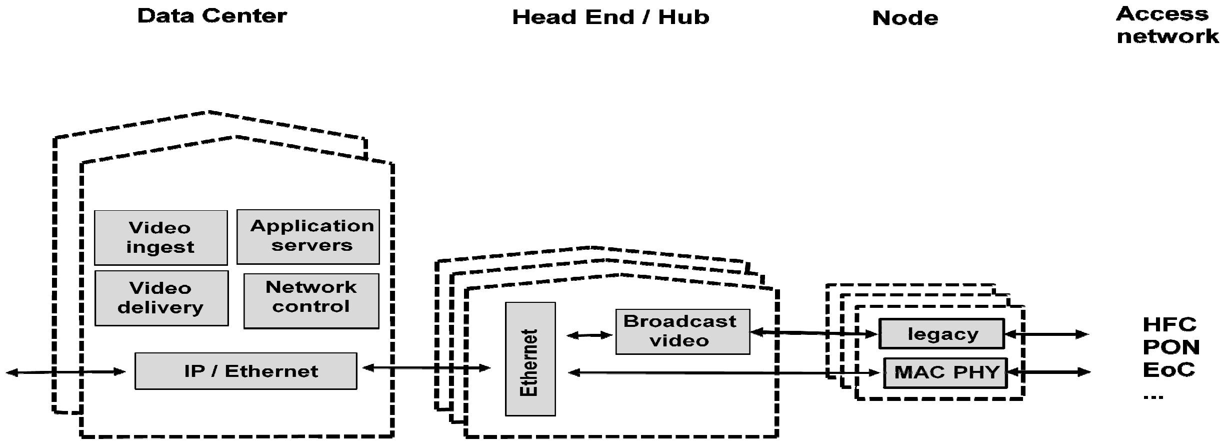


Figure 9: Stage 5 Narrowcast Equipment Removal

Stage 6

With the removal of traditional broadcast the head end in its current form can go away completely and be replaced by a data center, an Ethernet distribution hub and a simple node as shown in Figure 10. At this point the advantage of centralization, standard IP/Ethernet transport and isolation of HFC specific functions described in the earlier phases are now fully realized.

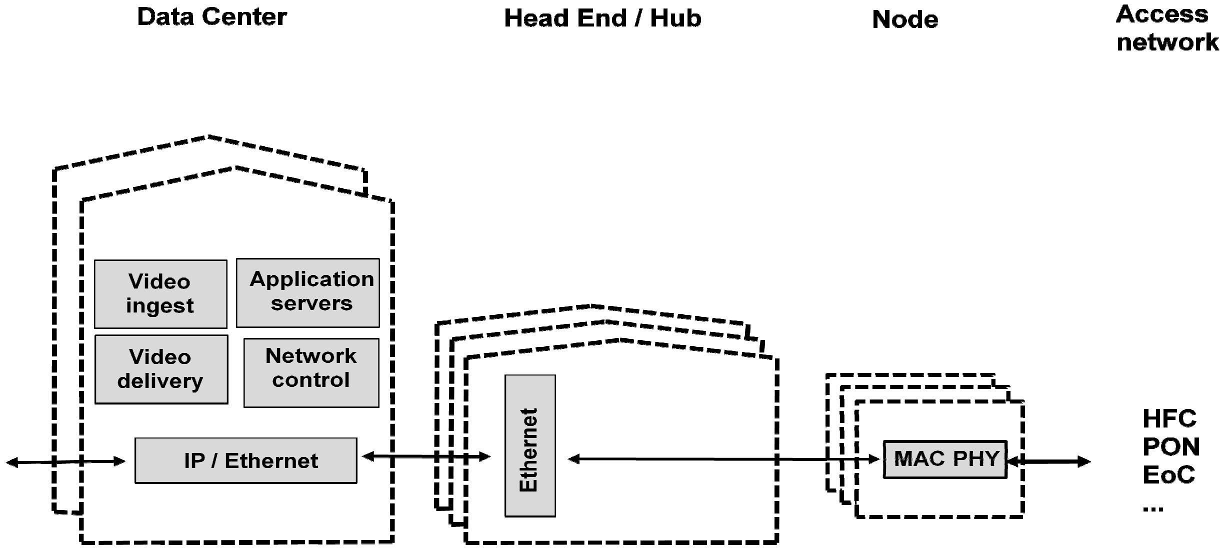


Figure 10: Stage 6 Broadcast Equipment Removal

Moving the MAC and PHY functions from the head end to the node allows the use of standard Ethernet optics and enables distributed processing but results in a more intelligent outside plant architecture. Operators who do not wish to take this step and prefer to keep a simpler outside plant can elect to deploy the MAC-PHY components in the remote hub rather than the node as shown in Figure 11. They still retain the advantages of the move to the data center and a significant reduction in hub complexity. Readers interested in an in depth comparison of traditional and intelligent HFC architectures are referred to [HFCDFC].

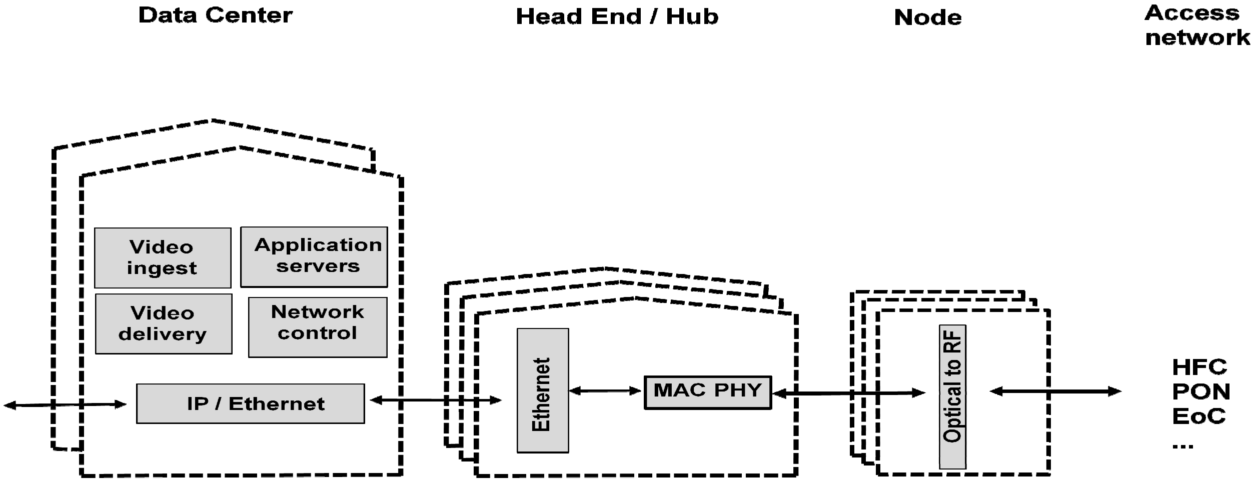


Figure 11: Passive HFC Architecture

DEPLOYMENT

The transition stages previously described illustrate a logical roadmap to the data center architecture. They are based on technology evolution but are essentially independent. Which stages are deployed by an operator will depend on their customer needs, timing, risk profile, budget and network architecture.

Table 1 shows a summary of the risks and benefits associated with each phase together with an indication of when it would be appropriate to be used.

Stage	**Change**	**Benefits**	**Risks**	**When Appropriate**
1	Move to CCAP platform	Increased density, lower cost, simplified combining	Minimal, well understood problem	Need to add high density narrowcast services in HE/hub
2	Decoupled MAC / PHY in head end	Processing and port scaling separated, simpler redundancy, simpler combining	DOCSIS MAC/PHY timing split.	Gain HE/hub benefits without touching node
3	PMD+ PHY move to node	Ethernet Digital optics to node	Power & cooling in node	Consolidate small HE/hub; retain existing core platforms
4	MAC to node	Data center to node links all Ethernet for narrowcast traffic	Power & cooling in node	Standardize on Ethernet transport to the node to set up for stages 5 & 6
5	Narrowcast removed from HE	Head end space savings	minimal	Consolidation from HE/hub to data center for lower OPEX
6	RF broadcast removed from HE	HE becomes simple switching center or is removed	minimal	Consolidate or retire HE/hub

Table 1: Risks and Benefits of Each Stage

CONCLUSION

Moving functions from the head end or distribution hub into a data center has many advantages and has the capability to provide significant capital and operational savings. To transition to a network architecture which can take full advantage of this move is not trivial but can be achieved through a series of stages as technology evolves and service needs demand. The transition stages described are largely independent and any given operator can select the transition path best suited to their specific needs.

REFERENCES

[CCAP1]	J. Salinger, “Proposed Next Generation Cable Access Network Architecture”, SCTE Conference on Emerging Technology, 2009.
[CCAP2]	J. Salinger, “Understanding and Planning CMAP Network Design and Operations”, SCTE Cable-Tec Expo, 2010.
[CCAP3]	J. Finkelstein, J. Salinger, “IP Video Delivery using Converged Multi-Service Access Platform (CMAP)”, SCTE Canadian Summit, 2011
[D-CCAP]	John Ulm, Gerry White New Converged Access Architectures for Cable Services, NCTA 2011 Spring Technical Forum
[HFCDFC]	M. Emmendorfer, S. Shape, T. Cloonan & Z. Maricevic Examining HFC and DFC (Digital Fiber Coax) Access Architectures, SCTE Cable -TEC Expo 2011
[M-CMTS]	Cablelabs DOCSIS® Specifications — Modular Headend Architecture (MHA)
[MSIPD]	John Ulm, Gerry White Arch & Migration Strategies for Multi-screen IP Video Delivery, SCTE Canadian Summit 2012
[OPENF]	www.openflow.org
[SAND]	Global Internet Phenomena Report Fall 2011; Sandvine

ABBREVIATIONS AND ACRONYMS

CCAP	Converged Cable Access Platform
CDN	Content Delivery Network
CMTS	DOCSIS Cable Modem Termination System
COTS	Commercial Off The Shelf
CPE	Customer Premise Equipment
DOCSIS	Data over Cable Service Interface Specification
DRM	Digital Rights Management
DVR	Digital Video Recorder
DWDM	Dense Wave Division Multiplexing
EAS	Emergency Alert System
EoC	Ethernet over Coax
EPoC	EPON over Coax
EPON	Ethernet Passive Optical Network
EQAM	Edge QAM device
FSM	Finite State Machine
Gbps	Gigabit per second
HFC	Hybrid Fiber Coaxial system
HSD	High Speed Data; broadband data service
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
MAC	Media Access Control (layer)
Mbps	Megabit per second
MPEG	Moving Picture Experts Group
MPEG-TS	MPEG Transport Stream
nDVR	network (based) Digital Video Recorder
OTT	Over The Top (video)
PHY	Physical (layer)
PMD	Physical Medium Dependent (layer)
PON	Passive Optical Network
RF	Radio Frequency
STB	Set Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VOD	Video On-Demand
VoIP	Voice over IP
WDM	Wave Division Multiplexing

ARCHITECTUAL APPROACHES FOR INTEGRATING SP Wi-Fi IN CABLE MSO NETWORKS

Rajiv Asati, Distinguished Engineer, rajiva@cisco.com
Sangeeta Ramakrishnan, Principal Engineer, rsangeet@cisco.com
Rajesh Pazhyannur, Technical Leader, rpazhyan@cisco.com

Abstract

Cable MSOs have an enticing opportunity with Wi-Fi residential and business services.

In this paper, we discuss the common requirements, challenges (that Cable MSOs face) and necessary architecture (that MSOs could use) for integrating SP Wi-Fi in Cable MSO networks to support both residential and hotspots use-cases. This paper also qualifies various architectural approaches for network transport in the context of DOCSIS access along with the time-to-market perspective, so as to enable MSOs to quickly capitalize on this opportunity.

1. INTRODUCTION

Wi-Fi is a pervasive & proven access technology that is commonly used by Homes and Enterprises around the world, and its usage by Service Providers (SPs) is gaining traction as well. SPs can use Wi-Fi to deliver one or more of the triple-play services (e.g Video, Voice, Data) to the customers indoor and outdoor, and enhance the customer/user experience (by allowing mobile consumption of content as well as access to data).

In fact, SPs, particularly, Mobile SPs have been leveraging Wi-Fi for better cost-efficiency and QoE. As the number of mobile devices keeps growing exponentially, it is expected that the Mobile network traffic would keep growing exponentially as well (studies have predicted a 18-fold increase in mobile data traffic in the next 5 years, as illustrated in `Figure 1`).

There will be 5B mobile devices and 2B M2M nodes
Global mobile traffic wil
grow 26X to 6.3 EB/mo
Video will be 66% of all mobile traffic by 2015
© 2011 Cisco and/or its affiliates. All rights reserved.
Cisco Cor

Source: Cisco Visual Networking Index (VNI) Global Mobile Data Forecast, 201

Unfortunately, most mobile SPs do not have enough licensed radio spectrum to accommodate this increase. Given that a large amount of traffic is consumed indoors (in homes, offices, public-spaces like hotels, café's, etc), where Wi-Fi connectivity is much more widely available than cellular, the usage & focus on Wi-Fi to offload traffic from cellular networks has greatly increased. In fact, 'Mobile Data Offload & Onload Video Whitepaper (published by Juniper Research in April 2011) predicts that Wi-Fi usage for mobile traffic offload could exceed ~1EB / month by 2015. This is illustrated in `Figure 2`.

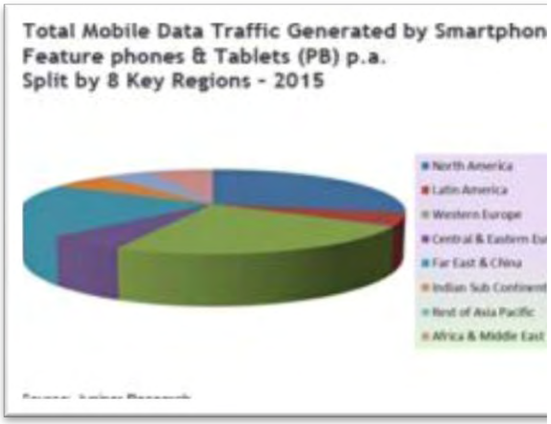


Figure 2 Mobile Traffic offload Prediction

Needless to say, Mobile SPs would need to acquire sites for installing Wi-Fi based macro-cells, and hence, mobile SPs are increasingly motivated to rely on other SPs/Providers offering the Wi-Fi based solutions.

Cable MSOs have a fantastic opportunity with Wi-Fi. In this paper, we discuss the common challenges (that Cable MSOs face) and necessary architecture (that MSOs could use) for integrating SP Wi-Fi in Cable MSO networks to support both residential and hotspots use-cases. We also qualify various architectural approaches for network transport in the context of DOCSIS access along with time-to-market perspective, so as to enable MSOs to quickly capitalize on this opportunity.

2. SP Wi-Fi: MSO REQUIREMENTS / CHALLENGES

SP Wi-Fi primarily refers to an 802.11 Wi-Fi system deployed and managed by a Service Provider (SP) for public access (aka community access) to its network for services such as High Speed Data Internet service. Public Access means that Wi-Fi is available to the customers of the SP and/or partner SPs and/or any customers. SPs may provide managed (and sometimes hosted) Wi-Fi services to other service providers (e.g. Mobile SPs).

SP Wi-Fi differs from general Wi-Fi e.g. Enterprise Wi-Fi (or Residential Wi-Fi) in three key aspects:

1. **Scale** – The number of APs and user clients tends to be very large – thousands to millions.
2. **Carrier Grade** – The high-availability and manageability aspects tends to be of carrier class (e.g. 5 9's)
3. **Multi-Vendor** – The existence of multiple vendor devices is expected – warranting the usage of standards based end-to-end architecture.

2.1 Use-Cases

SP WiFi architecture should be flexible enough to enable Cable MSO to serve one or more the following deployment use-cases:

1. **Residential** (Indoor) –re-use the Wi-Fi APs that are integrated with the (SP managed) residential gateways to provide public access Wi-Fi. In this case, the AP is located indoor (in a residential customer home).
2. **Metro** (Outdoor) –deploy Wi-Fi APs outdoor in public places to provide public access Wi-Fi. In this case, the APs are typically mounted on aerial cable strands, street-poles, roof-tops etc.
3. **HotSpot / SMB** (Indoor) –re-use the managed Wi-Fi service to SMBs such as coffee shops, bookstores, retail-stores etc., having 10s or 100s of employees, for both private and public access WiFi.
4. **HotSpot** (Outdoor) –deploy large concentration of APs in a relatively small area such as stadium, amphitheaters, parks etc. having large number of users in that area. The APs are usually located outdoor to offer public access Wi-Fi.

5. **Wholesale / offload** – allow partners' customers to access the Wi-Fi services, and/or backhaul mobile operators' customers traffic over the MSO infrastructure. In this case, the APs are located indoor and outdoor.

2.2 Access Point / 802.11 Radio

Access Point (AP) is the most fundamental element in the SP WiFi architecture. Hence, the AP requirements must be carefully assessed. The following are some of the key considerations for the Wi-Fi AP:

1. Coverage: refers to AP's range to = what throughput upto what distance. Coverage determines the number of APs required to cover a certain area. Naturally, 802.11n radio on AP is preferred for optimal coverage.
2. Capacity: refers to the maximum number of clients that AP can concurrently support/associate. Some prefer to define capacity in terms of maximum number of active users that can be supported with each user guaranteed a minimum throughput. Capacity directly influences the number of APs required to cover a certain area (e.g. the number of APs are determined by capacity requirements rather than coverage).
3. Interference Management: refers to AP's capability to continuously select the best radio channel (through constant monitoring since startup) while managing the radio interference so as to get the best radio performance. The interference could be generated by other Wi-Fi APs or by non Wi-Fi sources such as Bluetooth, DECT phones, Microwave etc. Naturally, techniques such as Beamforming to improve the signal strength received by the client, interference identification for reporting etc. become important.
4. Dual radio– refers to AP supporting simultaneous usage of 2.4GHz and 5GHz. This is particularly important for APs that are used for creating private and public WLANs. This should be controllable by the MSOs.

2.3 Security

Security is one of the most-pressing issues, as security threats such as snooping, Eavesdropping, session hi-jacking, session side-jacking, evil twin attack etc. expose the insecurity in WiFi networks that rely on open SSID. Hence, it is important to have secure SSID/WLAN.

Note that most SP Wi-Fi deployments have not used secured SSID because of lack of support on clients for EAP methods and/or complexity in distributing and managing user-security credentials. Hopefully, this will change with Hotspot2.0 recommendations. Please see more details on this here [Hotspot2.0].

Additionally, in case of residential SP WiFi, the AP must support at least one private WLAN/SSID for the residential customer's usage, and at least one public WLAN/SSID for public usage, for security reasons.

In summary, SP WiFi architecture should include user authentication and cryptography (e.g. WPA-2 Enterprise), as well as separate control and management of public and private WLANs so as to pave the way for 'Secure WLANs'.

2.4 Inter-Operator Roaming

It would be desirable to let the users use other MSOs' or SPs' Wi-Fi networks to get one or more services (such as high speed data connectivity to the Internet) when the users are roaming [Wi-Fi-Roam]. However, how would the customer's device know the right SSID (assuming more than one SSIDs) on the partner Wi-Fi network? If the users knew the right SSID, they may have to manually login and get authenticated so as to use partner Wi-Fi network. This is deemed not only inconvenient to the user, but also as a lost opportunity for the MSOs to influence users' network selection.

Once authenticated, then depending on the mobility requirement, home network or the partner network should assign the IP address to the user client device. If the roaming users managed to use partner Wi-Fi network, then they may get limited time before they are asked to re-authenticate, causing them another source of inconvenience. Lastly, as MSOs allow the roaming users, appropriate billing ruleset, Lawful Intercept etc. have to be enforced. Of course, this all assumes the MSOs to have struck the roaming agreements with other MSOs & SPs.
To address this challenge, IEEE 802.11u could be necessitated. Please see more details on this here [Hotspot2.0].

2.5 Mobility

Mobility is defined in many different ways, resulting in many different requirements. However, MSOs may not find all the mobility requirements to be important and/or applicable. A brief summary of mobility requirements is provided below:

- Fast Roaming: enables AP-to-AP handover user re-authenticate the user. Specifically, the re-association procedures are performed in parallel with key negotiation procedures, as per IEEE 802.1r.
- Micro-Mobility: In deployments with a small number of APs in a site (such as bookstore, restaurant) there is need to support mobility to reduce adverse impact on end user experience as they roam within the site. In most scenarios, when user walks out of the site, they will lose Wi-Fi coverage. Reconnecting to Wi-Fi in another location/site would typically result in users getting a different IP address.
- Macro-Mobility: In deployments where there is large contiguous area covered by Wi-Fi (such as outdoor APs) there is need for end users to maintain IP address as they roam between Wi-Fi APs. In such cases, the solution may need tunnels between centralized Wi-Fi aggregators (WLC, CMTS, MAG, etc) to provide this form of mobility
- Inter-Vendor Mobility: As mentioned earlier, SP Wi-Fi deployments tend to comprise network elements e.g. APs from different vendors, hence, it is important to ensure that mobility works between different vendors' APs. Further, in some scenarios, the vendors may provide overlapping Wi-Fi coverage.
- Inter-Technology Mobility: A significant portion of Wi-Fi devices are likely to have a cellular (3G/4G radio) as well. In some cases, it may be desirable to provide mobility as users roam between radio-technologies (between Wi-Fi and Cellular). Such mobility can be provided by using client based mobility mechanisms (Mobile IP, DSMIPv6) or network based mobility mechanisms (such as PMIPv6)..

While many of the above requirements may be reasonable, it is worth noting that continuous Wi-Fi coverage is a prerequisite of any form of mobility. Hence, mobility may not be possible everywhere or applicable, requiring careful justification.

2.6 Traffic Separation

As SP WiFi traffic is transported over the MSOs network infrastructure, traffic separation capabilities in the network especially on the access (e.g. DOCSIS) side will become critical.

2.6.1 Separation of HSD subscriber's traffic from SP Wi-Fi traffic

Most operators have bandwidth caps and tiers of service deployed whereby each subscribers' traffic is separately measured (for bandwidth cap purposes) and QoS is applied to ensure the traffic complies to the tier of service the user has subscribed to (example, 6Mbps down, 1Mbps up). Once the cable modem deployed at a business or home, is enabled for SP Wi-Fi, operators will want to ensure that the SP Wi-Fi users' traffic does not count towards the HSD subscriber's limits. Given that in DOCSIS the Service Flow is the unit on which accounting and QoS is applied, the architecture needs to ensure that the SP Wi-Fi traffic is mapped to a different service flow than that of the subscriber's HSD service flow. This mapping needs to be done both in the Upstream and Downstream directions.

An implicit challenge here is that needing specific US and DS classifiers may result in having unique CM config file for each modem. The chosen architecture must address this challenge.

2.6.2 Separation of Services per Fiber Node

The previous section discussed the separation of a single HSD subscriber's traffic from the SP Wi-Fi users attached to the same CM/AP. Additionally operators may want to ensure that a certain amount of bandwidth is set aside for HSD use versus SP Wi-Fi use across the entire Service Group. This would ensure that one service on an aggregate doesn't crowd out the other service on a Service Group. It would also be beneficial if any unused bandwidth provisioned for one service was made available for the other service to use as needed.

DOCSIS provides the Bonding Group construct which can be used to provide such a service separation between the two services. By using overlapping bonding groups across a set of RF channels, and steering HSD service flows to one Bonding Group and the SP Wi-Fi service flows to the other bonding group, operators can achieve such separation. Depending on how much bandwidth an operator wishes to set aside for each service, they can configure the bonding groups appropriately to achieve their goals.

2.7 Network Transport

SP WiFi services may need to be deployed over various types of access networks e.g. DOCSIS/HFC, EPON/Fiber etc. that are present in MSO networks. For example some operators are considering offering business services over EPON. The overall architecture chosen for deployment will need to be such that they are easily deployable across different access technologies. Hence the Access Point itself will need to support various backhaul technologies such as DOCSIS, EPON etc.

For utmost cost-effectiveness, it would be desirable to leverage the IP or MPLS (or 802.1 based carrier Ethernet) network transport that is already used by MSOs for other services. In fact, many MSOs have converged their networks (or on the path to do so) and been using MPLS technology for various services. The key is to choose the network transport that yields the simplification of SP WiFi architecture while satisfying other SP WiFi requirements that are important to the MSO.

2.8 Provisioning & Management

In particular, the WiFi APs should be automatically configured without needing any manual intervention for utmost cost-effectiveness (given the expected scale).

Thankfully, both DOCSIS cable modem and eDOCSIS[1] device already allows auto-

[1] An eDOCSIS device consists of an embedded DOCSIS cable modem (eCM) and one or more embedded Service/Application Functional Entities (eSAFEs) such as eAP, eRouter, eSTB, eMTA etc. There are already various vendors' eDOCSIS devices

configuration of cable modem (and DPOE allows auto-configuration of ONU) and integrated AP. Moreover, eDOCSIS device, by definition, has a single software image for the entire device.

However, if the chosen SP WiFi architecture requires each modem to rely on a unique config file, then it could become a provisioning challenge (as MSOs generally use a few cable modem config files across tens of thousands or millions of modems. This challenge can be solved if template based cable modem config file generation method is used.

For residential SP Wi-Fi deployments in particular the number of APs may well be as high as the number of deployed cable modems. Hence being able to provision at scale is critically important.

Needless to say that CMTS provisioning should not be needed on a per modem basis.

In summary, seamless integration of the WiFi provisioning (e.g. AP provisioning) into the existing provisioning infrastructure is going to be required for possible auto-provisioning of APs.

2.9 Subscriber Management

Like other services, SP WiFi services will also require subscriber management. This may include capabilities such as bandwidth accounting, quality of service, legal intercept etc. Such services will require a policy enforcement engine that is subscriber aware and learns the policies to be applied from a policy management system. All SP WiFi traffic will have to be routed through such a policy enforcement engine in order to provide the above-mentioned subscriber services.

Subscriber management could occur centrally in which case all traffic needs to be routed to the Subscriber Management Gateway. Different options are available to achieve this, and are discussed in more detail in the Transport Network section 4.1.

It is worth noting that for HSD services, such subscriber management capabilities are applied at the CMTS, hence no requirements to route HSD traffic to any other central entity really exist in MSO networks.

2.10 IPv6

Given the IPv4 address exhaustion becoming a reality for many MSOs & SPs sooner or later and given that SP Wi-Fi would involve 10,000s of APs and millions of users, it is imperative to have IPv6 in SP Wi-Fi usage from day 1. This means that IPv6 should be used not only for addressing users, but also for the underlying infrastructure (e.g. APs, CMTSs, PEs, etc.) irrespective of any IP tunneling is used or not. In other words, both user and AP addressing should be done using IPv6.

While using IPv4 is an option, MSOs would end up requiring many more bandaids (e.g. Carrier Grade NATs) to make it work in a large-scale environment, thereby negatively impacting CAPEX and OPEX associated with SP Wi-Fi.

2.10 Monetization

Once the basic SP Wi-Fi services (e.g. high speed data) get rolled out for the purposes such as customer retention, MSOs may increase the focus on monetization. This would require the architecture to be flexible enough to allow intelligent network to help with advanced services such as advertising, remote monitoring/security etc.

3. SP Wi-Fi ARCHITECTURE

(including 802.11n Wi-Fi Cable Gateway devices [Wi-Fi-GW]) in MSO deployments.

The SP Wi-Fi architecture needs to be flexible enough to satisfy some or all of the requirements (described in section 3) in an incremental & modular way. Such a flexibility would be an important trait to MSOs, since not every MSO would deem every requirements applicable to them day 1.

The SP Wi-Fi architecture needs to be flexible enough to satisfy some or all of the requirements (described in section 3) in an incremental & modular way. Such a flexibility would be an important trait to MSOs, since not every MSO would deem every requirements applicable to them day 1.

This section provides a simplified overview of SP Wi-Fi architecture, and focuses on the architectural approaches for transporting SP Wi-Fi traffic through the transport network while hinting at their flexibility. The Figure 3 below illustrates a high-level SP Wi-Fi architecture:

A SP Wi-Fi architecture illustrated above contains one or more of the following elements:

1. Wi-Fi Access Points: The Wi-Fi Access Points may be either embedded with a cable modem (as in outdoor or residential) i.e. eDOCSIS device (also referred to as Cable Wi-Fi Gateway) or deployed separately from the cable modem (as in many indoor hotspots).

2. Access Network: This is the DOCSIS based HFC network (or EPON or Ethernet based Fiber network) comprising CMTS or CCAP, Fiber Nodes, and CMs (or ONUs) providing network connectivity to/from the AP. The CMTS terminates DOCSIS connections from the cable modems as well as connects to the metro/aggregation Network.

3. Metro/Aggregation Network: The

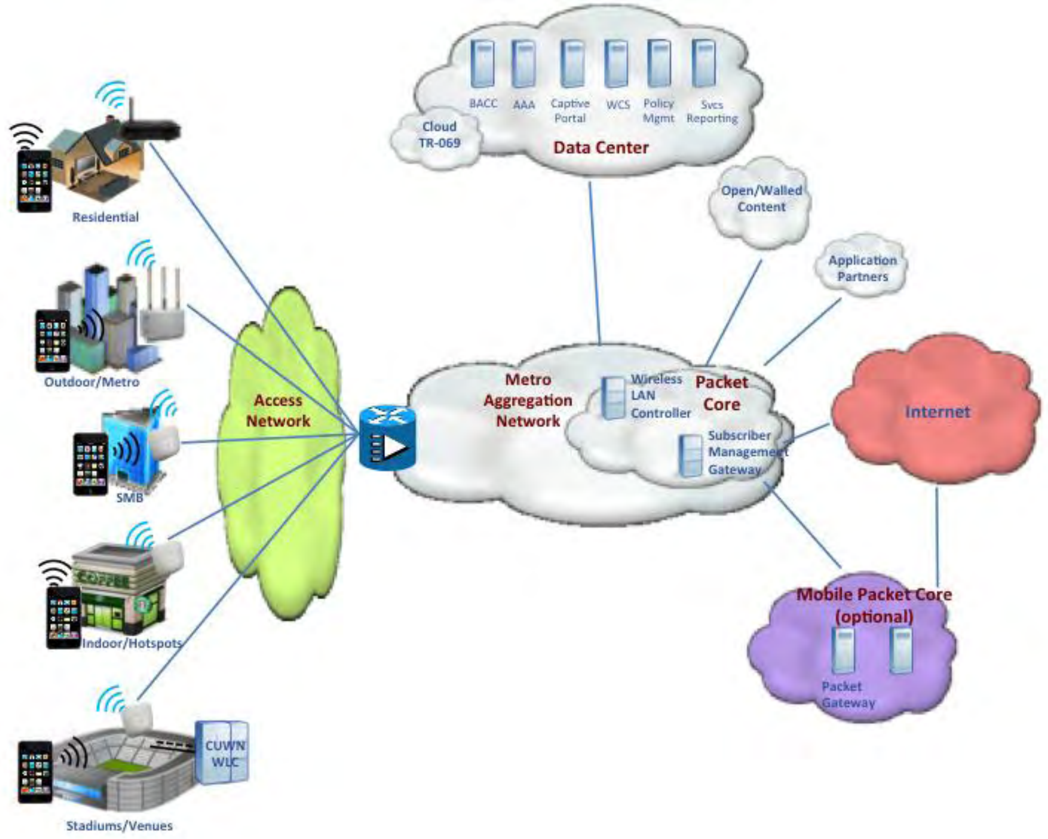


Figure 3 SP Wi-Fi Architecture (simplified)

network that CMTS uses to ultimately connect the users to the internet or partner networks or the open/walled-garden content. There may also be a regional and/or backbone network (not shown in the figure) between the metro network and internet. Metro network is usually an IP or IP/MPLS network (or sometimes a layer2 Ethernet/bridged network).

4. Wireless LAN Controller (WLC): The WLC is a centralized point of control and management of Wi-Fi APs using CAPWAP protocol (IETF RFC 5415). It tunnels data plane (user) traffic to/from the AP using the CAPWAP data plane tunnel (Please see section 3.1.1.2). It is part of the Wi-Fi packet core. It is worth pointing out that not all Wi-Fi APs are based on CAPWAP. Specifically, residential APs (i.e. eDOCSIS device) are not based on CAPWAP. This is better illustrated in the next section.

5. Subscriber Management Gateway: The Subscriber Management Gateway (dubbed as the centralized entity in this paper) is an IP point of attachment that functions as a Policy Enforcement Point (PEP). Specifically, the gateway is responsible to maintain user awareness and enforce of the relevant QoS settings, bandwidth limits, accounting, DPI, etc. The gateway is also referred to as Intelligent Services Gateway (ISG). It is part of the Wi-Fi packet core.

   It is worth pointing out that the Subscriber Management Gateway function could be implemented on the CMTS.

6. Data Center: The Service Network containing elements such as BAC, AAA, DNS, DHCP, Policy Servers and OSS/BSS elements providing network management and service management

7. Mobile Packet Core: This is optional, but it is needed for ensuring inter-technology (3G to Wi-Fi, say) or inter-domain mobility. This includes 3GPP specific elements such as PDN Gateway etc. pertaining to cellular network.

3.1 Network Transport Architecture

Wi-Fi AP connects wireless user devices to each other and/or to a wired network. In general[2], Wi-Fi AP is a layer2 bridge device that bridges Wi-Fi user devices' Ethernet frames between 802.11 wireless network (WLAN) and wired network (LAN). (One could relate AP to a Cable Modem, which is also a layer2 bridge device, but it bridges wired user devices' Ethernet frames between Ethernet network (LAN) and DOCSIS network).

Due to subscriber management requirements described in section 2.9, the traffic from the Wi-Fi Access Points will need to be routed to a centralized entity located on the wired network for subscriber management. The subscriber management capability may reside on the WLC, ISG, MAG or even the CMTS depending on the chosen architecture.

This means that the Wi-Fi user device must have layer2 connectivity upto that centralized entity through the AP, even if AP and the centralized entity are multiple hops away from each other and reachable via the underlying network. If the underlying network

[2] A non-bridging AP will allow the association of wireless user clients, but will not allow connecting to a wired network.

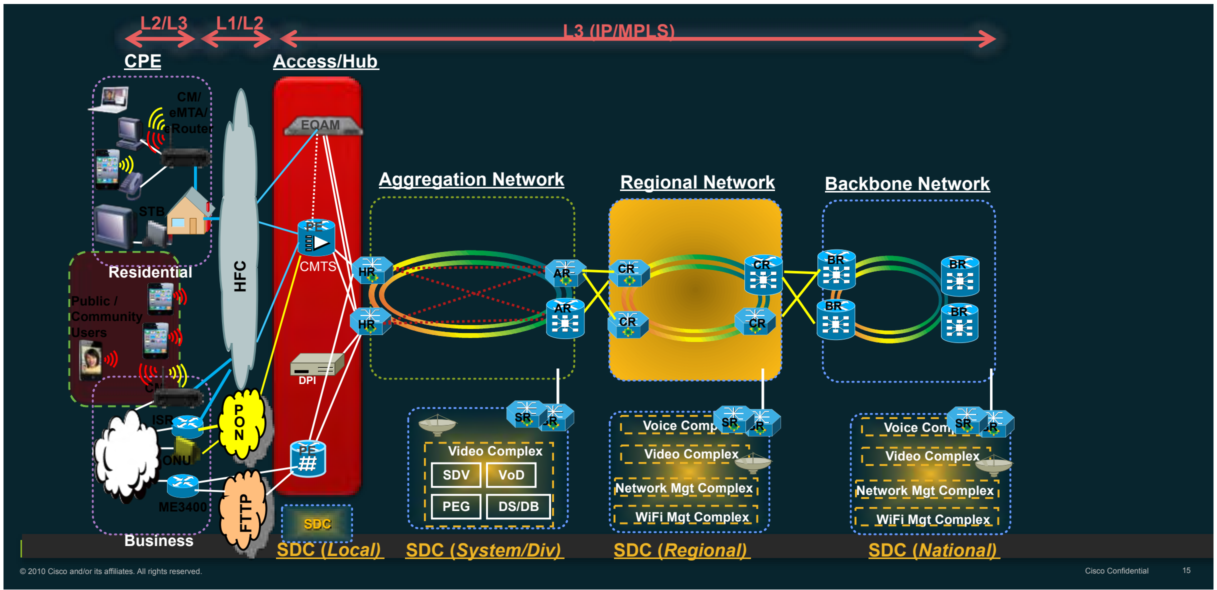
is a layer2 network (i.e. Ethernet bridged network), then it is relatively simpler to ensure the needed layer2 connectivity between user devices and the centralized entity. However, if the underlying network is a layer3 network (e.g. IP or IP/MPLS network comprising routers), then it can get complicated, depending one the chosen architectural approach (there are number of architecture approaches, as discussed later in this section).

Before we discuss various architectural approaches, it is important to put the MSO network in the perspective. The underlying network in the context of a cable MSOs is commonly a layer3 network in which CMTS (or CCAP) presents itself as the layer3 next-hop (as well as layer2 next-hop) to the user devices behind the standalone modems (e.g. CM, ONU) or embedded modems [eDOCSIS] (i.e. eCM) acting as the bridge.

the underlying network infrastructure must facilitate the bidirectional connectivity between the Wi-Fi user device and the centralized entity acting as the first IP next-hop, wherever that entity is located. This can be done in number of ways, based on the chosen architecture and requirements.

This section discusses such architectural options while keeping Cable MSOs' network infrastructure in mind. While this section focuses on DOCSIS access, it is well applicable to EPON access given the DPoE relevance. The following network transport architectural approaches are qualified for backhauling SP Wi-Fi traffic:

1. IP tunneling from AP
2. BSoD L2VPN
3. BSoD L3VPN



A reference Cable MSO network high-level diagram (not showing SP Wi-Fi elements) is shown in Figure 4.

As discussed earlier, if the underlying network is a layer3 network (e.g. IP or IP/MPLS network comprising routers), then

While each of the above architectural approaches are described in detail in the subsequent sections, the `Figure 5` below briefly illustrates them with their data plane specifics and how they relate to one of key AP capabilities:

There are number of options within this particular architectural approach that leverages IP tunneling from AP itself so as to tunnel the Wi-Fi traffic (either at layer2 or layer3) through the network.

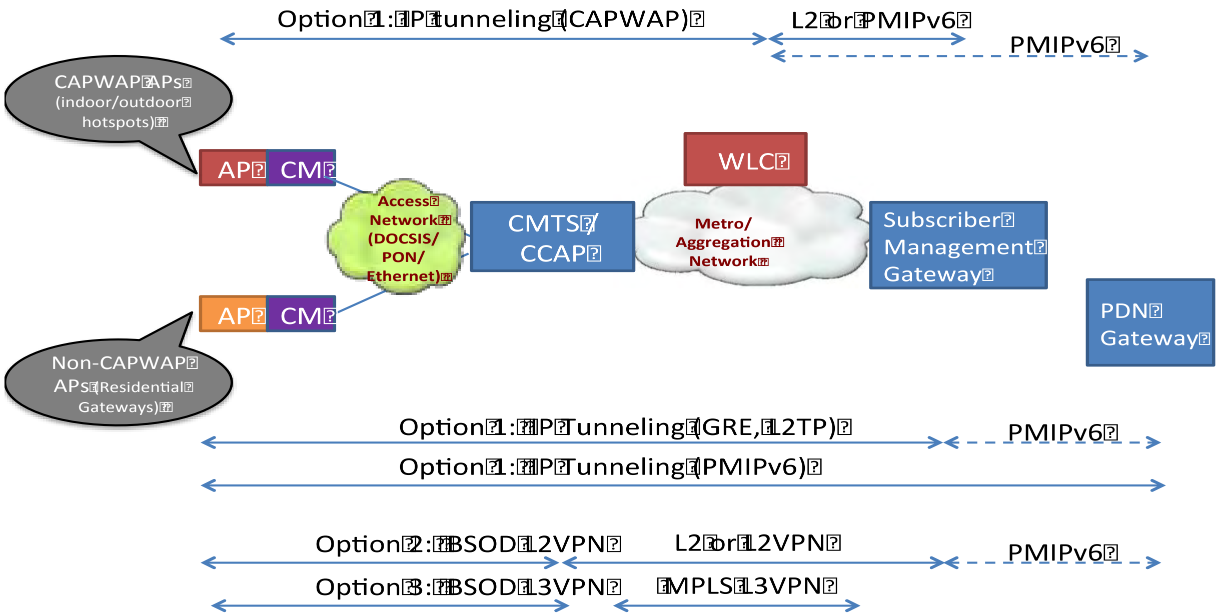

Option 1: IP tunneling (CAPWAP)
L2 or PMIPv6
PMIPv6
CAPWAP APs (indoor/outdoor hotspots)
AP
CM
WLC
Access Network (DOCSIS/ PON/ Ethernet)
CMTS/ CCAP
Metro/ Aggregation Network
Subscriber Management Gateway
AP
CM
PDN Gateway
Non-CAPWAP APs (Residential Gateways)
Option 1: IP Tunneling (GRE, L2TP)
PMIPv6
Option 1: IP Tunneling (PMIPv6)
Option 2: BSOD L2VPN
L2 or L2VPN
PMIPv6
Option 3: BSOD L3VPN
MPLS L3VPN

Figure 5 Network Transport Architectural Approaches - Data Plane

- CAPWAP APs: The traffic to/from AP is IP tunneled to the WLC using CAPWAP. The traffic between the WLC and Subscriber Gateway is a L2/802.1Q. PMIPv6 usage is optional.

- Non-CAPWAP APs: The traffic to/from AP is either tunneled over the network (option 1) or forwarded natively (option 2 or 3). PMIPv6 usage is optional.

The next section discusses each of the above network transport architectural options in details.

3.1.1 IP Tunneling from AP

3.1.1.1 PMIPv6

The architectural approach here is to build an over-the-top IP tunnel between AP and a remotely located centralized entity, using GRE over IP. In this approach, the data plane comprises "IPv4|v6 over GRE over IPv4|v6 over Ethernet [over DOCSIS (or PON)]" in the last-mile access and "IPv4|v6 over GRE over IPv4|v6" (over MPLS, if existed) in rest of the network (upto that centralized entity).

PMIPv6 is well standardized at the IETF [RFC5213] and [RFC5844]. PMIPv6 involves Mobility Access Gateway (MAG) and Local Mobility Anchor (LMA). LMA is defined to be the topological anchor point i.e. home agent for the Mobile Node's (e.g. Wi-Fi user device's) IP prefix(es) and manages MN's binding state via MAG. MAG manages mobility-related signaling for the MN that is attached to its access link. It is responsible for tracking the MN's movements to and from the access link and for signaling to the LMA.

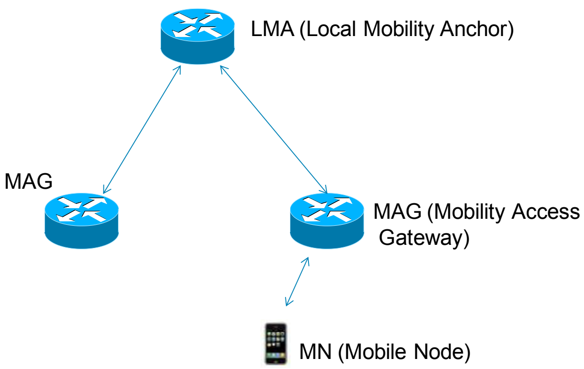
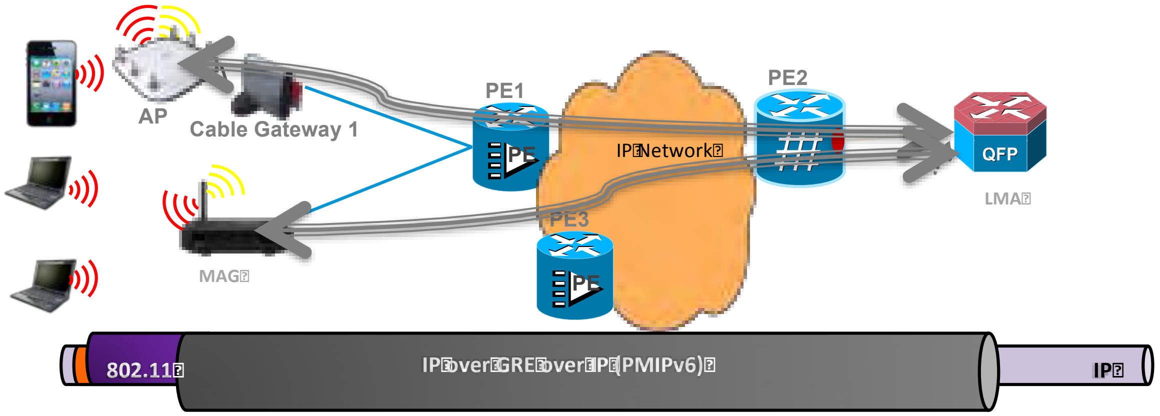


Figure 6 PMIPv6 Components

Error! Reference source not found. above illustrates PMIPv6 components.

While GRE over IP is commonly used tunnel mode, PMIPv6 also allows for other tunnel modes such as 'Ethernet over IPv6 over IPv6', Ethernet over UDP over IPv4 etc.



PMIPv6 is the only protocol that is claimed to qualify SP WiFi (with 802.1x/EAP) as the 'trusted non-3GPP access' and ensure mobility in every scenario.

Using PMIPv6 based architectural approach, an AP (acting as the MAG) uses PMIPv6 protocol messages to inform the LMA about the Wi-Fi user device (e.g. Mobile Node) getting attached. This allows AP/MAG and LMA to install (or update) the corresponding forwarding entries for the IP address assigned to the Wi-Fi user device. AP/MAG terminates user's layer2 and sends/receives user's IP traffic over the PMIPv6 tunnel. In other words, AP/MAG acts as the IP next-hop/gateway for the Wi-Fi user. While the Wi-Fi user is connected to AP/MAG at layer2, its IP address is anchored the LMA. This allows IP mobility, when the Wi-Fi user roams and changes AP/MAG attachments.

Figure 7 illustrates PMIPv6 tunneling applicability for SP Wi-Fi in sample MSO network topology.

It is important to highlight that instead of enabling PMIPv6 (MAG function) at the AP (as shown in this particular approach), it can be instead enabled on ISG, WLC or CMTS (as shown in other architectural approaches e.g. BSoD L2VPN) in an incremental manner for mobility.

The advantages of this approach are – (a) scales extremely well, (b) provides IP mobility for all scenarios, (c) integrates with 3GPP based cellular network

The disadvantages of this approach are – (a) requires MAG function as well as user management/control on AP/Modem – increased complexity on residential modems/gateways, (b) requires unique config file per modem for DS classification, (c) subjected to fragmentation and reassembly on last-mile access, (d) prohibits 5-tuple classification for QoS in the network, (e) results in sub-optimal multicast replication (e.g. network capacity wastage) if multiple user devices consume the multicast content

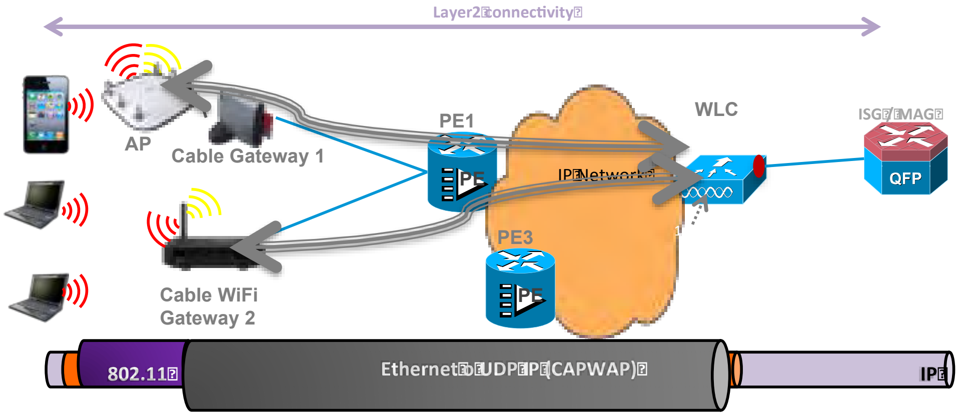
3.1.1.2 CAPWAP

The architectural approach here is to deliver the WiFi 802.11 traffic to a remotely located centralized entity e.g. Wireless LAN Controller (WLC), using UDP over IP. In this approach, the data plane comprises users' "Ethernet over UDP over IPv4|v6 over Ethernet [over DOCSIS (or PON)]" in the last-mile access and "Ethernet over UDP over IPv4|v6" (over MPLS, if existed) in rest of the network (upto WLC). UDP port is a well-known port 5247.

CAPWAP is well standardized at the IETF [RFC5415] and [RFC5416].

CAPWAP is a de facto protocol for Control and Provisioning of APs, and extensively used in most SP Wi-Fi deployments use-cases.

Figure 8 illustrates CAPWAP tunneling applicability for SP Wi-Fi in sample MSO network topology:



Using this approach, an AP establishes a CAPWAP tunnel (i.e. UDP over IP tunnel) with WLC (e.g. centralized entity). The 802.11 frames sent by the user device are forwarded by AP over the CAPWAP tunnel to WLC, which decapsulates the CAPWAP header and forwards the user device' IP packet using IP forwarding lookup. If the IP destination of the packet is another WiFi user device, then the IP packet is encapsulated in the 802.11 header and placed on the CAPWAP tunnel towards the appropriate AP. If the IP destination of the packet is on the wired network, then the IP packet is forwarded as usual.

The returning traffic gets subjected to the IP forwarding lookup, and gets placed on the appropriate CAPWAP tunnel, which is terminated at the AP. AP then delivers the 802.11 frames to the WiFi user device.

CAPWAP provides fragmentation and reassembly as per the path MTU discovery done by both AP and WLC, and allows for optional encryption using DTLS. CAPWAP also allows for PMIPv6 integration, as/if/when desired. This means that PMIPv6 elements (e.g. MAG and LMA) can incrementally be introduced, in which the MAG function can be enabled at the WLC.

The advantages of this approach are – (a) provides network administrators with a structured and hierarchical model to control & configure the APs, (b) controls hand-offs between AP during user roaming = foundation for mobility (c) works with layer2 or layer3 network, (d) allows 802.11 link-layer control, (e) works with NAT

The disadvantages of this approach are – (a) CAPWAP is not deemed useful for the residential APs, (b) network capacity wastage due to unnecessary multicast replication at WLC may happen if multiple user devices consume the multicast content

<u>3.1.1.3 GRE</u>

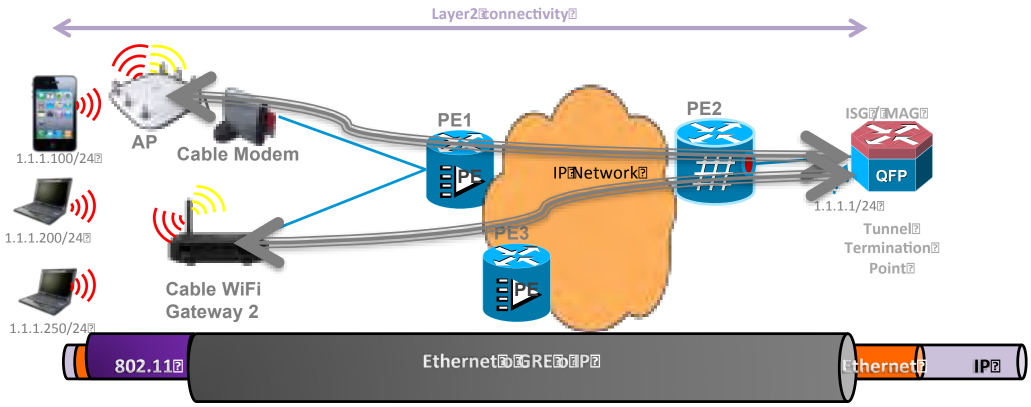
The architectural approach here is to build an over-the-top IP tunnel to deliver the Wi-Fi user device's Ethernet traffic between AP and a remotely located centralized entity (i.e. tunnel termination entity), using GRE. This approach requires IP connectivity between AP and the centralized entity. In this approach, the data plane comprises users' "Ethernet over GRE over IPv4|v6 over Ethernet [over DOCSIS (or PON)]" in the last-mile access and "Ethernet over GRE over IPv4|v6" (over MPLS, if existed) in rest of the network (upto that centralized entity).

> While Ethernet over GRE over IP usage is not well known or used, it is standardized at the IETF [RFC1771].

`Figure 9` illustrates GRE tunneling applicability for SP Wi-Fi in sample MSO network topology.

Using this approach, an AP establishes a GRE tunnel with the remote L2TP tunnel concentrator (e.g. centralized entity) and sends/receives Wi-Fi user device's Ethernet frames, over GRE (over IP) tunnel. It is important to note that GRE doesn't require a control channel and can be set up in a stateless manner without requiring any tunnel configuration.

The <u>advantages</u> of this approach are – (a)



maintains simplicity on AP or Gateways (b) scales well (if stateless tunneling is used, (c) maintains subscriber management/control at the remotely located centralized entity (e.g. tunnel termination point) based on IP, (d) provides IP mobility natively within the Layer 2 domain, (e) can integrate with PMIPv6 (by having the MAG function on the tunnel termination point) to provide macro-mobility.

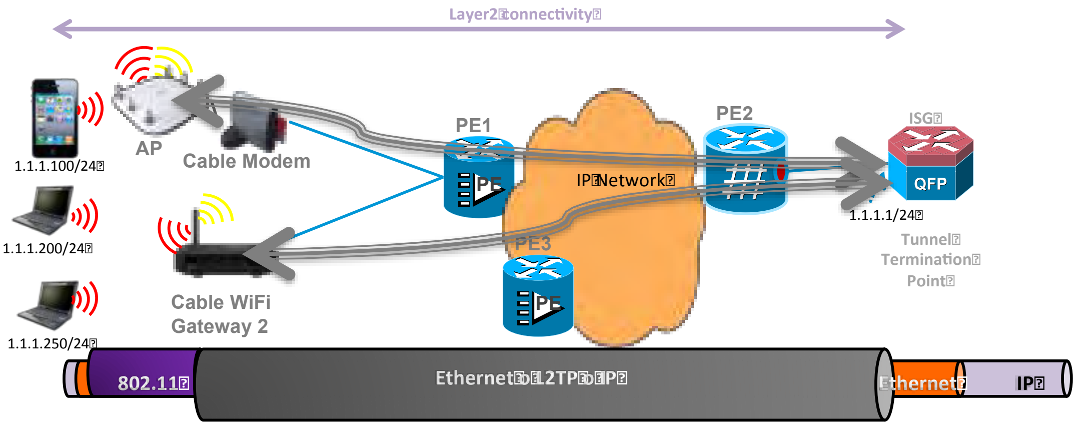
The disadvantages of this approach are – (a) does not integrate with 3GPP and doesn't provide mobility in all scenarios, (b) requires unique config file per modem for DS classification, (c) relies on IP tunneling, (d) subjected to fragmentation and reassembly on last-mile access, (e) prohibits 5-tuple classification for QoS in the network, (f) results in sub-optimal multicast replication (e.g. network capacity wastage) if multiple user devices consume the multicast content

3.1.1.4 L2TP

The architectural approach here is to build an over-the-top Layer 2 circuit (over IP network) to deliver the Wi-Fi traffic (e.g. Ethernet frames) between AP and a remotely located centralized entity (i.e. tunnel termination entity), using Layer 2 Tunneling Protocol (L2TP). This approach requires IP connectivity between AP and the centralized entity. In this approach, the data plane comprises users' "Ethernet over L2TP over IPv4|v6 over Ethernet [over DOCSIS (or PON)]" in the last-mile access and "Ethernet over L2TP over IPv4|v6" (over MPLS, if existed) in rest of the network (upto that centralized entity).

L2TPv2 is standardized at the IETF [RFC2661], whereas L2TPv3 is standardized at the IETF [RFC3931].

Figure 10 illustrates L2TP tunneling applicability for SP Wi-Fi in sample MSO network topology.



Using this approach, an AP establishes an L2TP tunnel with the remote L2TP tunnel concentrator (e.g. centralized entity) and sends/receives Wi-Fi user device's Ethernet frames, over L2TP (over IP) tunnel. It is important to note that L2TP requires a control channel to establish the tunnel.

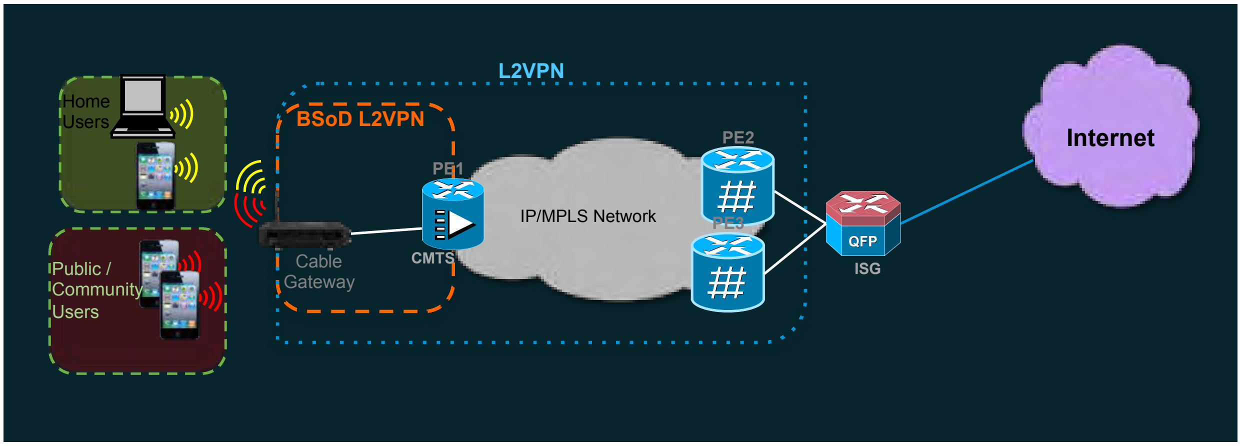
This architectural approach allows for PMIPv6 integration, as/if/when desired by the MSO to achieve mobility between Wi-Fi and Wi-Fi as well as cellular and Wi-Fi. This means that PMIPv6 elements (e.g. MAG and LMA) can incrementally be introduced in the

MSO network, in which the MAG function can be enabled at the L2TP tunnel concentrator.

The advantages of this approach are – (a) has its own control channel, (b) can make use of a cookie for added security

The disadvantages of this approach are – (a) does not scale (beyond few thousand tunnels), (b) requires unique config file per CM for proper DS classification, (c) does not integrate with 3GPP and doesn’t provide mobility in all scenarios by itself,

3.1.2 BSoD L2VPN



The idea in this architectural approach is very simple – use Layer 2 VPN to deliver the Wi-Fi traffic to a remotely located centralized entity at layer2 (without requiring any IP lookup). In this approach, the data plane comprises Ethernet [over DOCSIS (or PON)] in the last-mile access and Ethernet over MPLS (or just Ethernet) in rest of the network (upto the centralized entity).

Thankfully, Layer 2 VPN is a well known and well used option in many MSO deployments already, given that CableLabs standardized the Layer 2 VPN over DOCSIS in form of BSoD L2VPN [BSODL2VPN] and enabled many MSOs to use Layer 2 VPN to serve business customers with Metro Ethernet services (e.g. MEF (E-LINE, E-LAN, E-TREE), TLS etc.) when the VPN sites are attached to the HFC access. It is becoming quite useful for other purposes such as traffic separation for different services.

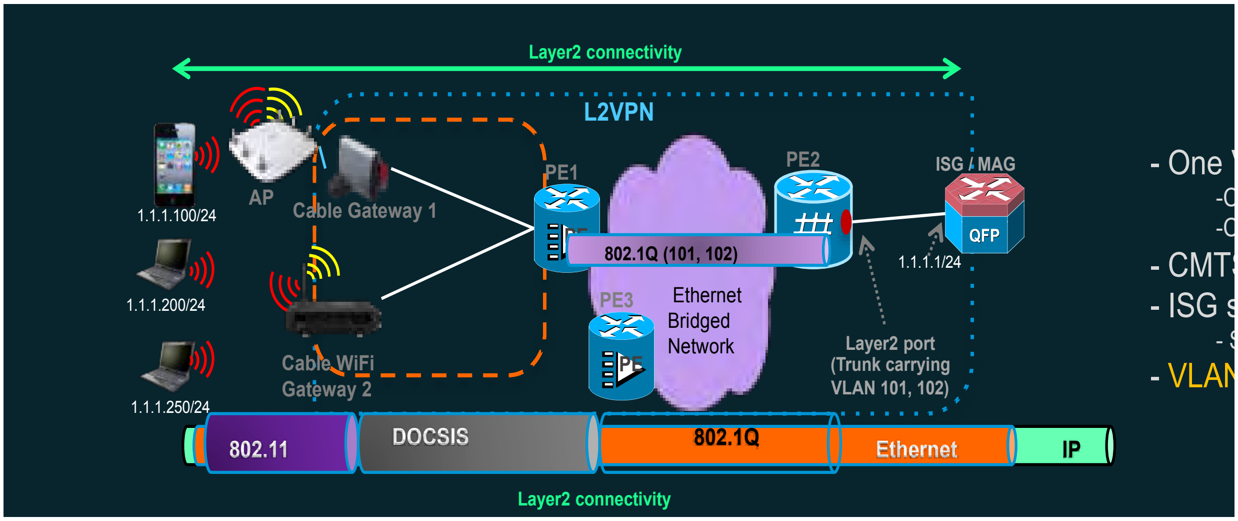
`Figure 11` illustrates L2VPN applicability in sample MSO network topology. It is important to note that the service-flows used for SP Wi-Fi (e.g. Public/Community users) are different from the ones used by the residential users. This automatically allows for traffic separation and IP prefix/address assignment separation between SP Wi-Fi users and residential users (throughout the network).

Using BSoD L2VPN, a CM is able to classify the upstream traffic (received from the AP) using SSID (in case of embedded CM) or VLAN (in case of standalone CM) present in the Ethernet frames, and forward the traffic over a particular DOCSIS service-flow (e.g. impose DOCSIS Header on the received Ethernet frame) to the CMTS. A CM is also able to forward the downstream traffic (received from the CMTS on a particular DOCSIS service-flow) to the AP (e.g. remove DOCSIS header and retrieve Ethernet frame).

Using BSoD L2VPN, a CMTS is able to forward the upstream traffic (received from the CM) on its uplink e.g. NSI towards the centralized entity, after removing the DOCSIS header and imposing an 802.1Q or 802.1AD or MPLS header, as per what MSO chose (and set in the config file). CMTS is also able to forward the downstream traffic (received from the network/centralized entity) after removing the 802.1Q or 802.1AD or MPLS header, to the Cable Modem on a particular DOCSIS downstream service-flow. It is important to highlight that the downstream Classification can be done by the CMTS without needing any CM config file dependency.

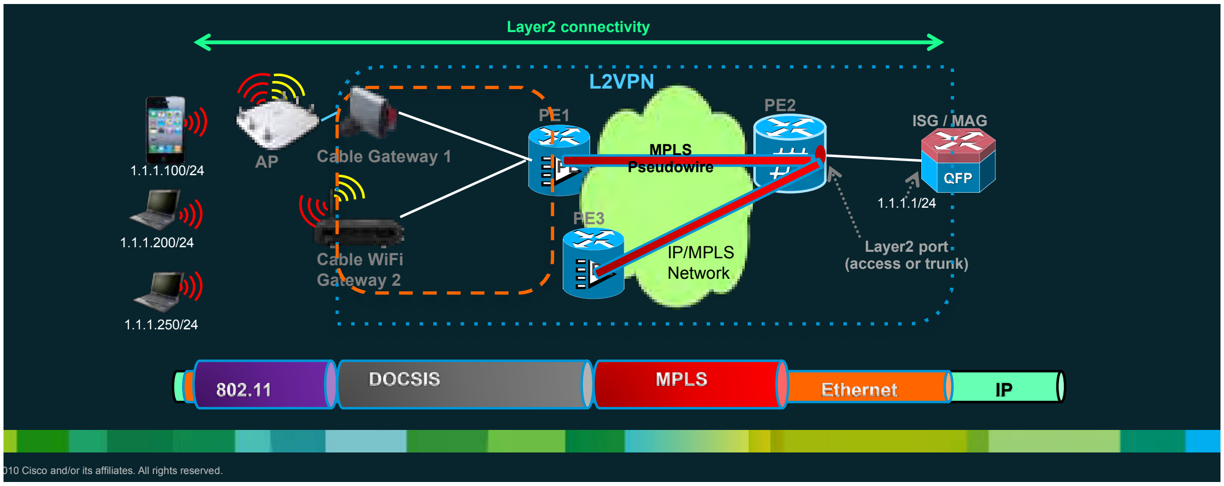
`Figure 12` illustrates using BSoD L2VPN using 802.1Q encapsulation variant. The figures below illustrate using BSoD L2VPN using MPLS encapsulation variants.

BSoD L2VPN does not require any tunneling from AP or CM, resulting in zero overhead on DOCSIS RFI, hence, avoiding any fragmentation/reassembly possibility, and also resulting in leveraging what's already supported in deployed MSO networks.



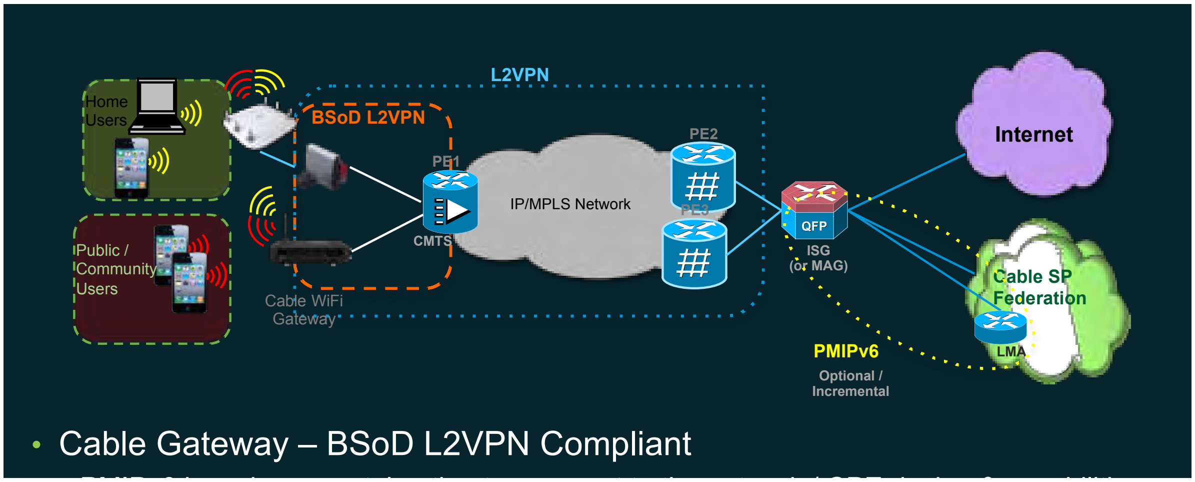
The CM config file includes TLVs that describe the mapping of one or more SFs with L2VPN designated for SP Wi-Fi. The config file does not need anything per-modem or AP specific to ensure the DS classification of the SP Wi-Fi traffic.

BSoD L2VPN with 802.1Q encap requires one VLAN per CM (if using P2P L2VPN) or one VLAN per network (if using P2MP L2VPN) for SP Wi-Fi.
BSoD L2VPN with MPLS encap



requires one MPLS pseudowire per CM (if using P2P L2VPN) or one MPLS pseudowire per CMTS (if using P2MP L2VPN) for SP Wi-Fi.

What's really nice about this architectural approach is that it allows for PMIPv6 integration, as/if/when desired by the MSO to infuse mobility during Wi-Fi and Wi-Fi as well as cellular and Wi-Fi handoff. This means that PMIPv6 elements (e.g. MAG and LMA) can incrementally be introduced in the MSO network without changing the existing L2VPN setup, as illustrated in Figure 14:



- Cable Gateway – BSoD L2VPN Compliant

The advantages of this approach are: (a) Works in the existing deployments, (b) downstream classification is possible without any config file dependency, (c) Separate traffic management for SP Wi-Fi users and residential users, (d) common config file pertaining to SP Wi-Fi for the CMs with P2MP L2VPN, (e) Seamless mobility in all scenarios is possible with PMIPv6 integration, as/if necessary, (f) requires no fragmentation/reassembly on the last-mile access = better data-plane throughput

The disadvantages of this approach are: (a) unique CM config file per modem if P2P L2VPN is used (note thathe upcoming BSoD L2VPN specification changes (CableLabs work underway) will no longer require unique CM config file, thanks to the dynamic discovery of remote PEs), (b) dynamic SF (e.g. DSx) support may not be available, (c) does not integrate with 3GPP and doesn't provide mobility in all scenarios.

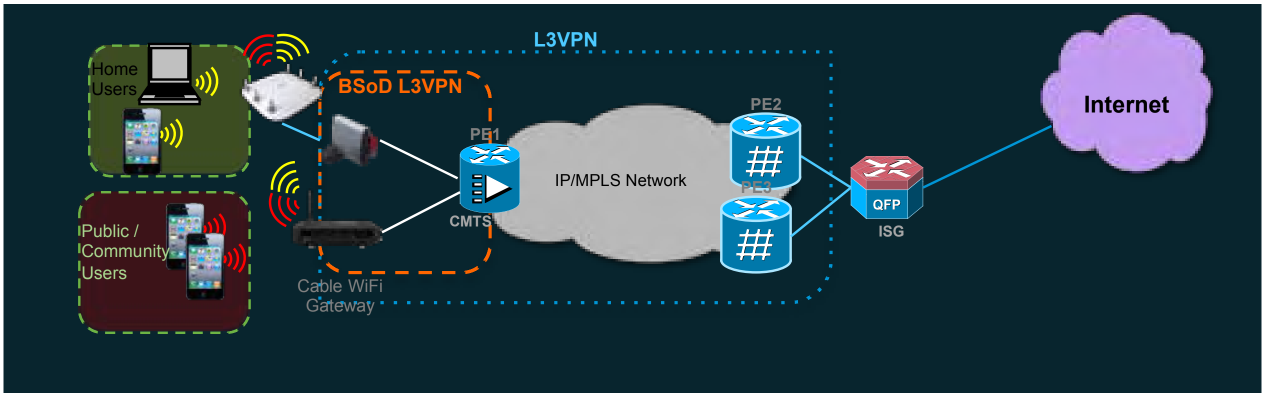
3.1.3 L3VPN

IP/VPN [RFC4364] is one of the most used technologies in SP networks (Wireline or Mobile) for internal purposes (e.g. network virtualization) and/or external purposes (e.g. Business L3VPN service).

This architectural approach allows the CMTS to terminate layer2 and use Layer 3 VPN to deliver the Wi-Fi traffic to remotely located centralized entity at layer3. In this approach, the data plane comprises 'IP over Ethernet over DOCSIS (or PON)' in the last-mile access and 'IP over MPLS' in rest of the network.

CableLabs standardization of L3VPN is underway (IP/VPN working group).

Figure 15 illustrates L3VPN applicability in sample MSO network topology.



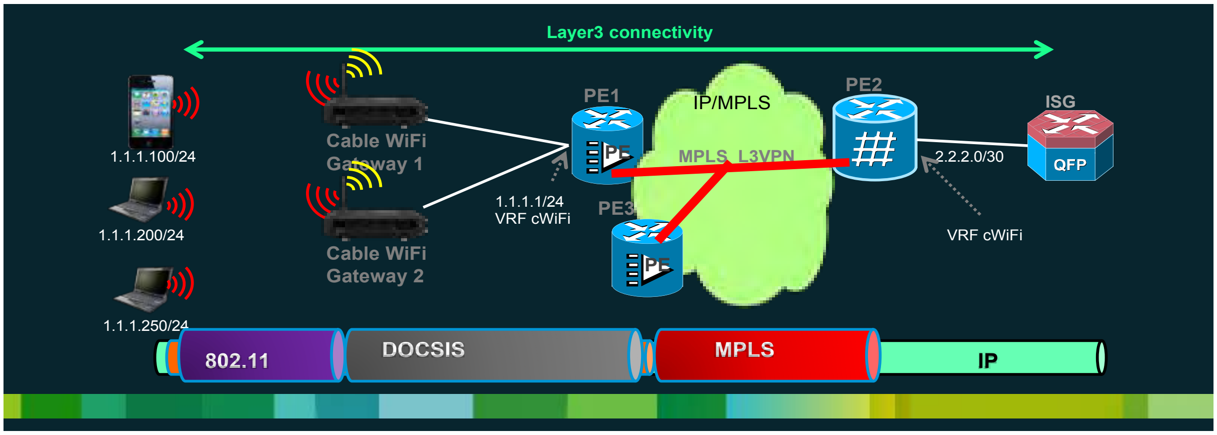
It is important to note that the service-flows used for SP Wi-Fi (e.g. Public/Community users) are different from the ones used by the residential users. This automatically allows for traffic separation and IP prefix/address assignment separation between SP Wi-Fi users and residential users.

A CM is able to classify the upstream traffic (received from the AP) using SSID (in case of embedded CM) or VLAN (in case of standalone CM) present in the Ethernet frames, and forward the traffic over a particular DOCSIS service-flow (e.g. impose DOCSIS Header on the received Ethernet frame) to the CMTS. A CM is also able to forward the downstream traffic (received from the CMTS on a particular DOCSIS service-flow) to the AP (e.g. remove DOCSIS header and retrieve Ethernet frame).

Using IP/VPN, a CMTS is able to forward the upstream traffic (received from the CM) on its uplink e.g. NSI to the network (or towards the centralized entity, if present), after removing the DOCSIS header and imposing an MPLS header. A CMTS is also able to forward the downstream traffic (received from the IP/MPLS network or centralized entity) after removing the MPLS header, to the Cable Modem on a particular DOCSIS downstream service-flow. It is important to note that the downstream Classification can be done by the CMTS without needing any CM config file dependency (e.g. per-CM or per-AP classifier).

The CM config file includes TLVs that describe the mapping of SFs with L3VPN designated for SP Wi-Fi (e.g. cWi-Fi in the figure above).

Figure 16 illustrates the data plane utilized when IP/VPN is used for SP Wi-Fi.



The advantages of this approach are: (a) CMTS could become the per-user policy enforcement point (with or without MAG function), (b) common config file pertaining to SP Wi-Fi for the CMs, (c) downstream classification is possible without any config file dependency, (c) the Wi-Fi traffic could follow the IP routing right from the CMTS, if needed, ((d) Wi-Fi users can be served by any DHCP server, (e) dynamic SF (e.g. DSx) support is available, (f) Seamless mobility is possible if the Wi-Fi user gets handed-off between APs served by the same CMTS

The disadvantages of this approach are: (a) Seamless Mobility is not possible all the time, since IP address preservation can not be guaranteed upon AP hand-off from one CMTS to another (without some additional complexity), (b) does not integrate with 3GPP, (c) cablelabs standardization not completed yet

Like L2VPN, L3VPN does not require any tunneling from AP or CM, resulting in zero overhead on DOCSIS RFI, hence, avoiding any fragmentation/reassembly possibility, and also resulting in leveraging what's already supported in deployed MSO networks.

3.1.4 Future Possibilities

In the previous section, although transport options are discussed as three discrete options, there are various other ways to achieve the requirements set out earlier. For example the benefits of PMIPv6 can be derived without the tradeoffs of tunneling by implementing the MAG in the network. Of course such an architecture brings its own set of tradeoffs. Similarly if subscriber management is implemented at the edge of the network it may eliminate the need for L2VPN/L3VPN architectures that are used to route traffic to a centralized entity. Such advanced architectures and solutions are outside the scope of this paper and are not discussed in any further detail here.

3.1.5 Comparison of Transport Options

The table below compares the three architectural approaches for network transport:

Table 1 Comparison of Various approaches

		IP Tunneling (from AP)	L2VPN	L3VPN
1	CableLabs Standardized	No	Yes	In progress[3]
2	Available	No[4]	Yes	Yes
3	Data Plane (Last-Mile Access)	User Ethernet frame over GRE|L2TP over IP over Ethernet over DOCSIS	User Ethernet frame over DOCSIS	User Ethernet frame over DOCSIS
4	Data Plane (Network)	User Ethernet frame over GRE|L2TP over IP (over MPLS)	User Ethernet frame over .1Q or .1AD or MPLS	User IP packet over MPLS
5	Overhead on Last-Mile Access	Yes	No	No
6	Requires Unique CM config file per Modem	Yes	Yes/No	No
7	User Awareness	ISG, MAG	ISG, MAG	CMTS or ISG
8	CMTS/CCAP Uplink/NSI needs?	IP	802.1Q Trunk, or IP/MPLS	IP/MPLS
9	DOCSIS Upstream Classifier?	IP Address	SSID or VLAN tag	SSID or VLAN tag
10	DOCSIS Downstream Classifier?	IP Address	MPLS label or VLAN tag	MPLS label or VLAN tag
11	DOCSIS Fragmentation & Reassembly (on CMTS, CM)	Yes	No	No
12	5-Tuple[5] based Classification by CMTS	No	Yes	Yes
13	5-Tuple based Classification by other routers	No	No	Yes
14	Mobility (WiFi-WiFi)	Yes	Yes[6]	Yes/No[7]
15	Mobility (WiFi-Cellular)	Yes	Yes	No
16	Accounting/DPI/LI possible at CMTS?	No	Yes	Yes

[3] CableLabs Standardization progressing in IPVPN Working Group

[4] Except L2TP, none of the IP tunneling variants seem to be available at the moment on the Modem / Gateway

[5] 5-Tuple = Src IP, Dest IP, Proto, Src Port, Dest Port

[6] May Require PMIPv6 Integration

[7] Seamless mobility as long as AP handoff doesn't change the CMTS.

4.0 CONCLUSION

A number of network transport options for SP Wi-Fi are discussed in this paper. Some of them are already deployed, whereas some of them are being considered for deployment.

The architectural options that help simplify the SP Wi-Fi architecture and harvest network intelligence would provide not only the cost-effectiveness, but also enable monetization opportunities. Monetization is where the next

REFERENCES

[Wi-Fi-Roam]CableLabs Wi-Fi Roaming Architecture and Interfaces Specification
[Wi-Fi-GW] CableLabsWi-Fi Requirements for Cable Modem Gateways
[eDOCSIS] CableLabseDOCSIS Specification
[DPOE] CableLabsDOCSIS Provisioning of EPON Specification 1.0
[BSODIPVPN] CableLabs IP VPN
[BSODL2VPN] CableLabs Business Services over DOCSIS Layer 2 VPN Specification
[Hotspot2.0]WFA
http://www.cisco.com/en/US/solutions/collateral/ns341/ns524/ns673/white_paper_c11-649337.html.
[RFC4364] IETF BGP/MPLS IP Virtual Private Networks (VPNs)

ABBREVIATIONS

AP	Access Point
BSOD	Business Services over DOCSIS
CAPWAP	Control and Provisioning of Wireless Access Points
CM	Cable Modem
CMTS	Cable Modem Termination System
DPI	Deep Packet Inspection
GRE	Generic Routing Encapsulation
LI	Legal Intercept
LMA	Local Mobility Anchor
MAG	Mobile Access Gateway
PMIPv6	Proxy Mobile IPv6
WLC	Wireless LAN Controller

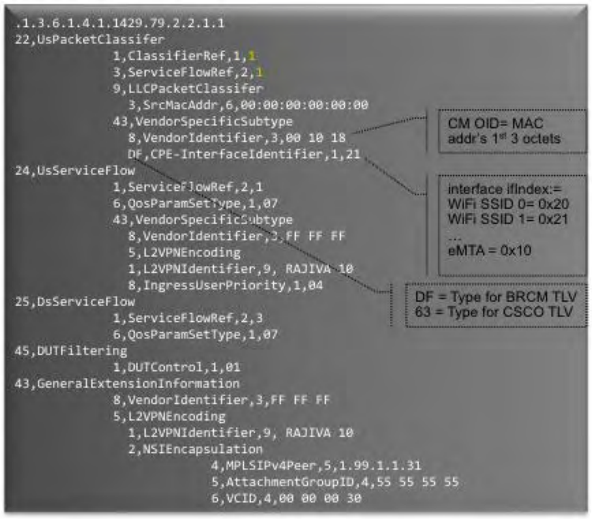
ACKNOWLEDGEMENTS

The author wishes to thank colleagues at Cisco for help and support in writing this paper.

APPENDIX

SP Wi-Fi using BSoD L2VPN – Sample Config file

A sample eCM config file for BSoD L2VPN having SSID-SF mapping is shown below



```
.1.3.6.1.4.1.1429.79.2.2.1.1
22,UsPacketClassifer
            1,ClassifierRef,1,1
            3,ServiceFlowRef,2,1
            9,LLCPacketClassifer
              3,SrcMacAddr,6,00:00:00:00:00:00
            43,VendorSpecificSubtype
              8,VendorIdentifier,3,00 10 18
              DF,CPE-InterfaceIdentifier,1,21
24,UsServiceFlow
            1,ServiceFlowRef,2,1
            6,QosParamSetType,1,07
            43,VendorSpecificSubtype
              8,VendorIdentifier,3,FF FF FF
              5,L2VPNEncoding
              1,L2VPNIdentifier,9, RAJIVA 10
              8,IngressUserPriority,1,04
25,DsServiceFlow
            1,ServiceFlowRef,2,3
            6,QosParamSetType,1,07
45,DUTFiltering
            1,DUTControl,1,01
43,GeneralExtensionInformation
            8,VendorIdentifier,3,FF FF FF
            5,L2VPNEncoding
              1,L2VPNIdentifier,9, RAJIVA 10
              2,NSIEncapsulation
                        4,MPLSIPv4Peer,5,1.99.1.1.31
                        5,AttachmentGroupID,4,55 55 55 55
                        6,VCID,4,00 00 00 30
```

CM OID= MAC addr's 1st 3 octets

interface ifIndex:=
WiFi SSID 0= 0x20
WiFi SSID 1= 0x21
...
eMTA = 0x10

DF = Type for BRCM TLV
63 = Type for CSCO TLV

Optimizing Wireless Networking of Wi-Fi and LTE

Marty Glapa, Amit Mukhopadhyay
Bell Laboratories, Alcatel-Lucent

Abstract

The proliferation of smart devices such as iPhones, DROIDs, tablets and others has resulted in huge increases in data traffic across cellular networks. These devices support multiple wireless technologies including 3G, 4G LTE and Wi-Fi. The massive adaptation of these devices can enable Wi-Fi to play a significant role in addressing the 3G/4G mobile data networks' increasing capacity requirements. Wireline and cable operators can both provide Wi-Fi offload for wireless operators. In this paper, we show how to optimize the performance and cost of heterogeneous networks comprised of cellular and Wi-Fi technologies.

INTRODUCTION

Most smartphones, tablets and PCs in today's world support Wi-Fi technology. While historically there have been hesitations on the part of wireless operators to embrace Wi-Fi as a complimentary technology to cellular, developments over the last several years in 3GPP interoperability have been breaking down the barriers (see, for example 1, 2). It's no longer an "either/or" discussion or debate but rather a complimentary use of both cellular and Wi-Fi technologies by operators to provide their end-users with optimized access to a rich set of services. Note that some Wi-Fi network deployments may include applications that drive Wi-Fi only traffic and not cellular traffic.

In this paper, we deal with the following key questions, which have not been commonly addressed, to the best of our knowledge:

- How much traffic can potentially be offloaded to Wi-Fi networks? This helps in sizing the Wi-Fi as well as the cellular network, since the latter now needs to carry only the remaining load.
- How do we overcome the well-known problems of interoperability between Wi-Fi and cellular networks, e.g., user authentication and admission control, mobility between the two networks, interference issues, guaranteed Quality of Service (QoS), etc.
- What are the best locations to deploy Wi-Fi hotspots? Access Point footprints are quite small compared to macro cellular footprints and deploying and clustering Access Points at the right locations, especially in high-traffic areas, is critical to the service provider for getting the most out of their investment and maintaining a consistent coverage footprint for nomadic users.
- How does the economics of combining Wi-Fi and cellular networks compare with the cellular network alone? A smart combination of Wi-Fi and cellular networks can keep the costs down and yet satisfy traffic demands.

Cable operators are in an excellent position to leverage their networks not only for using Wi-Fi as an extension of fixed access broadband services, but also in partnering with wireless service providers to use Wi-Fi and offload cellular network traffic.

We present a model to assess Wi-Fi offload potential in a network, based on applications, user behavior, etc. We show techniques of creating traffic density maps and identifying high traffic areas. Finally, we present a techno-economic model to compare the network options.

CONSUMERS, DEVICES AND TRAFFIC DRIVE "INTERWORKING"

A Wireless World

Nearly every mobile device in the foreseeable future will support multiple wireless technologies including 2G and 3G[1], 4G LTE[2] and Wi-Fi[3]. Wi-Fi plays a significant role in addressing the 3G/4G mobile data networks' increasing capacity requirements. Wi-Fi, in addition to providing a wireless extension for fixed wireline broadband, has emerged as a way to gain alternative connection to the 3G/4G cellular network services while off-loading data traffic from its radio access network (RAN). The complimentary use of LTE and Wi-Fi in providing wireless services enables the network operator to balance network and transport costs, while providing the consumer with services to meet their bandwidth needs.

Forces Driving Traffic Explosion

While early days of mobile data traffic primarily consisted of applications such as occasional web browsing, running search engines or instant messaging, today's mobile data traffic is dominated by richer applications such as video streaming, social networking and large file transfers. In many markets, voice and SMS traffic is being replaced by various web-based applications. Looking into the future, the five main applications for mobile data are considered to be cloud computing, different types of streaming, back-up and storage, full motion gaming and video communications.

There are several factors that are combining to trigger the mobile data explosion. On one end of the spectrum are technology factors like advancements in wireless technologies as well as end user devices. At the other end, cloud-based applications are encouraging social networking behaviors that were unthinkable of only a couple of years ago.

While early cell phone devices were not ideally suited for data communications, the introduction of QWERTY keyboards was the first game changer. Touch-screen phones brought on another round of evolution along with dramatic improvements in human-machine interfaces and software applications, all triggered by advancements in computing power and storage that can be packed in a small form factor. While PC data consumption on mobile networks remains high, the data usage by hand-held devices/tablets has been increasing sharply.

Figure1 below provides Bell Labs' projection of data traffic over the next several years.

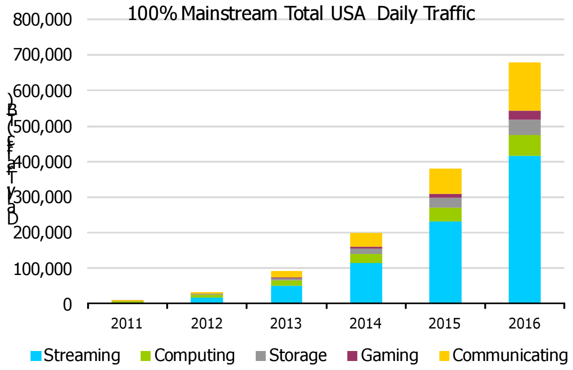


Figure1: Mobile Data Projections

Wi-Fi and LTE Applications

At the highest level, cellular networks can be used both as a mobile broadband solution, such as making a video call riding a train or bus, and as a non-mobile broadband solution, such as sitting in the backyard watching a video clip. Wi-Fi, on the other hand, is primarily used today as an extension to fixed broadband solutions. Table 1 below

[1] 2nd and 3rd Generation wireless standards that use licensed spectrum for Wide Area Networks.

[2] Long Term Evolution, a 4th Generation (4G) wireless standard that uses licensed spectrum for Wide Area Networks.

[3] A wireless technology that uses unlicensed spectrum for Local Area Networks.

illustrates the common types of applications vis-a-vis technologies.

Applications	Cellular	Wi-Fi
Fixed	Yes	Yes
Nomadic	Yes	Yes
Mobility	Yes	Very limited

Table 1: Applications & Technologies

Additional discussions describing key characteristics of cellular and Wi-Fi technologies are provided below.

Cellular

- 3G/LTE enables a high speed data connection to services when a user is mobile, in a fixed location, or when Wi-Fi is not available in a wireline broadband extension (fixed location) scenario.
- Cell site serving areas of several Km, depending on antenna height, location and geography; coupled with complex robust mobility algorithms; these help facilitate effective mobility hand-offs at vehicular speeds.
- A comprehensive security framework maintains secure connections and enables fast handoffs.

Wi-Fi

- An extension of wireline broadband via radio for “the last 100m”. This includes the use of Wi-Fi hotspots in public locations, homes and enterprises.
- Data offload of licensed spectrum RAN networks using radio for “the last 100m” and offloading to broadband wireline connections.
- Nearly every mobile device and broadband modem today has built-in Wi-Fi capabilities. Many devices today can automatically search for available hotspots or can even themselves serve as Wi-Fi hotspots for other Wi-Fi devices.

Using Wi-Fi in Real-Time Mobile Applications

There are major challenges of using Wi-Fi in a real-time mobile solution in an uncontrolled public environment. These are:

- Interference – Wi-Fi uses unlicensed spectrum; a limited number of overlapping channels and uncoordinated neighboring Access Point deployments and spectrum used by competitive providers or even residential or enterprise users. This can result in interference, which in turn can limit capacity, mobility and service continuity.
- Mobility – Wi-Fi is intended as a short range wireless solution. Mobility is limited to slow pedestrian speeds. Wi-Fi mobility is defined in IEEE standards 802.11r and is generally supported within major vendor products. Not all vendors have implemented 802.11r and it is not clear whether 802.11r will be required for Wi-Fi Alliance certification. IETF is also involved in defining Wi-Fi mobility with RFC3990. Mobility at vehicular speeds is impractical due to small wireless coverage areas of Access Points, the challenges associated with hand-offs and admission control, and a lack of algorithms needed for service continuity. There is also CAPWAP, which is an IETF standard defined in RFC3990, which addresses mobility.
- Admission control (in the form of resource management) – at the time of a session handover from one Access Point to another Access Point. In the worst case, it would be similar to starting a new session for best effort traffic, though there is separate signaling that is used in 802.11 for resource reservation. Major vendors are addressing this and some may be providing seamless handoffs.
- Re-association with the target Access Point – requiring a large number of roundtrips for authentication. Security throughout mobility events, like handoffs, is not maintained, and has to be fully re-

established. 802.11r may help in reducing the number of roundtrips for the delay.

- Radio resource management granularity limits the ability to share channels between many users. This limitation is generally not noticeable in fixed and nomadic applications; it is an impediment for dense use mobile applications.
- Propagation characteristics at 2.4GHz ISM band are subject to significant interference; at 5.1GHz, signal strength fades away rather quickly resulting in smaller cell ranges and the device eco-system is still developing.

Addressing Key Wi-Fi Challenges

3GPP, working with other industry bodies, has developed two fundamental approaches for integrating Wi-Fi with cellular technologies. Figure 2 shows the architecture where the cellular operator has no control over the Wi-Fi Access Point, and Figure 3 shows the architecture when the operator has full control over the Access Point.

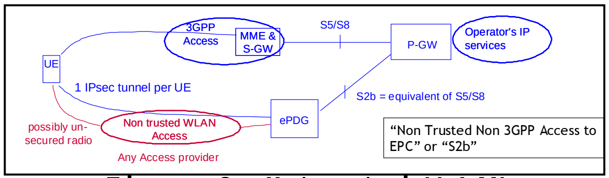


Figure 2: Untrusted W-LAN

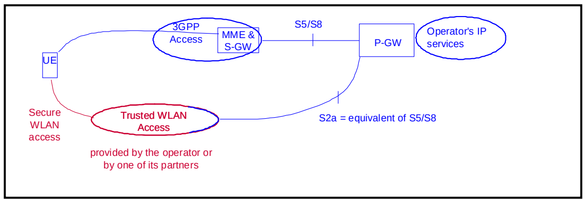


Figure 3: Trusted W-LAN

Additional developments in 3GPP continue in the form of initiatives like Access Network Delivery Selection Function (ANDSF) where the cellular network assesses the quality of experience in the Wi-Fi and Cellular networks for given applications and based on policies, may switch the user from one technology to the other. 802.11u also defines another way to achieve this. HotSpot 2.0 and Wi-Fi Alliance activities not only enable Wi-Fi roaming among operators but also open the doors for further integration between Wi-Fi and cellular networks.

OPTIMIZING WI-FI AND LTE NETWORKS

Traffic Offload to Wi-Fi

A significant part of mobile data traffic is considered nomadic and not necessarily mobile, thus making many cellular users amenable to Wi-Fi offload. It is likely that Wi-Fi offload may grow today from roughly 22% of traffic in North America to over 30% within the next four years. The amount of offload will depend upon various factors like residential broadband penetration; ubiquity of public Wi-Fi hotspots, mobile data tariffs, and technology evolution for seamless Wi-Fi-cellular integration etc., the potential for offload could be greater than 70% as seen from various studies in certain international markets.

Optimizing Wi-Fi Hot Spot Locations

Wi-Fi offers good user throughput in an interference-free environment. Figure 4 provides a comparison between different technologies, based on 3GPP simulations. The Wi-Fi value is based on typical environment for today's 802.11b/g deployment with 20 MHz channels. Pure 802.11n environment is expected to achieve 50 Mbps+ average user throughput with a 40 MHz channel.

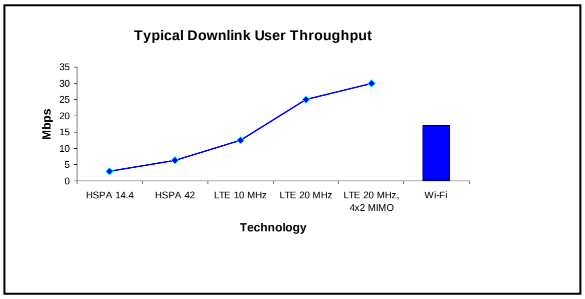


Figure 4: Throughput Comparison

However, Wi-Fi Access Points have small coverage areas, compared to macro cells. A comparison between technologies of capacity per unit area of coverage for typical dense urban environment is shown in Figure 5 below. While the cell range for a typical macro cell is 1.2 – 1.5 km in an urban environment, typical Wi-Fi Access Point range is around 30m and generally not too much more than 100m. Typical downlink sector throughput for 3G HSPA is around 6.7 Mbps whereas for 20 MHz LTE, it can be around 30 Mbps. For 802.11g, typical downlink user throughput is around 17 Mbps.

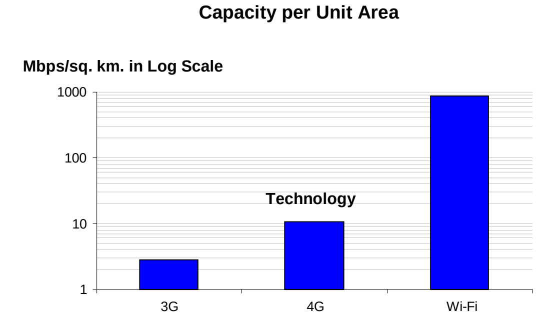


Figure 5: Unit Area Capacity

The challenge, thus, becomes how to enable maximum traffic offload Wi-Fi hotspots. Bell Labs analysis from various real networks has shown that a relatively large volume of mobile data traffic (50% - 60% or more) is often contained in a relatively small geographical area (10% - 15%) under a macro cell coverage area. Figure 6 shows the relationship between geographical area and amount of traffic during busy hour in a macro cell in a large North American city – in this particular example, only about 8% of the geographical area contains 60% of mobile data traffic.

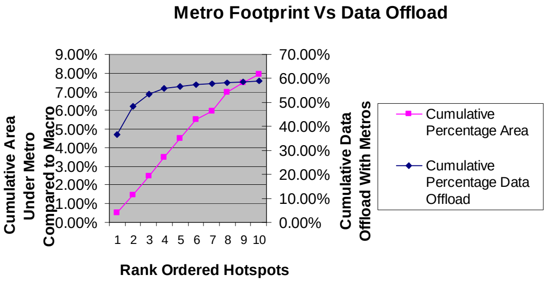


Figure 6: Traffic Density

To help address this challenge, industry techniques have been developed to create traffic density maps. This helps make a well-informed decision on placement of the Wi-Fi Access Points.

Techno-Economic Analysis

While the cost of a consumer Wi-Fi Access Point is almost negligible compared to the cost of cell site equipment, the cost of carrier grade Access Points is significantly higher than the cost of consumer Access Points. First, environmental hardening and security costs add significantly to capital expenses. Additionally, ongoing costs of backhaul and site rental significantly impact the Total Cost of Ownership (TCO). But overall, carrier grade Access Point costs are lower than macro cell site equipment.

Whether Wi-Fi deployment is economical or not depends upon a wide range of factors, including technical as well as commercial factors. In Figure 7 below, we provide a simple normalized cost comparison, using a subscriber’s monthly usage as a reference.

The cost points are used from a typical large wireless operator in Europe. The reference coverage area is the footprint of a macrocell in a large European city. It may be noted that the cost points for the Wi-Fi Access Points are for environmentally hardened network elements as required for outdoor deployment, which are significantly higher than indoor Access Points.

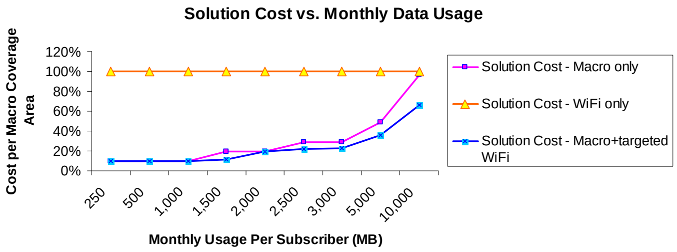


Figure 7: Technology Cost Comparison

The figure clearly shows that covering the entire macro footprint with Wi-Fi Access Points is an impractical solution. A macro-only solution is suitable for low data usage per subscriber but as the traffic per subscriber increases, macro complemented with targeted Wi-Fi deployment becomes the cost-optimal solution.

SUMMARY

Wi-Fi and LTE each have their own set of applications, but are most importantly complimentary:

- Wide area coverage with full mobility using licensed spectrum base stations with higher power and operating via higher towers (e.g., LTE) as a compliment to lower power unlicensed spectrum street level or campus environment deployments (Wi-Fi).
- Coverage and capacity limited network design that is independent of local deployments of other WLANs where the design can be impacted negatively if another WLAN is deployed nearby.
- Effective offloading of data traffic from congested cellular networks can be achieved by transporting this traffic over wireline and Wi-Fi facilities while enabling the user to enjoy the rich applications provided by these networks.
- Roaming capabilities and common authentication methods using the Wi-Fi network are adopted and certified by the Wi-Fi Alliance4.
- Careful identification of dense traffic areas and locating Wi-Fi Access Points at those locations is a key to efficient cellular-Wi-Fi integration.

Cable operators can leverage their networks not only for using Wi-Fi as an extension of fixed access broadband services, but also in partnering with wireless service providers to use Wi-Fi for offloading cellular network traffic. Optimizing the performance and cost of heterogeneous networks comprised of cellular and Wi-Fi technologies is critical for performance, customer satisfaction and cost management.

ACKNOWLEDGMENTS

The authors would like to thanks our Bell Labs colleagues Perminder Dhawan, Mohamed El-Sayed, Bob Novo and Carlos Urrutia-Valdés for providing valuable references and critique of this document.

REFERENCES

1. 3GPP TS 22.234, Requirements on 3GPP system to Wireless Local Area Network (WLAN) interworking
2. 3GPP TS 23.234, 3GPP system to Wireless Local Area Network (WLAN) interworking; System description
3. 3GPP TS 23.402, Architecture enhancements for non-3GPP accesses
4. 3GPP TS 24.312, Access Network Discovery and Selection Function (ANDSF) Management Object (MO)

[4] HotSpot 2.0 standardization effort and certification is expected in mid 2012.

WI-FI NETWORKS IN HFC CABLE NETWORKS - HOW TO UNDERSTAND AND MEASURE USER QUALITY PROBLEMS IN A WIRELESS ENVIRONMENT

HUGO RAMOS, NET Servicos SA

An initial effort to try to bring the quality of service of our carrier grade high speed data node based service from cable network, to a cell wireless network, presenting the concept of WUEQi (Wireless User Experience Quality index).

ABSTRACT

In this paper, is presented the concept of WUEQi, an index like a MOS index, to give ability for cable operators that are deploying Wi-Fi networks to have better control of this technology to deliver services. Looking basically to give to the users a better user experience and to the operator a good way to plan the expansion of the Wi-Fi network.

Keywords
Wi-Fi, Wireless networks, CATV, Strand mount AP, average power, EIRP, antennas, user experience, DOCSIS.

1. INTRODUCTION

Today with the dissemination of a new generation of very powerful mobile devices such as smartphones and tablets the way of our society communicate and consume information is changing, creating an exponential rise in the importance of mobility services, with the anything, anytime, anywhere concept. With this in mind a lot of cable MSO are dealing with a new and different ways to deliver mobile services to their customers with the deployment of wireless networks over the cable HFC plant.

With the proliferation of mobile devices the bandwidth requirements in the traditional mobile networks has also change pushing these mobile operators to find new ways of deliver these services with adoption of new technologies and that search can be a very good opportunity for the cable industry as well, since the deployment of this networks over the cable HFC plant can be very fast and efficient.

One technology that is getting a very big attention of the market right now is Wi-Fi. The Wi-Fi standard today is undoubtedly the most heavily used wireless technology in terms of number of devices that is capable of use the technology and as the amount of data traffic transmitted over networks using this wireless technology. The technology is very mature and in any report that you have access to, it is set that around the world, millions and millions of devices such as smartphones, tablets, netbooks, notebooks, TVs, STBs, home gateways, even cars and refrigerators has embedded and certified chipsets with this technology and this number is growing, actually skyrocketing each day. That is why deploy a Wi-Fi network is a very attractive option to any convergent operators to give fixed broadband customers the mobility that they look forward.

But unlike from the common sense thinking that says the Wi-Fi technology is a very mature technology, this standard is not mature yet to deploy a service provider type of network. This is very easily pointed by the fact that to gain the scale of number of devices, this standard utilizes non license frequencies and since it is inherent to a wireless network (different to a cable HFC that is confined cable network) this type of solution deals with devices that are moving, unstable customer demands and different device power transmission that changes the way of get to know, understand and solve the problems that affect the most important thing at the end that is the user experience.

Since motivators of this deployments, in cable could be retention and churn reduction, 3G offload to traditional mobile operators and location based services, user quality experience is became very important to the SP to provide the service and the use of analytics, throughout the WUEQi, could be the answer to try to have a little control to the chaos of delivering Wi-Fi mobile services to a customer that is used to have the quality of service of our carrier grade high speed data node based service from cable network.

This paper is organized as follows. In the next section we present the advantages of deploy a Wi-Fi network over cable HFC plant. In section 3.Important concepts about Wi-Fi networks deployment (Link budget calculation). Section 4. Show the challenges to deliver good and control user experience to the chaos of a non-license frequency wireless network, the problems about different link budgets with different devices and the concept of WUEQi.

2. ADVANTAGES OF WI-FI NETWORK IN A CABLE ENVIRONMENT

There are three big obstacles that service providers which intend to deploy a Wi-Fi network will probably face, named: reliable powering the access point units, make the backhaul link with Internet traffic to the APs and acquiring mounting sites. These three obstacles can make a huge difference in deploying big networks more efficient. HFC cable networks are already design and build with reliable power supply that can feed the Wi-Fi access point, of course if the access point is capable of handle the power from HFC plant and there is some units at the market that can not only handle but also being mounted in strand, solving the second issue. Also in a cable network the IP connectivity can be delivered in the same cable that is connected to feed the AP throughout the high-speed data DOCSIS network that is already in place. With this issues solved in HFC networks, cable operators will can quickly and easily deploy this wireless networks in the existing plant.

So as shown above, the implementation of a Wi-Fi network in an HFC plant is faster, easier and more efficient but something that we are not used to work and it is a big problem to control is the wireless access part mainly in deal with average transmitted power (TX power from a device) that is going to be translate to coverage and differ from the DOCSIS very much since the cable modem is easily controlled and does not move around. This problem is going to reflect in a bad user experience

3. IMPORTANT CONCEPTS ABOUT WI-FI NETWORKS DEPLOYMENT

3.1 Link Budget

A lot of questions are important when designing a Wi-Fi network such as:

1. What type of devices I want to provide the service?
2. What type of services I want to deliver?
3. What type of environment I am going to deploy the network?

So, with this in mind you can find out how to calculate the link budgets to provide services to the devices that you chose taking in consideration the environment and the services that you want to provide.

In our case we are going to consider 3 devices to provide the services, a typical notebook, a typical table and a typical Smartphone. Of course in a real scenario we are not going to see typical devices but real information that can be get from devices providers, FCC test and in our case ANATEL(Brazil Telecom Agency). Our environment will be a very dense and populated metropolis with a lot of buildings and reflections.

To determine the link budgets we should consider the usage of some models. For a free space, line of sight environment, the propagation law would result in square law decay. i.e. for each doubling of distance the power drops by 6dB. The free space model is only an idealized model and the real world is not line of sight. There are objects in the environment such as trees, buildings, poles that signals reflect from and with this, the signal can add and subtract to produce a signal that decays faster than the square law.

Below we show a typical configuration of some device that we are going to consider calculating the link budgets.

	AP	Typical Notebook	Typical Tablet	Typical Smartphone
Average Power (A) dBm	21	16	17.5	16.0
Antenna Gain (B) dBi	5	6	2.0	0.0
EIRP (A+B) dBm	26	22	19.5	16.0

Table 1. Typical power of devices

Using some models we can show below the link budget in an urban and very dense city environment.

Link Budget between an AP and a typical notebook:

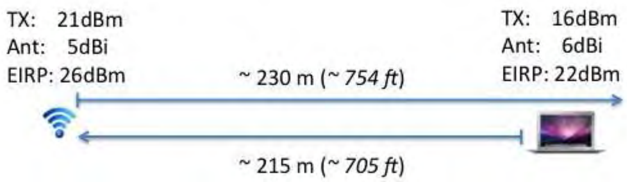


figure 1. Link budget – AP and notebook

So in this case the link budget is limited in 215 meters from the notebook to the AP.

Link Budget between an AP and a typical tablet:

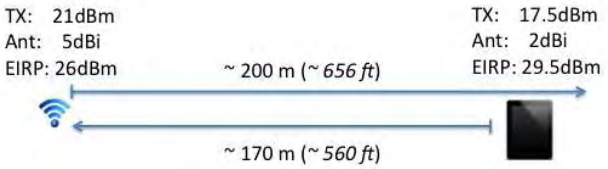


figure 2. Link budget – AP and tablet

So in this case the link budget is limited in 170 meters from the tablet to the AP.

Link Budget between an AP and a typical smartphone:

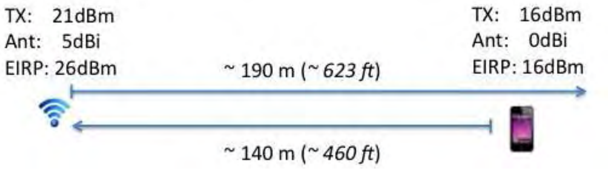


figure 3. Link budget – AP and smartphone

So in this case the link budget is limited in 140 meters from the smartphone to the AP.

This will show to us that the limit of the link budget will be defined by the transmission from device to the AP in upstream direction rather than in downstream direction since the average power transmitted in a downstream direction is bigger than the upstream direction.

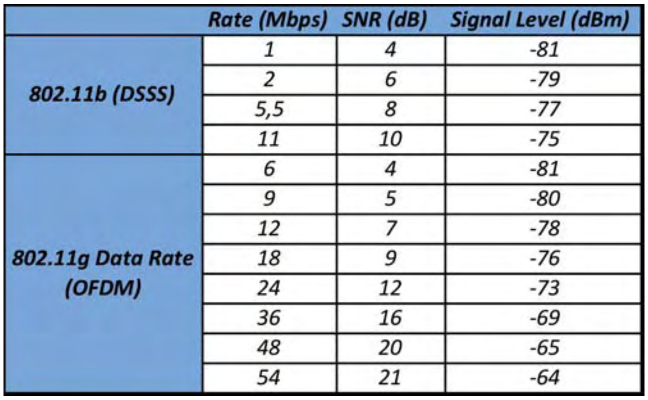
3.2 Relation between type of services and SNR

Since the Wi-Fi is a system that uses adaptive modulation, the modulation nominal rate is dependent of the SNR of the link budget that is the difference from the noise floor (typically -90dBm in our city) to the signal that are receive by the device and the access point. Of course this SNR will make our user experience better or worse so we have to consider this in our index.

Below we show a table of typical SNR versus the type of service that is possible to use.

Service	SNR
Video 99% sucess	>= 35 dB
Voice 99% success	>= 25 dB
Web surfing 99% success Voice 90% sucess	>= 20 dB
Email – Web surfing 90% success	>= 7 dB

Table 2. Tested Services vs. SNR



	Rate (Mbps)	SNR (dB)	Signal Level (dBm)
802.11b (DSSS)	1	4	-81
	2	6	-79
	5,5	8	-77
	11	10	-75
802.11g Data Rate (OFDM)	6	4	-81
	9	5	-80
	12	7	-78
	18	9	-76
	24	12	-73
	36	16	-69
	48	20	-65
	54	21	-64

Table 3. Theoretical table data rate vs. SNR

4. THE CHALLENGES TO DELIVER GOOD USER EXPERIENCE

Since the limit of the link budget will be defined mainly by the transmission from the device to the AP some situations can occurs like a smartphone see the network and could not associate (getting a fail association status a the system) or have a small SNR, having a bad user experience, in a place where a notebook can associate.

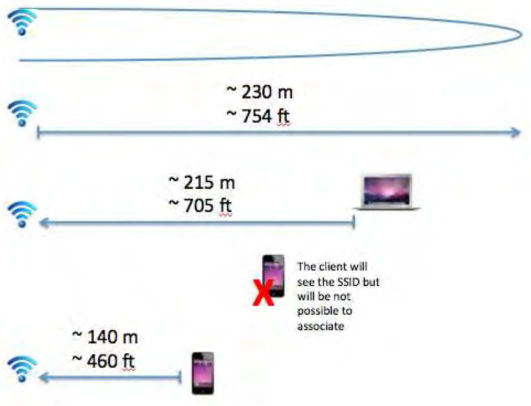
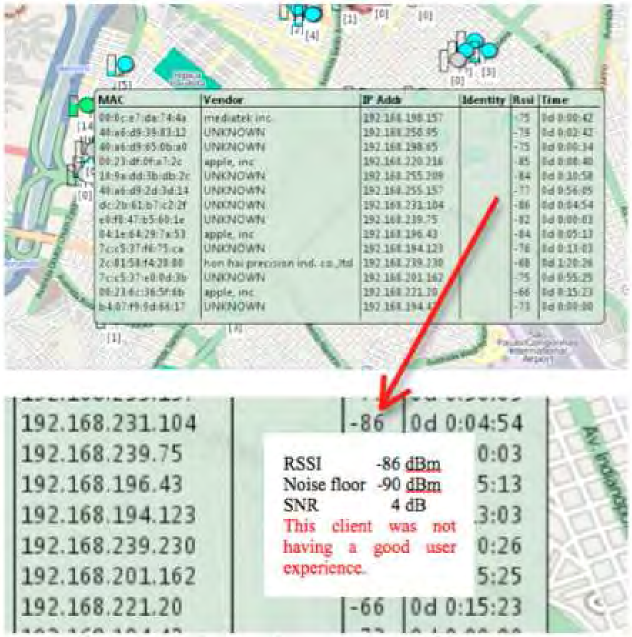

~ 230 m
~ 754 ft
~ 215 m
~ 705 ft
The client will see the SSID but will be not possible to associate
~ 140 m
~ 460 ft

figure 4. Association problem



MAC	Vendor	IP Addr	Identity	Rssi	Time
00:0c:e7:da:74:4a	mediatek inc	192.168.198.157		-75	0d 0:00:42
40:a6:d9:39:83:12	UNKNOWN	192.168.250.95		-78	0d 0:02:42
40:a6:d9:65:0b:a0	UNKNOWN	192.168.198.65		-75	0d 0:00:34
00:23:df:0f:a7:2c	apple, inc	192.168.220.216		-85	0d 0:08:40
18:9a:dd:3b:db:2c	UNKNOWN	192.168.255.209		-84	0d 0:10:58
40:a6:d9:2d:3d:14	UNKNOWN	192.168.255.157		-77	0d 0:56:05
dc:2b:61:b7:c2:2f	UNKNOWN	192.168.231.104		-86	0d 0:04:54
e0:f8:47:b5:60:1e	UNKNOWN	192.168.239.75		-82	0d 0:00:03
84:1e:64:29:7a:53	apple, inc	192.168.196.43		-84	0d 0:05:13
7c:c5:37:f6:75:ca	UNKNOWN	192.168.194.123		-76	0d 0:13:03
2c:81:58:f4:28:80	hon hai precision ind. co.,ltd	192.168.239.230		-68	0d 1:20:26
7c:c5:37:e0:0d:3b	UNKNOWN	192.168.201.162		-75	0d 0:55:25
00:23:6c:36:5f:6b	apple, inc	192.168.221.20		-66	0d 0:15:23
b4:07:f9:9d:66:17	UNKNOWN	192.168.194.4		-73	0d 0:00:00

RSSI -86 dBm
Noise floor -90 dBm
SNR 4 dB
This client was not having a good user experience.

figure 5 – Real network management with customers

The figure 5 above show a particularly customer that was not having a good user experience, what we notice was that this customer could associate, was able to get an IP address (192.168.231.104) but he was not having a good throughput. In this example, we noticed that in most cases the algorithm of the device shown at the screen two of three bar of signal giving the erroneous impression to the user that it has sufficient signal strength to associate or have a good user experience but unfortunately this is not true.

In other cases we also notice another customers with less SNR was not be able to get an IP address because the throughput was too low and another that could not accomplish to associate and generate a trap registering that he had a fail association.

Because of this less control of the CPE and the type of the business models delivered where the product is offered free of charge to the subscriber as an extension of their fixed broadband service, there is a natural tendency to think that if the client can not connect to the Internet in a place where the Wi-Fi network is broadcast, there is no problem because the customer will get another form of connection to solve their connection issue. This will cause a bad user experience to the client and will surely threatened one of the most important asset that we have that is our brand and also cause a money issue in the project since one of the reason of deploying this network is to retain the customer.

Clearly we need to seek technical "thermometers" so we can monitor the user experience in the wireless network service, not only depending from customer surveys to measure the quality of the service, and try to deliver a wireless services at the same level of quality that our industry is providing carrier-grade services to our customers. That is the difference of being called a carrier grade Wi-Fi service provide

4.1 Wireless User Experience Quality index

Since this problems of low SNR and fail association could occurs, we need to find a way of get, register and understand this data to solve problems like user experience, find places where the signal coverage is not good enough, control the network and know if the network is getting better or worse in time and also with this understanding make the upgrade and make the coverage bigger with a more cost effective way because you with those information can understand better where your customers need coverage. That why we present the concept of WUEQi, that is a index that is a scale of 1 (bad) to 5 (excellent) user quality of experience, similar to a MOS type index.

Variable #1 – SNR of the connection – S
The SNR variable has a weight of 2 in the total.

Margins	**S**
< 7 dB	1.0
7 dB < S < 15 dB	4.5
15 dB < S < 25 dB	4.9
> 25 dB	5.0

Table 4. “S” variable

Variable #2 – Fail Association - FA
The Fail Association variable has a weight of 3 in the total.

Margins	**FA**
< 3 fail association in a AP per hour	5.0
3 < Fail Association per hour < 10	2.5
10 < Fail Association per hour < 20	1.0
> 20 Fail Association per hour	0.5

Table 5. “FA” variable

Variable #3 – Get an IP address - I
The fail getting IP Address per hour variable has a weight of 3 in the total.

Margins	**I**
< 3 fail getting IP address in a AP per hour	5.0
3 < fail getting IP address in a AP per hour < 10	2.5
10 < fail getting IP address in a AP per hour < 20	1.0
> 20 fail getting IP address in a AP per hour	0.5

Table 6. “I” variable

Variable #4 – Medium throughput of the connection - M
The medium throughput of the connection has a weight of 2 in the total.

Margins	**M**
M < 6 Mpbs	1.0
6 Mbps < M < 9 Mbps	2.5
9 Mbps < M < 12 Mbps	4.75
12 Mbps < M < 18 Mbps	4.8
18 Mbps < M < 25 Mbps	4.85
25 Mbps < M < 50 Mbps	4.9
50 Mbps < M < 75 Mbps	4.95
M > 75 Mbps	5.0

Table 7. “M” variable

Using the formulas show below we can find the WUEQi.

$$MT = \frac{\sum M}{\sum users(1h)}$$

$$ST = \frac{\sum S}{\sum users(1h)}$$

$$WUEWQi = \frac{2xMT + 2xST + 3xI + 3xFA}{10}$$

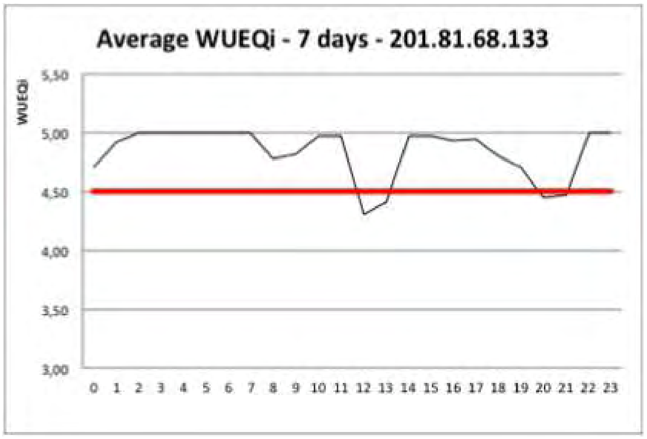
We can calculate from the WUEQi of an AP to the WUEQi of the network using a simple mean formula:

$$WUEWQi(network) = \frac{\sum WUEQi(allAP)}{\# AP}$$

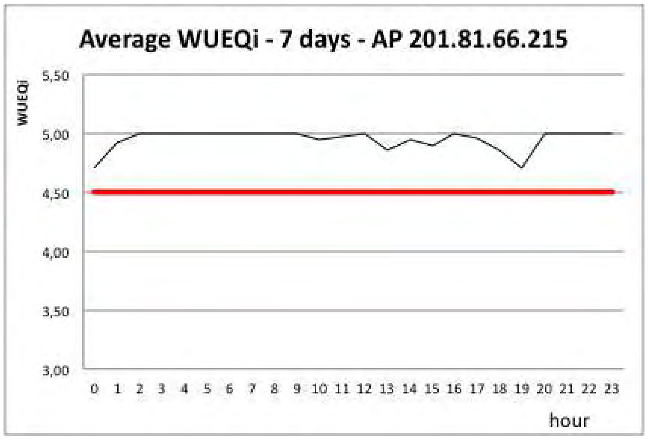
4.2 Using the index

With this index we can identify which access points are facing problems with bad user experience and make site surveys to implement more coverage in the places that the customer are demanding. With this data you can make better planning for the growth of the network, more efficient and cost effective. This index can also keep the track of changes of the network keeping history of the network grade.

The case showed below is an AP that was facing problems with a bad user experience and has to have a survey and make the coverage better.



The case showed below is an AP that has a good comportment



5. CONCLUSION

This work does not intend to be the final word in measuring the quality of service of a Wi-Fi network but a starting point of a more controlled network and a point of discussion. As a service provider, we must have the concern with the quality of any service that our companies are delivering. When we were deploying this network we faced with a lot of our users (testers) saying that in some places they could look the network and connect well, but in other places they found the network associate and could not use the Internet or found and could not associate, so with this concepts we could not only measure the network quality but more important get to know where to focus our effort to make expansion to grow our coverage and serve our customer better.

REFERENCES

[1] Kemisola Ogunjemilua, John N. Davies, Vic Grout and Rich Picking, "An Investigation into Signal Strength of 802.11n WLAN" Centre for Applied Internet Research (CAIR) Glyndŵr University, University of Wales, Wrexham, UK.

[2] Friis H.T.,(1946) Proc. IRE, vol. 34, p.254. 1946

[3] Garg, V.K., (2007). Wireless Communications and Networking, Morgan Kaufmann PublishersGast M., (2002), 802.11

Wireless Networks: The Definitive Guide, O’Reilly Media Incorporated.

[4] BELAIR Presentation - BelAir SNMP Overview – January 2011

[5] BELAIR Link Budget Presentation - 2011

[6] CISCO - Cisco ClientLink: Optimized Device Performance with 802.11n

[7] CISCO Datasheet - Cisco Aironet 1550 Series Outdoor Access Point

Video Calling Over Wireless Networks

David Urban
Comcast

Abstract

Video calling over wireless networks has become increasingly popular as wireless networks become faster and more reliable and devices with cameras and video calling applications become more ubiquitous.

Measurement and analysis of video calls over home, outdoor and cellular wireless networks have determined the criteria for making a successful call over a wireless network. Signal strength, packet loss, jitter and round trip delay are critical parameters. Video calling over wireless networks is shown to be practical, provided that the critical parameters are met.

INTRODUCTION

Video calling is a technology that has been around for quite some time but has never caught on as much as one might think. As the saying goes, video calling is the technology of the future and always will be.

Bell Labs began building experimental prototypes in 1956 culminating in the 1964 New York World's Fair demonstration of the Picturephone service [1]. By 1969, the transition from voice calling to video calling appeared to be at the threshold. Looking back at the Picturephone service of the 1970s, it is interesting and instructive to find that many of the standards and specifications are similar to those used today. The analog bandwidth of the black and white video picture was 1 MHz with an interlaced 250 lines refreshed at 30 frames per second. The screen size was 5.5 x 5". When digitized to be transported a distance greater than 6 miles the combined video and audio signal was 6.3 Mbps.

So have things changed appreciably enough to suspect that this might just be the time that video calling really catches on? There is ample reason to think that the answer is yes. Many factors have fallen into place to make video calling more feasible than ever before.

Many people now carry around smart phones with a front and back camera, video calling software and a data connection fast enough for video calling. This means that when initiating a video call, one needn't count on the other side being at their computer with attached web camera and logged into a video calling application.

Televisions can be made into high definition video conferencing solutions with convenient and inexpensive add on products such as video cameras with built in microphones and small computer appliances to run the video calling application. Again, this avoids the inconvenience of having to fire up your computer, plug in your webcam, and open and log in to your video conferencing software. Any time you are watching television you can make or receive a video call.

A video call on a large screen television set can be much more enjoyable than using a computer. Several members of the family can participate while sitting on the couch rather than crouching around a small computer screen. And the 720P resolution of a 32 inch or larger diagonal flat screen television provides a much better viewing experience than a notebook computer or smart phone screen can.

Successful video calling requires a network connection with high data rate, low latency and jitter, and negligible packet loss.

Broadband connections are becoming more common and the performance keeps improving, making video calling more practical. This is true for both fixed and mobile networks. Video codecs have and continue to improve and work is being done specifically for video grade wireless distribution.

While there are still some impediments to video calling such as high cost and the lack of simple, intuitive and reliable user interfaces, many hurdles to successful video calling have recently been cleared and the remaining obstacles are trending toward resolution. In the 1960s and 1970s video calling moved from a laboratory curiosity to an ambitious but ultimately disappointing large scale national project. Then in the 1980s and 1990s video conferencing remained a niche application mainly for big businesses. With the advent of the personal computer and broadband residential network connectivity video calling has become an increasingly popular method to stay in touch with family and friends. The big transition occurring today is the move from video calling on desktop and notebook computers to video calling on smart phones, tablets and televisions. This transition makes wireless home and public network performance and reliability more important than ever. Table 1 shows some video call data rates.

Video Call Type	Tx Mbps	Rx Mbps	Video Quality
1080P WiFi	5	5	excellent
720P WiFi	1.5	1.5	excellent
3-Way WiFi	1	1	good
Smart Phone	0.5	0.5	fair
3G Cellular	0.2	0.2	poor

Table 1. Typical Data Rates of video calls

VIDEO CALLING PARAMETERS

A video call can be at times amazing when it works perfectly while at other times the experience can be frustrating when things go wrong. The elements of a video call include two parties, each with a video camera, video display, audio microphone, and audio speaker. Each party needs a computing device to run a video calling application and the devices need to have a network connection to establish the call, send the video and audio streams, and end the call.

For a two-way video call, a video stream will be sent from the video camera and another video stream will be received for the video display. Likewise, an audio stream will be sent from the microphone and another audio stream will be received by the speaker.

The audio and the video must be synchronized. A delay from the video camera of one user to the video display of the other user can be distracting. For example, if a caller wishes to show an object by putting it in front of the camera but gets no reaction from the other side, this can be confusing. Then the other side finally comments but long after the object has been removed from the camera view. This distracts from the real time interactivity of the video call.

Disconnects, long reconnections, poor video quality, long delays, lack of video and audio synchronization, freezing of the video, brief distortions of the video display, screen refresh and resolution issues; these are the problems that make video calling frustrating. A successful video call requires a good network connection on both ends, good processing power in the CPU running the application, a good video calling application, a good camera and display on both ends.

The key network connection parameters necessary for a successful video call include data rate, packet loss, jitter, delay, and relays. The data rate will be dependent upon the screen size and resolution. A video call on a 1080P LCD television will have a data rate of 10 Mbps whereas a video call on a 4.3 inch diagonal screen smart phone will have a data rate of 1 Mbps.

Video calling applications often report call technical information. Among the reported parameters are jitter, packet loss, send packet loss, receive packet loss, round trip time, and relays. Relays can be used to work around firewalls and other networking issues that prevent a direct UDP connection between the two video callers. Relays in general are undesirable since they often prevent HD video calling. The video and audio streams are sent as UDP packets.

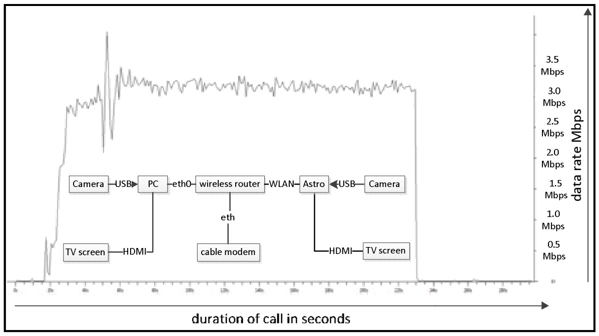


Fig.1 Data Rate and Block Diagram of 720P Video Call

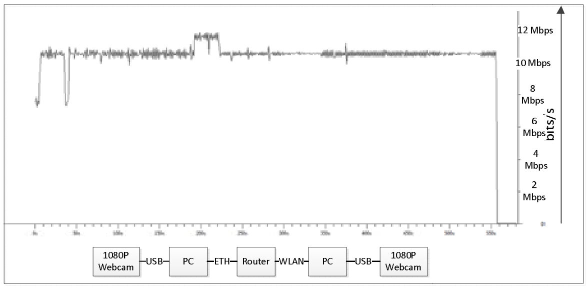
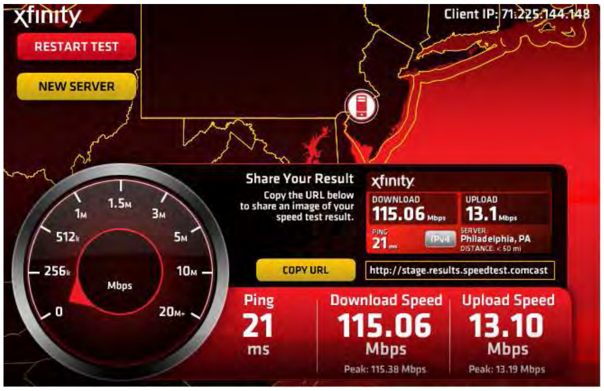


Fig. 2 1080P video call data rate

Wireshark was used to record the packets during a video call. The data rate during the call is shown in Fig.1 along with a block diagram of the test set up. Both the camera and the display were capable of 720P operation. The upstream and downstream data rate was measured to be 1.5 Mbps for a total of data rate of 3 Mbps. The video call quality was excellent. Packet analysis shows that the data protocol was UDP with packet size around 1400 bytes. In this particular test the video and audio streams were sent between devices on the same local area network. Most video calls will span a wide area network adding additional challenges for a successful video call.

Figure 2 shows a video call with 1080P video resolution. In this case the data rate is much higher at 10 Mbps, 5 Mbps for each video stream. Setting up a 1080P video call can be a bit tricky. You'll need a 1080P video camera and display at both ends, video calling software the supports 1080P resolution at both ends, and network connectivity supporting 5 Mbps UDP traffic in both the upstream and downstream direction. Residential broadband connections that support this high upstream data rate have only recently been offered. Figure 3 shows a speed test for a cable modem connection capable of supporting 1080P video calling.



Client IP: 71.225.144.148

Share Your Result
Copy the URL below to share an image of your speed test result.

DOWNLOAD 115.06 Mbps
UPLOAD 13.1 Mbps
PING 21 ms
IPv4
SERVER: Philadelphia, PA
DISTANCE: < 50 mi

http://stage.results.speedtest.comcast

Ping 21 ms
Download Speed 115.06 Mbps Peak: 115.38 Mbps
Upload Speed 13.10 Mbps Peak: 13.19 Mbps

Fig.3 Broadband Connection speed for 1080P video call.

The user experience of a 1080P video call is remarkable. The picture is clear and sharp on a big screen television and the live fast action response to motion is impressive.

Large screens with high resolution benefit from very high continuous data rates during a video call; however, many video calls involve smart phones which have much smaller screens that do not need such high data rates. Figure 4 show the data rate measured during a video call using a smart phone. The smart phone network connectivity is over a wireless home network.

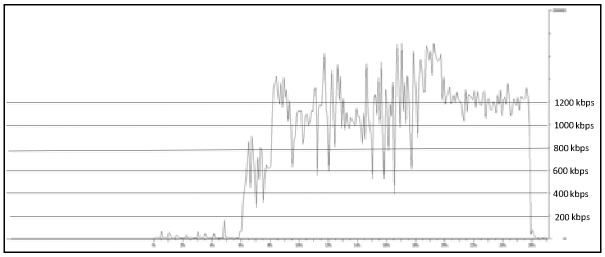


Fig. 4 Video Call using a smart phone with WiFi

The data rate measured about 1.2 Mbps and the quality of the video was good. Notice that the data rate is much less consistent than the previous plots with the bit rate over time being very choppy. This is due to several factors including the wireless home network link, the smart phone CPU processing speed and memory, and the impact of the wide area network. For the video call in figure 4 two cable modems were used so that the video and audio streams had to traverse the HFC network.

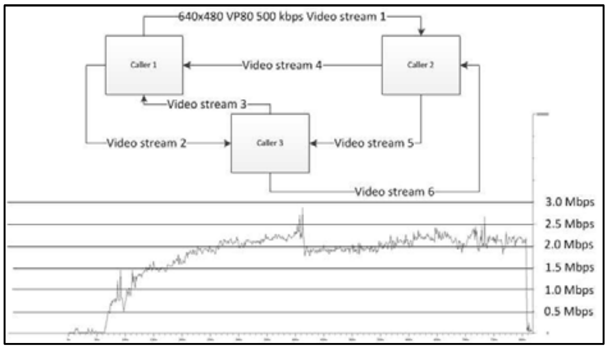


Fig. 5 Three-Way Video Call

A video call can be made between three or more parties. For a three way video call, a video caller sends two video streams and receives two video streams. The video callers' display typically shows the two received video streams side by side on the screen with a small caption of the video send stream. With this format the video caller can see both of the people he is calling as large as possible and still monitor what the other parties see of him. Since two video streams must share the display, the resolution and bit rate of a single video stream is reduced, i.e. one cannot display two 1080P video streams on a single 1080P video display. Testing 3-way video calls, the video send and receive streams were found to be 640x480 with VP80 codec at 30 frames per second and a bit rate around 500 kbps. A video caller participating in a 3-way call will thus send two 500 kbps video streams and receive two 500 kbps video streams for a total data rate of 2 Mbps as shown in figure 5.

VIDEO CALLING OVER WIRELESS NETWORKS

Characteristics of the wireless network

The making of a successful video call requires network connectivity with low packet loss, low latency, and low jitter and must support UDP data rates between 1 and 10 Mbps depending on the screen size and video quality requirements. Several wireless networks were tested to gauge their performance against the demands of video calling.

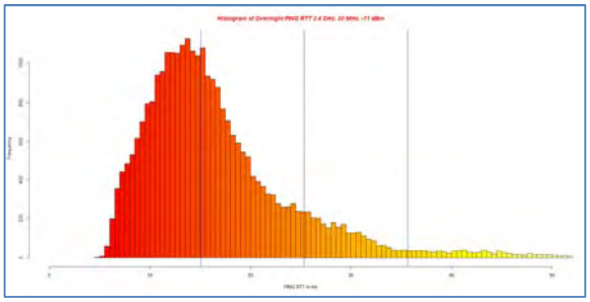


Fig. 6. Histogram of Overnight PING RTT 2.4 GHz, 20 MHz, -71 dBm from 0 to 50 ms

Fig. 6 shows a histogram of the round trip time, RTT, measured while sending PING packets between a wireless client and a wireless access point of a home wireless local area network, WLAN. The x-axis is the PING RTT from 0 to 50 ms and the y-axis is the number of occurrences. As indicated by the vertical lines in figure 6, the median RTT was found to be 15 ms, the first standard deviation above the median was 25 ms and the second standard deviation above the median was 35 ms.

The wireless access point was set to 2.4 GHz channel 8 with a 20 MHz channel width. Both the STA and the AP were IEEE 802.11n with dual stream capability. The wireless access point was set to B/G/N Mixed wireless mode. The beacon interval was set to 100. The RTS threshold was set to 2347. The guard interval was set to 800 ns. The STA and AP were separated by 36 feet and one floor and two walls of a residential home. The PING tests were taken over a 12 hour period. The x-axis of the plot is the PING round trip time measured in ms. There are three vertical lines shown on the graph, from left to right these lines are the median, first standard deviation, and second standard deviation, respectively. The receive level measured by the STA was -71 dBm.

The results indicate that the latency and jitter of a wireless home network have much more variability than a wired network over the course of time. This can be due to signal fading and interfering signal sources. The statistical distribution of the round trip time of packets between the AP and the STA are well within the requirements of a video call. A round trip time of 40 ms will support a high quality video call. On the histogram of round trip time measured in ms, 40 ms is beyond the second standard deviation and thus is a rare occurrence.

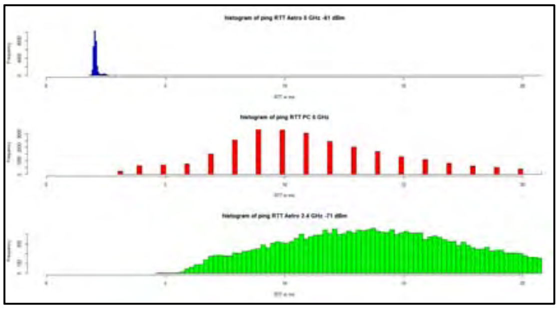


Fig. 7 Histogram of PING RTT for 5 GHz -61 dBm(top), 5 GHz at same location with PC (middle), and 2.4 GHz -71 dBm (bottom) from 0 to 20 ms

Figure 7 shows test results of the distribution of PING round trip time in ms over the course of a 12 hour test period. There are three different test conditions, the top blue graph is the PING RTT distribution over a 12 hour test period of a 5 GHz wireless home network connection with a receive level of -61 dBm. The computer used for this test was a small form factor LINUX device with built in WiFi client. For this test the AP and the STA were separated by one wall and 12 feet. 5 GHz band with the AP and STA in close proximity results in a much lower median round trip time latency of 2 ms with no significant measurements greater than 5 ms.

The middle red distribution of figure 7 shows another 5 GHz test taken in the same location and same time as the top distribution but using a different wireless client station. The middle test results used a notebook computer with built in wireless card and antennas. This RTT distribution shows a median of 10 ms with most RTT measurements fewer than 20 ms.

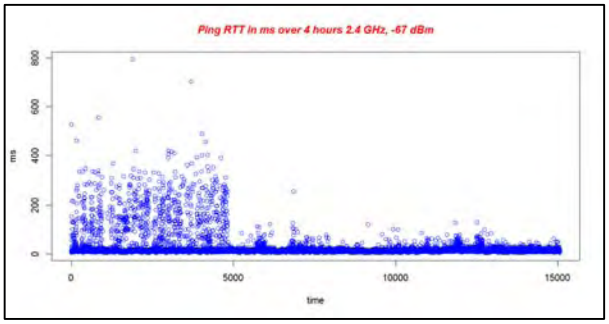


Fig. 8 Plot of PING RTT in ms over 4 hours at 2.4 GHz, -67 dBm, y-axis is RTT in ms from 0 to 800ms.

Both test results are good and well within the requirements for a successful video call. But why would two tests, both wireless clients using the same channel at the same time, both in the same location, give such different results? It turns out that it was not due to hardware differences between the two stations since subsequent overnight tests revealed that by slight manipulations of the antenna positioning one could reverse the results.

As illustrated in figure 8 the PING round trip time can change abruptly in time and these changes can last for hours at a time. This could be due to other applications sharing the spectrum or even to the movement of people or objects within the home.

The antenna patterns of notebook computers and small form factor devices with built in antennas will have nulls due to internal obstructions. By slightly repositioning the computers and devices one can place the nulls in more or less advantageous a location and this can influence the PING RTT results.

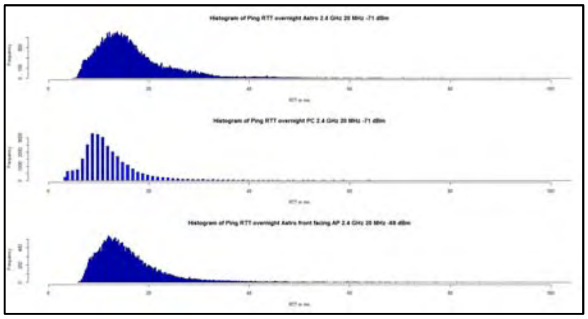


Fig. 9 Histogram of PING RTT overnight at (top) 2.4 GHz, 20 MHz, -71 dBm, LINUX, (middle) 2.4 GHz, 20 MHz, -71 dBm, WINDOWS, (bottom) 2.4 GHz, 20 MHz, -68 dBm, LINUX with antenna facing AP, x-axes are RTT in ms from 0 to 100.

The bottom green PING RTT overnight test distribution in figure 7 was taken using 2.4 GHz with the STA receive level of -71 dBm due to a larger separation distance between AP and STA of 36 feet with one floor and two walls. As expected the PING RTT distribution is much larger than the test using 5 GHz at closer AP to STA separation distance with a median of 14 ms and a significant number of round trip times having latency greater than 20 ms. The 2.4 GHz and -71 dBm receive level overnight PING test shows performance that is well within the bounds for a successful video call but as we will later see this is at the threshold of successful video calling operation.

One can expect variety of latency and jitter distributions for wireless home networks. Other examples are shown in figures 9 and 10. Figure 9 shows three tests taken in the same location, all at 2.4 GHz with 20 MHz channel width. The difference between the three plots is due to slight differences in antenna positioning. Figure 10 shows a comparison of 2.4 GHz performance versus 5 GHz at a 24 foot AP to STA distance with one wall in between. At close range 5 GHz proved to have consistently lower round trip times, however, figure 10 shows that at farther distances and more wall attenuation 2.4 GHz operation can have lower round trip time than 5 GHz.

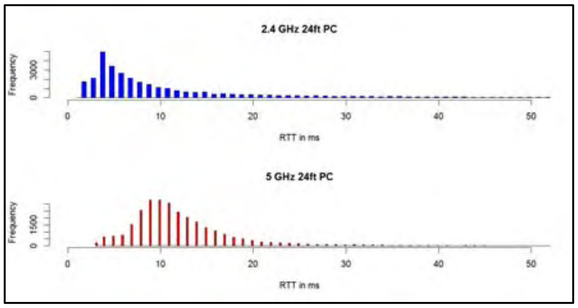


Fig. 10 Histogram of PING RTT 2.4 GHz vs 5 GHz and 24 ft AP to STA distance

In general wireless connections will be worse than wired connections in this regard. It is important to note that some routers handle IP video and peer to peer stream video better than others. As a rule of thumb, using 5 GHz at very close distance will give more consistent performance for video calling than 2.4 GHz at far separation distances as illustrated in figure 7. If you are having trouble making a video call using a wireless home network connection, then slight variations in antenna positioning of either the client station or access point can make significant performance improvements. Changing the RF channel, channel width, guard interval, and mode may also be experimented with to fix problems.

Figure 11 shows the statistical distribution of the call technical information reported by the video calling application. A small form factor Linux computer was used to run the video call application. The video camera and display at both ends of the video call were 720P and the video call data rate was 3 Mbps. The wireless network connection used 2.4 GHz with a 36 feet AP to STA separation distance with one floor and two walls in between. The video calling software has an option to report call quality technical information which includes a measure of network packet loss, roundtrip time, and jitter. These statistics are used to adjust the call quality to account for network connectivity issues. If these parameters degrade, then the video calling application will adjust by lowering the video quality such as adjusting the resolution from 1280x720 to 640x480. Adaptive bit rate streaming is a technique to provide the best video quality for given network limitations.

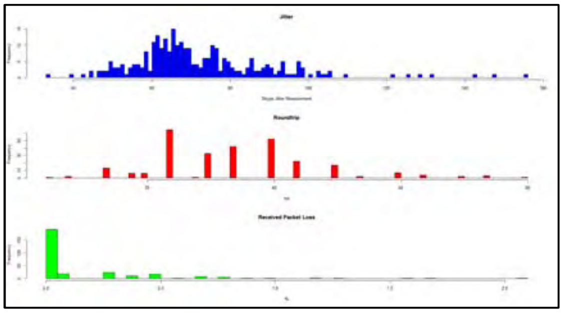
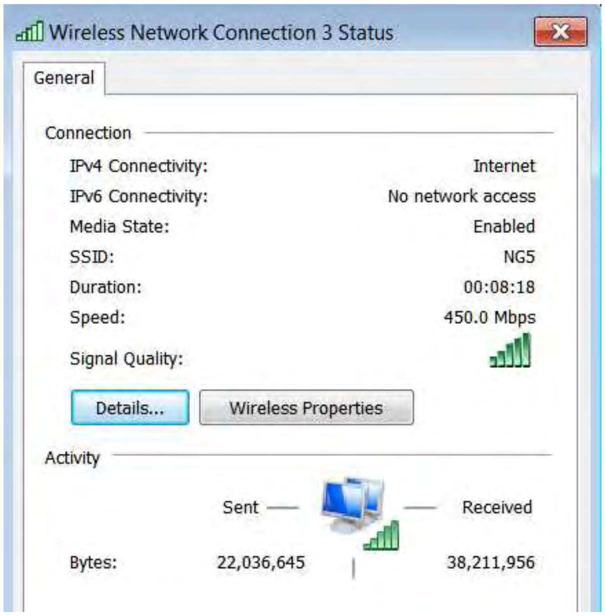


Fig.11 Video Call Quality Technical Information Top Histogram is Jitter from 40 to 160, Middle Histogram is Roundtrip time from 0 to 80 ms, Bottom Histogram is received packet loss from 0 to 2%

The call quality technical information was saved to a text file during the video call. A PERL language program was written to filter out the three parameters of interest into an array suitable for statistical analysis using the R statistical programming language. The analysis shows that the packet loss throughout the video call remained low at less than 0.5%. The round trip time varied significantly with a noticeable amount of measurements as high as 60 ms. The jitter measurement also varied significantly during the course of the call. Despite the variations, test results show that the wireless network was able to support a 720P video call with good reliability.



Wireless Network Connection 3 Status

General

Connection

IPv4 Connectivity:	Internet
IPv6 Connectivity:	No network access
Media State:	Enabled
SSID:	NG5
Duration:	00:08:18
Speed:	450.0 Mbps
Signal Quality:	

Activity

	Sent	Received
Bytes:	22,036,645	38,211,956

Fig. 12 3by3 AP and 3by3 STA reporting 450 Mbps modulation and coding scheme

Wi-Fi packet analysis of a 1080P video call

A video call was set up between two callers with one caller using a wireless home network. Both the access point and the client station of the wireless home network were capable of three stream operation. The highest data rate of the wireless home network was 450 Mbps. By carefully positioning the access point and the client station in close proximity and applying some tricks such as using cookie sheets to create reflections it was possible to get the client wireless software to report 450 Mbps as shown in figure 12.

However, during the video call the highest data rate achieved was 324 Mbps. The data rate of 324 Mbps has three spatial streams, a guard interval of 800 ns, a 40 MHz channel width, and MCS 21 64-QAM with 2/3 rate binary convolutional coding. The method to calculate the data rate of 324 Mbps is shown in equation 1. The details behind these calculations can be found in [2], [3],[4].

$$R = \frac{3(streams) * 6\left(\frac{bits}{sc}\right) * \left(\frac{2}{3}\right) * 108(sc)}{(3.2 + .8)(\mu s)} = 324\ Mbps\ [1]$$

With three transmit and three receive antennas in the access point and the client station there are nine paths between the transmit antennas and the receive antennas as shown in figure 13.

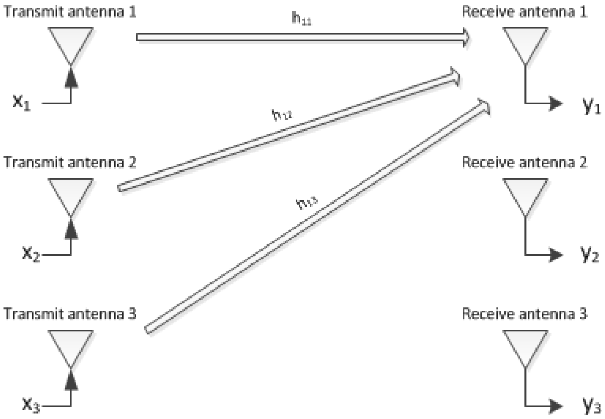


Fig 13. 3x3 MIMO Block Diagram

The output signals of the receive antennas, y_i with i={1,2,3}, is equal to the input signals of the transmit antennas, x_i, times the complex path loss of the nine paths between transmit and receive antennas, h_{ij}, as shown in figure 13. The relationship between the output signal of the three receive antennas and the path loss between the antennas and the input signals to the three transmit antennas can be expressed as a matrix equation [2]. If the H matrix of equation [2] can be inverted then it is possible to calculate the input signals by measuring the output signals and multiplying by the inverse of the H matrix. The determinant of the H matrix is zero if all of the elements are the same. The inverse of the H matrix is proportional to the inverse of the determinant. Thus, if all the elements of the H matrix are identical the determinant will go to zero and the inverse will blow up to infinity and it will not be possible to determine the input signals with knowledge of the output signals and path characteristics. Multiple streams can only work if there are differences, most desirably phase differences, between all of the nine paths between antennas. Spreading the antennas apart spatially is one method to increase the phase differences between the paths. However, with compact access points and particularly with compact client stations the amount of spatial separation is limited. Here, 5 GHz operation has an advantage over 2.4 GHz operation since for a given spatial separation, electrically in terms of wavelengths the separation between antennas is greater at 5 GHz than 2.4 GHz. The most effective and desirable method to create differences between the paths is reflections. A multipath rich environment with many reflected signals is the best for realizing multiple streams of data. In equation [1] the data rate from a signal antenna is multiplied by 3 since each of the 3 transmit antennas are transmitting an independent data stream.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\ [2]$$

Each subcarrier of the OFDM symbol is 64-QAM modulated so that a subcarrier is mapped to 6 bits. The bits are the output of a binary convolutional coder that inputs 2 data bits and outputs 3 coded bits. Thus, each OFDM subcarrier is mapped to 4 data bits as reflected in equation [1].

The channel width observed for this test video call for this packet was 40 MHz. A 40 MHz 802.11n signal consists of 128 subcarriers. Subcarriers at the channel edges and center are nulled to form a guard band and prevent DC offset. Some subcarriers are used as pilots to allow for frequency acquisition and carrier lock. This leaves us with 108 data subcarriers for a high throughput 802.11n data packet as reflected in equation [1]. Equation [1] reveals that each OFDM symbol for the observed packet has 1,296 data bits or 162 data bytes.

Finally, in order to determine the data rate we need to know the symbol time. The 128 modulated subcarriers that comprise the OFDM symbol are converted to a time domain representation using an inverse fast Fourier transform, IFFT. The 128 point IFFT is a transform with 128 complex frequency domain input numbers and 128 complex time domain output numbers. A digital to analog conversion, ADC, is required to turn the complex numbers into a real waveform capable of being upconverted to a carrier frequency to excite an antenna current in order to form an electromagnetic wave that can radiate from the transmit antenna to the receive antenna. The clock of the ADC determines the channel width of the analog time domain waveform. The channel width will be 40 MHz if the sampling interval is 25 ns. The formula is shown in equation [3] with W being the channel width in Hz and τ is the sampling interval of the time domain waveform in seconds.

$$W = \frac{1}{\tau} \quad [3]$$

With 128 subcarriers turned into 128 time domain samples by an IFFT and applying a 25 ns sampling interval, an OFDM symbol can be transmitted in 3.2 µs. This is sometimes referred to as the useful symbol rate. In theory OFDM symbols could be sent every 3.2 µs, however, in practice a guard interval in the form of a cyclic prefix is added so that the OFDM symbols are sent at an interval that is longer than the minimum possible. This eliminates inter-symbol interference that results when the receiver is hit with two different symbols at the same time due to reflections. As long as the guard time is longer than the time delay of the largest reflection then the receiver can ignore the guard time and demodulate the useful symbol time without inter-symbol interference. For 802.11n OFDM symbols the guard interval, GI, can be either 800 ns or 400 ns. A GI of 400 ns is referred to as a “short guard interval.” For the packet analyzed in the example the guard interval was 800 ns or 0.8 µs. The total OFDM symbol time is the sum of 3.2 µs and 0.8 µs for a total symbol time of 4 µs. This is shown in equation [1]. Now the data rate can be calculated by dividing the bits per OFDM symbol by the symbol time, in this case the data rate is 324 Mbps.

The frame length of the QoS data packet was 1395 bytes so that 9 OFDM symbols carrying 162 bytes each are needed to send the packet data payload. A burst of 9 OFDM HT symbols in this example lasts 36 µs since the OFDM symbol time for normal guard interval is 4 µs. Thus, over the 36 µs period of the 9 OFDM HT symbols the data rate is 324 Mbps.

When digital video signals are sent from a cable headend to a receiving set top box, the 256-QAM modulated 6 MHz wide signal transmits 38 Mbps continuously, a 100% duty cycle. This is not the case for wireless local area network transmissions. Since the medium is shared between uplink and downlink transmissions, amongst other users of the wireless home network, amongst co-existing wireless home networks, and amongst other spectrum users such as microwave ovens, cordless phones, remote controls, and sensors, a 100% duty cycle is not possible. Data is sent in bursts with short time frames and these bursts require a preamble in order to be received. The preamble is needed in order for the receiver to acquire carrier lock, understand the basic parameters of the packet, so that demodulation of the payload symbols can be made accurately.

All packets must be preceded with a preamble. The first part of the preamble is a short training field made up of 12 subcarriers. The short training field is 8 µs long. The short training field consists of 12 subcarriers. Figure 15 shows the subcarriers of the short training field measured by a vector signal analyzer. The short training field is followed by and 8 µs long training field and then a 4 µs signal field.

The transmission of a video packet of from the AP to the STA requires a sequence of packets as shown in figures 14 and 15. First, the AP sends a request to send message to the STA. The STA responds with a clear to send message. A QoS Data packet is sent from the AP to the STA. Finally, a block acknowledgement message is sent from the STA to the AP. This process is repeated continuously throughout the video call for both uplink and downlink transmission.

The request to send packet reported a length of 16 bytes and a data rate of 24 Mbps in its signal field, labeled SIG in Fig. 14. This is a legacy packet and thus has a 20 µs preamble consisting of an 8 µs short training field, STF, an 8 µs long training field, LTF, and a 4 µs signal field. The symbol period is 4 µs, a symbol has 48 data subcarriers mapped to 2 data bits so each symbol carries 96 bits or 12 bytes of data. The RTS packet is 28 µs and the request is to transmit a packet sequence with 224 µs duration.

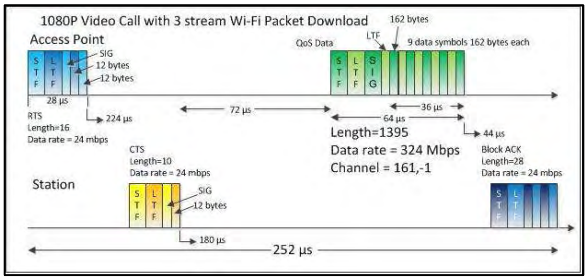


Fig. 14 WiFi Downlink video packet sequence

Following the RTS from the access point to the station will be a clear to send, CTS, response from the station to the access point. The clear to send packet has a length of 10 bytes and a data rate of 24 Mbps with a channel of 161. As with the RTS, the CTS packet has an 8 µs short training field, followed by an 8 µs long training field, followed by a 4 µs signal field, followed by 4 µs OFDM data symbols carrying 12 bytes of data. Since the CTS field length is 10 bytes only one OFDM data symbol is needed.

The CTS packet time duration is 24 µs. The CTS signal field reports that the duration from the end of the CTS to the end of the packet sequence is 180 µs. By taking the difference between the reported duration by the RTS packet and the CTS packet, we calculate the time duration from the end of the RTS packet to the end of the CTS packet of 44 µs. Thus there is a gap of 12 µs from the end of the RTS packet to the beginning of the CTS packet allowing for time for the access point request to be made and the client station response to be sent.

After the access point makes a request to transmit data and the client station responds with a clear to send signal then the QoS data packet can be sent from the access point to the client. Once the QoS packet has been sent by the access point and received by the client station then the client station sends a block acknowledgement back to the access point.

So in this example packet sequence measured during a 1080P video call, 1395 data bytes were transmitted over a 252 µs time period. The data rate accounting for the signaling and overhead is 1395 bytes divided by 252 µs which is 44 Mbps. Since the 1080P video call requires a sustained 10 Mbps data rate, the duty cycle of 324 Mbps data rate QoS data packet sequences during a 1080P video call is 23%.

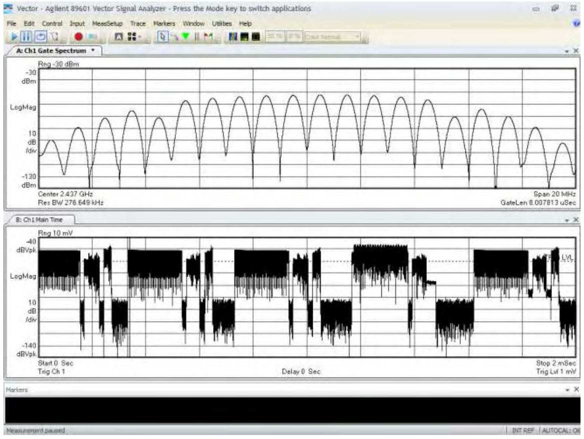


Fig. 15 Spectrum Analysis of WiFi video call

Taking a look at the distribution of the data rate of QoS Data packets reveals that many of the data packets were sent at a lower data rate than 324 Mbps. During the entire 1080P video call 3 stream operation was only utilized a small percentage of the time. All in all, 392,129 packets were analyzed. Of all of the downlink QoS data packets 9.42% had a data rate of 324 Mbps utilizing 3 stream operation. The majority of downlink QoS data packets operated at a 2 stream data rate of 243 Mbps representing 74.92% of the downlink QoS data frames. On the uplink 85.59% of the QoS data frames had 2 stream 270 Mbps while only 0.4% of uplink packets used 3 streams at a data rate of 324 Mbps or higher.

Plotting the histogram of the packet lengths during the video call shows statistically the anecdotal observation made by looking through the packet decodes. The RTS, CTS, QoS Data, Block ACK sequence with a 1400 byte UDP data burst is repeated throughout the video call. This packet sequence dominates the WiFi traffic during the video call. This is illustrated by the histogram shown in figure 16. Packet sizes are either less than 40 bytes or around 1400 bytes. The small byte size packets are signaling messages, RTS, CTS, and Block ACK. The packet sizes concentrated around 1400 bytes are video packets.

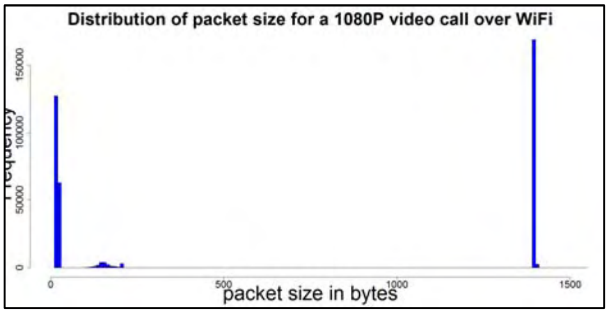


Fig. 16 Video call packet sizes are either small or large

Much of this analysis of the WiFi packets during a video call is focused on allowing the calculation of duty cycle, the percent of the time the application needs to use the RF spectrum. The reason that this is so critical is that WiFi uses unlicensed spectrum and thus any application must be judged based upon how well it will work while sharing the spectrum with other devices and applications. It is not a valid excuse for wireless LAN equipment and applications to claim that poor performance is due to a “noisy” environment. By “noisy” it is meant that other users of the spectrum are preventing the equipment or applications from working. However, equipment and applications using unlicensed spectrum must be designed to work in a shared spectrum environment. Users of unlicensed band equipment and applications must not expect performance levels that can only be realized with unshared spectrum. Even licensed band spectrum suffers considerable interference from adjacent cells and from spectral spillover from harmonically related or adjacent spectrum bands. So even licensed band equipment and applications must be designed to operate in the presence of fading and interference.

WIFI PACKET ANALYSIS OF A 720P CALL

A video call was set up with both callers having a 720P camera and display. One of the computers used a wireless home network connection. The wireless home network used

2.4 GHz channel 8 with a 20 MHz bandwidth and both the AP and the STA had 2 stream capability. The distance between the wireless access point and the wireless client station spanned about 36 feet, two floors and three walls of a residential home. The received signal strength level at the wireless client station ranged between -68 and -72 dBm. The video call lasted about 6 minutes and 30 seconds.

The video calling software reported call quality technical information. The video send stream was 1280 by 720 H.264 at 30 frames per second with a 1522 kbps data rate. The video receive stream was 1280 by 720 H.264 at 30 frames per second with a 1507 kbps data rate. The call technical information reported 0 relays indicating that the UDP traffic flowed directly between the two callers without intermediate nodes. The set up and data rate are shown in figure 17. The video call experience was excellent during this test. One video caller has an Ethernet 1 Gbps connection while the other video call uses a wireless home network connection with challenging RF signal conditions.

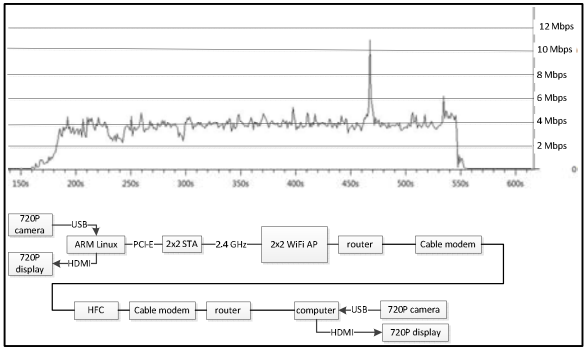


Fig.17 Video Call Set Up and Data Rate 720P 2.4 GHz.

During the video call an Airmagnet WiFi analyzer was used to capture the wireless local area network traffic. In all, 408,803 WiFi packets were captured and used for the statistical analysis of the call. The WiFi analyzer packet capture data file was saved as a Wireshark file and Wireshark analysis was used to create the IO data rate graph. The Wireshark data was then exported to a text file and a Perl program was written to extract and calculate a data array consisting of the burst time of the 408,803 packets. The R statistical programming language was then used analyze the distribution of the WiFi burst times.

The summation of the burst time of all the WiFi packets was 92.208012 seconds. Since the call lasted 6.5 minutes or 390 seconds, the percentage of time that the video call computer wireless station was either transmitting or receiving was 23.6%. In other words, the duty cycle of the 720P video wireless home network was found to be about 25%, one quarter of the time. If four such video calls were made utilizing the same wireless spectrum then we would expect conflicts due to 100% spectrum utilization.

The mean burst duration was calculated to be 225 µs. The median burst duration was found to be 40 µs indicating that many of the bursts were of short duration such as RTS, CTS, and Block ACK signals. The standard deviation of the burst times was 380 µs.

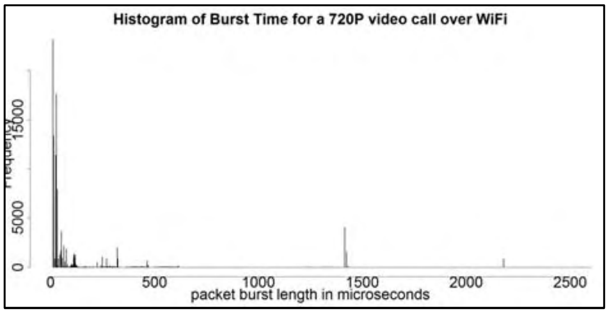


Fig 18. Histogram of the WiFi Burst Duration during a 720P video call.

The histogram of the burst durations is shown in figure 18. The bursts that last longer than 2 ms are beacons.

Figure 19 shows the histogram of packet burst time duration from 20 µs to 1 ms. The spreadsheet in Table 2 shows the percentage of packets for each possible data rate of transmission. With this histogram and spreadsheet it is easy to identify the main data rates used for sending video packets of about 1400 bytes. The three most prominent data rates are 13, 19.5, and 52 Mbps with burst durations of about 950, 640, and 260 µs, respectively.

Data Rate	Burst length	Burst Time	RX	TX	RX	TX
Mbps	bytes	microseconds	frames	frames	%	%
1	16	156	453	9,008	0.19%	5.31%
2	16	92	366	0	0.15%	0.00%
5.5	1495	2204	1	0	0.00%	0.00%
6.5	1495	1868	375	5,633	0.16%	3.32%
11	16	40	90,123	46,470	37.93%	27.40%
12	16	40	36,099	0	15.19%	0.00%
13	1495	948	1,159	27,950	0.49%	16.48%
19.5	1495	644	8,375	52,457	3.52%	30.93%
24	16	36	29,534	14,328	12.43%	8.45%
26	1495	488	7,615	10,199	3.20%	6.01%
39	1495	336	12,127	1,425	5.10%	0.84%
52	1495	260	50,095	1,041	21.08%	0.61%
58.5	1495	236	2	34	0.00%	0.02%
65	1495	212	17	49	0.01%	0.03%
78	1495	184	1,284	676	0.54%	0.40%
104	1495	144	1	338	0.00%	0.20%
117	1495	132	0	10	0.00%	0.01%
			237,626	169,618		

Table 2. 720P video call data rates and burst time

There are a couple points of interest in this analysis. First, although both the wireless access point and wireless station have two transmit and receive antenna chains and are thus capable of dual stream operation, the data rate rarely goes above 65 Mbps and most video packets are being sent at a data rate lower than 65 Mbps. This is significant because a single antenna wireless station lacking dual stream capability will max out at 65 Mbps for 20 MHz channel width and normal guard interval. Under these circumstances, the single antenna client station is at no disadvantage compared with a multi-antenna client. In fact, with only one antenna chain the power consumption is reduced and there is less physical footprint to pick up on board interference. The late Steve Jobs was noted for his passion for simplicity and functionality. He demanded products that worked and were a pleasure to use. Long battery life and comfortable operating temperature trumped the fastest Mbps claim on the outside of the box. This is reflected in mobile products that for the most part use a single antenna design with 20 MHz channel width and normal guard interval.

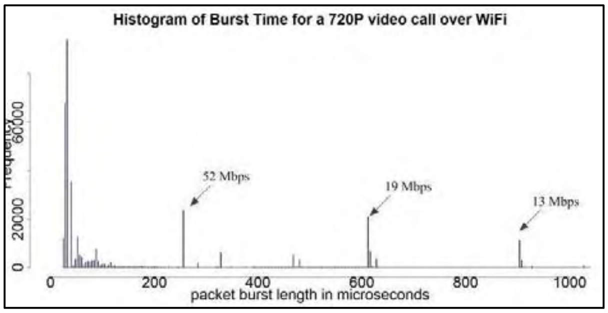


Fig. 19 Histogram of the WiFi Burst Duration up to 1 ms.

The second thing to note is that this test set up is operating at the threshold of a successful 720P video call. A significant portion of packets are operating at 13 Mbps having burst duration of almost a millisecond. This is good enough for a 720P call and as we’ve seen only a quarter of the RF spectrum is utilized for this application, meaning that co-existence with other applications is reasonable. However, any lower modulation and coding schemes than this and the 720P video call will not work. Once operation goes below the 13 Mbps data rate bursts, the video calling software will reduce the video quality due to packet loss and jitter measurements. And this will be particularly noticeable if any competing traffic or applications are sharing the spectrum.

Video Call with a smart phone over a wireless home network

A video call was set up between a PC and a smart phone. The display of the smart phone had a 4.3 inch diagonal and the video camera was 1080P. The smart phone connected over WiFi 2.4 GHz to a home wireless gateway with integrated WiFi and cable modem. A speed test application run on the smart phone measured a latency of 29 ms, a download speed of 5294 kbps and an upload speed of 7968 kbps. The gateway and the smart phone were separated by 36 feet one floor and two walls.

The other end of the call was a PC with a 1080P video camera and display connected to the Ethernet interface of a router and a cable modem. Two different cable modems were used in this test so that the video call packets would have to traverse the HFC network to the CMTS. The same CMTS terminated both cable modems in this test. The data rate and block diagram of the test is shown in figure 20.

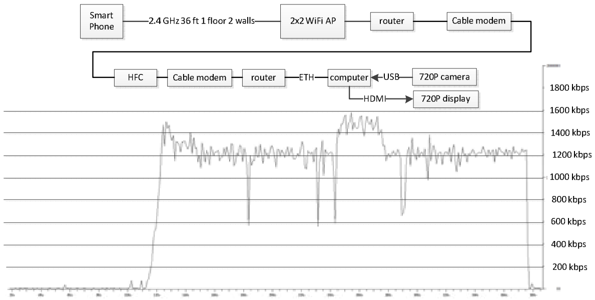


Fig. 20 Block Diagram and Data Rate of video call of smart phone with WiFi network connection.

The call technical information reported by the video calling software was monitored during the call. The number of relays was 0. The roundtrip time was 19 ms. The jitter was 69. The packet loss was 0.1%. The call lasted for 380 seconds or about 6 minutes.

The video send stream was 640x480 at 15 frames per second with H264 coding and 549 kbps bit rate. The video receive stream was 320x240 with H264 coding at 14 frames per second and a 605 kbps bit rate.

The number of packets captured for analysis was 60,348. The traffic protocol was UDP. The average data rate during the video call was 823 kbps and the average packet size was 648 bytes.

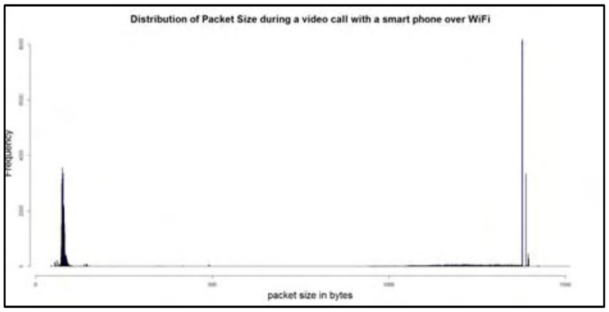


Fig. 21 Distribution of Packet Sizes during a video call using a smart phone with wifi network connectivity, x axis is packet byte size from 0 to 1500

Figure 21 shows the distribution of packet sizes during the video call using a smart phone with WiFi network connectivity. The packets are either very large or very small. The UDP video packets are typically about 1400 bytes whereas the WiFi signaling packets are typically less than 20 bytes in length. This explains the barbell type distribution of packet sizes.

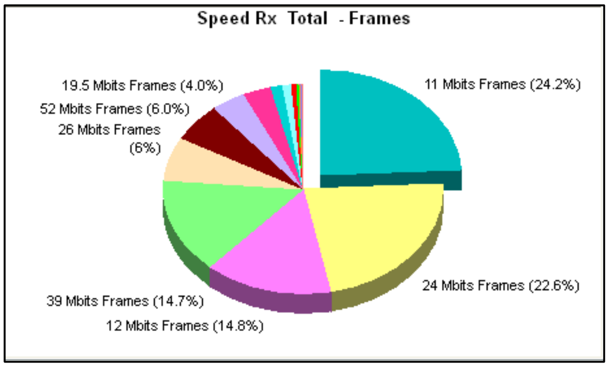


Fig. 22 Smart Phone over WiFi video call

Figure 22 shows the percentage of frame types with various data rates. A WiFi packet analyzer was used to create the pie chart. The majority of the packets had a data rate of 11 Mbps representing 24.2% of all WiFi packets sent. These packets are signaling packets, typically RTS,CTS, or Block ACK packets with short lengths of 16, 10, and 28 bytes respectively. 22.6% of the frames were 24 Mbps which are also signaling frames. The largest percentage of data carrying frames was the 39 Mbps frames representing 14.7 percent of the total number of frames. The 39 Mbps frames carry the large UDP video packets of about 1400 bytes of payload data. 14.8% of the frames were 12 Mbps. 6% of the frames were 26 Mbps. 6% of the frames were 52 Mbps. 4% of the frames were 19.5 Mbps.

During this video call using a smart phone with WiFi connectivity the WiFi analyzer captured WiFi network packets, the output was saved as a text file and a PERL program was written to calculate the burst duration of the 130,226 packets captured based upon the data rate and the byte size. The video call lasted 142.413 seconds and the period of time that the WiFi client was either transmitting or receiving was found to be 27,290,218 microseconds. Dividing the latter number by the former allows us to calculate that the utilization factor of the wireless spectrum during the video call was 19.2%. A histogram of the burst times is shown in figure 23. The predominate data rate of 39 Mbps for video packets of 1400 bytes which has a burst time of 336 microseconds is clearly indicated in the histogram. All in all it has been determined that a video call over a wireless home network using a smart phone has a data rate of 1 Mbps and uses up about one fifth of the wireless channel capacity.

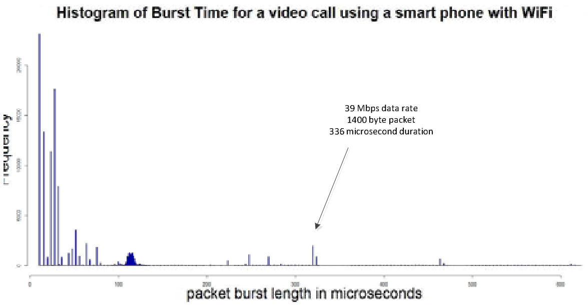


Fig. 23 Smart Phone over WiFi video call

Video Call over Cellular Wireless Networks

Video calls can be made over both 3G and 4G cellular networks. Here 3G networks refer to CDMA based networks and 4G networks refer to OFDM based networks since the characteristics pertinent to video calling varies considerably between these two multiplexing techniques. From a standards body standpoint, and from a service marketing standpoint, the use of the terms “3G” and “4G” is much more complex and nuanced and outside the scope of this paper.

Packet analysis was performed on a video call using a 3G cellular network lasting 1589 seconds or about 26 minutes. The video call quality was poor and the call dropped and re-established many times during the conversation. Still, the 3G end of the call was in the beautiful Florida Keys and the overall video calling experience was satisfying, refreshing to see a warm beach on a sunny day while huddled inside to avoid a cold grey Philadelphia winter. Video calling using a smart phone with a 3G data connection can be quite good at times as long as there is some tolerance for occasional disconnects, screen freezes, and fuzzy video.

Packets were captured with Wireshark on a PC with an Ethernet connection. The PC established a video call with another PC using a 3G cellular data card. The number of packets captured was 116,477. The average packet size was 295 bytes and the average data rate was 173 kbps.

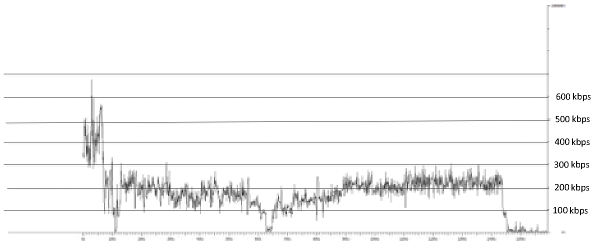


Fig. 24 Data Rate Measured During 3G video call

Figure 24 shows the data rate measured throughout the video call over the 3G cellular network. The data rate peaks at about 500 kbps, shows two steep drop offs where the call was lost, and otherwise runs at about 200 kbps.

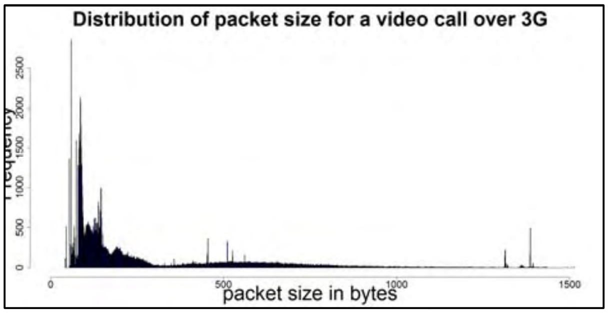
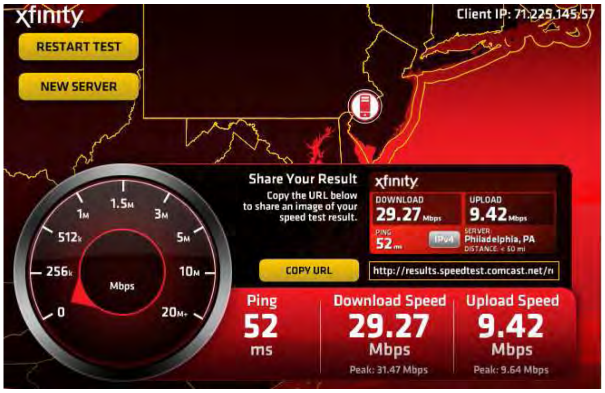


Fig. 25 Distribution of Packet Sizes of 3G video call.

The Wireshark packet analysis was exported to a text file, a PERL program was written to create an array of all the packet sizes for statistical analysis with the R statistical analysis tool. The resulting histogram is shown in figure 25 with the x-axis being the packet size in bytes ranging from 0 to 1500 bytes. By comparing the distribution of packet size between the 3G cellular network with that of the wireless home network, one notices that the packet sizes are generally much smaller when making a video call using the 3G network when compared to using a home WiFi network.



Client IP: 71.229.145.57

Share Your Result

Copy the URL below to share an image of your speed test result.

DOWNLOAD 29.27 Mbps
UPLOAD 9.42 Mbps
PING 52 ms
SERVER Philadelphia, PA
DISTANCE < 50 mi

http://results.speedtest.comcast.net/r

Ping 52 ms
Download Speed 29.27 Mbps Peak: 31.47 Mbps
Upload Speed 9.42 Mbps Peak: 9.64 Mbps

Fig.26 Speed test of 4G 700 MHz 10 MHz FDD pair

With the introduction of 4G networks having much higher data rates, and much lower latency and jitter, video calling over cellular networks will become better and more reliable. Like WiFi, 4G networks use OFDM which has a guard band in time to reduce inter-symbol interference as compared to a rake receiver or some type of adaptive equalizer used in CDMA networks. The adapter equalizer techniques used for single carrier wideband systems work well at times but require knowledge of the channel impulse response and so have difficulty under rapidly changing multi-path conditions. OFDM with a much simpler guard time inter-symbol interference mechanism can work even under rapidly changing multi-path conditions.

As the channel width increases the impulse response gets more complicated, requiring more taps for an adaptive equalizer and more calculations to respond to changes in multi-path conditions. This limits the channel width of CDMA based systems. The channel width used in the 3G video call of figure 24 and 25 has a 2 MHz channel width using CDMA. There are also 3G CDMA networks with 5 MHz channel width.

4G OFDM systems can operate at increased channel widths of 10 MHz, which gives them higher data throughput. Figure 26 shows the speed test results for a 4G network operating in the 700 MHz spectrum band with two frequency division duplexed, FDD, 10 MHz channel width signals. The download data rate is 29 Mbps and the upload data rate is 9 Mbps with 52 ms latency. These data rates, if maintained throughout the course of the call, are sufficient for 1080P video calling. One caution, cellular networks tend to be used in cars, buses, trains, or even when walking around and while moving throughout a geographical area the data rate will vary significantly and even switch from 4G to 3G and 2G coverage areas. So it is unlikely to always maintain these speeds while moving. In the area where video call testing was performed for this paper, 4G coverage was not available so video call testing and analysis was performed using a 3G network.

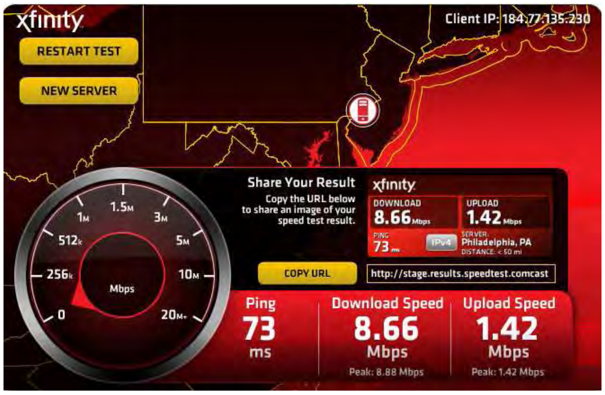
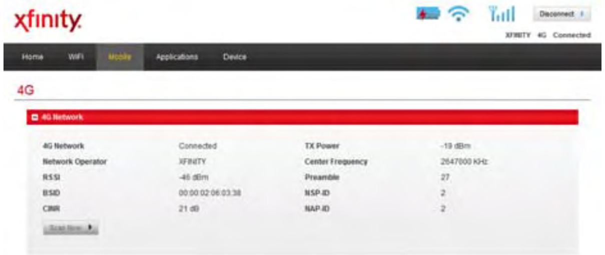


Fig. 27 4G network 2.5 GHz, 10 MHz TDD

Figure 27 shows the speed test results of a 4G network operating in the 2.5-2.7 GHz spectrum band with a 10 MHz channel width using time division duplexing, TDD. The measured upload speed was 1.4 Mbps and the download speed was 8.6 Mbps with 73 ms latency. While the upload data rate was not high enough to support a 1080P video call, it was close to the 1.5 Mbps upload speed required for a 720P video call. The upload data rate of 1.4 Mbps is enough for a 500 kbps video send stream of a smart phone video call and the 8.6 Mbps download data rate has lots of room to support a 500 kbps receive video stream from a smart phone video call.

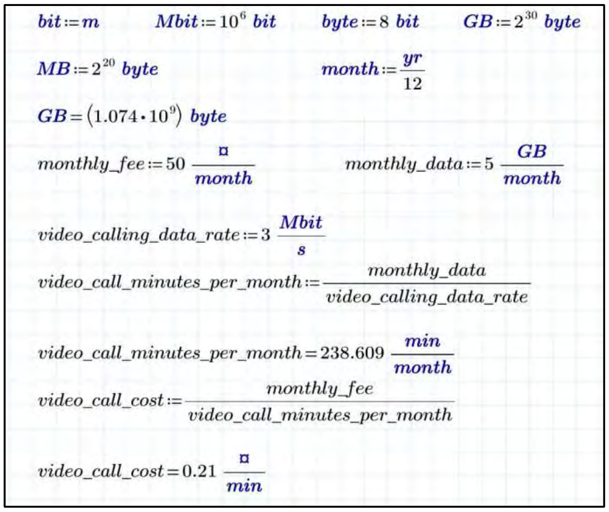
Figure 28 shows the parameters for the speed test results shown in Figure 27. The center frequency of operation is 2.647 GHz. The received signal strength is a very high -46 dBm indicating very good RF signal conditions and probable operation in close proximity to a base station. The carrier to interference and noise ratio was 21 dB. The transmit power was -19 dBm, the transmit power can go up as high as +20 dBm if the attenuation of the RF signal to the base station is high.



4G Network	Connected	TX Power	-19 dBm
Network Operator	XFINITY	Center Frequency	2647000 KHz
RSSI	-46 dBm	Preamble	27
BSID	00:00:02:06:03:38	NSP-ID	2
CINR	21 dB	NAP-ID	2

Fig. 28 4G Network 2.5 GHz, 10 MHz TDD

While 4G networks have the technical capability to make video calls under good RF conditions, will the costs impede usage? Figure 29 shows some calculations that translate a typical 4G data plan into some metrics familiar to many who in the past have bargain shopped for long distance plans based upon cents per minute or cellular phone plans based upon monthly minutes of talk time. The plan analyzed is a cellular data plan with 5 GB for $50 per month. If a video call is assume to have a data rate of 3 Mbps, the data rate measured for a 720P video call, then a video call is 21 cents per minute. At one point in time 10 cents a minute for a long distance plan seemed like a good deal. Cellular data plans with monthly minutes of talk time tend to be priced about 9 cents per minute. In terms of monthly talk time if one was to switch from voice calling to video calling, $50 per month with a 5 GB data cap gives one 238 minutes



$bit := m \qquad Mbit := 10^{6}\ bit \qquad byte := 8\ bit \qquad GB := 2^{30}\ byte$

$MB := 2^{20}\ byte \qquad month := \frac{yr}{12}$

$GB = \left(1.074 \cdot 10^{9}\right)\ byte$

$monthly\_fee := 50\ \frac{\text{¤}}{month} \qquad monthly\_data := 5\ \frac{GB}{month}$

$video\_calling\_data\_rate := 3\ \frac{Mbit}{s}$

$$video\_call\_minutes\_per\_month := \frac{monthly\_data}{video\_calling\_data\_rate}$$

$video\_call\_minutes\_per\_month = 238.609\ \frac{min}{month}$

$$video\_call\_cost := \frac{monthly\_fee}{video\_call\_minutes\_per\_month}$$

$video\_call\_cost = 0.21\ \frac{\text{¤}}{min}$

Fig.29 Calculating the cost of a 4G video call

of talk time with video calls at 3 Mbps. Voice cellular plans in this price range tend to offer about 450 minutes of talk time. So with these parameters, making video calls rather than voice calls tends to cost about twice as much. Of course, the assumed bit rate of the video call is the critical parameter. If the video calls were all 10 Mbps 1080P then it would be very expensive, 70 cents per minute and only 70 minutes of monthly talk time. However, on the other hand many folks may be quite content with making video calls on the road with a smart phone operating at 1 Mbps, in this case the cost per minute is 7 cents with 715 minutes of monthly talk time. These last numbers are roughly equal to the cost of cellular voice calls. So if you have a smart phone and a 4G data plan, live it up, make a video call instead of a voice call.

CONCLUSION

Many things have come together recently to encourage the use of video calling. Broadband connections in the home are faster than ever. Many homes have wireless home networks to connect mobile devices. In the past providing Internet connectivity to your television may have been inconvenient due to the lack of a nearby CAT-5 outlet. The wireless home network takes away the inconvenience. More and more people carry smart phones with WiFi and 3G or 4G network connectivity and these smart phones have front and back cameras and video calling application software.

Tests of video calls have shown that a 1080P video call can run at a symmetrical 10 Mbps, with the video send stream at 5 Mbps and the video receive stream at 5 Mbps. An excellent quality 720P video call on a large screen television set can run at 3 Mbps total data rate. 3-way video calls tend to run at about 2 Mbps. Smart phone video calling with a 4.3 inch diagonal display runs at about a 1 Mbps data rate over a home WiFi network. Video calling using a 3G cellular network runs at about 200 kbps with lower video quality and reliability.

Video signals are sent in packets of about 1400 bytes. Wireless home networks supporting video calls tend to have very concentrated packet size distribution around 1400 bytes and 20 bytes, representing the video packets and signaling packets, respectively. A typical packet sequence of a video call over WiFi lasts about 250 µs and consists of a request to send signal, a clear to send signal, the data packet of 1400 bytes, and a block acknowledgement signal.

Tests were performed of a 720P video call and a 1080P video call whereby all of the WiFi packets were captured. The captured packets contained information on byte size on the wire and data rate of the modulated burst. Accounting for the preamble length, the time length of each packet transmission was calculated and statistically analyzed. With the duration of the video call and the transmission time of each packet during the call determined, the percentage of time that the wireless home network was used by the video calling application was determined. For a 1080P video call under ideal RF conditions, the duty cycle was found to be 25%. For a 720P video call under threshold RF conditions the duty cycle was 20%. This indicates that most wireless home networks could support no more than 4 to 5 simultaneous video calls and that even during a video call over a wireless home network there is still over 75% of the capacity available for spectrum sharing.

Finally, the distribution of packet sizes of a video call using a 3G network was measured and analyzed. The packet size distribution shows that packet size in general was not as large when making a video call over a 3G network. The video calling application adjusted to the higher latency and packet loss of the 3G network in order to make the call while sacrificing quality. The speed of 4G networks was measured and reported indicating that 4G networks do support the data rates required for high quality video calling, at least under ideal RF conditions. The cost of video calls on 4G networks was analyzed and it was found that with today's pricing plans, video calls of very high quality are more expensive than voice calls but not prohibitively so, while lower quality video calls on a smart phone screen today cost about the same as most common voice plans.

Wireless network connectivity is a crucial factor in encouraging the use of video calling. The signal strength is a critical indicator, wireless home networks should have signal strength indication of -60 dBm or higher for reliable video calling. Signal strength of -70 dBm for a wireless home network connection was found to be at the threshold of operation for a successful video call. The use of 5 GHz band can work better than 2.4 GHz band but only at close proximity. It was found that during a video call with a 3x3 AP and 3x3 client at 5 GHz in close proximity that 3 stream operation was very rare. It was also found that during a video call with a 2x2 AP and 2x2 client at 2.4 GHz at a 36 foot AP to STA separation distance that 2 stream operation was very rare.

REFERENCES

[1] Dorros, Irwin, "Picturephone", Bell Laboratories Record, Vol. 47, Number 5, May/June 1969, pp. 136-141.

[2] Urban, D., Albano, C., Devotta, D., "Delivering DOCSIS 3.0 Cable Modem Speeds over the Home Network", SCTE Cable-TEC EXPO'10, October 2010.

[3] Urban, D., Albano, C., Ong, I., Gilson, R.,"Designing a reliable wireless home network in a residential environment to optimize coverage and enhance application experience", SCTE CABLE-TEC EXPO '11, November 15-17 2011 Atlanta Georgia.

[4] IEEE Std. 802.11n-2009, "Amendment 5: Enhancements for Higher Throughput.

ABBREVIATIONS

AP Wireless local area network access point
STA Wireless local area network client station
OFDM orthogonal frequency division multiplexing
GI guard interval for OFDM
CDMA code division multiple access
HFC Hybrid Fiber Coaxial Cable network architecture
CM cable modem
RTS request to send WLAN signal
CTS clear to send WLAN signal
Block ACK Block Acknowledgement WLAN signal
WLAN Wireless local area network
LTE Long Term Evolution 4G cellular network
WiMAX type of 4G cellular network
4G OFDM based cellular network
3G High speed CDMA cellular network
RTT round trip time in ms
HT High Throughput WiFi mode
MIMO multiple input multiple output antennas
IFFT inverse fast Fourier transform
FFT fast Fourier transform
TDD Time Domain Duplexing
FDD Frequency Domain Duplexing

ACKNOWLEDGEMENTS

The author wishes to thank colleagues at Comcast and especially Chris Albano, Doug Jones, Erami Botchway, Sam Chernak, Chris Bastian, Jorge Salinger, and Neil Spence for help and support in writing this paper.

ARCHITECTING THE DOCSIS® NETWORK TO OFFER SYMMETRIC 1GBPS SERVICE OVER THE NEXT TWO DECADES

Ayham Al-Banna
ARRIS Group, Inc.

Abstract

The paper analyzes various options to increase the capacity of HFC networks in order to meet the capacity demands over the next two decades. A smooth migration plan is proposed to enable MSOs offering beyond than 1Gbps US service. A High-split prototype system is built and initial results are introduced.

1. INTRODUCTION

The current architecture of Hybrid Fiber Coaxial cable (HFC) networks along with the exponential growth in bandwidth demand are placing the cable Multiple Service Operators (MSOs) at competitive disadvantage as they face capacity limitations. These limitations may preclude the MSOs from satisfying the customers' demands if not properly addressed.

In order for the MSOs to maintain their business and offer more services at faster speeds (e.g., services to business customers, IPTV fans, gamers, etc.), they need to immediately start brainstorming, architecting, and upgrading their networks in ways that will meet the pressing bandwidth demands. This process requires taking smart and gradual steps toward the goal system architecture, which will support beyond than symmetrical 1Gbps service.

Multiple factors need to be considered while going through the system and plant migration: cost, network architecture, spectrum allocation, operational issues, technical challenges, headend equipment (e.g., Converged Cable Access Platform (CCAP) compatible?, servers scale?, etc.), customers Quality of Experience (QoE), etc. The list goes on! Not only do MSOs have to think about the above factors as they prepare their networks for future services, they also need to think thoroughly about the appropriate sequence of steps to take such that an optimal architecture is achieved. The optimal architecture can be defined as a flexible network topology that results in maximum capacity and minimum cost over extended periods of time.

This paper is organized as follows. Section 2 describes the traffic growth trends based on recent real data. Several multiple factors that play heavily in the decision process of network migrations are briefly described in Section 3. Section 4 lists and analyzes the available options to extend the US BW to offer 1Gbps service. A sample plan that offers *smooth* migration steps to result in an optimal network architecture, which offers symmetric 1Gbps architecture and multi-gigabit system in the future, is described in Section 5. Section 6 concludes the paper.

2. RECENT TRENDS IN BW DEMAND

The traffic demand has been growing exponentially for the last 30 years. Different applications and services appeared at different times over the last three decades to ensure that the traffic growth stays on track! Among many, business services, gaming, and IPTV make today's motivation for guaranteed traffic growth for the next few years. The constant traffic growth over the past three decades is shown in Fig. 1, which shows the maximum DS rate per subscriber over cable networks [1]. This curve is sometimes referred to as the Nielsen curve for Cable networks.

Recent data obtained from different MSOs shows similar growth pattern for the average traffic on their networks. In particular, Fig. 2 shows the DS BW Average Cumulative Growth Rate (CAGR) per subscriber for three different MSOs over the past couple of years. Note that the CAGR value for all MSOs is more than 50% per year. Figure 3, on the other hand, depicts the US BW CAGR per subscriber over the past two years for two different MSOs. Observe from the figure that the CAGR averaged over both MSOs results in an US BW growth rate of about 30% per year.

The data in Fig. 2 and Fig. 3 shows that while some MSOs may observe slow growth rate on their networks, other MSOs observe larger growth rates. Additionally, the cumulative traffic growth averaged over all MSOs for the past two years agrees with the traffic growth trend observed for the past thirty years as was shown in Fig. 1.

From Figs. 1 through 3, the average DS and US BW per subscriber CAGR is shown to be >50% and 30%, respectively, for the past thirty years. Therefore, it might be reasonable to assume that the traffic growth will maintain the same trend in the future. In subsequent analyses in this paper, where we focus on the US BW problem in HFC networks, we assume that the US CAGR is at 30% on average.

Given that the US CAGR is at 30%, the question at hand is: when will the current 5-42MHz spectrum be totally consumed and therefore an upgrade of some sort is necessary? The answer to that question not only depends on the US CAGR, but also on the value of the maximum offered subscriber rate (Tmax) today. Between the US CAGR and Tmax values offered today, it will be straightforward to predict when the current US spectrum runs out of capacity, which will be shown later in this section.

The maximum offered rates have been published recently [2]. Table 1 shows DS and US Tmax values currently offered in North America. Note that the table lists Tmax values offered by different industries (Cable and others). Tables 2 and 3, show DS Tmax values offered by different MSOs in Europe. Observe that some European MSOs offer higher rates than their counterparts in North America. In particular, the maximum DS Tmax currently offered in Europe is 360Mbps by Zon Multimedia (See Table 3). The current offering of Zon for DS Tmax and US Tmax shows a constant ratio of 15 between the rates. Therefore, the US Tmax value offered by Zon is assumed to be around 24Mbps. Note that this is close to the 20Mbps US Tmax being offered by Videotron in North America.

One important point to observe from Table 1 is that the maximum Tmax service is offered by Verizon, which is not a cable MSO! Therefore, in addition to customer traffic demand, Table 1 clearly shows the other side of the equation that pushes cable MSOs to add capacity to their networks: Competition!

With the assumptions that the US CAGR is 30% and the current offered US Tmax value is 24Mbps, the next step is to calculate the time when the current US spectrum runs out of capacity. Given a certain US Tmax value per subscriber, some MSOs might consider providing a total capacity of 1.5*Tmax in order to offer service with adequate Quality of Experience (QoE) to their subscribers. Other MSOs might choose other factors that are different from 1.5*Tmax (e.g., 2*Tmax). For the analysis in this paper, we assume that a capacity of 1.5*Tmax is required in order to offer good QoE service for customers with Tmax as the maximum rate per subscriber.

Figure 4 shows the extrapolated growth of US Tmax per subscriber using the above assumptions. Note that the current US spectrum (5-42MHz) capacity is assumed to be around 133Mbps. This is because the total BW of 37MHz may not be completely usable at the highest possible modulation order (some channels can potentially run at QAM256 while others will run at QAM16). Also, strong FEC is assumed for the same reason (some parts of the spectrum are very clean while others are really challenging). The combination of moderate order modulation order (QAM64) and strong FEC (code rate = 0.75) compensates for noisy channels, unusable spectrum, and spectrum that used for services other than data). The total capacity of 133Mbps might be close to what MSOs can achieve in the real-world from the 5-42MHz spectrum. You may notice that this number is a little higher than the more conservative estimates that have been published earlier by the ARRIS' team (the author included) [3], which assumed a total capacity for the 5-42MHz to be around 118Mbps. Upcoming sections in this paper, where comparisons between the capacities of different split options is provided, will refer to the past capacity work and will point out that the estimates might be a little conservative and therefore can be slightly increased. In all cases, the total capacity always depends on the plant condition and MSO's usage plan for the spectrum. Observe, however, that the difference in capacity numbers (15Mbps) due to different assumptions is not significant given the Tmax CAGR growth rate shown earlier.

Observe in Fig. 4 that the current US spectrum runs out of Tmax capacity just before year 2017. This corresponds to service offering of Tmax~90Mbps. Note that 1Gbps US Tmax service will be required around year 2026, if not earlier. One may realize that it not too early to start planning for network architecture updates and migration strategies in order to offer capacities that satisfy the projected traffic demands over the upcoming years.

3. PLAYING FACTORS IN HFC NETWORKS MIGRATION

This section briefly describes the various factors to be taken into consideration while going through the system and plant migration process in order to meet the capacity demands over the next two decades. Not only these elements need to be studied thoroughly, but also the interaction between them needs to be analyzed carefully. The interaction happens because some elements depend on others, where the choice of some elements affects the choice of others. There might be no one ideal solution for all MSOs. However, different MSOs may have different optimal solutions depending on their position from the factors listed below.

3.1. Network Architecture

Both the components composing the network and network topology affect the performance heavily. The number and characteristics of amplifiers, line extenders, bridgers, taps, and other passive devices affect both signal loss and noise. The characteristics of some of these equipment also define the operational BW where signals can be transmitted in the DS or US direction. The type and length of coaxial cables (trunk and drop) affect the signal loss too. The length and type of fiber links as well as the features of the optical transmitter and receiver also affect the performance.

How deep the fiber node in the plant affects the performance. For example, longer cascades results in more attenuation, noise, and worse filters roll-offs, which impair the signals transmitted around band edges. Shorter cascades on the other hand, result in better network performance.

Networks topology needs to be analyzed frequently because the plant topology changes over time as MSOs update their network to expand the capacity of their networks. The change in network topology may affect various customers differently depending on the location of the customer relative to the network update. Specifically, Fig. 4 shows an example of N+5 network topology. After node segmentation and splitting, the network topology becomes as shown in Fig. 5. Note that it is sometimes difficult to balance the number of subscribers between new fiber nodes as apparent from Fig. 5, which affect the capacity per subscriber. Also, the example in Fig. 5 is a good illustration to the node splitting and segmentation process whose output does not guarantee that new nodes have the same cascade length. Figure 5 shows that the resultant nodes possess different lengths (i.e., different number of cascaded amplifiers behind the fiber nodes). Again, this affects the attenuation, noise, and therefore capacity.

The number of cascaded amplifiers behind a fiber node has declined over the years. Some MSOs estimate the current average of their networks to be at N+5 (to N+6)[1]. The current network topologies along with the limited US spectrum (5-42MHz in the USA, 5-65MHz in Europe) place a tight limit on the capacity that can be offered by today's networks and therefore gradual sequential upgrades will be necessary to cover the demand as well as competition over the next two decades!

3.2. Spectrum Allocation

This is a critical topic because it touches many areas. The choice of which split to choose for the US spectrum (mid-split, high-split, top-split) comes as a result of studies of technical feasibility, which analyzes the technical challenges and offered capacity associated with the implementation of each split option. Besides cost, operational aspects are affected depending on the chosen split. For example, affected operational parts include: reclaiming/reallocating analog TV channels, moving DS spectrum, capping DS BW, transition bands (guard bands between DS and US), addressing the Out-Of-Band (OOB) signaling of Set-Top Boxes (STB), etc. This factor (spectrum allocation) is studied in more details in later sections of this paper.

3.3. Operational Issues

Various Operational issues are to be addressed when network migration occurs. Depending on the network architecture and the update to occur, operational aspect that can be affected include: Analog channels reclamation and reassignment, specifying spectrum for DOCSIS and digital channels, addressing STB OOB signaling, DOCSIS and Video management, network maintenance process (depending on equipment being in the headend or in the headend and FN together), network reliability and availability (again, related to equipment being in headend or in headend and FN). Observe that placing more intelligent equipment in the FN introduces higher risk in terms of network availability and reliability. Some of these operational aspects will be addressed in later sections of this paper.

3.4. Technical Challenges

The technical aspects of any solution or proposed network update must be studied thoroughly. The technical study results in recommendations regarding feasibility, cost, capacity estimates, and implementation requirements. For example, the feasibility of certain US spectrum split is a function of the signal attenuation experienced on that split. Another example of how technical studies are important is that understanding the different noise and channel impairments, which exist

[1] The total number of actives behind a single FN is currently estimated to be around 30.

on different parts of the spectrum and how they can be mitigated via different PHY and MAC technologies, will affect the proposed solution requirements, capacity, efficiency, and cost. A technical evaluation of different capacity-expanding options is included later in this paper, where a migration plan is proposed.

3.5. Headend Equipment

While network topology affects the system performance and offered services, headend equipment also plays a major role into that. The MSOs needs to make sure they specify requirements for products that can scale very well with the projected service offerings. This scale is related to number of channels as well as number of service groups, service group size, servers scale, management and scale of IP addressing scheme (IPv4 & IPv6), etc.

Additionally, not only scale is important, but also the architecture of the headend equipment should be chosen to minimize cost and maximize capacity. Available architectures include Integrated and modular. The MSOs need to make sure they place requirements that result in optimal system architecture in terms of capacity and cost.

On a side note, the Cable industry already started the effort of specifying the scale and requirements of the next generation network architecture, where different requirements were listed in the Converged Cable Access Platform (CCAP) specifications.

3.6. Quality of Experience (QoE)

Quality of Experience is one of the most challenging topics to be addressed. The problem with this topic is that it deals with the customer's perception about the service. The MSO has to collect various system and traffic parameters in order to analyze how the service offering is rated in the customer's eye. The MSOs normally works with system vendors on developing different algorithms and performance metrics that measure the satisfaction of the customers. In this kind of analysis, good questions to be addressed include: For how many seconds can the subscriber wait for a webpage to download? What is the webpage size that the customer is trying to download? How often is he online? How often does he jump between pages while online? What about games latency? What is the pattern of the traffic of a certain game? Does that apply to all games? Does statistical multiplexing help? If so, how does it interact with the number of bonded channels?, etc. The list can go forever!

In order to make sure that the customer has good QoE, the MSOs also need to understand how networks availability affects QoE. Additionally, the effect of the FN size, SG size, offered Tmax needs to be analyzed and understood. Then, the MSO may need to work with system vendors to create algorithms that manage latency and service flows priorities to result in best potential customer QoE.

3.7. Cost

This is the most important factor to consider when planning networks migrations. It is a function of all of the above factors. The goal of network migration is to offer adequate capacity at minimum cost. In many scenarios, the MSOs use the cost per unit of BW as a metric to decide between different proposed solutions. The cost of a certain proposal should take into consideration the investment protection provided by different solutions. It is instructive here to mention that backward compatibility can offer large cost savings, for most of the time, as it capitalizes on using the established base. In many cases, the savings exceed the added cost and complexity which occur when requiring that the new solution be backward compatible with the existing technology.

3.8. Next Steps & Sequence of Steps

There are many network topology options to consider when it comes to the plant migration. These options include: utilizing the available spectrum efficiently, expand the US spectrum, introduce new techniques for better spectral efficiency (like more efficient Forward Error Correction (FEC)), introduce new robust and more efficient PHY technologies (like Orthogonal Frequency Division Multiplexing (OFDM)), require backward compatibility for added enhancements, Go deeper with fiber, etc.

The decision of choosing particular options and the sequence of implementing the options depend on all of the above factors that need to be analyzed thoroughly. In particular, the available options listed above need to be evaluated technically, operationally, and financially. The purpose of this paper, in the next few sections, is to analyze these proposals from the technical point view to provide recommendations to the MSOs as they brainstorm about their network. The technical analysis will provide implications regarding the cost of different solutions. Some options will also be evaluated from the operational point view.

4. OPTIONS TO ACHIEVE 1GBPS IN THE UPSTREAM

This section lists and analyzes the different options, from which the MSOs can choose when planning networks updates in order to produce the goal network architecture. Along with the analysis, technical and operational challenges that may appear throughout the migration process will be exposed and addressed.

4.1. Utilizing the Available BW Efficiently

The utilization of the current 5-42MHz spectrum is far from efficient. In particular, there are portions of the spectrum that are not used at all, while other parts are used inefficiently such that the obtained capacity is way less than what can be potentially offered by that part of the spectrum.

The DOCSIS3.0 has many tools and features in order to help the MSOs achieve the best capacity out of their US spectrum [4] [5] [6] [12]. Some of these parameters include:

- Multiple access technologies (e.g., Advanced Time Division Multiple Access (ATDMA) and Synchronous Code Division Multiple Access (SCDMA)). SCDMA can be very helpful in fighting impulse noise in the lower part of the 5-42MHz spectrum.
- Center frequency selection
- Symbol rate range (0.16 – 5.12 Msymbol/sec)
- Modulation orders (QPSK, 8QAM, 16QAM, 32QAM, or 64QAM)
- Reed-Solomon Forward Error Correction (RS-FEC) to correct up to 16 bytes
- Codeword size selection
- 24-tap pre-equalization
- Long preambles up to 1536 bits
- Ability to adjust to longer/more powerful Preambles
- Proprietary noise mitigation techniques
  - Ex: Ingress Noise Cancellation
- ATDMA Interleaving…
- SCDMA Interleaving
- SCDMA de-spreading
- SCDMA spreading
- SCDMA Trellis Coded Modulation (TCM)
- SCDMA Maximum Scheduled Codes (MSC) feature
- SCDMA Selective Active Codes (SAC) feature
- Channel bonding (MAC layer feature used for PHY layer noise mitigation)
- & Many Many others (Last Codeword Shortened (LCS), max burst size,

scramble seed, differential encoding, etc.)

Detailed analysis of utilizing the above tools and optimizing the spectrum usage can be found in [4] [5] [6]. The abundance of parameters and the flexibility in choosing their values makes it a challenge to optimize them to result in the best spectral efficiency. Therefore, automated tools can be used to measure the different types of noise and also search the solution space of all the parameters and choose the optimal ones that result in the best spectral efficiency. For example, Fig.7 shows that the spectrum can have different types of noise in different portions. Therefore, the automated algorithm shown in Fig. 7 captures the noise in the channel and specifies the best modulation profile and channel parameters that result in the best spectral efficiency. Any automated algorithm needs to have the flexibility to specify constraints for the optimal solution. This is highly desired especially if the MSO does not want to use certain parameters or want to specify certain range for specific parameters. An example of that is shown in Fig. 8, where the algorithm can accept multiple constraints and then searches the constrained solution space to find the optimal parameters that result in the best spectral efficiency.

4.2. Segmenting and Splitting Nodes

Examples of node splits and segmentations were provided in previous sections. The process of node split and segmentation helps in many ways:

- Less Noise funneling as a result of reducing the number of subscribers per node or service group. Lower noise translates to higher SNR and therefore increased capacity.
- Less attenuation because: the deeper the node is, the shorter the coaxial cable becomes, and therefore less signal attenuation is introduced. The lower attenuation translates to higher SNR and therefore increased capacity.
- More average capacity per subscriber. This comes as a natural result of reducing the number of subscribers per node or service group.
- Less contention for BW. Again, this is a natural result of reducing the number of subscribers per node or service group. The reduction in BW contention makes the assumption of requiring 1.5Tmax (or 2Tmax) of capacity to offer Tmax service more reasonable.

Since node splits and segmentations offer all of the above benefits and increased capacity, one may think of performing this process infinitely as demand increases. This, in fact, can be a good approach! However, the cost of node splits rise exponentially every times they are to be performed because the number of resultant nodes doubles after every node split operation. Therefore, there will be a time, when performing the next node split operation will cost more than changing the US spectrum split or laying fibers all the way to the homes or and therefore the natural step after those many node split operations becomes Fiber To The Home (FTTH). This will then make the most reasonable decision from cost point view and also offers multiple times of capacity that may actually be needed by that time.

4.3. Adding More US Spectrum

At some point in the future, the MSOs will need to add more US spectrum to their networks to provide enough capacity to meet the traffic demands. Adding more US spectrum can take many forms: mid-split (5-85MHz), High-split (5-200MHz, 5-238MHz, 5-300MHz, etc.), and top-split (placing US spectrum above the current DS BW). This is shown in Fig. 9 and Fig. 10.

The above splits can be classified into two categories: diplex category (mid-split and high-split), and triplex category (top-split). In particular, in the diplex category, there is only one transition band in the spectrum which separates the US spectrum below the transition band and the DS spectrum above the transition band as shown in Fig. 9. The triplex category, on the other hand, contains two transition bands separating the US and DS spectra as shown in Fig. 10. Specifically, in the triplex architecture, the lower part of the spectrum is used by US traffic, which is followed by the first transition band that is followed by the DS spectrum. The second transition band sits above the DS spectrum and separates it from the US spectrum at the top.

In order for the MSOs to have enough capacity to offer 1Gbps Tmax service and beyond, they will need to move to either high-split or top-split as a goal architecture. This is because mid-split does not offer enough capacity and also MSOs may choose to move from sub-split to high-split directly (instead of going through mid-split) in order to save on the cost of plan upgrade. In particular, the move from sub-split to high-split directly avoids the need to touch the plan multiple times. Other MSOs, however, might choose to go through the mid-split step in order to avoid addressing the OOB STB signaling issue for few years, which allows them to phase out these STBs before moving to high-split architecture.

There are multiple advantages and disadvantage for both the top-split and high-split options. Some of the advantages of the top-split option are:

1. It does not interfere with the OOB STB signaling (frequency range is 70-130MHz).
2. The DS spectrum layout does not need to change. No video channels are affected.

On the other hand, there are several disadvantages for the top-split option including:

1. High signal attenuation, which results in reduced total capacity and inefficient spectrum usage (analysis shown later).
2. More expensive than the high-split option [3].
3. Requires two transition bands which translate to wasted capacity.
4. Requires large bandwidth for the top transition band. In general, the bandwidth of the transition band depends on the frequency of the band. Since the top transition band occurs at high frequency, the transition band bandwidth will be large and this translates to more wasted capacity.
5. Places a cap on the growth of DS spectrum. Once the US spectrum is placed on the top of the DS spectrum, there will be no room to expand the BW of the DS spectrum. Any future growth for the DS will be very challenging because it has to be on the top of the US spectrum and therefore results in these exact disadvantages of wasted capacity (if that option is ever feasible).
6. Requires high modem transmit power for reliable transmission (still at lower capacity).
7. Requires changing all actives to introduce the second transition band.

The high-split architecture, on the other hand, has various advantages including:

1. Offers the highest system capacity (analysis shown later).
2. Less signal attenuation.
3. Single transition band is required.
4. The transition band is narrow because it happens at low frequency.
5. Offers the cheapest solution [3].
6. Does not place a limit on the growth of the DS spectrum.

7. Leverages some of the existing HFC components like laser transmitters and receivers as some of them do support the high-split BW.
8. Offers some backward compatibility because the current DOCSIS3.0 specifications have the US DOCSIS defined from 5-85MHz. This capability already exists in the hardware of various CMTS and modem equipment.

Some of the disadvantages of the high-split option are:

1. It interferes with the OOB STB signaling.
3. It affects the layout of the DS spectrum because the bottom part of the DS spectrum is chewed by the new US spectrum. Some modifications to the DS spectrum layout and channel assignments need to occur.
2. Requires changing all actives to move the current transition band to a higher frequency.

As mentioned above, one of the challenges introduced by the high-split architecture is addressing the OOB STB signaling scheme. There are different scenarios for addressing this issue including:

1. Some MSOs do not have this issue because they have IP or DOCSIS STBs deployed as opposed to legacy STBs which require the signaling in the frequency range 70-130MHz.
2. Phase-out legacy STBs out of the plant. Some MSOs use 9 years as turn-over time for their STBs. Therefore, if the MSOs plan to move to the high-split option in the future and start planning accordingly, the legacy STB problem may not be an issue. The MSOs still have at least 5 years before they need to make any change with the spectrum from a Tmax perspective. This was illustrated in Fig. 4, where the offered Tmax capacity by the current 5-42MHz spectrum runs out of steam around 2017, when the MSO can offer about 90Mbps (assuming that a required channel capacity of 1.5Tmax to offer Tmax service). Note, however, that if the MSOs assume 2Tmax capacity is needed to offer Tmax service, the 5-42MHz spectrum will be consumed (from Tmax point view) one year elarier, namely in 2016, enabling the MSOs to offer Tmax service of ~70Mbps by then. The date, when the capacity of the 5-42MHz spectrum is consumed, can be pushed further in the future if spectrum is used more efficiently via optimizing modulations profiles parameters (shown in earlier sections) and introducing DOCSIS enhancements (will be explained in later sections).
3. Use up-conversion and down-conversion techniques to move the STB signals to higher frequencies beyond the high-split limit. Several approaches are available to perform this, where each approach has its own advantages and disadvantages. The discussion of these solutions is outside the scope of this paper.

Extensive analysis has been done by the ARRIS' team (the author included) to compare different split options from cost and capacity point view [3]. The detailed analysis in [3] is summarized here for convenience. This analysis shows that the high-split option is the most economical solution that offers the highest capacity.

The assumptions used in this analysis are kind of conservative because it was assumed that parts of the spectrum are completely unusable (which may not be the case in most plants). Also, the analysis defines the capacity to be the available DOCSIS3.0 bonding capacity offered by the spectrum. In other words, the analysis does not assume channels used for legacy devices or spectrum monitoring to be part of the available capacity. Specifically, only 22.4MHz was assumed to generate the capacity numbers for the 5-42MHz spectrum. This was rationalized by the different items listed in Table 4. Others assumptions used for this analysis are shown in Tables 5 and 6, while the analysis results are shown in Fig. 11. As mentioned earlier, these numbers can be slightly increased because the assumptions were a little conservative. However, this may not change the course of actions that the MSOs need to do to augment their networks because the difference is insignificant compared to the CAGR of US Tmax.

As can be seen from the above analysis, the high-split option makes the best potential choice for the US spectrum as MSOs plan to upgrade their networks to offer adequate capacity for the required Tmax offerings. Therefore, ARRIS has built a high-split prototype system to mimic the example real-world N+3 network architecture shown in Fig. 12. The real prototype setup is show in Fig. 13. In Fig. 13, all of the active HFC components are ARRIS-made and modified and support 200MHz high-split operation.

The purpose of this effort is to characterize the system and identify any potential limitations or hurdles that may appear as a result of transmitting US signals using the high-split system. The ultimate goal of this experiment is to develop and propose solutions to any identified challenges well before the time of real network migration has come. System analysis for the high-split setup in Fig. 13 has already started. Fig. 14 shows an initial Noise Power Ratio (NPR) curve measured at early stages of the experiment. Further analyses and experiments are still pending and the obtained results will be shared in future papers.

4.4. Introducing PHY Enhancements (Higher Order Modulations) for Better Spectral Efficiency

Introducing higher order modulation options for US transmissions can be a smart move to increase the offered capacity. Currently, the US part of DOCSIS3.0 can support up to QAM64 (or QAM128 with Trellis Coded Modulation (TCM)). Introducing higher order modulations like QAM256, QAM1024, and QAM4096[2] can help in achieving higher spectral efficiencies if/when the plants can support them. For the above modulation orders, QAM256 offers 33% more spectral efficiency than QAM64. QAM1024 offers 25% more capacity than QAM256, and QAM4096 offers 20% more capacity than QAM1024.

As mentioned earlier, node splits and segmentations can result in reduced signal attenuation and noise funneling. Both of these result in higher SNR values that enable the operation of higher order modulation profiles. DOCSIS3.0 noise mitigation toolkit can also help enable the use of higher order modulation orders. Additionally, the next two sections will explain few enhancements that can be added to the DOCSIS, which result in SNR gains that can enable the operation of high order modulation orders.

4.5. Introducing PHY Enhancements (New PHY) for Better Spectral Efficiency

Enhancements to the DOCSIS standard can go beyond offering higher order modulations. Adding modern transmission technologies to DOCSIS toolkit can increase the spectral efficiency. For example, the multi-carrier Orthogonal Frequency Division Multiplexing (OFDM) technology is one of the common PHY techniques used in many of the modern applications including the

[2] These are even-order modulations. Odd modulation orders can be proposed too for higher granularity.

European standard Digital Video Broadcast standard (DVB-C2) [7].

OFDM can be implemented efficiently using the Fast Fourier Transform (FFT) algorithm. Therefore, it requires less chip resources when compared to other transmission technologies with comparable noise immunity, which is one attractive feature that enabled OFDM to be used by different applications. OFDM is also known to have good immunity to various types of noise and channel impairments, which is enabled by the use of subcarriers that also results in long symbol duration, which helps the performance in the presence of impulse noise. The good noise immunity is another attractive feature of OFDM. The proposal to use OFDM (to be exact, Orthogonal Frequency Division Multiple Access (OFDMA)) for US DOCSIS is not a new concept in this paper. In particular, the author analyzed the performance of multi-carrier signals in the presence of HFC noise in 2009 [8] and also proposed the use of OFDM technology for US transmissions in DOCSIS back in 2010 [9].

This section analyzes the gain obtained from using OFDM for US transmissions in DOCSIS networks. The gain obviously depends on the assumptions and input parameters to the model. The analysis assumes an Additive White Gaussian Noise (AWGN) channel. Therefore, the analysis shown here is not an extensive or comprehensive analysis but only shows the gain obtained for one example scenario. More detailed analysis for the benefits of using multi-carrier signals can be found in [9]. Fig. 15 shows capacity estimates for an US single carrier DOCSIS signal and Fig. 16 shows an analysis for the capacity of 200MHz high-split system that uses OFDM. Comparing the results in Fig. 16 to those in Fig. 15, the gain resulting from using OFDM instead of Single carrier is about 2.6% or 0.129 bps/Hz of capacity improvement.

Observe that the increased capacity obtained from introducing OFDM as a new PHY technology for US transmissions is 2.6% when calculated at QAM256. Note that this value is highly dependent on the choice of the OFDM parameters, particularly the cyclic prefix code length and the preamble-to-burst-length ratio. The above improvement at QAM256 is equivalent to an additional 0.214 bits, which translates to 0.63dB of SNR gain. Although the gain may not be very large, some MSOs may choose to use OFDM for US transmissions in order to utilize the US spectrum in the most efficient way and also to use the noise and impairment immunity of OFDM to provide reliable transmissions in harsh plant conditions. In fact, the gain provided by the use OFDM can increase significantly when other parameters are used and also when different noise types (other than AWGN) and channel impairments exist on the channel [9].

Apart from the insignificant capacity improvement provided by OFDM when used with US DOCSIS transmissions shown in the above example, there are many benefits that can be drawn from using OFDM for US DOCSIS including:

1. Backward compatibility with US Channel Bonding: The MSOs can consider bonding across two different PHY technologies and therefore achieve the best possible spectrum utilization. This concept was originally introduced in [9].
2. Easy coexistence and smooth migration: The ability to turn on/off OFDM subcarriers makes it straightforward to accommodate legacy channels within the BW used for the new technology. The reader may be referred to [9] for more details.
3. Low Cost and Optimized Implementation [9]: The OFDM is based on the efficient FFT algorithm and is believed to result in simpler

implementation, which translates to lower cost.

4. Robust to noise and channel impairments: the OFDM is one of the most powerful PHY technologies in terms of its ability to fight different noise types and also mitigate interference [8] [10]. In fact, OFDM is used for wireless channels which are more challenging than the DOCSIS US channels because of multipath fading.
5. More Efficient US Bandwidth Utilization: the analysis above shows that OFDM can result in better spectral efficiency. The analysis assumed AWGN channel, where the results showed minor gain. The gain can be much larger when different noise scenarios and channel impairments exist on the plant [8] [9].
6. Load-Balancing: MSOs can choose to load-balance the traffic on the US between two different PHY technologies. This concept was originally introduced in [9] and also helps with backward compatibility.

One *potential* drawback of OFDM is increased latency. This can appear if the subcarriers width is selected to be very small, which results in increased symbol duration and therefore extended latency. If the subcarriers width is chosen in such a way that the OFDM symbols durations are similar or shorter than the SCDMA symbol durations used in DOCSIS, there will be no extra latency.

4.6. Introducing PHY Enhancements (New FEC) for Better Spectral Efficiency

Another enhancement that can be added to DOCSIS, which results in highly efficient spectral efficiency, is the use of modern Forwarded Error Correction Techniques (FEC). For example, Low Density parity Check (LDPC) codes are known to be much more efficient that the traditional Reed-Solomon codes (RS) codes that are currently being used in DOCSIS. The LDPC scheme was invented many years ago (in 1960's) by Gallager who, at the time, was working on his PhD thesis in MIT on this topic [11]. The LDPC error correction scheme was abandoned for many years because of its implementation complexity that needs high processing power. Recently, LDPC codes have been used in many applications including the DVB-C2 standard [7], which was enabled by the advances in processing platforms.

In order to evaluate the gain offered by the LDPC coding scheme, computer simulations were performed for a QAM signal with un-concatenated Reed Solomon (RS) to represent the current DOCSIS signals [12]. The results of these computer simulations (packet size = 250Bytes) are plotted in Fig. 17 along with other performance numbers for QAM LDPC FEC that are obtained from the published DVB-C2 standard [7]. Note that the above simulated numbers for RS FEC are close to the numbers derived from the J.83 Annex A, where RS FEC and not concatenated RS (RS with convolutional codes) [13] is used, and also similar to the DOCSIS US signals that use vanilla RS FEC. If comparison is to be made against concatenated RS FEC, one will find that the gain achieved by adding LDPC FEC is less because concatenated RS FEC performs better than vanilla RS FEC. Vanilla RS FEC was used in this analysis because it is what currently being used in DODCSIS US transmissions.

Observe that the gains in the QAM256 for the three plotted data points are 4.4dB, 5.1dB, and 7dB, depending on the code rate. Similarly, the gain ranges between 4.2dB and 5.5dB for the QAM64 case depending on the code rate. Therefore, the *average* gain between the LDPC numbers (from DVB-C2) and the RS S numbers (simulated DOCSIS RS FEC) is found to be 5.5dB and 4.85dB for the

QAM256 and QAM64 modulations, respectively. These average SNR gains of 5.5dB or 4.85dB translate to 1.83 bits and 1.62 bits of capacity improvement, respectively.

The above analysis used the *average* SNR gain obtained from using LDPC (the gain is function of the code rate). Therefore, one can be extra conservative and assumes a minimum gain or generous and assumes maximum gain depending on code rate usage on the target network. This paper uses the average gain in the analysis as a reasonable assumption.

4.7. Protecting the Established Base via Backward Compatibility and/or Coexistence

Backward compatibility and coexistence are critical tools to attain investment protection for the established base. As mentioned above in the new PHY proposal, backward compatibility and coexistence can be achieved easily using the OFDM PHY technology if selected as a new PHY for future DOCSIS US transmissions. Several aspects of backward compatible features are offered by OFDM: backward compatibility with US channel bonding across different PHYs, coexistence via the ability to turn on/off subcarriers of OFDM, and load balancing between the legacy and new PHY channels [9].

4.8. Going Deep with Fiber

FTTH is still way in the future! With the current offered capacities and the various available options for MSOs to augment their networks to result in increased capacity, there will be so many years before the MSOs will need to go down the FTTH path.

In fact, gradual migration steps that the MSOs do normally get them smoothly toward FTTH. For example, node splits and segmentations process gets the node closer to the subscribers' homes, which makes it easy and more economical to jump to FTTH at some point in the future. By then, the required capacity will be high (multi-gigabits) and the move to the FTTH will come in the right time. This is one of the beauties of cable networks that they offer the opportunity for timely investments, where spent money and resources are actually used. This is opposed to investing in FTTH, where a large amount of money and resource is spent to offer capacities that are not needed yet.

4.9. Capacity Analysis Summary

This section summarizes the capacity analyses that were introduced in previous sections of this paper. We will start with the analysis from section 4.3, where expanding the US spectrum was proposed. We will use the estimates from that section [3]. Assuming QAM256, the net offered capacity by 200MHz high-split was found to be 855.6Mbps, while the net offered capacity by 238MHz high-split was found to be 999.5Mbps (1Gbps).

Section 4.6 showed an average SNR gain of up to 5.5dB using LDPC alone for the QAM256 case. Additionally, section 4.5 showed additional SNR gain of 0.63dB as a result of using OFDM. Therefore, the total gain introduced by using OFDM and LDPC, compared to the current US DOCIS technology, can be 6.13dB. This gain is equivalent to 2.04 bits. Therefore, the capacity of the 200MHz and 238MHz high-split systems will be as follows:

1. 200MHz High-split: 855.6/(200-5) = 4.3877 bps/Hz. Adding 2.04 bits will increase the above spectral efficiency (calculated at QAM256) to: 4.3877*(8+2.04)/8= 5.51bps/Hz (net capacity is 1.073Gbps).
2. 238MHz High-split: 999.5/(238-5) = 4.2897 bps/Hz. Adding 2.04 bits will increase the above spectral efficiency (calculated at QAM256) to:

4.2897*(8+2.04)/8= 5.38 bps/Hz (net capacity is 1.254 Gbps).

Assume that reduction in noise and signal attenuation that results from multiple node splits and segmentations, as well as optimizations of modulations and channel parameters, result in conservative gain estimate of 3dB (equivalent to one additional bit). Therefore, the capacity of the 200MHz and 238MHz high-split systems is increased as follows:

1. 200MHz-High-split: 5.51*(10.04+1)/10.04 = 6.05bps/Hz (net capacity is 1.18 Gbps).
2. 238MHz-High-split: 5.38*(10.04+1)/10.04 = 5.92bps/Hz (net capacity is 1.378 Gbps).

Since the offered channel capacity is well above 1Gbps in both of the above high-split architecture, one may argue that 200MHz high-split (with offered capacity of 1.18Gbps) is enough to offer a service with Tmax= 1Gbps. Although we assumed earlier that MSOs might choose to require 1.5Tmax of channel capacity to offer a Tmax service, one may suggest that a channel capacity of 15% more than the 1Gbps Tmax value is enough. The rationale behind that is that:

1. After so many node splits and segmentations, the number of subscribers per service groups drops exponentially. This reduces the chances that two subscribers will ask for BW at the same time.
2. When the Tmax value is really large (Tmax = 1Gbps), US bursts from subscribers consume very little time and therefore contention drops significantly. In particular, data transmissions from any single subscriber may not take an extended period of time and therefore will not likely affect other customers that are about to transmit their content. It is, therefore, likely that all customers attain the desired Tmax rates for their service.

Some MSOs may choose to be more cautious and decide to use 238MHz high-split option as a target US spectrum. After all, it is expected that either of the high-split options 200MHz (net capacity of 1.18Gbps) or 238MHz (net capacity of 1.254Gbps) will be able to offer a service with Tmax=1Gbps and beyond.

5. SAMPLE MIGRATION PLAN TO REALIZE SYMMETRIC 1GBPS SYSTEM AND BEYOND!

DOCSIS scales very well! It offers just-in-time steps for plant upgrades, where money and resources that are spent will actually be used. This is opposed to investing in FTTH before it is needed; following that path may lead to a large amount of money and resource being spent to offer capacities and capabilities that are not needed yet.

This section proposes smooth migration steps that MSOs might consider taking when upgrading their networks as they move into the future. These steps offer just-in-time investments that are necessary to offer the needed capacity that meets traffic demands. A natural consequence of these gradual steps is that they will likely occur over many years, with the end goal of migrating to a FTTH architecture when it is truly required. By migrating to FTTH at the right time, this approach will avoid upfront investments that will not be actually used until much later. Based on traffic engineering studies, the need for a FTTH architecture appears to be needed only when traffic demands require much more bandwidth than is provided by DOCSIS or DOCSIS variants. This condition appears to be many years down the road, so the economics of the upgrade process to FTTH can probably be deferred until that time.

As explained earlier, there are many steps and options to take in the process of network migration. One proposed sequence of these steps is given below:

1. **Step 0: Use the available spectrum efficiently.** Section 4.1 addressed this topic. For more details, refer to [4] [5] [6].

2. **Step 1: Node segmentations and splits.** This was covered in Section 4.2.

3. **Step 2: Add more BW.** This is divided into two categories:
    a. **CATEGORY 1 of STEP 2: Expand the US spectrum using High split as goal architecture.** This can be done in a single step to save on upgrade costs or via passing through Mid-split to gain more time to avoid the OOB legacy STB signaling problem. This topic was covered in Section 4.3.
    b. **CATEGORY 2 of STEP 2: Enhancements to DOCSIS.**
        i. <u>Higher order modulations</u>. This is viable because less noise and attenuation as a result of multiple noise segmentations / splits as well as other DOCSIS enhancements mentioned below. Section 4.4 covered this topic.
        ii. <u>New FEC (e.g., LDPC)</u>. This provides several dBs of SNR gain over RS. Section 4.6 addressed this topic.
        iii. <u>New PHY (e.g., OFDM)</u>. OFDM is an easy to implement technology that is robust against different types of noise. Section 4.5 covered this subject. For more details, refer to [9] [10]. Note that a new PHY may not be required because the capacity gain may be marginal as shown earlier. However, if MSOs would like to get the most out of the plant and use noise-robust technology, OFDM makes a good choice.
        iv. <u>Backward Compatibility</u>. This is a key item to maintain the increased offered capacity. Example is bonding across new and legacy channels. This was covered in Section 4.7. For more details, refer to [9].
    c. **NOTE:** The above categories of step 2 (items a & b) can be done in any order or even concurrently. This is a key feature to this proposal. In fact, some MSOs may choose not to go beyond category 1 if they think that it provides enough capacity. Others may jump to category 2 as it may line up better with the timing of their plans to expand the US spectrum. Others may go to both options concurrently (or consecutively) with a bold move to get the most capacity out of the plant.

4. **Step 3: FTTH.** way in the future. Natural step after many node segmentations/splits, which will enable MSOs to offer multi-gigabit service in DS and US. This was covered in section 4.8.

6. CONCLUSIONS

The paper studied different options available to the MSOs as they brainstorm to augment their networks for added capacity. The paper proposed using the current spectrum efficiently, performing node segmentations/splits, adding more US spectrum (Mid-Split/High-Split), and adding enhancements to DOCSIS (Higher order modulations/LDPC/OFDMA/Backward Compatibility for added features). A proposed sequence of gradual migration steps was included, which is deemed to carry the MSOs deep into the future with adequate offered capacity according to the provided analysis.

Since the high-split architecture was shown to make the best technical option for US transmissions, a description a high-split prototype system built by ARRIS was included in the paper. The prototype is aimed at studying and analyzing any potential challenges with the high-split proposal, which enables vendors to offer solutions for any problems or issues well before any mass deployment. Initial results for the prototype system were provided. Future papers are planned to share the results as more experiments are done and data is collected.

ACKNOWLEDGEMENTS

The author would like thank the following ARRIS colleagues for their contributions to this paper: Dr. Frank O'Keeffe, Dr. Tom Cloonan, and Mike Emmendorfer. Additionally, the author would like to thank the following ARRIS colleagues for their support in building the High-split prototype: Brent Arnold, Bill Dawson, Ron Miller, Doc Cummings, Bill Ward, Ameen Khalil, David Warner, and others.

REFERENCES

[1] Tom Cloonan, "On the Evolution of the HFC Network and the DOCSIS CMTS: A Roadmap for the 2012-2016 Era," Proceedings, SCTE 2008 Cable Tec-Expo (June, 2008).

[2] Alan Breznick, "Introduction: The Broadband Outlook", Light Reading Conference on Cable Next-Gen Broadband Strategies 2012 (March, 2012).

[3] Mike Emmendorfer, et. al., "Next Generation - Cable Access Network (NG-CAN): Examination of the Business Drivers and Network Approaches to Enable a Multi-Gigabit Downstream and Gigabit Upstream DOCSIS Service over Coaxial Networks", SCTE Canadian Summit, (March, 2012).

[4] Ayham Al-Banna, "DOCSIS3.0® Performance in the Presence of US HFC Noise", International Technical Seminar, SCTE-South America, (March, 2012).

[5] Tom Cloonan, et. al., "Novel CMTS-based Bandwidth Management Schemes Employing Congestion and Capacity Measurements with Throughput-Maximizing Adjustments for DOCSIS 2.0 Operation", SCTE Conference on Emerging Technologies, (January, 2005).

[6] Ayham Al-Banna, et. al., "DOCSIS® 3.0 Upstream Channel Bonding: Performance Analysis in the Presence of HFC Noise", SCTE-ET NCTA Conference, (April, 2009).

[7] Dirk Jaeger and Christoph Schaaf, "DVB-C2 High Performance Data Transmission on Cable – Technology, Implementation, Networks", (2010).

[8] Ayham Al-Banna and Tom Cloonan, "Performance Analysis of Multi-Carrier Systems when Applied to HFC Networks", SCTE-ET NCTA Conference, (April, 2009).

[9] Ayham Al-Banna, "WiMAX Links and OFDM Overlay for HFC Networks: Mobility and Higher US Capacity", 2010 Spring Technical Forum, NCTA-SCTE, (May, 2010).

[10] Ayham Al-Banna, “Multiple US PHY Technologies: Which Way to Take in Future HFC Networks?”, ANGA Cable Conference, (May, 2011).

[11] Robret Gallager, “Low Density Parity-Check Codes”, MIT Press, Cambridge, MA, (1963).

[12] CableLabs – “Data Over Cable Service Interface Specifications DOCSIS 3.0: Physical Layer Specification”, (October, 2010)

[13] Telecommunication Standardization Sector of ITU, “J.83: Series J: Transmission of television, Sound, programme and other multimedia Signals – Digital transmission of television Signals”, (April, 1997)

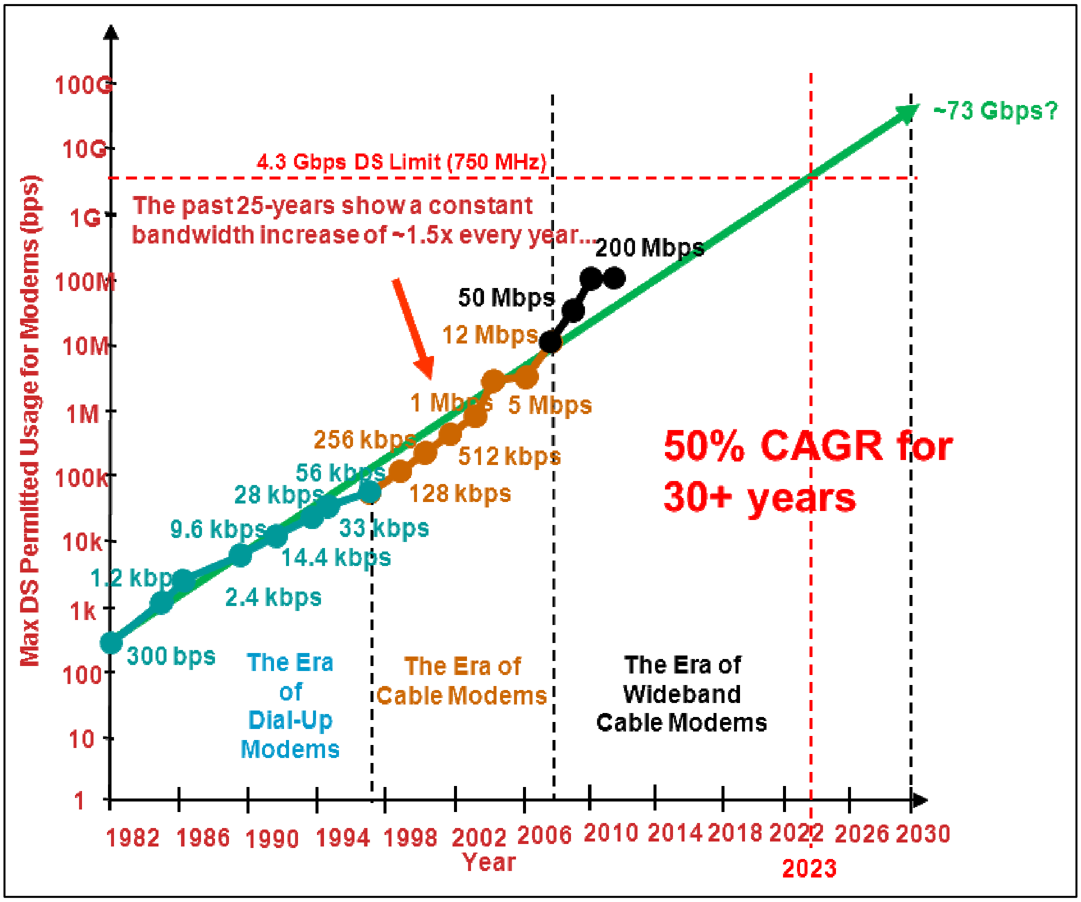


Fig. 1. The Nielson Curve for traffic growth over cable networks (Max. DS Usage/subscriber)

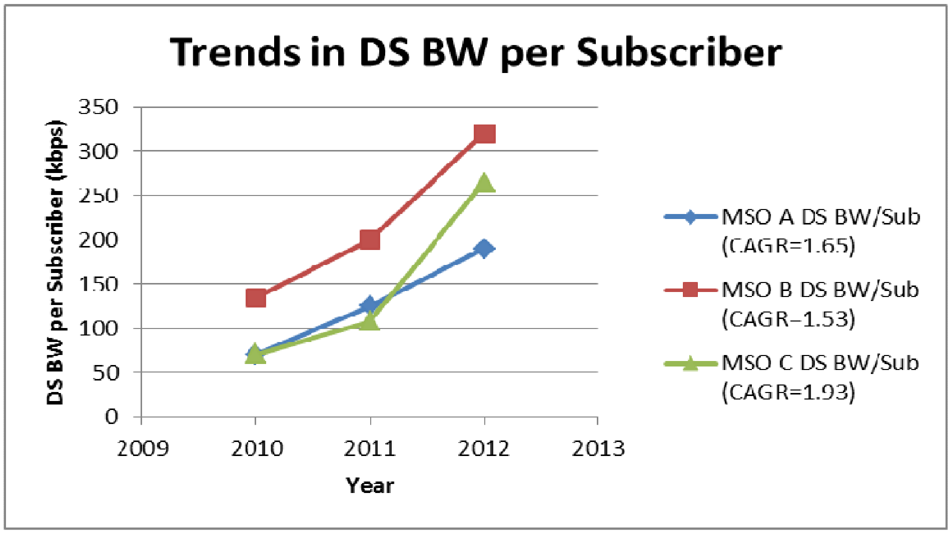


Fig. 2. CAGR of average DS BW per subscriber for three different MSOs over the past two years (>50% DS CAGR on average)

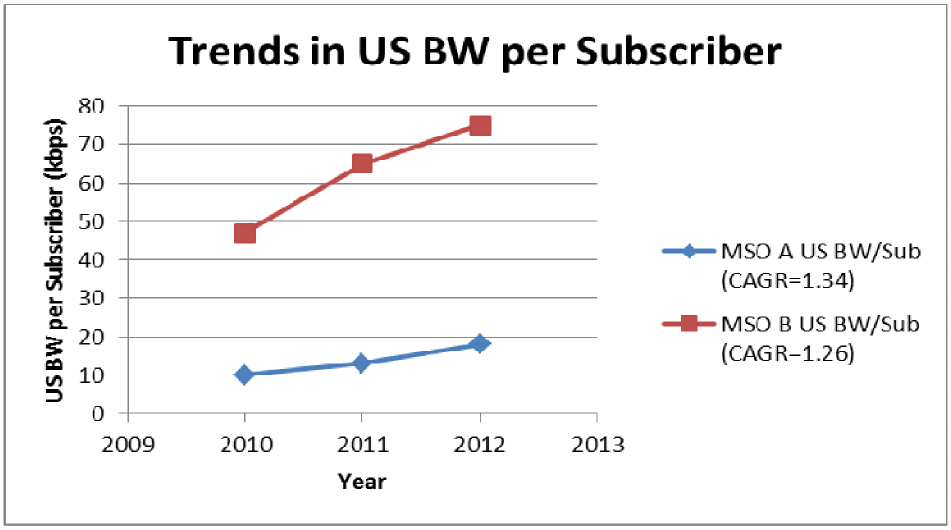


Fig. 3. CAGR of average US BW per subscriber for two different MSOs over the past two years (~30% US CAGR on average)

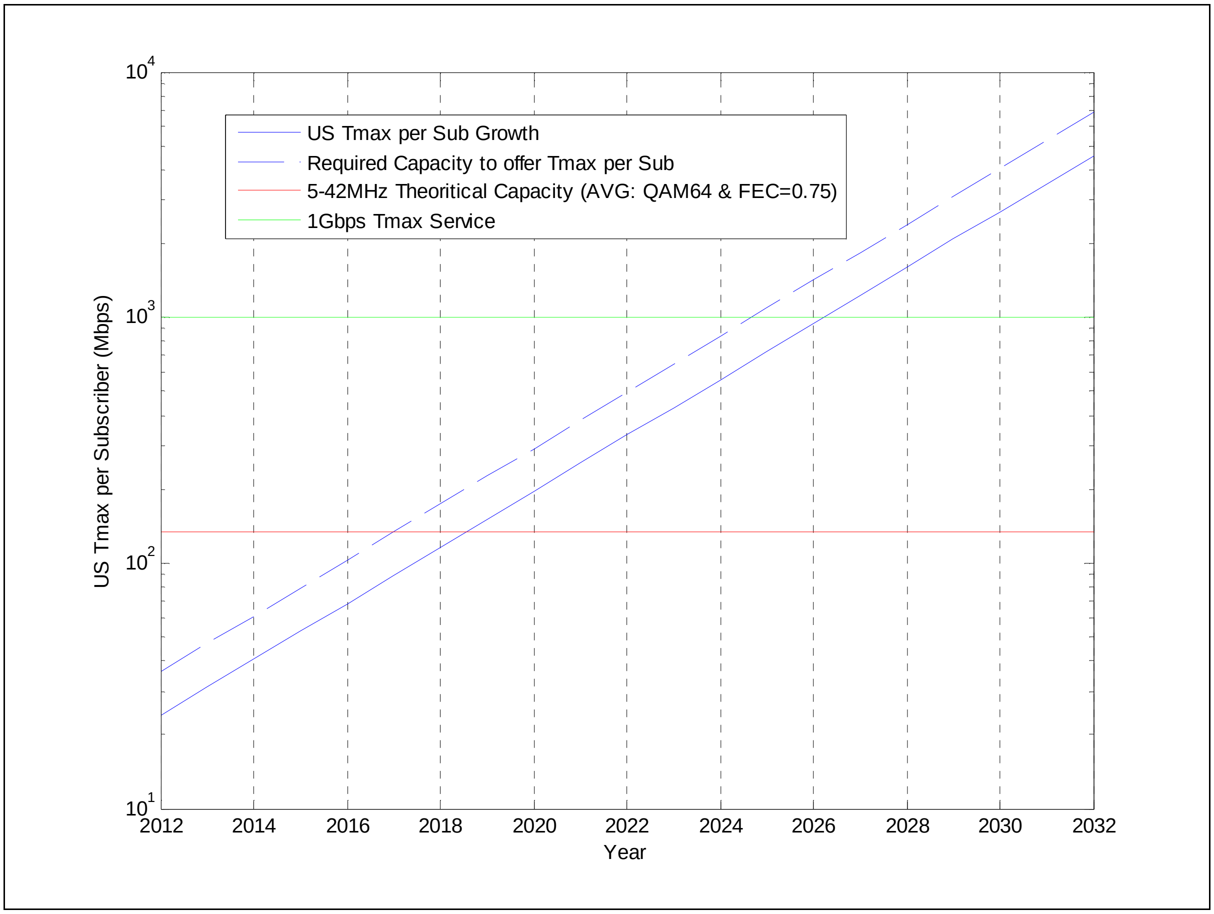


Fig. 4. US Tmax per subscriber growth over the next two decades (assuming CAGR = 30% & starting Tmax = 24Mbps per Subscriber in 2012)

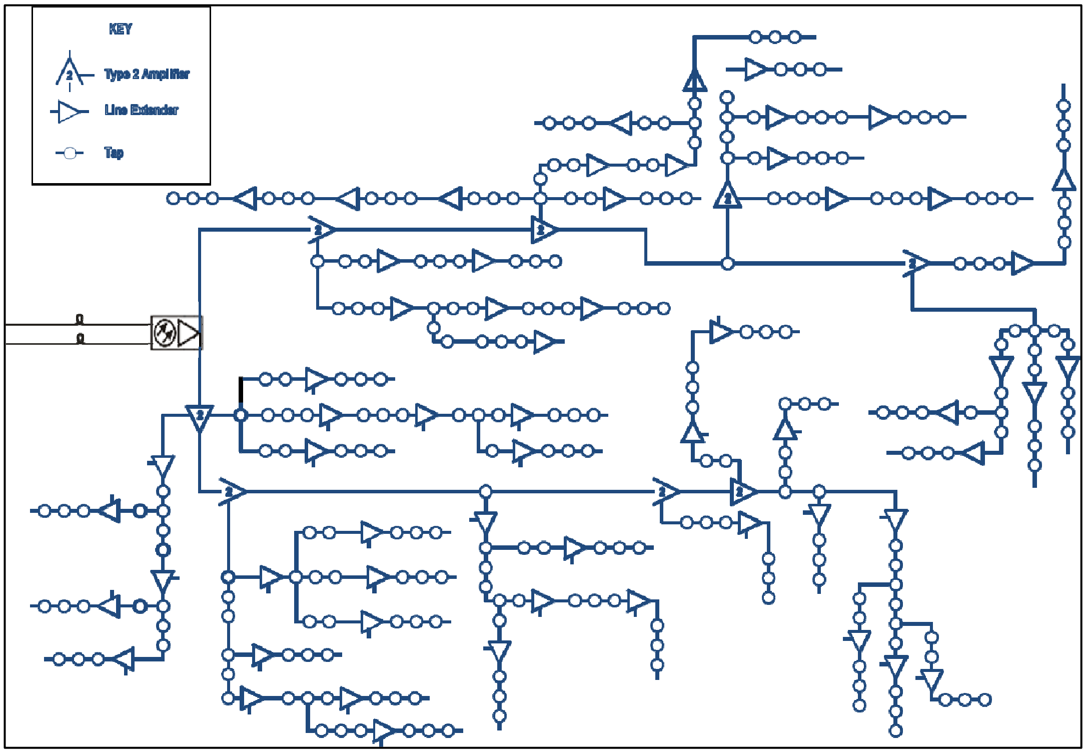


Fig. 5. Example of N+5 Network topology.

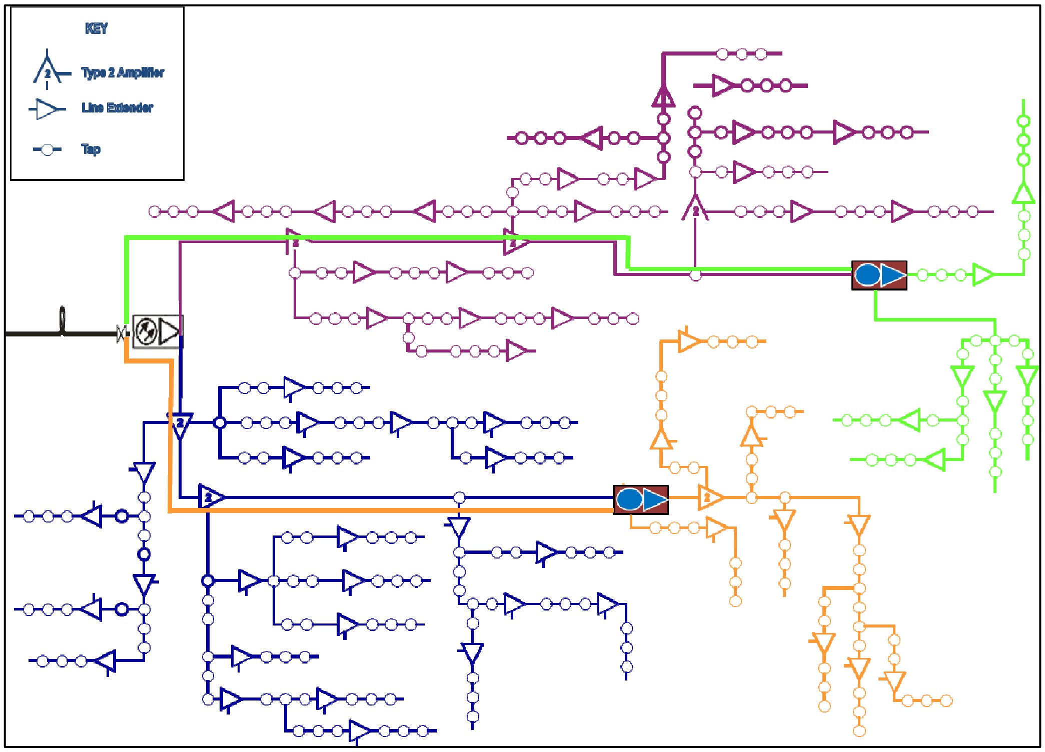
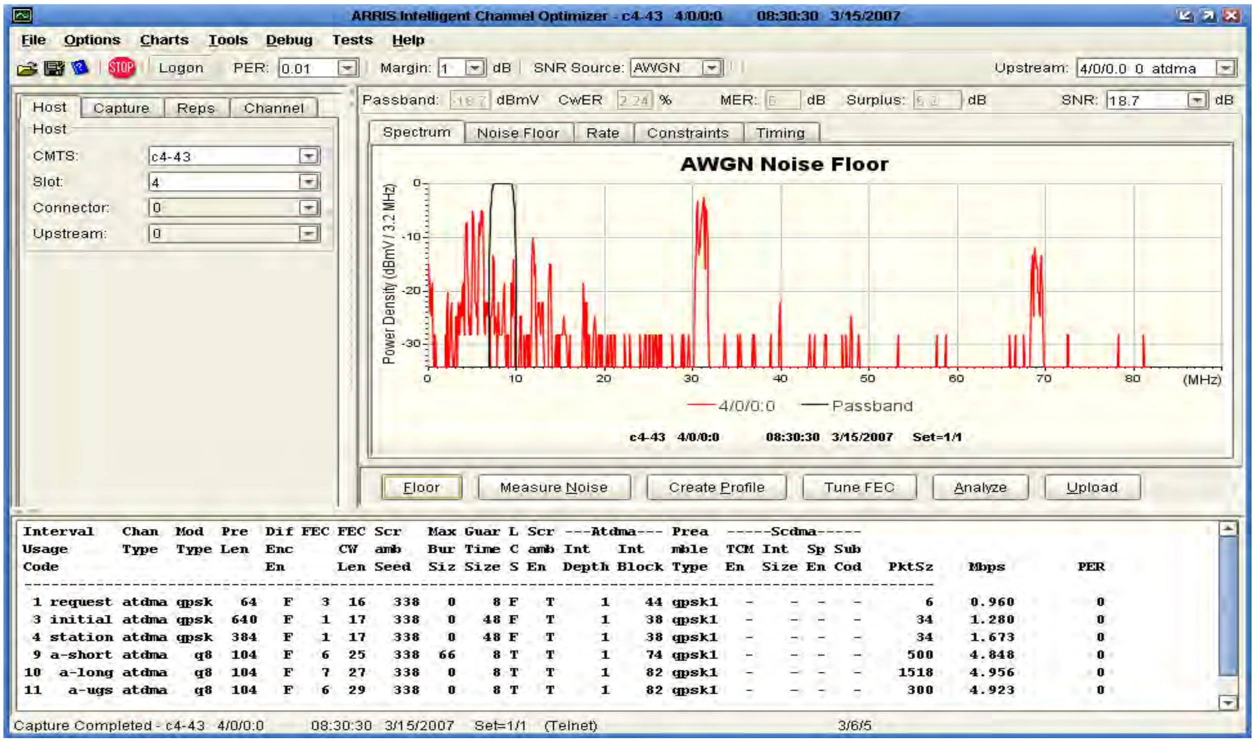


Fig. 6. Segmenting and Splitting the FN in the network example shown in Fig. 5.



ARRIS Intelligent Channel Optimizer - c4-43 4/0/0:0 08:30:30 3/15/2007

PER: 0.01 Margin: 1 dB SNR Source: AWGN Upstream: 4/0/0.0 0 atdma

Passband: -18.7 dBmV CwER 2.24 % MER: 6 dB Surplus: 6.2 dB SNR: 18.7 dB

Host

CMTS: c4-43

Slot: 4

Connector: 0

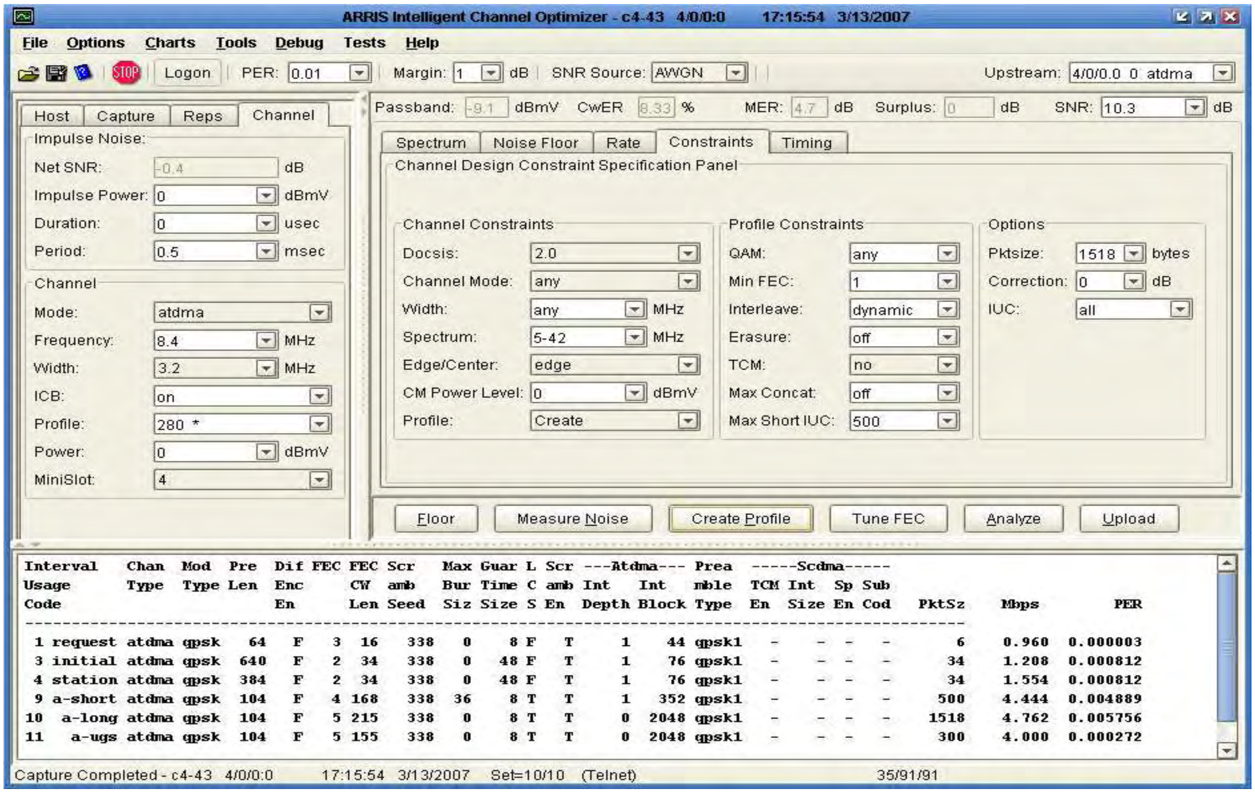
Upstream: 0

AWGN Noise Floor
Power Density (dBmV / 3.2 MHz)
0
-10
-20
-30
0
10
20
30
40
50
60
70
80
(MHz)
4/0/0:0
Passband
c4-43 4/0/0:0 08:30:30 3/15/2007 Set=1/1

```
Interval    Chan  Mod  Pre  Dif FEC FEC Scr   Max Guar L Scr ---Atdma--- Prea  -----Scdma-----
Usage       Type  Type Len  Enc     CW  amb   Bur Time C amb Int   Int   mble  TCM Int  Sp Sub
Code                        En      Len Seed  Siz Size S En  Depth Block Type  En  Size En Cod   PktSz   Mbps       PER
---------------------------------------------------------------------------------------------------
 1 request atdma qpsk   64   F   3  16   338   0    8 F   T     1    44 qpsk1  -    -  -   -        6   0.960         0
 3 initial atdma qpsk  640   F   1  17   338   0   48 F   T     1    38 qpsk1  -    -  -   -       34   1.280         0
 4 station atdma qpsk  384   F   1  17   338   0   48 F   T     1    38 qpsk1  -    -  -   -       34   1.673         0
 9 a-short atdma   q8  104   F   6  25   338  66    8 T   T     1    74 qpsk1  -    -  -   -      500   4.848         0
10  a-long atdma   q8  104   F   7  27   338   0    8 T   T     1    82 qpsk1  -    -  -   -     1518   4.956         0
11   a-ugs atdma   q8  104   F   6  29   338   0    8 T   T     1    82 qpsk1  -    -  -   -      300   4.923         0
```

Capture Completed - c4-43 4/0/0:0 08:30:30 3/15/2007 Set=1/1 (Telnet) 3/6/5

Fig. 7. Automation of optimizing the upstream modulation profile and channel parameters (choosing the best parameters for the noise that exists on the plant)



ARRIS Intelligent Channel Optimizer - c4-43 4/0/0:0 17:15:54 3/13/2007

PER: 0.01 Margin: 1 dB SNR Source: AWGN Upstream: 4/0/0.0 0 atdma

Passband: -9.1 dBmV CwER 8.33 % MER: 4.7 dB Surplus: 0 dB SNR: 10.3 dB

Impulse Noise:

Net SNR: -0.4 dB

Impulse Power: 0 dBmV

Duration: 0 usec

Period: 0.5 msec

Channel

Mode: atdma

Frequency: 8.4 MHz

Width: 3.2 MHz

ICB: on

Profile: 280 *

Power: 0 dBmV

MiniSlot: 4

Channel Design Constraint Specification Panel

Channel Constraints

Docsis: 2.0

Channel Mode: any

Width: any MHz

Spectrum: 5-42 MHz

Edge/Center: edge

CM Power Level: 0 dBmV

Profile: Create

Profile Constraints

QAM: any

Min FEC: 1

Interleave: dynamic

Erasure: off

TCM: no

Max Concat: off

Max Short IUC: 500

Options

Pktsize: 1518 bytes

Correction: 0 dB

IUC: all

```
Interval    Chan  Mod  Pre  Dif FEC FEC Scr   Max Guar L Scr ---Atdma--- Prea  -----Scdma-----
Usage       Type  Type Len  Enc     CW  amb   Bur Time C amb Int   Int   mble  TCM Int  Sp Sub
Code                        En      Len Seed  Siz Size S En  Depth Block Type  En  Size En Cod   PktSz   Mbps      PER
---------------------------------------------------------------------------------------------------
 1 request atdma qpsk   64   F   3  16   338   0    8 F   T     1    44 qpsk1  -    -  -   -        6   0.960  0.000003
 3 initial atdma qpsk  640   F   2  34   338   0   48 F   T     1    76 qpsk1  -    -  -   -       34   1.208  0.000812
 4 station atdma qpsk  384   F   2  34   338   0   48 F   T     1    76 qpsk1  -    -  -   -       34   1.554  0.000812
 9 a-short atdma qpsk  104   F   4 168   338  36    8 T   T     1   352 qpsk1  -    -  -   -      500   4.444  0.004889
10  a-long atdma qpsk  104   F   5 215   338   0    8 T   T     0  2048 qpsk1  -    -  -   -     1518   4.762  0.005756
11   a-ugs atdma qpsk  104   F   5 155   338   0    8 T   T     0  2048 qpsk1  -    -  -   -      300   4.000  0.000272
```

Capture Completed - c4-43 4/0/0:0 17:15:54 3/13/2007 Set=10/10 (Telnet) 35/91/91

Fig. 8. Automation of optimizing the upstream modulation profile and channel parameters (specifying constraints)

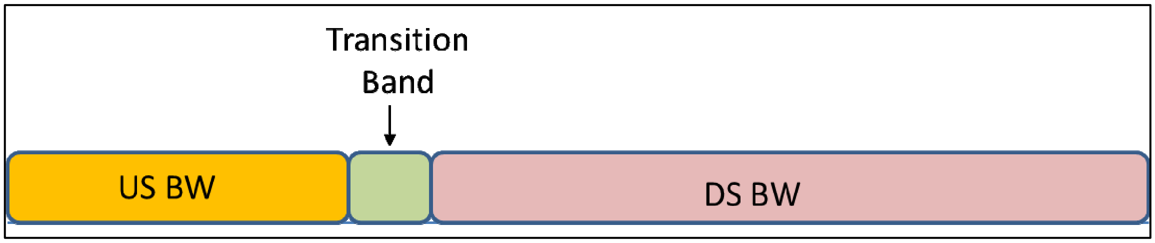


Fig. 9. Mid-Split and High-Split options for US spectrum usage

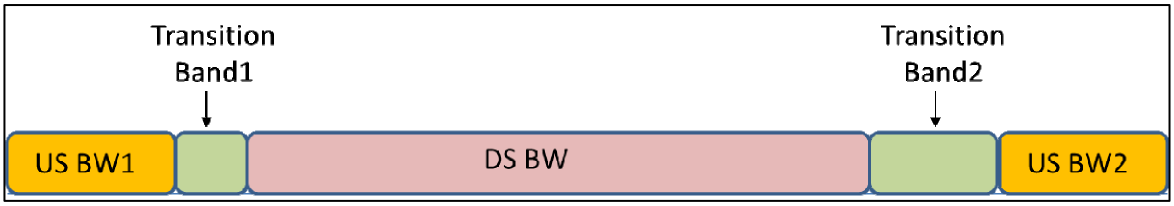
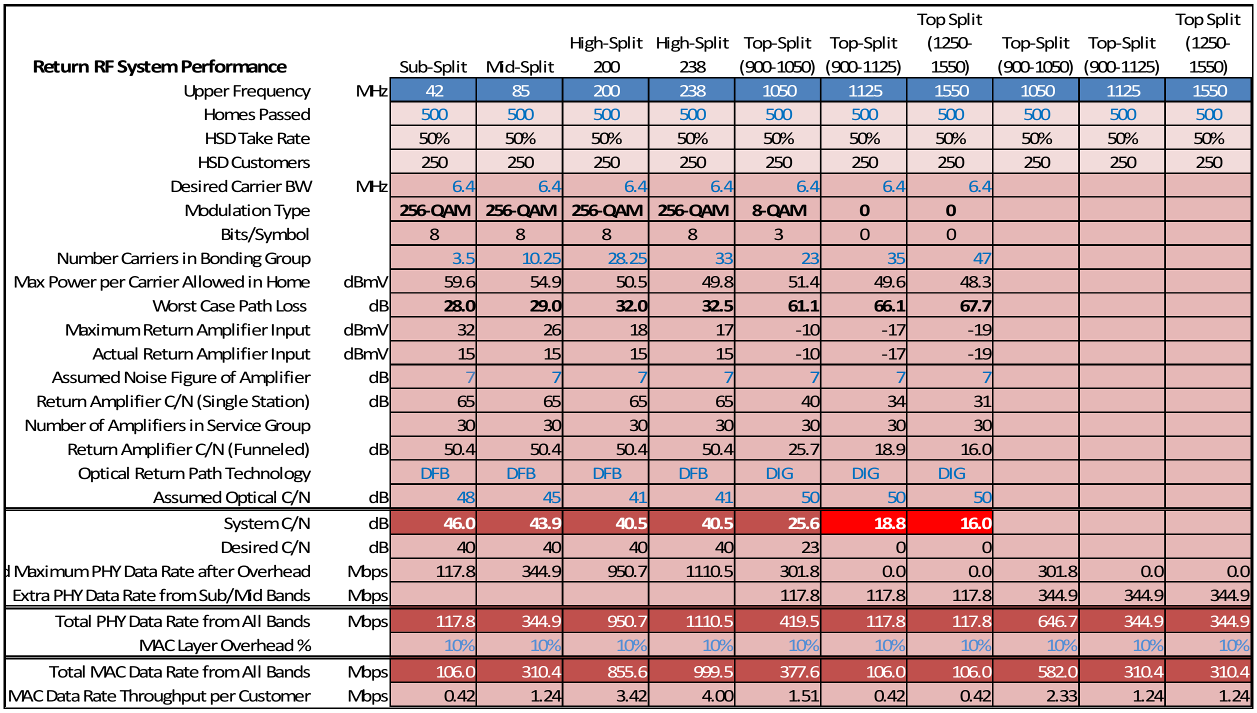


Fig. 10. Top-Split option for US spectrum usage



								Top Split			Top Split
				High-Split	High-Split	Top-Split	Top-Split	(1250-	Top-Split	Top-Split	(1250-
Return RF System Performance		Sub-Split	Mid-Split	200	238	(900-1050)	(900-1125)	1550)	(900-1050)	(900-1125)	1550)
Upper Frequency	MHz	42	85	200	238	1050	1125	1550	1050	1125	1550
Homes Passed		500	500	500	500	500	500	500	500	500	500
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	250	250	250	250	250	250	250
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4			
Modulation Type		**256-QAM**	**256-QAM**	**256-QAM**	**256-QAM**	**8-QAM**	**0**	**0**			
Bits/Symbol		8	8	8	8	3	0	0			
Number Carriers in Bonding Group		3.5	10.25	28.25	33	23	35	47			
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	50.5	49.8	51.4	49.6	48.3			
Worst Case Path Loss	dB	**28.0**	**29.0**	**32.0**	**32.5**	**61.1**	**66.1**	**67.7**			
Maximum Return Amplifier Input	dBmV	32	26	18	17	-10	-17	-19			
Actual Return Amplifier Input	dBmV	15	15	15	15	-10	-17	-19			
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7			
Return Amplifier C/N (Single Station)	dB	65	65	65	65	40	34	31			
Number of Amplifiers in Service Group		30	30	30	30	30	30	30			
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	50.4	25.7	18.9	16.0			
Optical Return Path Technology		DFB	DFB	DFB	DFB	DIG	DIG	DIG			
Assumed Optical C/N	dB	48	45	41	41	50	50	50			
System C/N	dB	**46.0**	**43.9**	**40.5**	**40.5**	**25.6**	**18.8**	**16.0**			
Desired C/N	dB	40	40	40	40	23	0	0			
d Maximum PHY Data Rate after Overhead	Mbps	117.8	344.9	950.7	1110.5	301.8	0.0	0.0	301.8	0.0	0.0
Extra PHY Data Rate from Sub/Mid Bands	Mbps					117.8	117.8	117.8	344.9	344.9	344.9
Total PHY Data Rate from All Bands	Mbps	117.8	344.9	950.7	1110.5	419.5	117.8	117.8	646.7	344.9	344.9
MAC Layer Overhead %		10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Total MAC Data Rate from All Bands	Mbps	106.0	310.4	855.6	999.5	377.6	106.0	106.0	582.0	310.4	310.4
MAC Data Rate Throughput per Customer	Mbps	0.42	1.24	3.42	4.00	1.51	0.42	0.42	2.33	1.24	1.24

Fig. 11. Analysis of different split options for the US spectrum in DOCSIS networks [3]

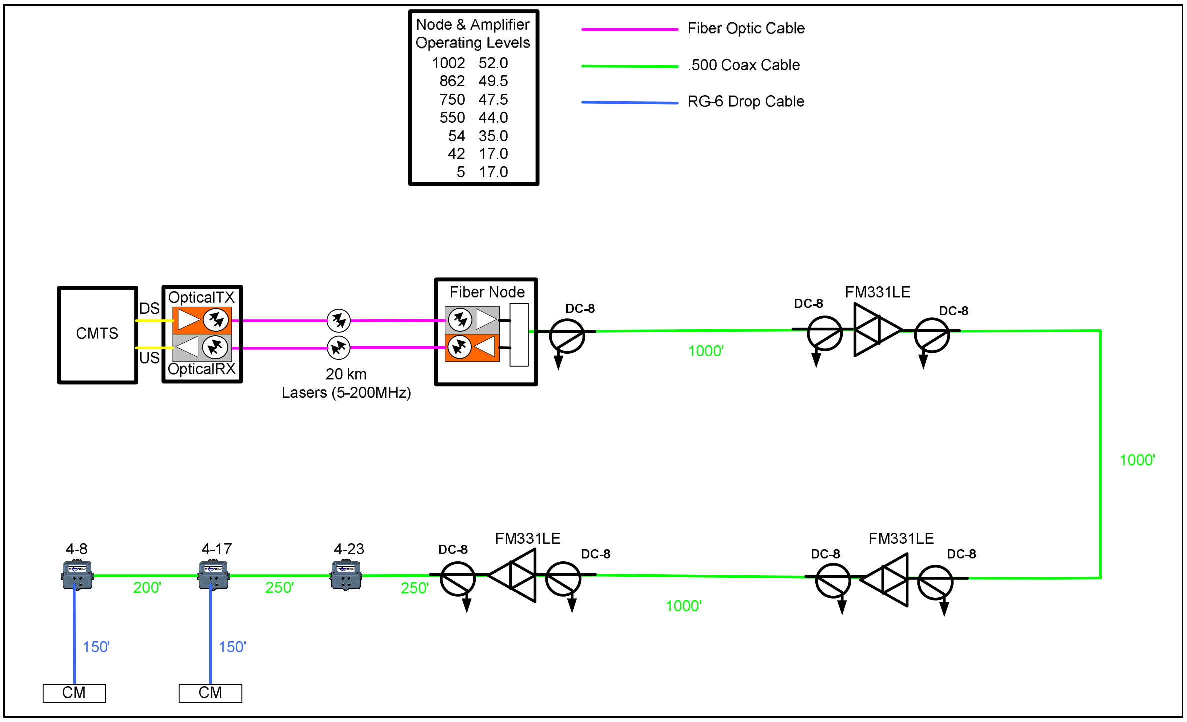


Fig. 12. Example setup for Real-world N+3 network architecture

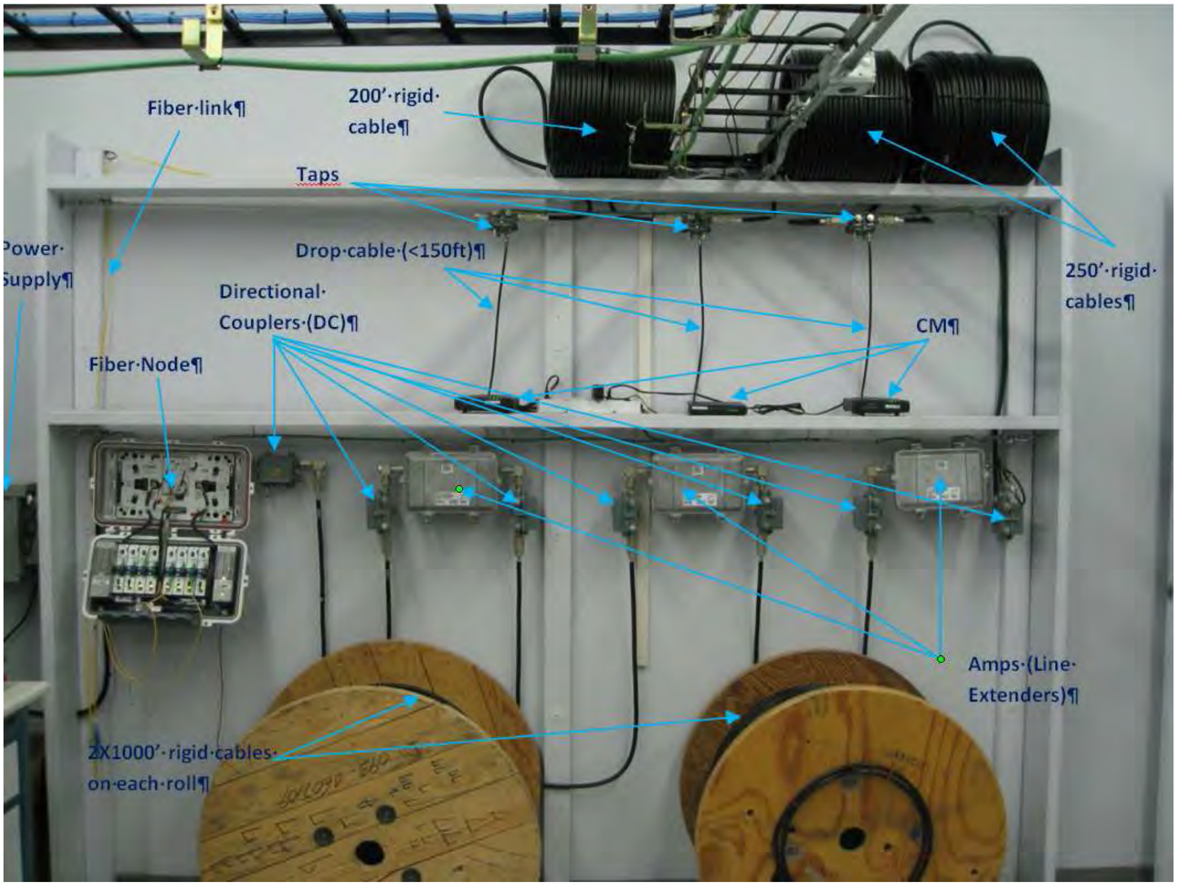


Fig. 13. ARRIS Implementation of high-split prototype architecture network to mimic the setup in Fig. 12 (laser Tx/Rx in the headend is not shown in the figure).

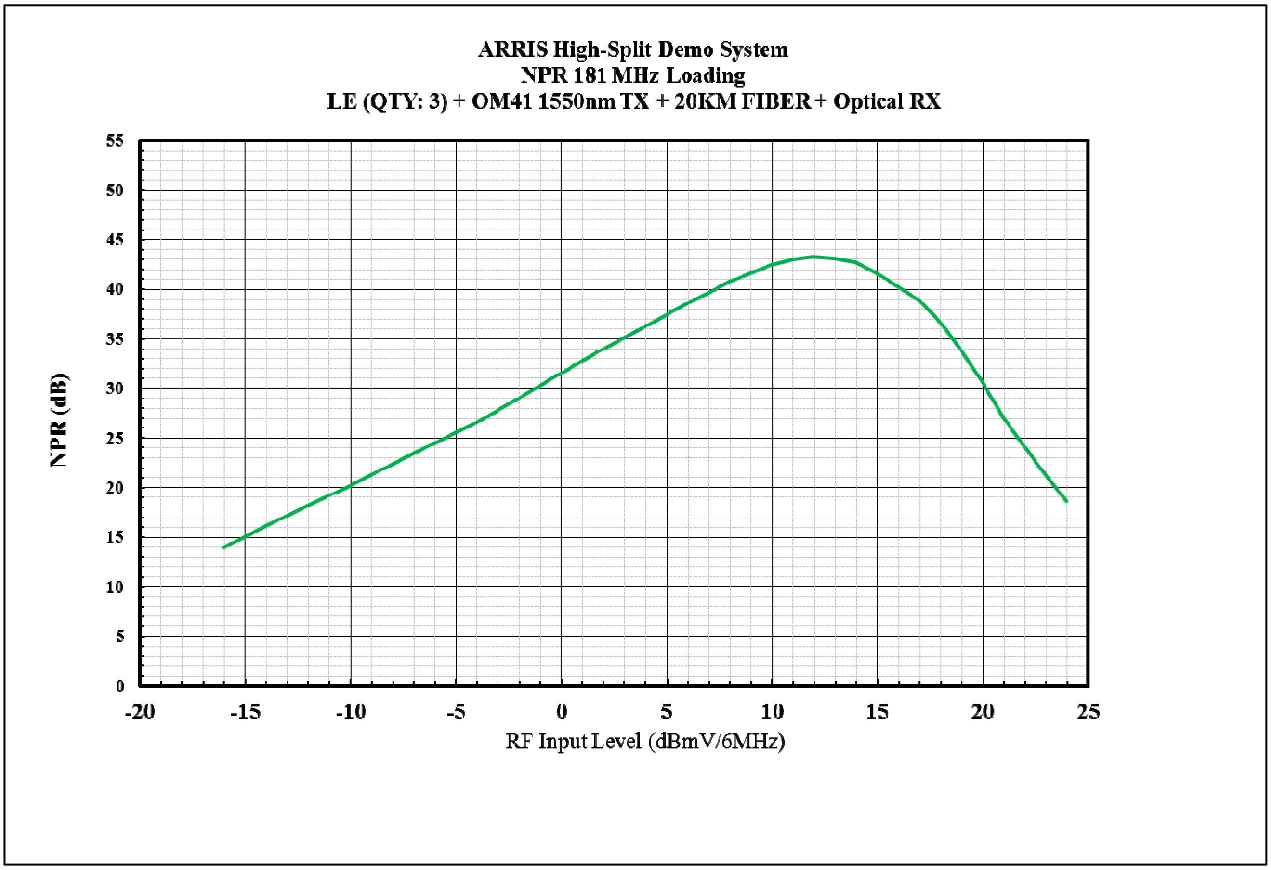
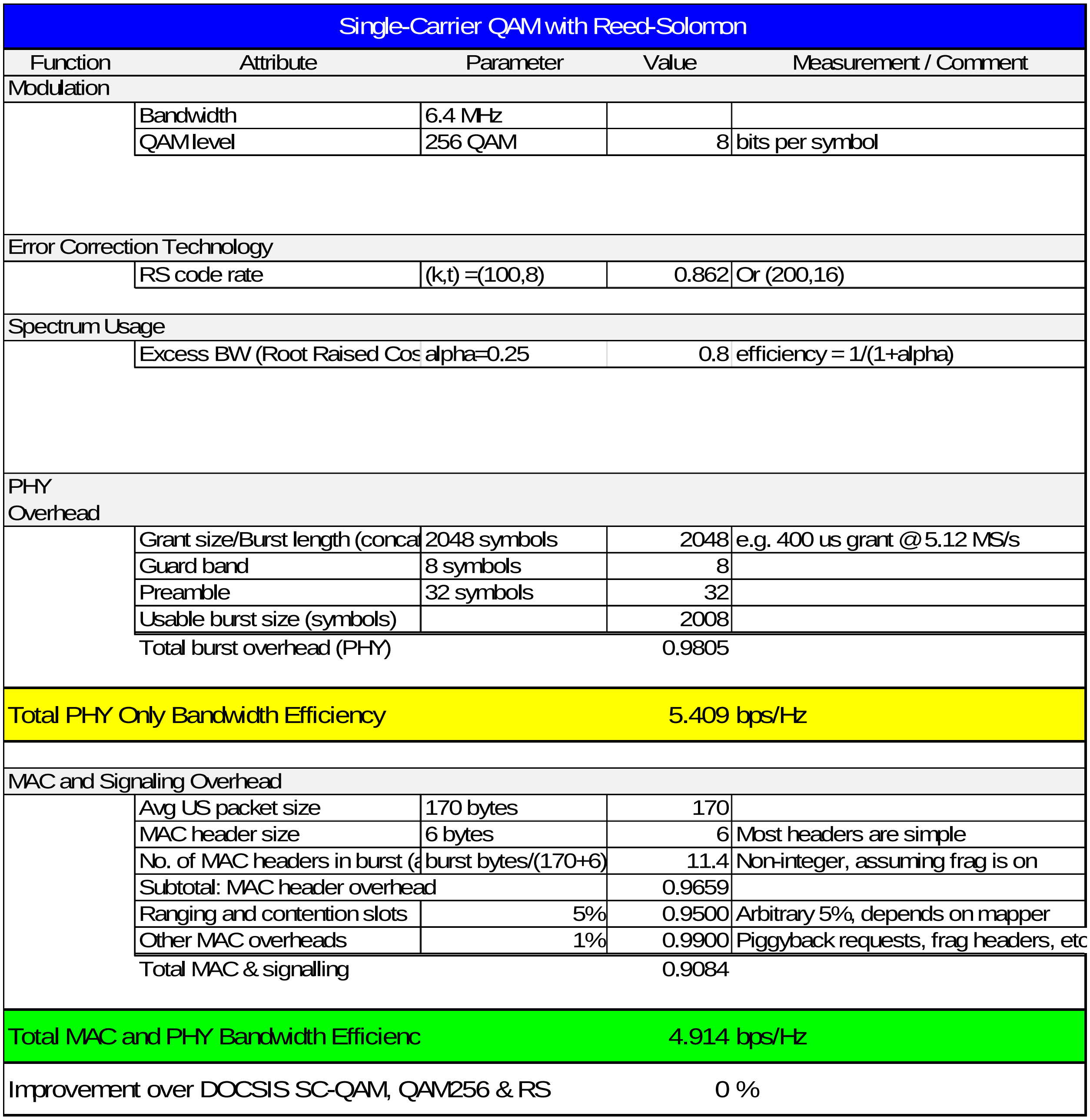


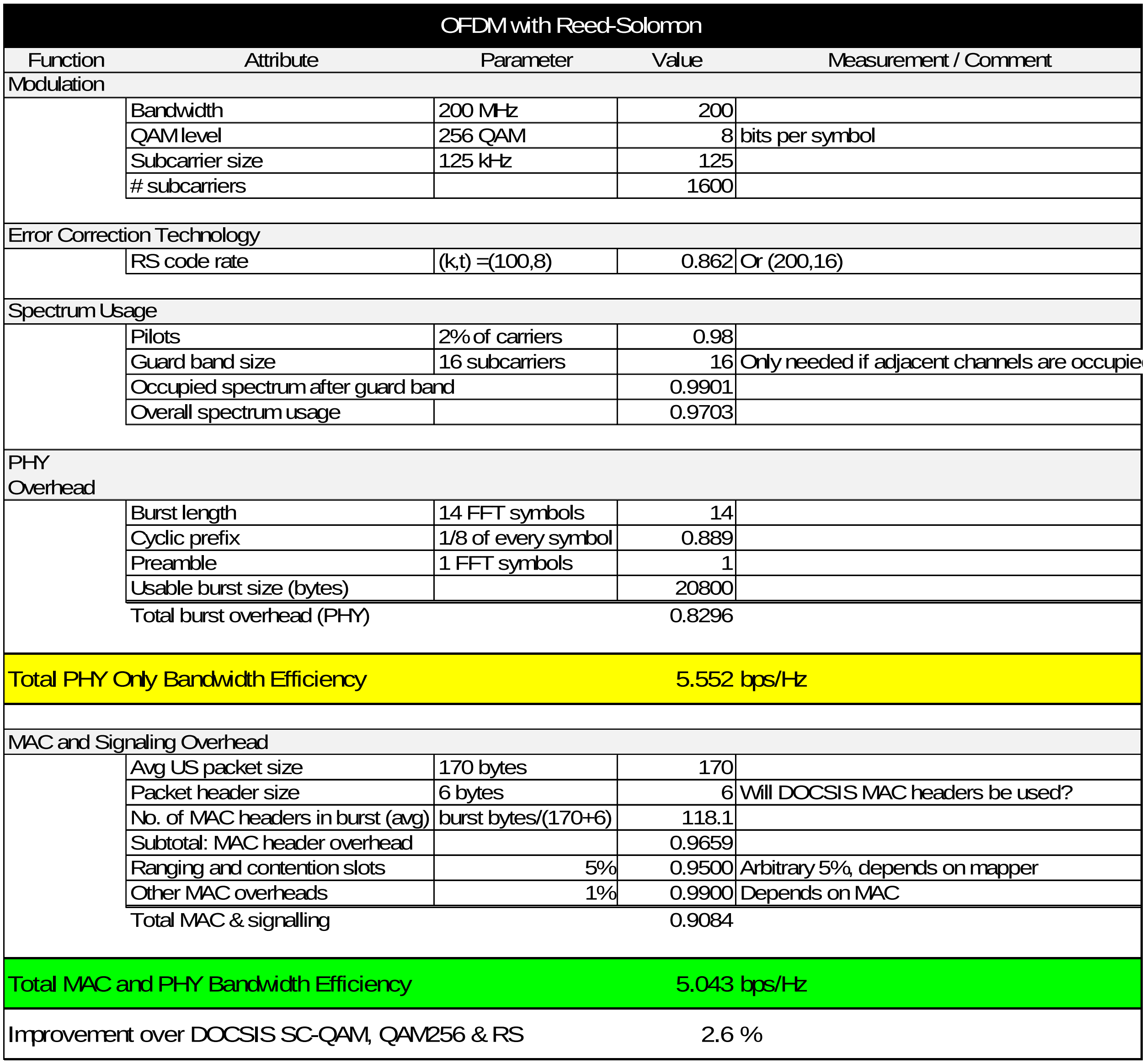
Fig. 14. An initial NPR curve for the plant setup shown in Fig. 13.



Single-Carrier QAM with Reed-Solomon

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	6.4 MHz		
	QAM level	256 QAM	8	bits per symbol
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Excess BW (Root Raised Cos	alpha=0.25	0.8	efficiency = 1/(1+alpha)
PHY Overhead				
	Grant size/Burst length (concat	2048 symbols	2048	e.g. 400 us grant @ 5.12 MS/s
	Guard band	8 symbols	8	
	Preamble	32 symbols	32	
	Usable burst size (symbols)		2008	
	Total burst overhead (PHY)		0.9805	
Total PHY Only Bandwidth Efficiency			5.409	bps/Hz
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	MAC header size	6 bytes	6	Most headers are simple
	No. of MAC headers in burst (a	burst bytes/(170+6)	11.4	Non-integer, assuming frag is on
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Piggyback requests, frag headers, etc
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficienc			4.914	bps/Hz
Improvement over DOCSIS SC-QAM, QAM256 & RS			0	%

Fig. 15. Capacity analysis for Single carrier DOCSIS signal



OFDM with Reed-Solomon				
Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	200 MHz	200	
	QAM level	256 QAM	8	bits per symbol
	Subcarrier size	125 kHz	125	
	# subcarriers		1600	
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Pilots	2% of carriers	0.98	
	Guard band size	16 subcarriers	16	Only needed if adjacent channels are occupie
	Occupied spectrum after guard band		0.9901	
	Overall spectrum usage		0.9703	
PHY Overhead				
	Burst length	14 FFT symbols	14	
	Cyclic prefix	1/8 of every symbol	0.889	
	Preamble	1 FFT symbols	1	
	Usable burst size (bytes)		20800	
	Total burst overhead (PHY)		0.8296	
Total PHY Only Bandwidth Efficiency			5.552 bps/Hz	
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	Packet header size	6 bytes	6	Will DOCSIS MAC headers be used?
	No. of MAC headers in burst (avg)	burst bytes/(170+6)	118.1	
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Depends on MAC
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			5.043 bps/Hz	
Improvement over DOCSIS SC-QAM, QAM256 & RS			2.6 %	

Fig. 16. Capacity analysis for OFDM signals when used for DOCSIS US transmissions

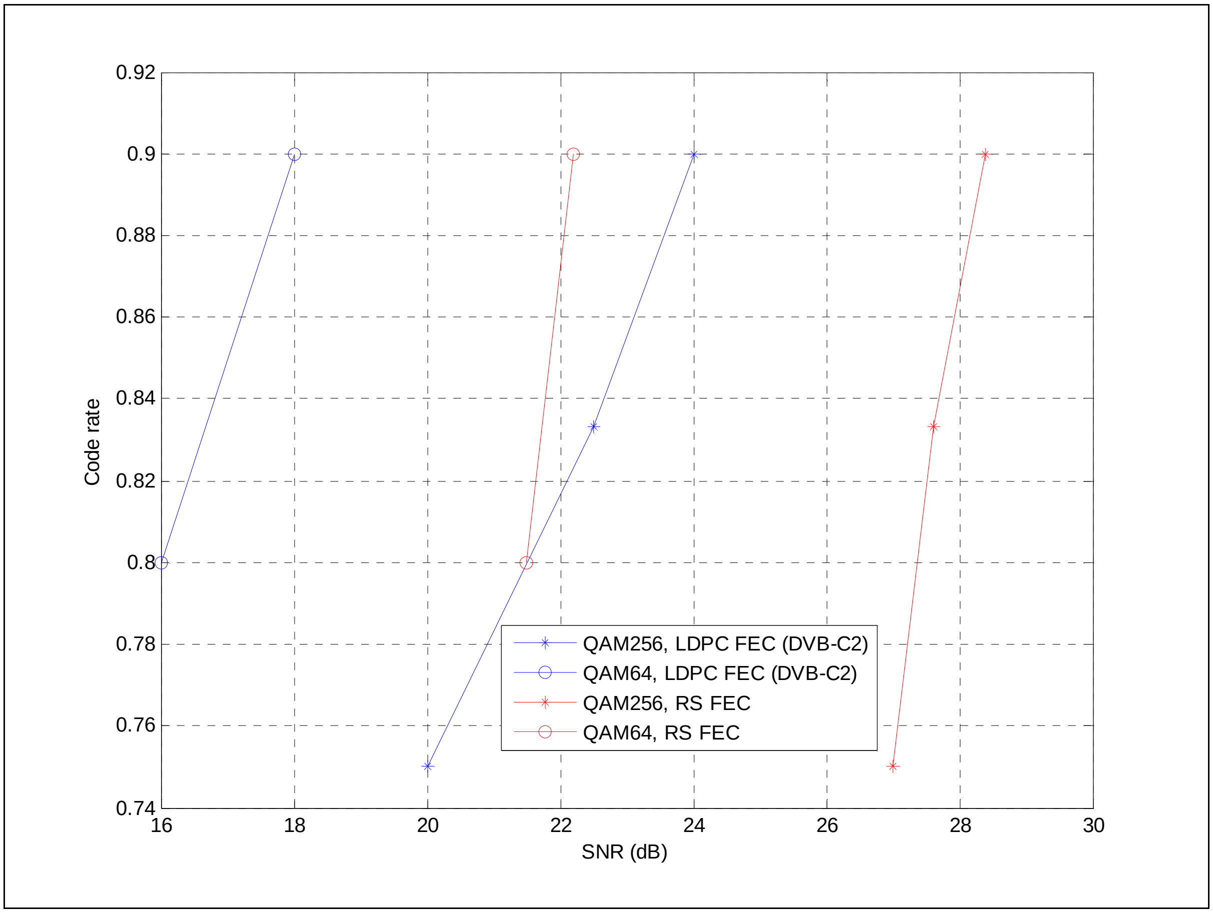
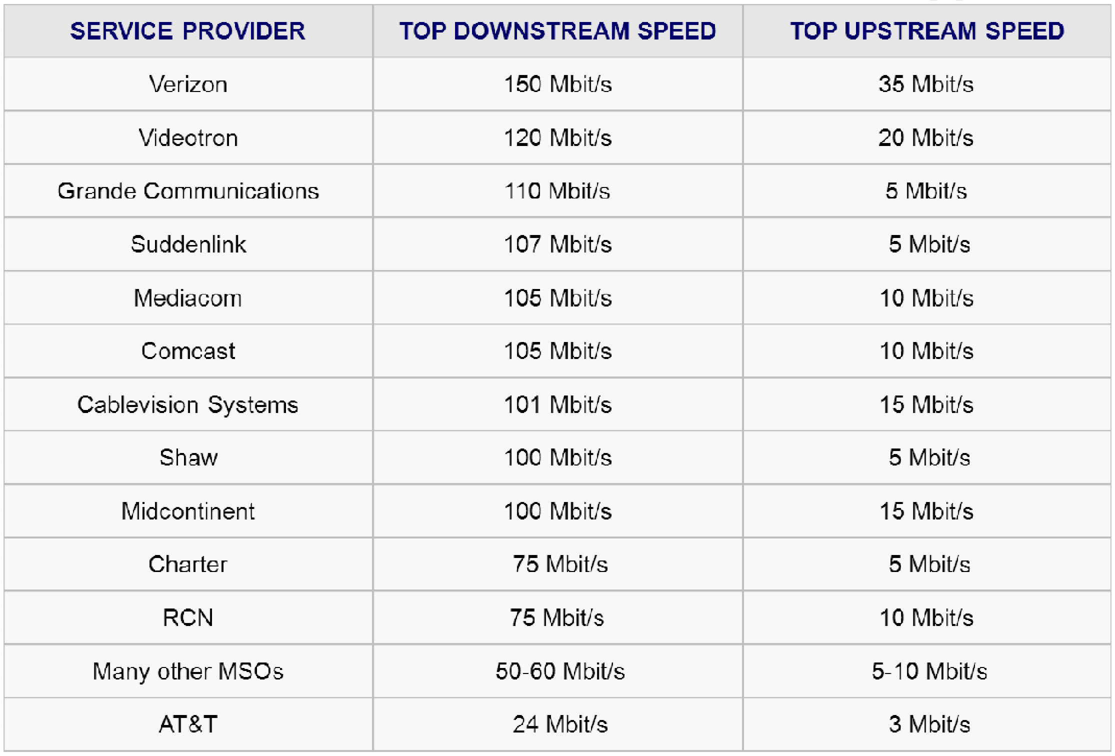


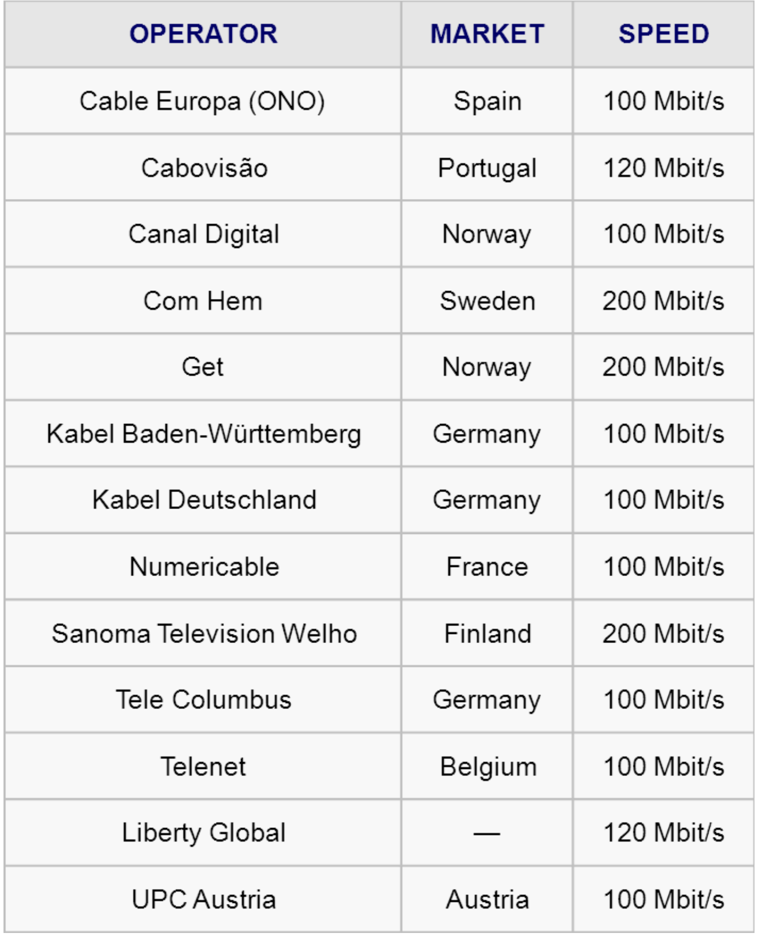
Fig. 17. Comparison between RS FEC (computer simulations) and LDPC FEC (from DVB-C2)

Table 1. Offered DS and US Tmax values in North America [2]



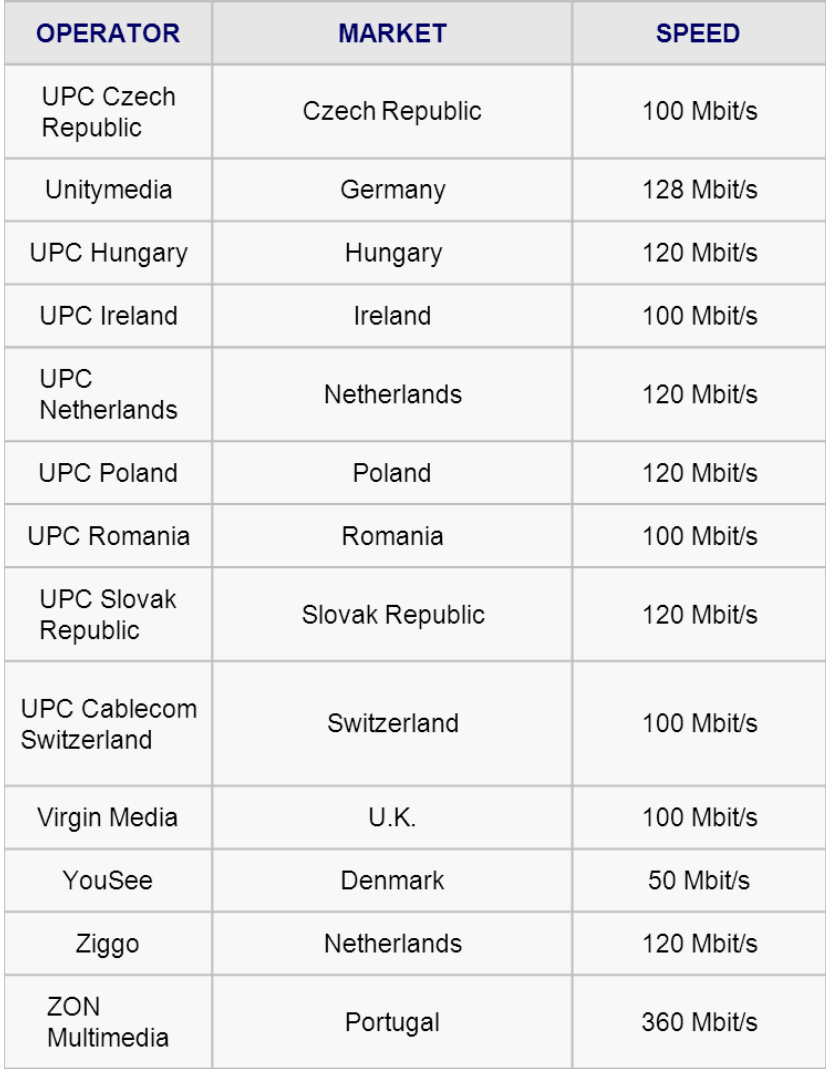
SERVICE PROVIDER	TOP DOWNSTREAM SPEED	TOP UPSTREAM SPEED
Verizon	150 Mbit/s	35 Mbit/s
Videotron	120 Mbit/s	20 Mbit/s
Grande Communications	110 Mbit/s	5 Mbit/s
Suddenlink	107 Mbit/s	5 Mbit/s
Mediacom	105 Mbit/s	10 Mbit/s
Comcast	105 Mbit/s	10 Mbit/s
Cablevision Systems	101 Mbit/s	15 Mbit/s
Shaw	100 Mbit/s	5 Mbit/s
Midcontinent	100 Mbit/s	15 Mbit/s
Charter	75 Mbit/s	5 Mbit/s
RCN	75 Mbit/s	10 Mbit/s
Many other MSOs	50-60 Mbit/s	5-10 Mbit/s
AT&T	24 Mbit/s	3 Mbit/s

Table 2. Offered DS Tmax values in Europe [2]



OPERATOR	MARKET	SPEED
Cable Europa (ONO)	Spain	100 Mbit/s
Cabovisão	Portugal	120 Mbit/s
Canal Digital	Norway	100 Mbit/s
Com Hem	Sweden	200 Mbit/s
Get	Norway	200 Mbit/s
Kabel Baden-Württemberg	Germany	100 Mbit/s
Kabel Deutschland	Germany	100 Mbit/s
Numericable	France	100 Mbit/s
Sanoma Television Welho	Finland	200 Mbit/s
Tele Columbus	Germany	100 Mbit/s
Telenet	Belgium	100 Mbit/s
Liberty Global	—	120 Mbit/s
UPC Austria	Austria	100 Mbit/s

Table 3. Offered DS Tmax values in Europe (Continued) [2]



OPERATOR	MARKET	SPEED
UPC Czech Republic	Czech Republic	100 Mbit/s
Unitymedia	Germany	128 Mbit/s
UPC Hungary	Hungary	120 Mbit/s
UPC Ireland	Ireland	100 Mbit/s
UPC Netherlands	Netherlands	120 Mbit/s
UPC Poland	Poland	120 Mbit/s
UPC Romania	Romania	100 Mbit/s
UPC Slovak Republic	Slovak Republic	120 Mbit/s
UPC Cablecom Switzerland	Switzerland	100 Mbit/s
Virgin Media	U.K.	100 Mbit/s
YouSee	Denmark	50 Mbit/s
Ziggo	Netherlands	120 Mbit/s
ZON Multimedia	Portugal	360 Mbit/s

Table 4. Assumptions about spectrum usage used in analyzing the capacity of 5-42MHz spectrum in [3]

Bandwidth	Description
37	Sup-split Upstream spectrum (5-42MHz)
-2	Assumed 2MHz as roll off (40-42MHz) being unusable
-5	Assumed that the noisy spectrum (5-MHz) to be unusable
-2	Legacy STBs
-2	Legacy Status Monitoring
-3.2	3.2MHz channel for legacy QAM16 DOCSIS
22.8	Possible spectrum for DOCSIS3.0 US channel bonding
22.4	Assumed value for capacity analysis

Table 5. Typical Fiber node assumptions used to compare different split options [3]

Item	Value	Unit
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	1000	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

Table 6. Express/distribution segments assumptions used to compare different split options [3]

"Express" (untapped) Segment Characterization	Unit	Sub-Split	Mid-Split	High-Split 200	High-Split 238	Top-Split (900-1050)	Top-Split (900-1125)	Top Split (1250-1550)
Upper Frequency	MHz	42	85	200	238	1050	1125	1550
Typical Maximum Cable Loss (Amp to Amp 70 deg F)	dB	6.5	9.2	14.1	14.8	35.7	36.9	43.3
Additional Gain Required for Thermal Control (0 to 140 deg F)	+/-dB	0.5	0.6	1.0	1.0	2.5	2.6	3.0
Total Reverse Amplifier Gain Required	dB	**6.9**	**9.8**	**15.1**	**15.8**	**38.2**	**39.5**	**46.4**
"Distribution" (tapped) Segment Characterization		Sub-Split	Mid-Split	High-Split 200	High-Split 238	Top-Split (900-1050)	Top-Split (900-1125)	Top Split (1250-1550)
Upper Frequency	MHz	42	85	200	238	1050	1125	1550
Worst Case Path Loss	dB	**27.9**	**28.9**	**33.1**	**33.5**	**63.0**	**68.0**	**69.9**
Path Loss from First Tap	dB	*27.9*	*28.9*	*31.0*	*31.0*	*42.2*	*44.6*	*44.8*
Distribution Cable Loss	dB	0.4	0.6	0.9	0.9	2.1	2.2	2.6
Tap Port Loss	dB	21.9	21.9	22.0	22.0	25.4	27.2	24.5
Drop Cable Loss	dB	2.1	2.9	4.6	4.6	10.1	10.4	12.2
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5
Path Loss from Last Tap	dB	*24.4*	*26.9*	*33.1*	*33.5*	*63.0*	*68.0*	*69.9*
Distribution Cable Loss	dB	4.0	5.7	8.8	9.2	21.2	22.0	25.8
Tap Insertion Loss	dB	7.9	7.9	9.2	9.2	16.7	18.7	17.9
Tap Port Loss	dB	6.9	6.9	7.0	7.0	10.4	12.2	8.5
Drop Cable Loss	dB	2.1	2.9	4.6	4.6	10.1	10.4	12.2
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5

EPoC Application & MAC Performance

Edward Boyd
Broadcom Corporation
Kevin A. Noll
Time Warner Cable

Abstract

Ethernet Passive Optical Network (EPON) systems have been successfully deployed worldwide for high-speed access networks. EPON uses the 802.3 Ethernet MAC over optical fiber to provide high-speed IP connectivity to the home or business. In November 2011, the IEEE 802.3 formed a study group [3] to study the feasibility of creating a coax cable physical layer (PHY) for the EPON MAC. With the Ethernet-Protocol-over-Coax (EPoC) PHY, cable operators can deploy high speed IP connectivity using the EPON MAC over optical fiber or coaxial cable. Key criteria for selecting and evaluating a PHY layer will be the application in which it is used and the MAC performance over the system.

The MAC layer performance over a Coax PHY layer will be different than an optical fiber PHY layer. Emerging interactive services and higher speed data links will require shorter delays than today's services over low speed links. In this paper, the bandwidth, buffering requirements, and delay over an EPOC network will be predicted for different deployment scenarios and physical layer technologies for the EPOC PHY. The impact of increasing the round trip delay will be considered in a comparison between EPON and EPOC with expected services requirements.

Introduction

EPoC provides a solution for Cable TV operators to provide fiber performance over a coax network or Hybrid Fiber Coax (HFC) network. By re-using the EPON OLT, EPoC promises common head-end or hub site equipment for both fiber and coax customers. There are many architectural choices for EPoC to connect the OLT to the Coax Network Unit (CNU) in the customer's home.

The IEEE 802.3 working group will define a new physical layer to operate on the coax cable. During this process, decisions will be made to achieve reliable performance, high efficiency, and low delay. This paper will explore a set of service requirements for VoIP and Metro Ethernet Forum (MEF) services operating on a potential EPoC implementation. The coax physical layer will require additional functionality that will add delay and increase the round trip time. The efficiency, buffer requirements, frame delay, and frame delay variation (jitter) will be considered for a range of round trip times to understand the impact to the operator. In a point-to-multipoint network like EPON or EPoC, the shared upstream contains the highest frame delay and frame delay variation. The upstream MAC layer differences with DOCSIS and bandwidth requesting mechanisms will be considered. This paper focuses on upstream traffic performance since it is the most challenging.

EPOC Architecture

There are several possible architectures that EPoC could follow. All are rooted at an

EPON OLT and have Coax Network Units (CNU) at the leaves. The variations exist in the outside plant configuration and the implementation of the electrical interface.

Direct Coaxial Connection

One possibility removes optical fiber from the link and attaches the coaxial cable directly to the OLT system. This approach, pictured in Figure 1, mirrors what is implemented with DOCSIS CMTSes today. In DOCSIS, the electrical interface is a coaxial cable secured to the CMTS (or Edge QAM) chassis with an F-connector. It is easy to imagine that an EPOC implementation would have the same electrical interface and F-connector mechanical attachment.

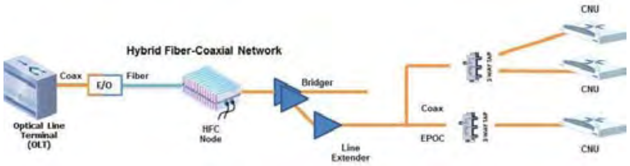


Figure 1: Direct Coaxial Connection

The practical application of this approach suffers from the fact that the bulk of coaxial plant is separated some distance from the hub site and connected via fiber optic cables. This means that the OLT would need to connect to a fiber optic link anyway. The development time and expense to develop a solution of this type is likely to be unproductive.

An alternate approach might carry the RF modulated EPoC signal over analog optics to an HFC node to be converted back to an electrical signal. This approach, however, does not provide the EPoC signal some easily realizable gains in the outside plant characteristics.

Baseband Signaling to Remote CMC

A more preferred architecture is one that uses baseband Ethernet or EPON signaling across the fiber plant. In this scenario, the hub site equipment might be (for example) an Ethernet switch containing WDM baseband optics connected to an OLT that is installed on the strand near an existing HFC node, or even in the HFC node. The OLT in this case could have a direct electrical connect to the coaxial cable and directly implement the EPOC PHY.

This architecture moves in a direction to reduce the use of expensive linear optics in the transmission path to support this type of application. However, the cost and operational complexity of installing an OLT in the outside plant is best avoided in most situations. In addition, for operators that already have EPON OLTs deployed in their hub sites, this approach is not a very effective use of capital.

A similar approach, and the one that is the focus of this analysis, uses the existing OLT and fiber plant to connect to an optical-electrical media converter that is installed in the coaxial plant. A typical configuration is shown in Figure 2.

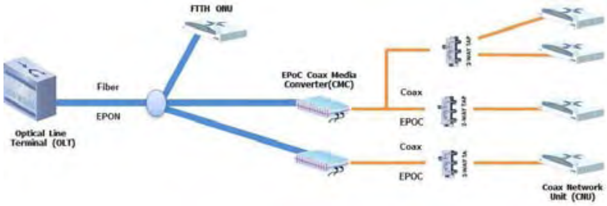


Figure 1: Baseband to Remote CMC

The EPON OLT provides the interface between the PON and external networks (the Internet, for example). It also is responsible for the well-known management functions in an EPON – admission control, station maintenance, scheduling upstream transmission, and other tasks. The role of the OLT in an EPOC network is no different than in an EPON and the CNUs appear to the OLT as if they are ONUs.

From the CMC to the CNU

As mentioned above, this chosen architecture, shown in Figure 2, requires an optical-electrical conversion. The implementation under study refers to this device as the Coaxial Media Converter (CMC). The CMC could be installed in or near an HFC node or somewhere closer to the subscriber. The CMC could be an Ethernet Switch or an Ethernet Repeater. The Ethernet Repeater could be a simpler and lower power device connecting the EPON optical PHY with the EPoC coax PHY. The Ethernet Switch would contain a bridge between an EPON MAC/PHY and an EPoC MAC/PHY.

Operators' Plant Characteristics

A coaxial cable plant, like any other transmission medium, has a set of characteristics that constrain the performance of the communication channel. The typical (but not exhaustive) list of physical-layer metrics includes signal-to-noise ratio or carrier-to-noise ratio, carrier-to-distortion ratios (Composite Triple Beat, Composite Second Order, etc.), carrier-to-interference ratio, group-delay, and micro-reflections. Each of these parameters varies based on operating frequency and bandwidth, so these two parameters must be specified as well.

Fully characterizing a coaxial cable-based network is a nearly intractable problem. Further complicating this is the variation in construction and operating practices from operator to operator and sometimes within a single operator's footprint. This study is focused on the MAC layer performance; therefore this study assumes that physical layer conditions are not a variable and circumstances allow the system to achieve the desired MAC signaling rates.

In addition to the physical-layer, the EPoC system will be expected to adapt to each operator's plant topological design and construction. The primary factors that characterize topology and affect capacity include the distance (which helps define loss characteristics and system timing constraints) from the CMC to the nearest and farthest subscriber, number of active subscribers, and offered subscription tiers (speeds).

Number of CNUs

The number of CNUs to be supported on the EPoC network needs to closely align with the number of active users on an HFC node today. This will help the operator avoid the cost of plant modifications required to deploy the EPoC system.

Today's HFC node-branch typically serves as few as 50 subscribers and as many as 500 subscribers (there are certainly cases where the node serves more or less than this). Based on this it is safe to require that the EPoC system support a similar range.

For the purpose of comparing EPOC performance to EPON performance, we should choose 32 CNUs. For the purpose of analyzing EPOC under conditions similar to today's average density this study will analyze network populations up to 512 CNUs and activity on up to 256 CNUs.

Distances

The propagation delay, that time required to transmit a frame across the coaxial cable plant, can have significant impact on the scheduler in the EPoC implementation. Therefore the distances spanned by fiber and coaxial cable in the network are an important parameter in the MAC performance.

Given the topology chosen for analysis – baseband EPON to a CMC located in or near an HFC node – we must consider two contributors to the distance from OLT to CNU. The first is the fiber from the OLT to the CMC. This distance can range from 0 meters (when the node is located in the hub site) to a typical maximum of 30km.

The second contributor to distance is the coaxial link from the CMC to the CNU. In an N+0 configuration the coaxial distance can range to around 150 meters. In an N+5 configuration with 1000-foot spacing the coaxial plant contributes about 1.7km to the total distance.

Subscription Tiers

Another factor in the system's ability to deliver traffic in a timely fashion is the speed tiers offered to subscribers. The typical Internet access service is a best effort service and ranges widely in offered data rates. A sampling of current offerings across the industry shows offered tiers 3x1Mbps (downstream x upstream bandwidth), 50x5Mbps, 60x6Mbps and as high as 100x10Mbps.

Operator Service Requirements

Operators offer many different services over their networks. Services include Internet access, Voice (VoIP), Video, Cellular Backhaul, Enterprise-class Ethernet circuits and more. Each service has its own set of network service level objectives.

Conveniently, there are two sets of specifications that can be referenced to cover the majority of these services and use cases. These specifications are the PacketCable specifications published by CableLabs and the MEF23 Implementation Agreement published by the Metro Ethernet Forum.

Packet Cable VOIP

MSOs provide packet cable VoIP service to residential and business subscribers. These are often a single line per home but multiple lines are possible, especially for business services customers.

Performance requirements for an access network supporting voice services are widely understood. Requirements specifications include packet loss, latency, and jitter. The major source of jitter in the EPON/EPOC network is scheduling the upstream transmission.

There are several sources of delay in the EPON/EPOC network. These include DSP processing and encryption, packetization, upstream transmission, and forwarding at the OLT.

Impairment	Value
Packetization Delay	20ms
Forwarding and Transmission Delay	< 10ms
Jitter	< 10ms

Table 1: VoIP Requirements

Table 1 summarizes these impairments and gives some typical tolerances in use by various service providers. In this analysis, we will assume that packet loss is trivial.

MEF 23H

The Metro Ethernet Forum defines a set of performance metrics that specify High, Medium and Low parameters that set the expectations for Ethernet services that traverse Metro, Regional, Continental, and Global distances (Performance Tiers). The general description of each performance tier (PT) is given in Table 2. In the context of this study, only the Metro PT is interesting and the EPON/EPoC network segment will generally only be a small portion of any one Ethernet service. The expected contribution of the EPON/EPOC link to the performance budget is expected to be small.

Performance Tier	Distance
PT1 (Metro)	< 250 km
PT2 (Regional)	< 1200 km
PT3 (Continental)	< 7000 km
PT4 (Global)	< 27500 km

Table 2: MEF Performance Tiers

Each PT definition includes a maximum frame delay (FD), mean frame delay (MFD) and a maximum inter-frame delay variation (IFDV).

The MEF 23 high quality service definition (H) is intended to carry delay sensitive traffic such as VoIP and financial trading transactions. These performance metrics for point-to-point delivery are summarized in **Error! Reference source not found.**.

Metric	Value
FD	≤10ms
MFD	≤7ms
IFDV	≤3ms

Table 3: MEF 23H Parameters [2]

MEF 23M

The MEF 23 medium quality service definition (M) is intended to carry traffic like Fax and network control traffic which are delay-sensitive but non-interactive. These performance metrics for point-to-point delivery are summarized in **Error! Reference source not found.**.

Metric	Value
FD	≤20ms
MFD	≤13ms
IFDV	≤8ms

Table 4: MEF 23M Parameters [2]

MEF 23L

The MEF 23 low quality service definition (L) is intended to carry Internet data service for business or residential where delay and jitter are not of any significant concern. These performance metrics for point-to-point delivery are summarized in Table 3.

Metric	Value
FD	≤37ms
MFD	≤28ms
IFDV	Unspecified

Table 3: MEF 23L Parameters [2]

EPoC System for Analysis

EPoC Sources of Delay

EPON Delays

In 1Gbps EPON, a round trip time of 250µs includes the propagation delay and physical layer delay for 20Km of fiber. The fiber propagation delay is about 100µs in each direction and 50µs covers the physical layer and synchronization delays in the OLT and ONU. For the analysis in this paper, the EPON round trip time of 250µs will be used as a baseline for comparison. EPoC bandwidth overhead (same FEC, 64/66) will be used on all RTT values so the difference is limited to the round trip delay.

EPoC Architecture

The MSO network has cable distances longer than the traditional TELCO network. While 20km may cover the entire network in EPON, EPoC will likely need to cover 30 km spans. The extended distance could add another 100µs of propagation delay.

EPoC PHY Functions

The EPoC PHY will require additional functionality to provide reliable performance when faced with burst or narrow band interference. A forward error correction (FEC) and interleaver will be selected to handle 25µs or more of burst error. The interleaver and FEC could add 400µs to 800µs delay to the round trip time.

Long symbol times of 20µs or 100µs will help combat multipath reflections. To gain better granularity, a block of symbols will be transmitted in selected carriers. Depending on the symbol and block size, an additional 400µs could easily be added.

Sharing Upstream & Downstream Frequency

Some operators like the option of using the same frequencies in the upstream and downstream in a Time Division Duplex (TDD) mode. While EPON is a full duplex protocol, half duplex operation to support TDD might be achieved by alternating between upstream and downstream transmissions in a fixed time block. To get reasonable efficiency on the upstream and downstream, a large block of transmission from each direction is needed. The larger block would be more efficient but it would add a significant amount of delay to the upstream and downstream. For example, an EPoC system that gave 1 millisecond of slot time to the upstream and 1 millisecond of slot time to the downstream would add 2 milliseconds of delay to the round trip time. The split between upstream and downstream maybe 50/50 or it might give a larger percentage to the downstream. In either case, the round trip time delay is the sum of the upstream block size and downstream block size. Small upstream block sizes would provide an additional restriction on the per-CNU upstream burst size. This paper will only consider the effect of the round trip time. An EPoC system using TDD would likely add 2 to 4 milliseconds of round trip time.

Switched or Repeated

The EPoC CMC provides a link between the optical fiber to an EPON OLT and the coax cable link to a CNU. The EPoC CMC could be defined as a switch or as a repeater.

An EPoC CMC Switch would contain an EPON ONU MAC layer connected to an EPoC OLT MAC layer through an 802.1D Ethernet Bridge. In this case, the access plant has two networks. The CMC will schedule and aggregate data from the CNUs and the OLT will schedule and aggregate data from the CMCs. The two layers of scheduling and aggregation allow for a more efficient use of the fiber. To determine the service delays, the fiber network frame delay and the coax network frame delay would be added together.

In an EPoC CMC Repeater, the EPON PHY and the EPoC PHY will be connected together in a fixed delay repeater. A single layer of scheduling and aggregation from the OLT handles upstream traffic. This system allows for a much simpler device but doesn’t provide the second level of aggregation so it will not get full utilization of the fiber network when multiple CMCs share an OLT port. In networks with large Coax plants, the fiber to the OLT would likely be point-to-point so there is no needed for aggregation on the fiber. When there are very few CNUs connected to each CMC coax segment, data from the CNUs could be aggregated to the fiber as if the CNUs are on the same coax plant. For example, four CMCs with 10 CNUs each could share an OLT port as a single 40 CNU network. The EPoC CMC Repeater does not require QoS buffers, classification, SLAs, or scheduling in the CMC.

For round trip delay analysis, only the CMC Repeater is considered in this paper. The CMC Switch performance can be determined by assuming 300µs less round trip delay on the CMC Repeater RTT time and adding a second system with the EPON delay of 250µs. For example, the FD results for a CMC switch could be determined from the CMC repeater results by the following equation.

FD-Switch(RTT) = FD-Repeater(RTT-250us) + FD-repeater(250us)

The IFDV would follow the same equation since the delay frame variation from the coax scheduling would be added to the fiber network. The total delay budget for the access plant must be shared between the coax aggregation and fiber aggregation to guarantee compliance. In all cases, the CMC Switch will add delay to the access plant because of the two stages.

Delay Summary

The EPoC system could have delays from 1ms to 6ms based on decisions made in the standard and architecture deployed by the operator. In the performance analysis, a selected set of round trip times will be used to analyze the performance impacts. In most cases, the delay would be different for upstream and downstream. To simplify the analysis, the round trip time will be divided evenly between upstream and downstream.

EPoC MAC Layer Performance

EPoC MAC Layer Differences

Packet Fragmentation

Like other Ethernet MAC solutions, EPoC does not support layer 2 fragmentation of packets in multiple flows [1]. Fragmentation in ATM and other networking technologies allow for improved Quality of Service on low speed links along with a large unit of granularity. EPoC will need to support variable packet sizes and burst sizes with a finer granularity. On higher speed links like EPoC, the value of fragmentation and reassembly is questionable for the additional complexity. Since QoS is measured by frame delay variation and maximum frame delay, QoS on cells (fragments of packets) is misleading for packet analysis. The scheduling of cells can increase the worst-case delay and frame delay variation since a packet could span multiple upstream bursts. Even though fragmentation is not supported in EPON and EPOC, this paper will consider the impact of fragmentation on the performance when appropriate.

Stateless REPORT Frame

EPON and EPOC use a REPORT frame to pass queue information from the subscriber side CNU to the operator side OLT. The REPORT frame is not a request for bandwidth. It identifies the depth of the queues at the time of generation [1]. REPORT frame values will only change when data moves in and out of the queue. It is the responsibility of the OLT to track what has been granted in the past. This method is commonly referred to as stateless bandwidth reporting since the CNU doesn’t hold state on the status of a bandwidth request. The CNU reports the queue size at the present time without regard to previous report frames.

DOCSIS systems use a stateful bandwidth request. The CM will generate a request for an upstream slot and it will not request for the same packets unless there is a timeout. The CMTS must grant the request or acknowledge it so the CM can update state on the request. A second request will not include the request in progress from an earlier request. The CM and CMTS must track the state of the request for the stateful system.

Stateful bandwidth requests were required for DOCSIS to support multicast bandwidth request slots. The multicast slots would only be used by a cable modem that hadn't already requested a bandwidth request. Stateful bandwidth requests are required for this function. EPON does not support multicast slots since the user count is lower and upstream bandwidth is higher. Performing a worst-case delay analysis is greatly simplified without multicast bandwidth request slots.

The stateless queue reporting of EPoC provides a simplification for a higher bandwidth upstream. It allows the CNU to avoid timers and long timeouts from a lost upstream request frame, downstream bandwidth acknowledge frame, or grant frame. In a stateless system, the polling interval determines the delay penalty for a lost upstream REPORT or gate frame. A timeout is considered in the delay penalty.

The REPORT frame provides a solution for reporting to frame boundaries. Since Ethernet doesn't support fragmentation, grants that aren't at frame boundaries will significantly decrease the efficiency. The REPORT frame contains one or multiple queue sets to define a queue's frame boundary at different thresholds. The queue set allows for the OLT to know a frame boundary at maximum size. A REPORT value for every frame in the queue would make a very large REPORT frame. The number of queue sets and maximum number of bytes can be configured with the SLA. For the analysis in this paper, a 4 queue set REPORT frame will be used. All 4 queue sets will have an equal limit. For example, queue set 1 will REPORT up to 4K bytes and queue set 2 will REPORT up to an additional 4K bytes. With a 4 queue set REPORT frame, the OLT can give 4 grants from a single REPORT frame before receiving the next REPORT frame. A smaller queue set will allow for smaller bursts and shorter delays for the upstream. Larger upstream queue sets will result in more efficient upstream bursts but longer delays.

Contention Slots

The EPoN MAC and EPoC system won't support contention or multicast slots. The lone exception to this rule is the discovery slot where multiple CNUs may respond. After discovery, all grants to an ONU or CNU will be unicast. Only one CNU or ONU will transmit in the slot. While the contention slots are useful in a large user network with many CMs, contention slots will prevent a smaller user network to reach high upstream data performance.

Since contention slots are not used in the EPoC based system, the worst-case delay is easier to determine and guarantee. It is also easier to show stable performance at close to or reaching 100% capacity.

The loss of contention bandwidth request slots also impacts the requirements for SLAs on the subscriber side. In DOCSIS, a cable modem will have an SLA to prevent it from over requesting bandwidth from the CMTS. The stateless REPORT frame of EPoC will only be sent by a CNU when requested by the OLT. The OLT has complete control over the

CNU for bandwidth granting and reporting so there is no need for an SLA on the CNU.

Piggybacking

REPORT frames can be sent in a single frame burst or as a frame in a longer burst with many frames. Since the REPORT frame contains the status of the upstream at the time of generation, it is normally sent as the last frame of the burst to exclude the frames in the burst. The OLT uses the force report indicator in the GATE frame to request a report frame in the burst. While a CNU could decide to send a REPORT frame in any burst, the normal practice is to send a REPORT frame only when requested by the OLT. The GATE frame with the force REPORT bit set is commonly referred to as piggybacking while the burst with only a REPORT frame in it is commonly referred to as a polling grant.

GATEs and MAPs

A MAP in DOCSIS provides a time slot description of the upstream with information for all stations. In EPoC, the GATE frame provides a unicast message to the CNU with a start time and length. In some cases, the MAP frame contains many grants over a significant portion of time. In the case of EPoC, the GATE frame will only contain a single grant to a single CNU. The GATE frame allows for up to four grants to the same CNU. In practice and in this analysis, a GATE will only contain a single grant. A MAP block delay or generation time does not exist for this reason.

Multiple LLIDs and Service Flows

A Cable Operator who provides multiple billed services to a single subscriber uses service flows to allow for different service level agreements. In EPON, the logical link identifier (LLID) provides a virtual point-to-point MAC connection between the OLT and CNU. A CNU with multiple LLIDs acts with multiple EPON MACs. With a MAC for each service, the OLT can monitor, enable, or grant the service independently of the other services on the CNU. By using multiple LLIDs, a cable operator can have multiple service flow like DOCSIS.

Activity based Polling

The large number of LLIDs or service flows on an OLT port will require a significant amount of bandwidth to query for status. Since service flows are often inactive for large residential systems, activity based polling can save bandwidth. Any service flow can have an active and inactive polling rate. The active polling rate would be much higher than the inactive rate. A simple example is a VoIP call where the active rate is used when a call is active and the inactive rate is used when no call is active. Activity can be determined by looking at the presence, rate, or type of frames on a link. The OLT system can determine the rules for activity and inactivity.

EPoC PHY Parameters for Analysis

The IEEE 802.3 will define overheads for the physical layer. Commonly suggested options for FEC and encoding burst overhead will be selected to get an estimate of the overhead. The Ethernet Frames will use the 64/66 encoding of 10G EPON and an 85% efficient LDPC FEC code. With these constant overheads, a fixed 20% overhead would be needed. A 1Gbps Ethernet MAC rate would require 1.2Gbps of Ethernet Line rate.

For bursts, a shortened FEC code word is allowed for end of bursts. A common burst overhead for EPON is 32 time quanta (time quanta are 16ns long) for sync time, 64 time quanta (TQ) for laser ON, and 64 TQ for laser OFF. At 1Gbps, the total burst overhead will be 1536 bits or 192 Bytes. EPoC will use the same burst overhead as EPON so the analysis can focus on performance differences due to round trip time. A larger EPoC burst overhead would reduce the performance and it should be considered in future analysis.

Packet Cable VoIP

Packet Cable VoIP service can be mapped to EPoC in a variety of ways. The most obvious is an unsolicited grant similar to DOCSIS. Another solution is a solicited granting based on polling.

For the analysis below, the G.711 codec will be assumed. Based on this code, a 218-byte packet will be generated every 20ms for each subscriber with an active voice call. A maximum FD and IFDV of 10 milliseconds will be required.

Unsolicited Grant Synchronization (UGS) Performance

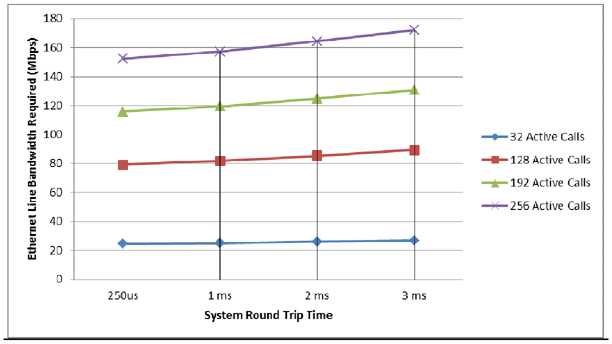
In the UGS solution, the EPoC system will establish two LLIDs. One LLID will carry signaling while the other LLID will carry the encoded voice. The encoded voice LLID will use unsolicited granting. Unsolicited granting is based on a timer at the OLT. A fixed size grant is given in a fixed time period. A REPORT frame with a non-zero queue set is not required for the grant generation. The signaling LLID will use solicited granting. Using activity based polling, the LLID will be polled at 17ms when the LLID is active and 100ms when inactive. The unsolicited granting could be enabled or disabled in the OLT based on the state of the voice call from observing the signaling channel.

The UGS slot will be sized large enough to carry a single 218 Byte Ethernet frame. The granting period of the UGS must guarantee a maximum delay of less than 10ms. The UGS slot is not aligned with the arrival time of the packet so the worst case scenario is an upstream frame just after the slot passed. The worst case delay of packet upstream will be the upstream transport delay plus the period of the UGS slot. The downstream delay does not factor into the UGS performance since the GATE is autonomously generated by the OLT. It is assumed that the worst case slot jitter from discovery slots is less than 500µs.

The period of UGS slots to a CNU must decrease with increased upstream delay. The period can be determined by subtracting the fixed delays from the worst case delay of 10ms. The equation below can be used to find the UGS period. As the UGS period decreases, the amount of upstream bandwidth consumed increases.

UGS-Period = MaxDelay – RTT/2 – SlotJitter

In the example scenario, each CNU will have a single VOIP session. The system is assumed to have 512 CNUs. The amount of Ethernet Line bandwidth required is shown for a different numbers of active voice calls and for different round trip times.



Graph 1: Required UGS Bandwidth

Graph 1 shows the bandwidth required for different round trip times and numbers of active voice calls. The System Round Trip Time of 250µs represents the performance of the all fiber 20Km EPON solution. The 1ms, 2ms, and 3ms show the performance of an

EPoC system with the corresponding total round trip time.

For UGS, the increase in bandwidth required due to longer round trip times is not significant for a small number of active calls. The increased RTT is more significant with 256 active callers.

The UGS efficiency is hurt by the single packet bursts. Additionally, the 20ms arrival time and sub-10ms delay will cause over half of the upstream slots to be empty.

Fragmentation or No Fragmentation

The UGS analysis assumes that EPoC does not allow fragmentation. Would fragmentation improve the capacity or decrease the delay? If packets were fragmented, they would need to wait for an additional UGS slot to be transported upstream. If the packet boundary and slots were miss-aligned, it would take up to 2 UGS slots for a frame to go upstream. In this case, the UGS slot would need to occur twice as often. The payload in the UGS slot could be divided in half. Since the overhead would double for the shorter interval, fragmentation would significantly increase the bandwidth required to transport the UGS flows.

Solicited Granting Performance

The UGS solution provides an adequate solution for transporting VoIP over EPON and EPoC. The UGS has the complexity of detecting the start and end of phone calls. UGS also requires a known packet interval and packet size. Additional phone lines at a CNU require more UGS flows or the complexity of detecting multiple phone calls in a single service flow. UGS is also not easily compatible with compressed voice or video conferencing. Solicited granting would greatly simplify the control and allow for other service options. Solicited is preferred if the performance is similar to UGS.

Solicited granting requires a REPORT frame to transmit upstream, a GATE frame downstream, and a data burst to be received upstream. The transport delay is therefore the downstream delay plus two times the upstream delay. Since data comes in asynchronous to the scheduler, the worst case delay should include the delay from simultaneous upstream slot requests from all active VoIP flows. For this analysis, we assume that VoIP is the highest priority. The polling period is the key parameter for the solicited solution. The following equation can be used to calculate the worst case delay.

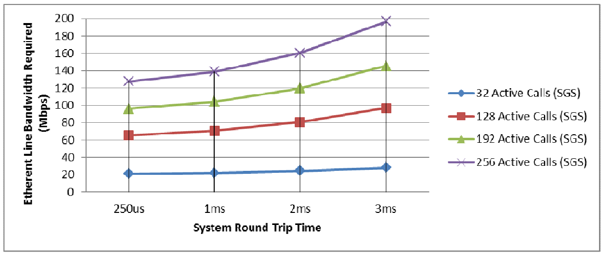
Tmax_delay = Tpolling + 2xTup + Tdown + Tall_service

The following equation solves for the polling interval.

Tpolling = Tmax_delay – 2xTup – Tdown – Tall_service

In the case of VoIP, piggybacking will not be used. While piggybacking would decrease the latency for arriving packets, it would not decrease the worst-case latency. In the case of the VoIP example, the packet spacing is larger than the maximum delay so a piggybacking would be useless to detect the next frame.

Graph 2 shows the bandwidth capacity required by the solicited VoIP solution.

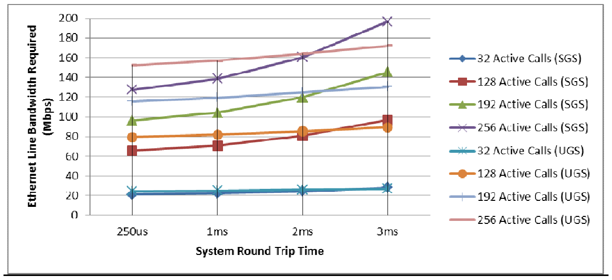


Graph 2: Required Solicited Bandwidth

The UGS bandwidth increase due to increased round trip time was much less than the solicited solution because of extra round trip in the delay equation. At a small number of active calls, the UGS shows little or no difference with a lower or higher round trip time.

The solicited solution is more efficient for the shorter round trip times. The solicited solution benefits from only granting data slots when a frame is present. As the RTT increases, the increasing polling rate to meet the maximum delay consumes more bandwidth than the wasted slot in the UGS solution.

When comparing the 250µs EPON data point, there is less than 10% increase in bandwidth to achieve the same delay performance if the round trip time is in the 1.5ms range. A 3ms round trip adds a 50% bandwidth penalty to achieve the same delays. It is clear that RTT delays beyond 3ms are unusable in the solicited.



Graph 3: UGS & Solicited Bandwidth

If the UGS and Solicited graphs are overlaid, it shows a cross over point between UGS and solicited around 2ms of round-trip time. A solicited solution is equal performance for a small number of users and it is better performance if the round trip time is less than 2ms. Since the solicited solution is more flexible for video or compressed content and simplifies controls, a lower round trip time that allows for efficient use of soliciting granting is preferred. For DOCSIS systems with many users and long delays, UGS must be used. For EPON systems with fewer users and shorter delays, solicited granting is clearly preferred.

Performance for MEF 23H

Requirement Overview

The MEF 23H service agreement is an example of a higher tier business or residential SLA. For the MEF 23H service, an IFDV of 3 milliseconds and a maximum FD of 8 milliseconds will be used as requirements. For the analysis, a 10Mbps-streaming load will be applied in the upstream direction. The 10Mbps load has a random packet size from 64 bytes to 1518 bytes.

Configuration to reach goals

With an unconstrained packet size, only solicited operation can be used since a UGS would require knowing the packet boundaries. In a system without fragmentation, the unknown packet boundary would be very inefficient. In a system with fragmentation, a packet spanning 2 grant slots would double the delay. In either case, UGS is not the preferred method.

Since the period of polling must be short to meet the IFDV requirement, there is no need to use piggybacking. While piggybacking may lower the average delay in some scenarios, it will not decrease the worst case IFDV or FD. Piggybacking would decrease the efficiency because of the additional REPORT frame in the burst.

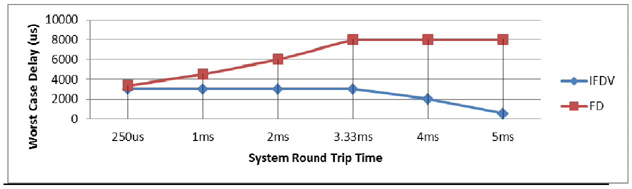
The IFDV is the critical constraint in this system. The IFDV in the upstream will be sum of the polling interval and the scheduler delay. A packet arriving just before the polling slot will have zero delay while a packet arriving just after the polling slot will wait the entire polling interval. The scheduler delay can be zero when only one CNU requests an upstream slot for shortest delay. The longest scheduler delay occurs when all CNUs need a slot at the same time. The maximum scheduler delay is number of CNUs times the maximum slot size.

The best efficiency can be found when the IFDV is equally split between polling and scheduler contention delay. For a 3ms IFDV, the scheduler delay of 1.5ms and a polling delay of 1.5ms are allowed.

Performance Analysis

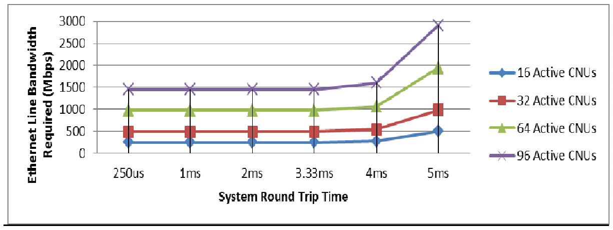
The maximum delay is defined by the same equation as the VOIP solicited grant example. It should be noted that this equation is the same as the IFDV plus the twice the upstream delay and downstream delay. The delay graph shows the relationship between the round trip time and the FD and IFDV. The bandwidth graph shows the best efficiency is found when the IFDV value is largest. A large IFDV allows for a lower polling rate and larger upstream data bursts which results in higher throughput.

At the EPON round trip time of 250µs, the maximum delay is far below the 8ms maximum. As constant delay is added for the round trip time increases, the IFDV and the efficiency remains the same. When the additional RTT delay causes the maximum delay to be exceeded, the polling period and burst size must be decreased. These decreases cause the bandwidth required to increase dramatically.



Graph 4: MEF 23H Delay

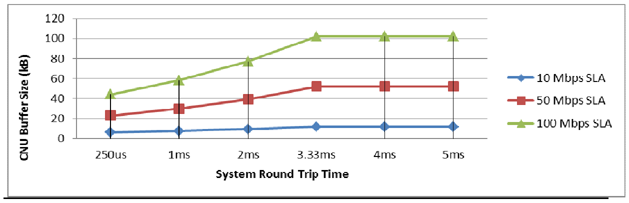
While the maximum delay increases for the RTT of 250µs to 3.33ms, delay is under the 8ms maximum and the efficiency is constant. If EPoC RTT delay is under 3.33ms, the MEF23H service can be supported without any additional bandwidth. Above 3.33ms, the penalty increases dramatically until the absolute limit of 5ms where the minimum polling period of 250µs is reached. At the 5ms limit, the bandwidth required to meet MEF 23H is more than double EPON at 250us.



Graph 5: MEF 23H Bandwidth

CNU Buffering Requirements

The additional delay will impact the buffering requirements for a CNU in the upstream direction. For a MEF 23H service, it is assumed that it is a guaranteed bandwidth without best effort data. The MEF 23M and MEF 23L will consider best effort data. With only guaranteed bandwidth to consider, the buffering required can be found by the multiplying the guaranteed rate by the frame delay (FD). Since the buffer is normally store-and-forward, 2000 bytes (the largest 802.3 packet size) is added.



Graph 6: MEF 23H Buffering

The results for MEF 23H buffer size required versus RTT has the same shape as the delay graph and the opposite shape of bandwidth graph. The increase in RTT increases the buffer size until the maximum delay is reached. After the maximum delay, additional RTT increases don’t change the buffer size but bandwidth for higher polling rate climbs. Graph 6 shows that while the increase in the EPON fiber RTT delay from 250µs to 3.33ms does not hurt the efficiency, it more than doubles the upstream buffering requirements in the CNU.

Performance for MEF 23M

Requirement Overview

For MEF 23M, a worst case IFDV of 8 milliseconds and FD of 20 milliseconds will be used. For the analysis, a 10 Mbps and 50 Mbps streaming load will be applied in the upstream direction. The load has a random packet size from 64 bytes to 1518 bytes.

Configuration to reach goals

The MEF 23M traffic patterns would not normally be a traffic pattern compatible with a UGS flow. Variable sized bursts of unknown packet sizes are best handled by solicited granting.

The longer FD and IFDV limit allow the use of piggybacking for better efficiency than the polling only solution used in MEF 23H. The polling timer will be reset by the generation of a polling burst or a piggybacked REPORT frame. For a bursting station with a polling period greater than or equal to the scheduler contention delay, no polling bursts will be requested.

There are 2 equations to determine the maximum delay. The first equation is based on a station that has been active but not bursting. Polling will detect the packet in this case. This scenario will be referred to as “burst detection”.

Tmax_delay = Tpolling + 2xTup + Tdown + Tall_service

The second equation is based on a CNU that is bursting and not reporting a zero length queue. In this case, the piggybacking will detect the packet arrival. This scenario assumes that the flight delay of 2xTup + Tdown is less than the time to service all stations. The scenario will be referred as “burst continuation”.

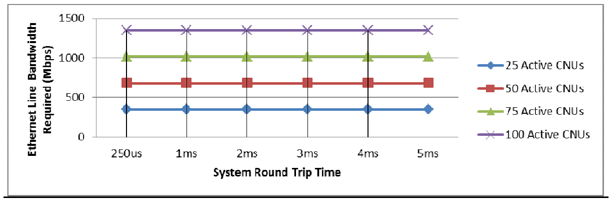
Tmax_delay = Tup + 2xTall_service

The worst case delay can be determined by taking the longer delay from the burst detection or burst continuation scenarios. For optimum performance, the polling interval should never be less than Tall_service and for best performance, they should be set equal. In this case, the burst continuation equation is not the worst case so the burst detection equation will be used for analysis. The Tpolling interval will be half the result of the maximum delay minus 2xTup + Tdown.

For a system mixed with higher priority services like MEF 23H, the Tall_service should include their burst interruptions. Tall_service should be the sum of all higher and same priority upstream slots. Since the MEF 23H IFDV is 3ms, the MEF 23M will assume no more than a 3ms disruption.

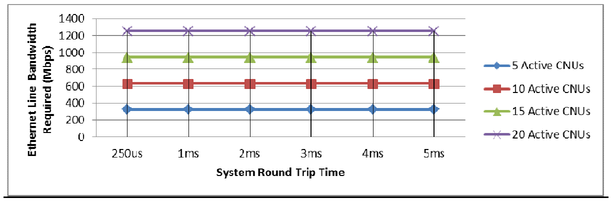
Performance Analysis

Graph 7 shows that as the round trip time increases no increase in the bandwidth required. Since the FD of 20ms is larger than the IFDV of 8ms plus 5ms RTT, there isn't a need to increase the granting rate. The bandwidth increase wouldn't occur in the MEF 23M until a RTT of around 12ms.



Graph 7: MEF 23M Bandwidth (10Mbps)

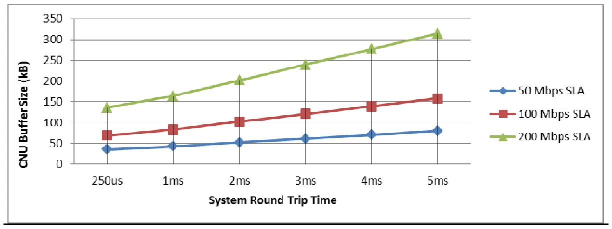
Graph 8 shows fewer active CNUs and therefore fewer bursts at a 50Mbps rate each. The charts show that the penalty for extended RTT is larger when there are more users and lower data rates. For a system with many users and higher data rates, the RTT doesn't have significant impact up to 5ms.



Graph 8: MEF 23M Bandwidth (50Mbps)

Buffering Requirements

While the efficiency of the system for MEF23M is constant from with the increased RTT, the buffering requirements on the CNU are not. The buffer required on a CNU to support the MEF 23M with data rates up to 200 Mbps would need to be more than double the EPON ONU. Graph 10 shows the buffering requirements to support average rates of 50, 100, and 200 Mbps.



Graph 10: MEF 23M Buffering

MEF 23L

Requirement Overview

The MEF 23L service agreement is an example of a best effort SLA. The MEF 23L specification contains a maximum FD requirement of 37 milliseconds. For the analysis, different data rate streaming load will be applied in the upstream direction. The load has a random packet size from 64 bytes to 1518 bytes.

Configuration to reach goals

For the same reasons as MEF 23M, a solicited granting with piggybacking will be used. A 37ms delay limit is very long for an EPON or EPOC system that is not oversubscribed. The RTT will be a small percentage of 37ms delay limit so it will not have a significant impact on the efficiency like the MEF 23M. The RTT will have a significant impact on the buffering requirements for a CNU to reach high bandwidth. In general, the MEF 23L needs to achieve high efficiency at a high data rate without requiring a large amount of upstream buffering.

The polling rate could be set for MEF 23L to 10ms and meet the FD requirement of 37ms. For the MEF 23L, different polling rates will be considered to balance efficiency with buffering requirements.

The burst detection condition will be considered for the same reason as MEF 23M.

The Tall_service delay is more difficult to determine at this priority level since many higher priority services could be active. The disruption from MEF 23H and MEF 23M services will be limited by the MEF23M IFDV of 8ms. Of course, this analysis assumes a round robin scheduler with guaranteed slots for lower priorities and shaping that streams the higher priority. Without these restrictions to the high priority, the delays to MEF 23L could be unbounded. Tpolling will be equal to Tall_service.

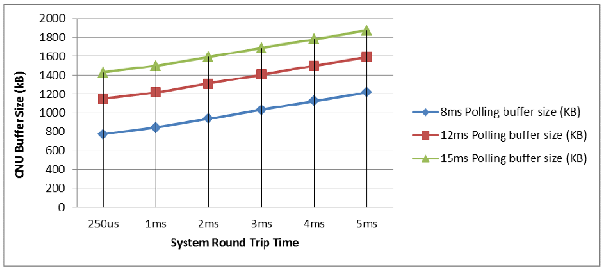
Tmax_delay = Tpolling + 2xTup + Tdown + Tall_service

The polling interval for the MEF 23L service can be determined subtracting the loop time and dividing by 2.
Tpolling = (Tmax_delay - 2xTup – Tdown)/2

To handle the disruption from high priority services, the MEF 23L polling rate shouldn't be set less than the 8ms IFDV of MEF 23M. For a 5ms delay, the maximum polling rate is just under 15ms. The analysis will look at polling rates from 8ms to 15ms.

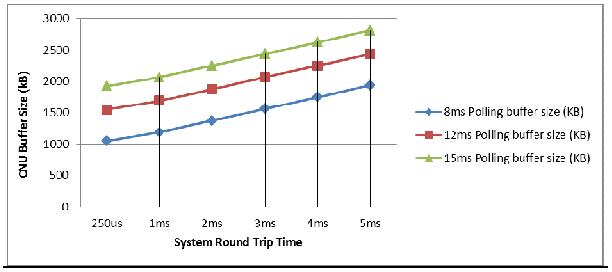
Performance Analysis

The MEF 23L buffer size is calculated for the different polling rates versus RTT. In Graph 11, the buffer size is considered for a sustained rate of 500 Mbps with 50% of the bandwidth taken by MEF 23M services. The buffer requirements are the maximum delay times the sustained input bandwidth.



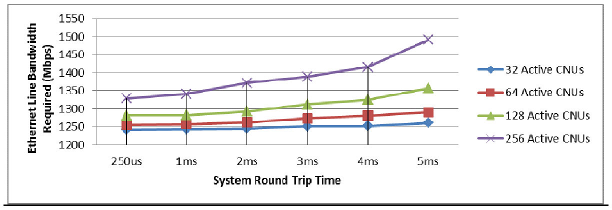
Graph 11: MEF 23L Buffer Size (500 Mbps)

Graph 12 shows the buffer size requirements for an empty system where a single MEF 23L CNU is bursting at 1 Gbps (100%) with no contention delay. Comparing Graph 11 and 12, it is clear that the worst case buffer requirement for MEF 23L is a single user with an SLA to reach maximum bandwidth. The buffer requirement decreases with more users sharing the upstream as the maximum data rate decreases.



Graph 12: MEF 23L Buffer Size (100% load)

Graph 12 shows that EPoC will require a significant amount of additional buffering (~1 MB) over EPON as the RTT time is increased. From Graph 12, the buffering requirements for EPON and the 5ms RTT are equivalent if the EPON system uses 15ms polling and the EPoC system uses 8ms polling. Since the buffering is a directly related to the delay, the EPON and EPoC would have the same delay as well. For a system with few CNUs, the penalty to compensate for RTT delay with polling will be small but a larger system will require significantly more bandwidth. Graph 13 shows the impact of increasing the polling rate to match the delay and buffer requirements of EPON. In the example for Graph 13, a fixed buffer size of 1.5 MB is used. The 1.5MB buffer represents the 12ms polling, 250µs RTT, and 25ms delay on Graph 12. Graph 13 assumes that the system will carry 1 Gbps of Ethernet traffic split evenly across the stations.



Graph 13: MEF 23L Bandwidth (100% load)

Graph 13 shows that the penalty for increased RTT multiples by the number of users. The system with 256 active users will have around a 15% penalty on bandwidth to match to match the EPON fiber based performance.

Conclusions

An EPoC PHY can be used by a cable operator in multiple network configurations. The EPoC PHY could be placed with an OLT in headend and operate over a traditional HFC network or the EPoC PHY could be placed in a CMC at a remote node and act as a switch or a repeater. The choice of architecture is dependent upon the individual operator's needs and plant design.

EPoC can provide a significant performance improvement over existing cable systems because of small service groups, a fast Ethernet MAC, and a single wide logical pipe. EPoC can provide VoIP, MEF 23H, MEF 23M, and MEF 23L services if the round trip time is low enough. RTT increases will impact the CNU cost dramatically if it requires an EPoC specific chip with more buffering than the standard EPON ONU. Increased polling rates can compensate for larger round trip times to certain limits and still meet MEF 23 requirements but bandwidth efficiency will be reduced. Going over 2ms, forces EPoC from a solicited VoIP into the less flexible UGS VoIP. At RTT's of 3.33ms and 5ms, some MEF 23 services become impossible. Solutions with a shorter round trip time will be more efficient and perform closer to the fiber solutions without additional hardware or bandwidth costs.

The IEEE 802.3 standard should seriously consider the round trip time impacts in selecting the solution. A solution that increases the bandwidth efficiency at the PHY layer by adding significantly delay could hurt the overall system efficiency.

References

[1] Glen Kramer, Ethernet Passive Optical Networks, McGraw-Hill, 2005
[2] MEF Technical Specification MEF 23.1, "Carrier Ethernet Class of Service - Phase 2"
[3] EPON Protocol over a Coax (EPoC) PHY Study Group. http://www.ieee802.org/3/epoc/index.html

Mission is Possible:
An Evolutionary Approach to Gigabit-Class DOCSIS

John T. Chapman, CTO Cable Access BU & Cisco Fellow,
Cisco, jchapman@cisco.com

Mike Emmendorfer, Sr. Director, Solution Architecture and Strategy,
Arris, Mike.Emmendorfer@arrisi.com

Robert Howald, Ph.D., Fellow of Technical Staff, Customer Architecture,
Motorola Mobility, rob.howald@motorola.com

Shaul Shulman, System Architect,
Intel, shaul.shulman@intel.com

Abstract

This paper is a joint paper presented by four leading suppliers to the cable industry, with the intent to move the industry forward in the area of next generation cable access network migration. To our knowledge, it is a first for four such suppliers to collaborate in this manner on a topic of such critical industry importance.

Cable operators are facing a rising threat associated with the limitations of today's 5 to 42 MHz return path. Constraints on capacity and peak service rate call for finding additional return spectrum to manage this emerging challenge.

We will explain how and why an approach based on the principle of an expanded diplex architecture, and using a "high-split" of up to 300 MHz, is the best path for operators to manage this growth. This includes considering the simultaneous expansion of the downstream capacity.

We will describe obstacles associated with legacy CPE in both Motorola and Cisco video architectures and propose solutions to these issues.

To use the reallocated HFC spectrum most effectively, we will consider an evolutionary strategy for DOCSIS and show how it capably meets the requirements ahead.

We will contemplate the application of new generations of communications technology, including a comparison of single-carrier approaches implemented today to multi-carrier techniques such as OFDM, including channelization options. We will consider higher order QAM formats as well as modern FEC tools such as LDPC.

We will discuss how these evolution alternatives can be harnessed to best extract network capacity. We will consider how evolution of the access architecture enables this new capacity, and how the end-to-end network components develop to support this growth.

In summary, we will present a strategy that preserves network investment, enables a versatile evolutionary path, and positions operators to create an enduring lifespan to meet the demands of current and future services.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The evolution of DOCSIS is bounded only by technology and imagination - both of which themselves are unbounded.

This white paper takes an in depth look into the technologies that are available to DOCSIS and then makes concrete recommendations on how DOCSIS should be taken to a new level of performance.

DOCSIS to Date

The original DOCSIS 1.0 I01 (Interim version 1) specification was released on March 26, 1997. DOCSIS technology has evolved very well since its inception over 15 years ago. Here are some of the interesting milestones from those first 15 years.

- 1997 Mar – DOCSIS 1.0 I01 released. Features basic data service.
- 1997 Dec – Cogeco has the first large scale DOCSIS 1.0 deployments
- 1999 Mar – First certified CM and qualified CMTS
- 1999 Apr – DOCSIS 1.1 released. Adds QoS.
- 1999 Dec – PacketCable 1.0 released. Adds voice over IP (VoIP)
- 2001 Dec – DOCSIS 2.0 released. Adds ATDMA and SCDMA.
- 2002 Feb – DSG released. Adds STB control channel to DOCSIS
- 2005 Aug – Modular CMTS (MHA) released. Shared EQAM between DOCSIS and video is added.
- 2006 Aug – DOCSIS 3.0 released. Adds bonding, IPv6, and multicast.

In the first phase of its life, DOCSIS focused on a moderately dense and complex MAC and PHY with a comprehensive set of features and services. DOCSIS now has a very rich and mature service layer.

If this was the first 15 years of DOCSIS, then what is the next 15 years of DOCSIS going to look like? How well will DOCSIS compete with other broadband technologies?

The Future Potential of DOCSIS

The next phase of DOCSIS will take it to gigabit speeds. DOCSIS needs to scale from a few RF channels within a CATV spectrum to being able to inherit the entire spectrum. And DOCSIS may not even stop there.

In the upstream, in an effort to get to gigabit speeds and beyond, DOCSIS needs to scale beyond its current 5 – 42 MHz (65 MHz In Europe) to multiple hundreds of megahertz. In the downstream, DOCSIS needs to extend beyond the current 1 GHz limit and set a new upper RF boundary for HFC Plant.

Table 1 shows where DOCSIS technology is today and where it is going.

Today, the deployed DOCSIS 3.0 cable modems have eight downstream channels (6 or 8 MHz) and four upstream channels (6.4 MHz). This provides an aggregate downstream data capacity of about 300 Mbps and an aggregated upstream data capacity of 100 Mbps.

Next year (2013), the market will see cable modems that have on the order of 24

Table 1 – The Future Potential of DOCSIS

	Parameter	Now	Phase 1	Phase 2	Phase 3
Downstream	Frequency Band	54 - 1002 MHz	108 - 1002 MHz	300 - 1002 MHz	500 - 1700 MHz
	Assumed Modulation	256-QAM	256-QAM	≥ 1024-QAM	≥ 1024-QAM
	Chan (or equiv)	8	24	116	200
	Data Capacity	300 Mbps	1 Gbps	5 Gbps	10 Gbps
Upstream	Frequency Band	5 - 42 MHz	5 - 85 MHz	5 - (230) MHz	5 - (400) MHz
	Assumed Modulation	64-QAM	64-QAM	≥ 256-QAM	≥ 1024-QAM
	Chan (or equiv)	4	12	33	55
	Data Capacity	100 Mbps	300 Mbps	1 Gbps	(2) Gbps

downstream channels and 8 upstream channels. DOCSIS 3.0 defines a mid-split upstream that takes the upstream spectrum up to 85 MHz and could contain at least 10 channels. That provides an aggregate data capacity of almost 1 Gbps in the downstream and 300 Mbps in the upstream.

The goal for the next generation of DOCSIS is to achieve 1 Gbps of data capacity in the upstream and to be able to scale to the full spectrum of the existing downstream. While the final spectrum plan has not been determined yet, an estimate would be a 5 Gbps down, 1 Gbps up system. That would maintain a 5:1 ratio between upstream and downstream bandwidth that is good for TCP.

As a stretch goal, there is additional spectrum above 1 GHz. If the downstream expanded into that spectrum, and the upstream spectrum was increased even further to keep the same 5:1 ratio, DOCSIS could become a 10 Gbps down and 2 Gbps up technology.   This would enable cable data capacity equivalent to next generation PON systems.

While the final choices for these numbers (indicated with "( )") still needs to be made, there seems to be at least three progressions of technology. Phase 1 upgrades the upstream to 85 MHz and takes advantage of technology available today. Phase 2 upgrades the upstream to 1 Gbps and the downstream to 1 GHz if it is not there already. Phase 3 extends the downstream to 1.7 GHz and gives a second boost to the upstream.

Now that we have established our goals, let’s look at how to achieve them.

2 CABLE SPECTRUM ANALYSIS

The spectrum allocation options should consider the impact to the overall end-to-end system architecture and cost. The solutions should also consider the timing of these changes as this may impact cost. The end-state architecture should be considered for this next touch to the HFC. We do not need to solve next decade's problems now, however we should consider them as part of the analysis.

The cable operator has several spectrum split options available and some are examined in this analysis. [33] [34] [35] Figure 1 below is an illustration of some of the spectrum split options; it also depicts a few other options, such as Top-split with Mid-split. In Figure 1, the Top-split (900-1050) options has a 150 MHz block of spectrum allocated for guard band between 750-900 MHz and 150 MHz block of spectrum between 900-1050 MHz for upstream.

2.1 Mid-split (85)

Overview

The Mid-split Architecture is defined as 5-85 MHz upstream with the downstream starting at approximately 105 MHz; this may also be referred to as the 85/105 split. The mid-split architecture essentially doubles the current upstream spectrum allocation

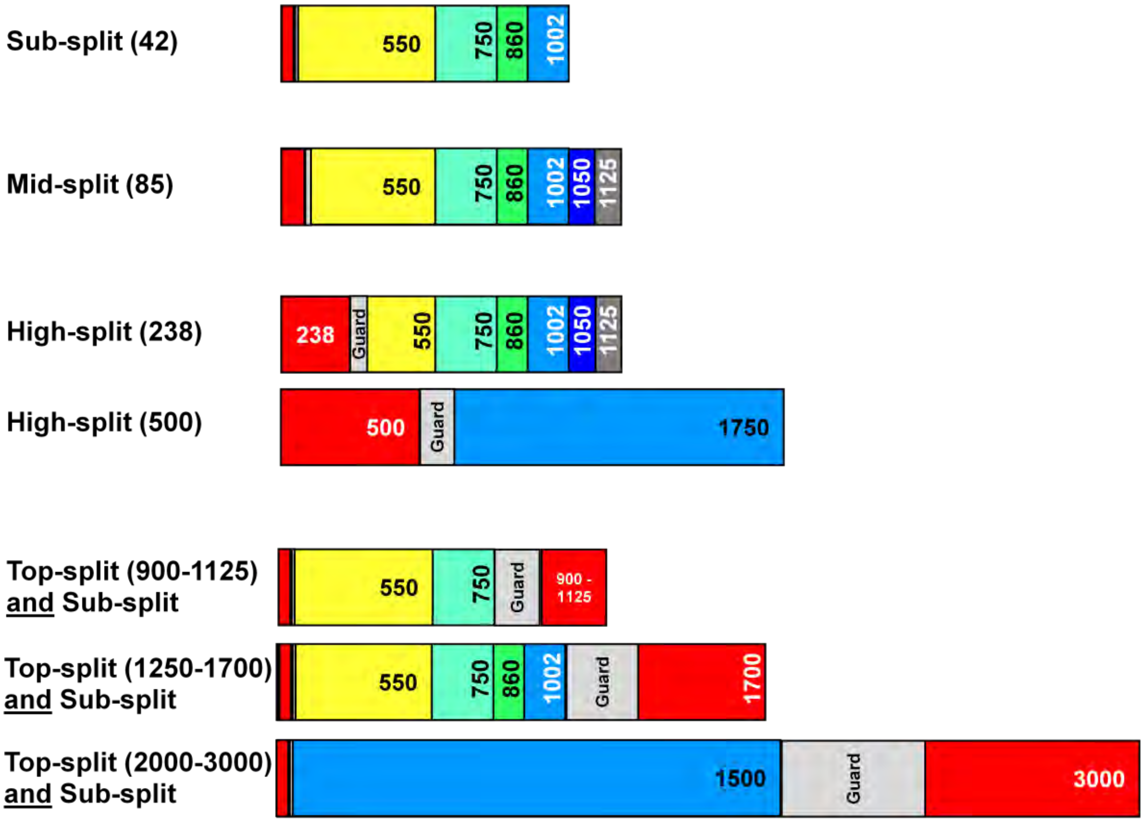


Figure 1 – Spectrum Allocation Options

however this may triple or even quadruple the IP based capacity.

The capacity increase in data throughput is a result of the high-order modulation and all of the new spectrum may be used for DOCSIS services, which is not the case with the sub-split spectrum that has generally accepted unusable spectrum and legacy devices consuming spectrum as well.

Pros

- Sufficient bandwidth to last nearly the entire decade
- DOCSIS QAM MAC layer capacity estimated at ~310 Mbps
- Avoids conflict with OOB STB Communications
- Lowest cost option
- High order modulation possible 256-QAM perhaps higher
- The use of 256-QAM translates to fewer CMTS ports and spectrum (using 64-QAM would require approximately 33% more CMTS ports and spectrum)
- DOCISIS systems already support this spectrum (5-85)
- MSOs that have already deployed DTAs (Digital Terminal Adapters) should strongly consider thing approach
- Some amplifiers support pluggable diplexer filter swap
- Some existing node transmitters and headend receives may be leveraged
- Does not touch the passives
- Upstream path level control is similar to the Sub-split (~1.4 times the loss change w/temp); Thermal Equalizers EQT-85 enables +/-0.5 dB/amp delta

Cons

- Impacts Video Service (in low channels)
- Reduces low VHF video spectrum
- Throughput of 310 Mbps is less than the newer PON technologies

Assessment

The selection of Mid-split seems like an excellent first step for the MSOs. This split option has little impact to the video services and does not impact the OOB STB commutations. This spectrum split may last nearly the entire decade, allowing time for the MSOs to assess future splits, if required, and the impacts to other split option at that time. The Mid-split appears to be an excellent first step. MSOs that have already deployed DTAs should strongly consider using this approach.

2.2 High-split (200, 238, or 500)

Overview

The High-split Architecture has generally been defined as 5-200 MHz with the downstream starting at approximately 250-258 MHz crossover for the downstream. However, we believe that a High-split (238) or even High-split (270) options should be considered, as this will have enough spectrum capacity to reach the desired 1 Gbps data rate, with reasonable PHY and MAC layer overhead removed. [33] [34] [35]

Also it is uncertain if the entire region of spectrum between 5-238 may be used as there could be legacy channels in service as well as frequency bands undesirable performance or usable for interference reasons. The use of High-split (500) has been mentioned as a possible long-term migration strategy if coaxial network want to offer the capacity of XG-PON1 systems.

In the case of 5-500 MHz our capacity targets assume a digital return HFC style optical connection and as will all architectures the paper model begins at a 500 HHP node to a 16 HHP node to determine capacity.

Pros

- High-split is far more predictable from an MSO deployment, operational, and service ability perspectives when compared with Top-split, as Top-split options have much tighter cable architecture requirements (refer to Cons of Top-split).
- Operates effectively at a typical 500 HHP node group using 256-QAM (see details in the sections later in this analysis)
- The use of 256-QAM translates to fewer CMTS ports and spectrum (using 64-QAM would require approximately 33% more CMTS ports and spectrum)
- High-split (238) using DOCSIS QAM reaches an estimated MAC layer capacity 1 Gbps
- However High-split (270) may be needed to allow for operational overhead
- High-split (500) at a 250 HHP through a 16 HHP optical node service group with digital return HFC optics is estimated to reach 2.2 Gbps DOCSIS QAM MAC layer capacity
- DOCSIS OFDM with LDPC may be able to use 2 orders higher modulation in same SNR environment
- Very low cost spectrum expansion option, especially considering similar capacity Top-split options (STB OOB cost was not considered in the analysis)
- The OOB STB problems will likely be reduced over time, and with the STB costs declining over time this will remove or reduce this issue to High-split adoption
- If DTAs are deployed or plan on being deployed High-split should be considered strongly, because DTA remove the Analog Video Service impact obstacle from High-split
- Lowest cost per Mbps of throughput
- Some existing HFC Equipment supports High-split like node transmitters and headend receivers
- DOCISIS systems already support some of this spectrum (5-85)
- Passives are untouched
- High-split provides sufficient upstream capacity and the ability to maximize the spectrum with very high order modulation
- High-split does not waste a lot of capacity on guard band
- Level control using Thermal Equalizers EQT-200 (~2.2 times Sub-split cable loss)
- Downstream could expand to 1050 MHz or even 1125 MHz perhaps using the existing passives

Cons

- Conflicts with OOB STB Communications if DOCSIS Set-top box Gateway (DSG) is not possible

- Takes away spectrum from Video Services (54-258 MHz or higher if the upstream stops at 238 MHz)
- Takes away spectrum from Video devices (TVs and STBs)
- Potentially revenue impacting because of spectrum loss supporting analog video service tier
- Downstream capacity upgrade from 750 MHz to 1 GHz to gain back capacity lost to upstream

Assessment

The use of high-split has several key challenges or cons listed above, and the major concerns include 1) the impact OOB Set-top Box communications for non-DOCSIS Set-top Gateways, 2) the analog video service tier and the simplicity of connecting to an subscribers TV to enable services, and 3) we takeaway valuable capacity from existing video devices like STBs and existing TVs.

However, if the deployment of High-split (238) is planned later in time, this may allow these older STBs to be phased out or redeployed to other markets. There may also be workarounds to enable high-split and keep the legacy OOB in place. The impact to the analog service tier is a major concern, this accounts for a large portion of how customers received video services.

If a customer is a digital video subscriber they likely have TVs, in fact likely more TVs, which are served with no STB at all, and receive a direct coax connection. This is a valuable service feature for the MSO. However, we do recognized that many MSOs are considering the deployment of DTAs to recover analog spectrum, if the MSOs do a full all digital service and have no analog, this will make a migration to High-split a stronger consideration.

Additionally, MSOs could expand to 1050 MHz or even 1125 MHz perhaps using the existing passives, this very important because the technical benefits of using the bandwidth around 1 GHz are superior for the forward path compared with placing the return approaching or above 1GHz, discussed in detail in this analysis.

If the main challenges with the use of High-split are overcome, this seems like the ideal location for the new upstream (technically). The economics are also compelling for High-split against the other split options considering just the network access layer.

If the STB Out of Band (OOB) and analog recovery need to be factored into to the High-split, the cost analysis will change, however these will continue to be phased out of the network. The costs to move analog services, which are non-STB subscribers, were not considered in the model. However many MSOs are already planning to use DTAs to reclaim the analog spectrum, this would make a migration to High-split more obtainable.

The High-split option may need to exceed 200 MHz and move to approximately 5-238 MHz to achieve a MAC Layer throughput around 1 Gbps. This would use the 22.4 MHz of spectrum in the existing Sub-split band and the new spectrum up to 238 MHz to allow thirty-three (33) 6.4 MHz wide DOCSIS 3.0 channels all using single carrier 256-QAM all in a channel bonding group.

2.3 Top-split (900-1125) Plus the use of Sub-split

Overview

A new spectrum split called Top-split (900-1125) defines two separate spectrum bands, which may use sub-split plus the new spectrum region of 900-1125 MHz for a combined upstream band. The total upstream capacity may be 262 MHz depending on the lower band frequency return selected and if the passives will allow 1125 MHz to be reached. The downstream would begin at either 54 MHz or 105 MHz and terminate at 750 MHz in the current specification.

All of these architectures will share a 150 MHz guard band between 750-900 MHz, this may vary in the end-state proposal however these defined spectrum splits will be used for our analysis. The placement of additional upstream atop the downstream has been considered for many years.

The Top-split (900-1125) approach may be similar to a Time Warner Cable trial called the Full Service Network in the mid 1990's, which is believed to have placed the upstream above the 750 MHz downstream. These are some of the pros and cons of Top-split (900-1125):

Pros

- Operates at a typical 500 HHP node group but with no more than QPSK (see details in the sections later in this analysis)
- Top-split with Sub-split DOCSIS QAM MAC layer capacity ~315 Mbps given a 500 HHP Node/Service Group
- Top-split with Mid-split DOCSIS QAM MAC layer capacity ~582 Mbps given a 500 HHP Node/Service Group (less than High-split)
- Top-split 900-1125 does operate at a 500 HHP node but may operate at not full spectrum and will only be able to utilize 24 channels at 6.4 widths.
- Top-split (900-1125) plus Sub-split using DOCSIS QAM has an estimated MAC layer capacity of ~932 Mbps given a 16 HHP Node/Service Group
- With Sub-split "no" video services, devices, and capacity is touched
- STB OOB Communications are not affected
- Estimated that most passives will not be untouched (only Top-split that avoids touching passives)
- Existing 750 MHz forward transmitters are leveraged

Con

- The absolute major disadvantage for Top-split is cable network architecture requirements to make the solutions possible and the demands to reach high data capacity push FTTLA.
- A major finding of this report found that the effects of noise funneling force smaller and smaller node service groups to increase data capacity regardless if this is a DOCSIS / HFC solution or Ethernet over Coax (EoC) solution
- FTTLA is really fiber to All Actives, this will increase the number of node (HFC or EoC) to approximately 30 times the level they are now to reach the capacity level that High-split can reach with just the existing 500 HHP node location

- High-split can work at a 500 HHP node and while Top-splits must reach 16 HHP (FTTLA) depending on spectrum/cable architecture more HHP or even less than 16 HHP to reach the equivalent data capacity, lots of dependencies.
- Top-split from deployment perspective can be a challenge different cable type and distances play a major role is the architectures performance even if FTTLA is deployed
- No products in the market place to determine performance or accurate cost impacts
- 16 HHP upstream Service Groups will be required to approach 1 Gbps speeds comparable to High-split (238)
- Spectrum Efficiency is a concern because of guard band (wasted spectrum) and lower order modulation (less bits per Hz) resulting in lower throughput when measured by summing the upstream and downstream of Top-split (900-1125) and High-split using similar spectral range.
- High-split has nearly 20% more capacity for revenue generation when compared to Top-split (900-1050) plus Mid-split at a 500 HHP node, this is because the guard band requirements waste bandwidth and low order modulation for Top-split
- Upstream is more of a challenge compared to using that same spectrum on the forward path
- Upstream is more of a challenge compared to using that same spectrum on the forward path (cable loss ~5x Sub-split, 2.3x High-split; ~+/-1 dB/amp level delta w/EQTs is unknown)
- Interference concerns with MoCA (simply unknown scale of impact but may affect downstream in same spectrum range)

Assessment

The major consequence of the Top-split approaches, which use frequencies that approach or exceed 1 GHz, will have significant network cost impacts when compared with High-split. The number of nodes will increase 30 times to yield same capacity of High-split.

However, the Top-split (900-1125) options are being considered because option keeps the video network "as is" when considering sub-split and has marginal impact if mid-split is used. The Top-split 900-11125 option has additional benefits in that the Set-top box out of band (OOB) challenge is avoided and this option does not touch the passives.

This Top-split is estimated to cost more than the High-split. However, not included in this analysis is an economic forecast of the cost for Top-split to reach 1 Gbps upstream capacity which is estimated to be a 16 HHP architecture, the analysis examined economics 500 HHP and 125 HHP node architecture.

The migration for FTTLA to achieve 1 Gbps, would be 16 HHP and require all amplifier locations, thirty (30) in our model, to be a node location and this will require unground and aerial fiber builds to all locations. The MSOs will just begin to evaluate this option against the others.

2.4 Top-split (1250-1550) with Sub-split Overview and Top-split (2000-3000)

Systems designed to leverage unused coaxial bandwidth above 1 GHz have been around for many years. New iterations of these approaches could be considered to activate currently unoccupied spectrum for adding upstream.

The primary advantages of the top split are operational considerations – leaving current service alone – and the potential of 1 Gbps capacity or peak service rates in unused spectrum. In theory, not interrupting legacy services makes an IP transition path non-intrusive to customers, although the plant implications likely challenge that assertion.

The Top-split (1250-1700) Architecture will be defined as part of the 1250 – 1750 MHz spectrum band. Top-split (2000-3000) In our analysis we limited the amount of spectrum allocated for data usage and transport to 450 MHz and defined the placement in the 1250–1700 MHz spectrum band.

The allocation of 450 MHz provides similar capacity when compared to the other split option. The main consideration for this Top-split option is that it avoids consuming existing downstream spectrum for upstream and avoids the OOB STB communication channel

2.4.1 Implementation Complexity

A key additional complexity to the top split is working the spectrum around or through existing plant actives, all of which are low-split diplex architectures. For top split, a new set of actives supporting a triplex, or a bypass approach, or an N+0/FTLA are necessary to make the architecture functional.

All of these are intrusive, and have heavy investment implications, with the latter at least consistent with business-as-usual HFC migration planning. The top split is best suited to N+0 due to the complexity of dealing with current plant actives as well as for link budget considerations. N+0 at least removes the need to developing new amplifiers for the cable plant.

By contrast, node platforms have been and continue to evolve towards more features, functions, and flexibility. Of course, N+0 can be leveraged as a high-performance architecture whether or not a top split is implemented – top split, however, practically requires it to succeed as an architecture.

The outside plant architecture is not the only architecture affected by the approach. With the emphasis on upstream loss and degraded SNR as a primary issue for top split, a top split also virtually demands a point-of-entry (POE) Home Gateway architecture.

The variability of in-home losses in today’s cable systems would seriously compound the problem if a top split CPE was required to drive through an unpredictable combination of splitters and amplifiers within a home.

The above issues apply to the case of Top-Split (900-1125) as well, but to a lesser degree with respect to RF attenuation and the inherent bandwidth capabilities of today’s passives.

2.4.2 Spectrum Inefficiency

The penalty of the triplex architecture in terms of RF bandwidth and capacity can be substantial. A triplex used to separate current downstream from new top split bandwidth removes 100-200 MHz of prime

CATV spectrum from use in order that a less capable band can be enabled.

This spectrum trade reduces the total aggregate capacity of the plant. Under the assumption used (MPEG-4 HD/IPV), approximately 90 channels of 1080i HD programming are lost to guard band loss in a top split implementation compared to a high split alternative.

A primary objective of an HFC migration plan is to optimize the available spectrum, extending the lifespan of the network in the face of traffic growth for as long as possible, perhaps even a "forever" end state for all practical purposes that is competitive with fiber. RF spectrum in the prime part of the forward band is the highest capacity spectrum in the cable architecture.

To architect a system that removes on the order of 100 MHz from use is a loss of significant capacity, as quantified above, and works against the objective of optimizing the long-term spectrum efficiency.

The above issues apply to the case of Top-Split (900-1125) as well, but to a somewhat lesser degree associated with the percentage of crossover bandwidth required – that number is slight lower when the top split band chosen is slightly lower.

Pros

- Top-split 1250-1700 with Sub-split DOCSIS QAM MAC layer capacity ~516 Mbps given a 125 HHP Node/Service Group
- Top-split 1250-1700 with Mid-split DOCSIS QAM MAC layer capacity ~720 Mbps given a 125 HHP Node/Service Group
- Top-split (1250-1700) plus Sub-split using DOCSIS QAM has an estimated MAC layer capacity of ~883 Mbps given a 16 HHP Node/Service Group
- Top-split (1250-1700) plus Sub-split using DOCSIS QAM has 716 Mbps MAC layer capacity of ~1.08 Gbps given a 16 HHP Node/Service Group
- With Sub-split "no" video services, devices, and capacity is touched
- STB OOB Communication is not affected
- Placing the upstream spectrum beginning at 1250 MHz and up allows for the expansion of capacity without impacting the downstream

Cons

- Much higher upstream loss = significantly more CPE power = lower modulation efficiency (less bps/Hz) for equivalent physical architecture
- Need to work around legacy plant devices incapable of processing signals in this band
- Altogether new CPE RF type
- New technology development and deployment risk
- Large lost capacity associated with triplexed frequency bands
- Bottlenecks downstream growth when used as an upstream-only architecture
- Let's elaborate on some of the key disadvantages identified above for an upstream top split
- Will operate at a typical 500 HHP node group but only capable of three of the

- 16 HHP Node and Use Mid-split and Sub-split spectrum meet the 1 Gbps capacity
- Highest cost solution compared with High-split and Top-Split (900-1050)
- The Top-split (1250-1700) with Sub-split cost more than High-split (200) and requires FTTLA
- No products in the market place to determine performance or accurate cost impacts.
- Return Path Gain Level Control: (cable loss >6x Sub-split, 2.8x High-split; +/-2 dB/amp w/EQTs is unknown)
- Interference concerns with MoCA (simply unknown scale of impact but may affect downstream in same spectrum range)

Assessment

The Top-split (1250-1550) with Sub-split is far more costly of High-split for the same capacity. The placement of the return above 1 GHz requires the passives to be replaced or upgraded with a faceplate change. There are approximately 180-220 passives per 500 HHP node service group.

A 500 HHP will not support Top-split 1250-1550, so the initial architecture will have to be a 125 HHP. However the requirements for higher capacity will force smaller node service group, which will add to the cost of the solution. The use of lower order modulations will require more CMTS upstream ports and more spectrum, which will impact the costs of the solution as well.

Additionally, the conditioning of the RF components to support above 1 GHz may add to the costs of the solution. However determining the financial impacts of performing "Above 1 GHz plant conditioning" is unknown and was not considered in the financial assessment found later in this report.

The economic estimate used for Top-split was for 500 HHP and 125 HHP node architecture. The migration for FTTLA to achieve 1 Gbps, would be 16 HHP and require all amplifier locations, thirty (30) in our model, to be a node location and this will require unground and aerial fiber builds to all locations. This was not provided in the analysis.

Lastly, there is a significant penalty to downstream bandwidth in the form of triplex guard band – on the order of 100 MHz of RF spectrum is made unavailable for use. In the case of Top Split (900-1125), the band eliminated consists entirely of prime, very high quality forward path spectrum.

If we consider the service and network capacity requirements for the upstream and downstream for the next decade and beyond, the cable industry should have sufficient capacity under 1 GHz, which is the capacity of their existing network.

2.5 Summaries for Cable Spectrum Band Plan

Continuing to leverage the current downstream and upstream spectrum will force operators to reduce service group size by using node splits and/or segmentation. This is ideal for MSOs that want to avoid re-spacing the amplifier network.

Additionally, spectrum changes will undoubtedly require service outages, because all the electronics and even passives (if above 1 GHz is selected) would have to be touched. Spectral changes may have higher service down time compared with node segmentation or node splits.

MSOs may want to consider spectrum expansion where node splits are costly. Depending on spectrum selection, the MSO could maintain large service group in the optical domain. In others words, the optical node could service a larger area and number of customers, if the MSO selects low frequency returns such as Sub-split, Mid-split, or High-split and if additional downstream spectrum is selected this will increase the length of time a optical node can support a given service group.

The channel allocation of video and data services will define the spectrum needs and node migration timing. Additionally, the service offering, such as network based PVR, will impact the spectral usages; thus drives toward more spectrum or smaller services groups.

There really are lots of levers that will drive the MSOs to changing spectrum and/or service group reductions, predicting with all certainty of how long a given network will last is greatly influenced by services and legacy devices that may need to be supported.

The legacy STB out of band (OOB) communications which uses spectrum in the High-split area will be a problem for this split options; however a mid-split as the first step will provide sufficient capacity for nearly the entire decade according to our service and capacity predictions. The thinking is that another decade goes by and the legacy STBs may be few or out of the network all together.

If the STBs still remain in service, another consideration is that these legacy STB may be retrieved and relocated to markets that may not need the advanced upstream spectrum options. Yet, another consideration is a down conversion of the OOB communications channel at the last amp or homes that have legacy two-way non-DOCSIS set-tops.

2.6 Spectrum Options, Capacity, and Timing Implications

We have discussed the Pros and Cons of the various upstream spectrum options. As discussed in Section 2.1, it is well-understood that a limitation of the 85 MHz mid-split architecture is that it cannot achieve 1 Gbps of capacity, at least not easily or in the near term. We will discuss upstream capacity itself in detail in Section 9.6 "Upstream Capacity".

While 85 MHz cannot achieve 1 Gbps of capacity, it is also not reasonable to jump to high-split in the near term because a plan must be in place to deal with the OOB channel, as shall be further described in Section 3.3.5 "Legacy OOB" and Section 3.4 "The Legacy Mediation Adapter (LMA)". As such, MSOs appear to be in a bind for handling upstream growth. Or, are they?

Let’s consider defining the 1 Gbps requirement for upstream data capacity. How would such a system fare in supporting long-term capacity requirements? We can easily quantify how this would help manage long-term traffic growth and compare it to examples like the 85 MHz Mid-Split.

This comparison is examined in Figure 2. It shows three threshold cases – 100 Mbps (A-TDMA only), 85 MHz Mid-Split (in this case, including use of S-CDMA), and the case of 1 Gbps of capacity, however we manage to achieve it (high-split or top-split).

Zeroing in on the red arrow identifying the gap between Mid-Split and 1 Gbps at 40% CAGR – very aggressive relative to 2011 observed growth rates – in each case with a node split assumed in the intervening years, we see that there exists about 2.5 years of additional growth. When we think of 1 Gbps, this intuitively seems odd. Why does migrating to Mid-Split buy a decade or more of traffic growth coverage, yet implementing a 1 Gbps system offers only a couple more years of survival on top of that decade?

This "linear" time scale on the y-axis is simply exemplifying how multiplicative compounding works. It is up to our own judgment and historical experiences to consider how valid it is to be guided by the compounding rules of CAGR originally identified by Nielsen, and if so what reasonable year-on-year (YOY) behavior assumption to assume.

However, the mathematical facts of CAGR-based analysis are quite straightforward: with CAGR behavior, it takes many YOY periods to grow from, for example, 5 Mbps services today, consuming or engineered for perhaps tens of Mbps of average return capacity, up nearly 400 Mbps or more. We will outline the data capacity possibilities for 85 MHz Mid-Split in Section 9.6, and then show a specific implementation in Section 7.1.2. However, once a 400 Mbps pipe has been filled, the subsequent annual steps sizes are now large. Because of this, consuming 1 Gbps is not many YOY periods of growth afterwards.

To demonstrate, we can calculate an example using 20 Mbps of average capacity satisfying demand today. At this aggregate demand, traffic can double four times and not eclipse 400 Mbps. It eclipses it in the 5th traffic doubling period. For ~40% CAGR (two years doubling), that's a total of ten years. For a CAGR of 25%, its about 15 years.

This is what Figure 2 is pointing out graphically. As such, relative to a solution that provides 1 Gbps, Mid-Split gets us through 80% of that lifespan under the assumption of an aggressive 40% CAGR and an intervening node split.

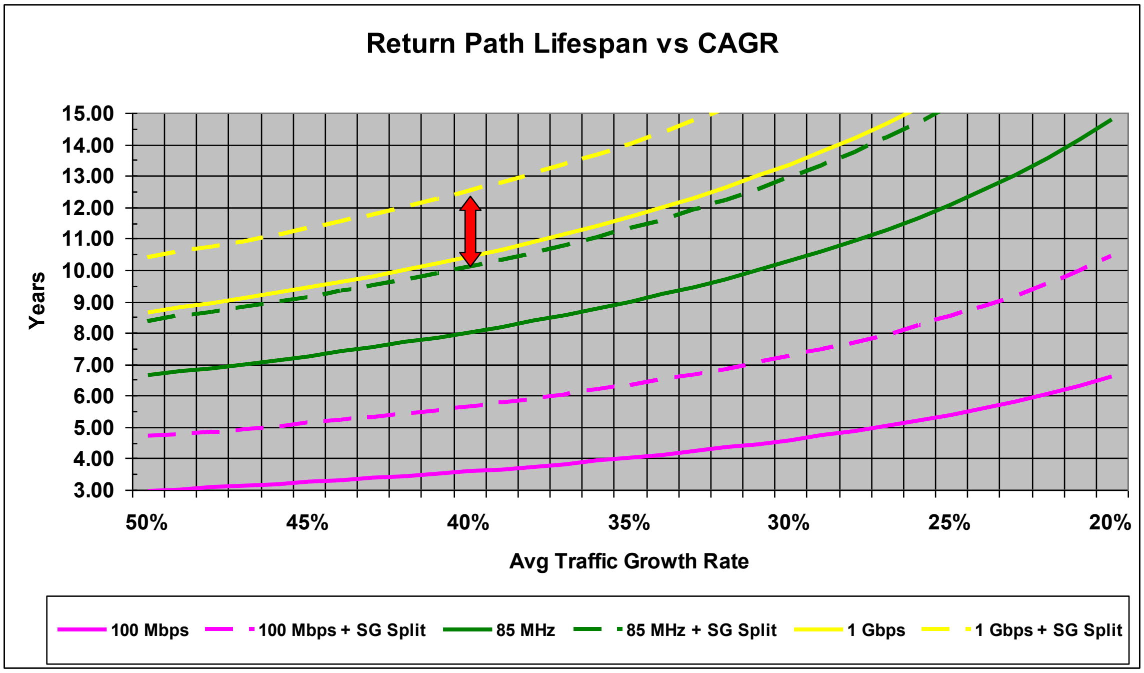


Figure 2 – Years of Growth: A-TDMA Only, 85 MHz Mid-Split, 200 MHz High Split

This Mid-Split vs. 1 Gbps lifespan analysis is an illustrative one in recognizing the long-term power of the 85 MHz Mid-Split. It provides nearly the same growth protection as a 1 Gbps solution would, if there even was one available. This means that the 1 Gbps requirement comes down to an operator's own considerations regarding the competitive environment, and whether a 1 Gbps market presence or service rate is important to their positioning for residential services.

3 SOLVING LEGACY ISSUES

3.1 Introduction

In order to significantly increase the upstream throughput in a DOCSIS system, more upstream spectrum is needed. That spectrum has to go somewhere. This white paper has examined multiple spectrum solutions and then different technology options within each spectrum solution.

Solutions are needed that allow an HFC plant to be migrated over to the next generation of DOCSIS without a full-scale replacement of subscriber equipment. Legacy and new equipment must co-exist in the same network.

The high level summary of the different spectrum solutions and their challenges is shown in Table 2.

This paper recommends mid-split and high-split as the best technical solutions. The attractiveness of top-split is that it interferes less with existing services. If the logistical problems of mid-split and high-split could be solved, then cable operators would be able to choose the best technical solution.

This section is going to specifically look at addressing the major logistical problems that the mid-split and high-split band plans face.

3.2 Summary of Operational Issues

Table 3 is a summary of the operational issue faced by each of the four upstream bandwidth solutions. This table is taken from [21].

There are several logistical challenges that are obstacles to the deployment of mid-split and high-split systems into an HFC plant that was designed for sub-split. The challenges include:

- Analog video
- FM band
- Aeronautical band interference
- Adjacent device interference
- Legacy OOB

Let's look at each one of these challenges in more detail.

Table 2 – Upstream Spectrum Comparison

Approach	Frequency	Comments
Sub-Split	5 - 42 MHz	Existing installed HFC plant. Add bandwidth with node splits.
Mid-Split	5 - 85 MHz	Technology available today with DOCSIS 3.0 CMTS and CM.
High-Split	5 - 200+ MHz	Best technical solution but challenging logistical solution
Top-Split	> 1 GHz	Tough technical solution but more attractive logistical solution

3.3 Analysis and Solutions

3.3.1 Analog Video

Problem Definition

There are many different channel plans in use around the world today. This white paper will choose the North American cable television plan as a specific example. This channel plan is defined in [20] and described in [18]. The upstream frequency cut-off is a maximum of 42 MHz. Some systems use a lower cutoff, depending upon the age of the system.

The downstream frequency range starts at 54 MHz. By convention, the analog

Table 3 – Summary of Operational Issues

Approach	Pros	Cons
Sub-Split	• All equipment already exists • No disturbance to spectrum • Simple	• Cost: Requires deeper fiber. • Cost: Requires more CMTS ports • Cannot hit peak rates over 100 Mbps of return path throughput
Mid-Split	• Supported by DOCSIS 3.0 equipment • Works with DS OOB	• All actives and some passives in HFC plant need to be upgraded • Cost about the same as high-split and only doubles the US throughput • Removes ch 2-6 of analog TV
High-Split	• Supports 1 Gbps throughput • Can co-exist with earlier versions of DOCSIS.	• All actives and some passives in HFC plant need to be upgraded • Does not work with DS OOB • New CM and CMTS components • Removes ch 2-36 analog TV • Removes FM band (issue in Europe)
Top-Split	• Leaves existing plant in place. • No impact to existing legacy customer CPE • Only customer taking new tiers would require new HGW CPE	• Requires triplexers • New active return path has to be built on top • High attenuation requires high RF power. Existing amplifier spacing may not be sufficient • Blocks expansion of downstream bandwidth directly above 1 GHz

channels are first in the spectrum followed later in frequency by the digital channels. The classic analog line-up is contained in channels 2 through 78 that occupy the spectrum from 54 MHz to 550 MHz. Within this spectrum are also channels 1 and 95 to 99.

The definition of the frequencies for a mid-split system has changed over the years. The mid-split for DOCSIS 3.0 is not exactly the same as legacy systems that used a return path upper frequency limit of 108 MHz ~ 116 MHz, with the downstream spectrum starting at 162 MHz~ 174 MHz (the actual frequencies varied among vendors).

The DOCSIS mid-split downstream frequency range starts at 108 MHz, which disrupts channels 1, 2-6 (54 MHz-88 MHz), and 95-97 (90 MHz-108 MHz) would be disrupted. A natural break point from a channel perspective would be to start the mid-split lineup at channel 14 a(120 MHz-126 MHz). If so, then channels 98-99 (108 MHz-120 MHz) would also be disrupted. Note that channels 7 through 13 (174 MHz-216 MHz) are located above channels 14 through 22 (120 MHz-174 MHz).

The upstream frequency range for high-split has not been chosen yet. If the high-split downstream frequency spectrum started at 300 MHz, then channels 1-36 and 95-99 would be lost.

Solutions

The first solution is to get rid of analog TV altogether on the cable spectrum. Any legacy TV that cannot receive direct digital QAM would have to be serviced with a digital transport adapter (DTA) or a conventional set-top box (STB). As radical as this idea may seem, several cable operators such as Comcast and CableVision are already free of analog channels on parts of their plants with plans to expand their no-analog foot print. The governments of many countries, including the USA, have already turned off most over the air analog broadcasts.

It costs money to retain analog channels. It is not that the money is spent on the analog channel equipment - which obviously is already paid for - it is that money needs to spent elsewhere to improve spectral efficiency. This may include plant upgrades, equipment upgrades or both.

Analog TV has only 5% of the efficiency of an MPEG-4 over IP video signal, yet analog TV typically occupies over 50% of the downstream spectrum. RF spectrum is always a scarce commodity, and this is a good example of where there can be a significant efficiency improvement.

The second solution would be to reduce the analog channels down to a smaller group of, say, 25 core channels. Then remap those analog channels into a higher channel space. For mid-split, only channels 2-6 need to be remapped. For high-split, it would be channels 2-36.

This may cause some channel confusion to the subscriber, but such a remapping trick has been done for high definition channels on STBs.

A semblance of continuity can be maintained by keeping the least significant digit the same. Remapping channel 2 to channel 62 is one example.

There are often contractual issues quoted, such as franchise agreements, market recognition, must-carry agreements, etc. These may have to be renegotiated. The driving force for doing so is a gigabit or more upstream speed. To the extent that these legal requirements are driven by the requirements of the community, then which is the bigger market - analog TV or an incredibly fast Internet access? The answer has to be a fast Internet service or there would not be a need to upgrade in the first place.

Finally, now that the government has shut down most over-the-air analog TV, the cable operators are the last service provider to have analog TV. The telco and satellite service providers are all digital.

There are two perspectives that can be taken on this. The first is that having analog TV makes the cable operators unique in being able to offer analog TV, and this differentiates them from all the other providers. The second is that the cable operators are the last to move to all digital, and that the other service providers may have more spectrum or resources as a result.

So, again, if the costs are equal, does analog TV with a lower Internet access speed beat out a competitor who has a significantly higher speed Internet service? What if the competitor is a fiber-to-the-home company with gigabit-per-second service?

The choice is somewhat obvious, but also very painful. It requires pain of some sort. But, the new upstream spectrum has to come from somewhere. Keeping analog TV spectrum indirectly costs money due to investment on alternative solutions.

3.3.2 FM Band

Problem Definition

The FM radio band is from 88 MHz to 108 MHz. There are two potential concerns.

The first concern is the loss of the ability for the cable operator to provide FM radio service over the cable system. This is not much of an issue in North America, but it is a concern in Europe and elsewhere.

The second concern is if interference generated by the HFC plant that might interfere with the FM band (signal leakage) or if the FM band might interfere with the with the HFC plant (ingress).

Solutions

As with analog TV, the easiest solution to the first requirement is to no longer carry the content. For Europe, this may require some regulatory work. The worst case would be to carry the FM band at a higher frequency on the HFC plant and down-convert it locally with the LMA. Refer to Section 3.4.

As far the HFC plant interfering with local FM reception, this should not be a problem. The capture effect of FM receivers [24] will most likely reject noise-like digital signals leaking from a cable network as a weaker signal. A strong FM signal might interfere with the upstream signal on the HFC plant. This can be mitigated with good plant shielding, ingress cancellation techniques, or OFDM noise/ingress mediation.

3.3.3 Aeronautical Interference

Problem Definition

The new CM will be transmitting at frequencies above 54 MHz at a higher power level than when the frequencies are transmitted as part of the downstream spectrum. The inherent leakage in the plant might be sufficient enough to cause interference with existing services.

For example, the frequencies from 108 MHz to 137 MHz are used for Aeronautical Mobile and Aeronautical Radio Navigation.

The radio frequency spectrum usage is shown in Figure 3. [23]

Specifically, the 108-118 MHz band has always been problematic because any CATV signal leakage here could interfere with aviation localizer (108-110 MHz) and VOR signals (110-118 MHz). Hence, sometimes channels 98 and 99 (also called A-2 and A-1) are not used to avoid this problem. The localizer is especially important, as it is responsible for providing the left/right guidance in an ILS approach; VORs are also important but more often used at longer ranges as navigation beacons.

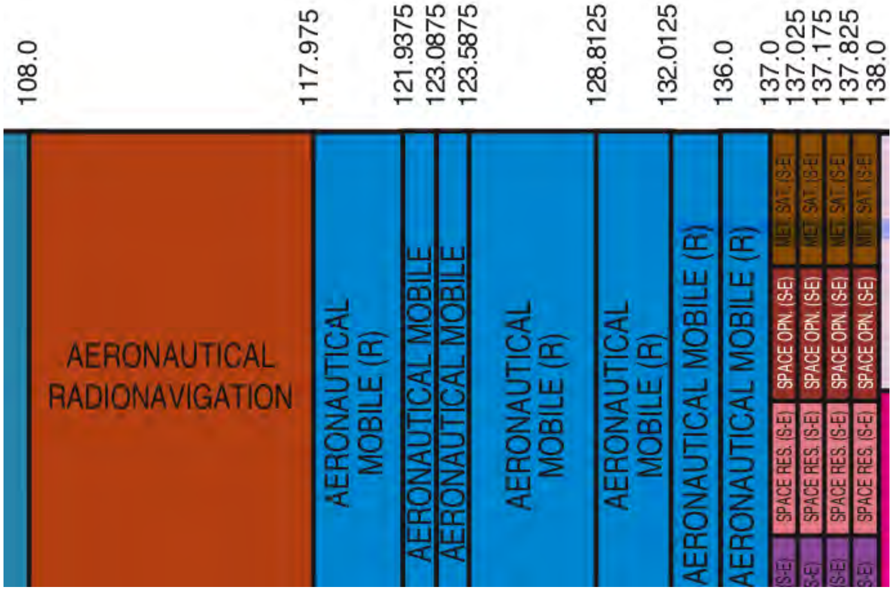


Figure 3 – Government Spectrum Allocation from 108 MHz to 138 MHz

There is also the 121.5 MHz aeronautical emergency frequency, and the 243.0 MHz distress (SAR) that may be of concern.

If the upstream spectrum expands above 300 MHz, another sensitive aviation band comes into play. The glideslope frequencies are in the 328-335 MHz band. The glideslope is the x-y counterpart to the localizer as it provides up/down guidance in an ILS approach.

Solutions

Research would have to be done to validate these concerns. If it is a problem, then the plant will have to be cleaned up to reduce this leakage. Some of this leakage may come from bad home wiring. That makes it even more important that the CM installation is done professionally.

In the absolute worst case, some or all of these frequencies would have to be avoided. The impact of that is that a larger upstream spectrum would have to be dedicated to DOCSIS. This would be a loss of up to 29 MHz or more in some networks.

Some of these interfering carriers are quite narrow. Current DOCSIS tools handles very narrow interferers better than modulated, but increasingly struggles as multiple interferers occupy a single carrier band. OFDM will be quite useful for notching these out.

This concern also existed 15 years ago prior to the deployment of DOCSIS. The plant did require cleaning up in many cases. It was done and the result was a more reliable HFC plant. So, it is doable, but must be planned and budgeted for.

3.3.4 Adjacent Device Interference (ADI)

Problem Definition

ADI refers to the situation where the operation of one device - such as a high-split cable modem - interferes with another device - such as a legacy TV or legacy set-top box. This is not an official abbreviation (yet). We are borrowing the concept from the term adjacent channel interference that describes a similar phenomenon, except ACI is in the frequency domain, and ADI is in domain of physical space.

For the sake of example, let's assume the high-split spectrum goes up to 230 MHz, and the downstream starts at 300 MHz.

Tuners in STBs and TVs in North American receive above 54 MHz with an expected maximum per-channel input power of +17 dBmV. Low-split and top-split can thus co-exist fine with legacy tuners. Mid-split and high-split systems carry RF energy in the upstream direction that is within the downstream operating range of the legacy STB and TVs.

If those devices are located near a CM that is blasting out energy above 54 MHz at levels approaching +57 dBmV (DOCSIS 3.0 max power for single 64-QAM), poor isolation and/or return loss in splitters and other devices could cause a significant amount of that upstream power to appear at the input connector of the legacy devices, which might saturate their RF input circuits, thus preventing the devices from receiving a signal at any frequency.

The typical North American legacy tuner has an output intermediate frequency (IF) centered at 44 MHz. If 44 MHz was applied to the input of a tuner with poor IF rejection, that signal might cause interference in the tuner, even through the tuner is tuned to another band. How much of a problem this is requires more research.

There is some evidence that shows that the sensitivity of the video signal to ADI decreases significantly as analog signals are replaced with digital. This is a somewhat intuitive conclusion, but validating data to this effect is important.

Solutions

So, what to do?

One solution is to put a filter in front of the legacy devices that filters out all content below the high-split cut-off frequency (85 MHz or 230 MHz in this example). But, is this filter needed in all cases? And where would the filter go? Let's look at this problem in more detail.

The general problem is best split up into two smaller scenarios:

- Impact within the same home as the new high-split DOCSIS CM.
- Impact to adjacent homes that do not have the new high-split DOCSIS CM

Same Home:

When a home is upgraded, the new DOCSIS CM will likely be installed as a home gateway (HGW). There are at least two scenarios. The first is a home with MPEG video STBs, and the second scenario is an all IP video home.

In the home that requires digital MPEG video, the HGW can receive the spectrum from the plant, filter the signal below 200 MHz, and pass the filtered spectrum into the home. The main filtering it is trying to achieve is from its own upstream transmitter. If the upstream transmitter is

+50 dBmV, the internal combiner has 20 dB of signal rejection, and the max signal level allowed is +15 dBmV, then the additional filtering has to provide 15 dB of attenuation. This filter could be located internal to the HGW or be an external inline filter in order to manage HGW costs.

For this to work, the HGW would have to be wired in-line with the home. That is not how CMs are installed today. CMs today are installed using a home run system. The drop cable from the street is split between the CM and the home. In this new configuration, the CM would have to have a return cable that then fed the home. This could add additional loss to the video path. However, it could be a workable solution.

In the home where there are only IP STBs, the downstream from the HFC plant does not have to be connected to the home. DOCSIS could be terminated at the HGW and the HGW would drive the coax in the house with MoCA. Video and data would be deployed with IP STBs that interfaced to the MoCA network.

The HGW becomes a demarcation point between DOCSIS and the cable plant on one side, and MoCA and the home network on the other side. Again, the CM would have to be in-line with the coax from the drop cable and the home. This does imply the need for a professional installation.

This is an interesting proposal in several ways. First, it solves the in home legacy tuner interference problem. Second, it isolates all the return path noise generated by the home network and prevents it from entering the HFC plant.

Adjacent Home

The other half of the problem is the impact to adjacent homes. While the installer has access to the home he is upgrading and has several options available to him, the home next door may not be part of the upgrade.

The energy from the new high-split CM would have to travel up the drop cable from the home, travel between the output ports on the tap plate, back down the drop cable to the next house, and then into the home network of the next house.

The easiest solution would be to set the new upstream power budget such that the signal would be sufficiently attenuated by the path described above so that it would not be a problem. This solution becomes harder when the customers are in a multiple-dwelling unit (MDU) where the coax drops may be shorter.

Worst case, in-line filters would have to be applied in-line with the drop cables of the adjacent home or within the adjacent home. Another approach is to put filters into the tap plate that serves an upgraded home and its adjacent homes. This would prevent the upgraded home from impacting the adjacent homes.

Thus, tap plates would only have to be replaced as part of a new deployment so the overall cost would be lower than having to replace them all at once. This assumes that the additional upstream path attenuation between taps on separate enclosures is sufficient.

As far as potential tuner sensitivity, the upstream spectrum could skip the frequencies from 41 to 47 MHz. This can be done, but it is a loss of 6 MHz of spectrum. The better plan is to make sure that the attenuation of the upstream signals into the downstream is sufficient that even 41 to 47 MHz is fine.

Summary

In summary, an external filter may not be needed. The HGW can be used to protect the upgraded home, although it has to be wired in line. The adjacent home should have enough attenuation from the drop cables and tap assembly. More caution may be needed in MDUs. An external in-line filter should be made available to fix the exception condition. Filtered taps may be good for dense situations such as MDUs.

3.3.5 Legacy OOB

Problem Definition

The out-of-band (OOB) channel is used on legacy STB to provide information to the STB and get information back. The OOB channel was used prior to the development of DOCSIS Set-top Gateway (DSG).

The downstream carrier is 1 MHz wide for SCTE 55-2 (Cisco) and approx 1.7 MHz wide for SCTE 55-1 (Motorola). Typical placement of center frequency is between 73.25 and 75.25 MHz as there is a gap between channels 4 and 5. The older “Jerrold” pilot (prior to Motorola/GI) was at 114 MHz. By spec [25], the STB must be able to tune up between 70 MHz and 130 MHz.

There is an upstream OOB carrier as well that is usually placed below 20 MHz.

CableCards are one-way and typically use only a downstream OOB channel.

There are no compatibility issues with the STB OOB channel and low-split or top-split. For mid-split, if the OOB channel can be placed above 108 MHz in the downstream spectrum then the problem is solved. This should work except for very old STBs that are fixed frequency. These STBs would have to be upgraded.

For high-split, this is probably the biggest issue. The 200+ MHz target cutoff for high-split is well above the 130 MHz upper end of the OOB tuner range.

Solutions

This is primarily a North American issue. In the rest of the world where legacy STB penetration with OOB is much lower or non-existent, and may not be a significant issue.

Of the STBs deployed in North America, many of the newer ones can actually tune to a frequency greater than 130 MHz because it was just as cheap to use a full spectrum tuner. Cisco estimates that > 70% of the Tier 1 installed base of Cisco STBs in 2015 would have this capability. (Further research is required. Software upgrades may be required.).

Then there is DSG. DSG is basically OOB over DOCSIS. Many of the deployed STBs have DSG built in but the DSG has not been enabled. Cablevision is an exception who has 100% DSG deployed, as does South Korea. So, DSG is proven to work.

It turns out there was a financial hitch with DSG. The original plan was add the STBs to an existing DOCSIS upstream channel. These upstream channels are engineered to be transmitted from the CMs on a home run cable. The STBs in the home have more attenuation, as they are deeper into the home coax network, so they are not always able to transmit onto an existing DOCSIS channel.

The solution is to use a separate QPSK DOCSIS channel. If this channel were the same modulation and power level as the existing OOB channel - which it would be - then if the OOB upstream worked, the

DOCSIS OOB upstream would also work. The problem is that this requires a dedicated carrier in the CMTS. This might be additional expense or the CMTS may not have the extra capacity. With newer CMTSs, there will be more upstream carriers available, so dedicating one carrier per port to DSG is a very reasonable solution.

It is also reasonable that any home that gets upgraded to a new high-split CM could also have their STBs upgraded to DSG compatible STB.

The OOB CableCard is easily replaceable and can migrate to DSG.

So that leaves STBs in North America, in non-upgraded homes, that are over 10 years old (by 2015), that can't tune above 130 MHz, that are non-DSG, and are not CableCards. That is really not a lot of STB. It could be around 0% to 10% of the STB population rather than the originally estimated 100%.

There is a motivation to replace these old STBs. They are beyond their capital write-down period. Further, these STB usually do not have the CPU or memory capacity required to run new applications. This means that new services cannot be sold to these customers.

Just to be on the safe side, there is a solution that does not require upgrading the old STB. That solution would be to put an inexpensive LMA behind legacy STB that provided an OOB channel. These LMAs would go inline with legacy STB. They would be cheap enough that they could be mailed out to customers who complain or are known to have specific legacy STBs.

If that does not work, only then a truck roll might be needed.

Summary

At first pass, the loss of the OOB channel seems like a major problem. However, by the time the next generation of DOCSIS is deployed, and with the variety of solutions, it is not really a problem at all.

Bear in mind that before the first high-split CM can be used in the new spectrum, the plant needs to be upgraded. But after the plant is upgraded, homes can be upgraded on a per home basis. This helps keep costs contained. Also, in a phased approach to upstream bandwidth expansion, a mid-split architecture may buy yet more time to eliminate or actively retire the older STBs.

This is a far better proposition than if all legacy STBs had to be replaced prior to upgrading the plant.

3.4 The Legacy Mediation Adapter (LMA)

In several of the plans to deal with legacy, there is a back-up plan that involves an in-line device that we will refer to as a legacy mediation adapter (LMA).

- The LMA could be used for generating and receiving an OOB signals.
- The LMA could be used for blocking upstream energy from entering the downstream.
- The LMA could be used to isolate the ingress originating from the home when the home no longer needs a return path internal to the home.
- The LMA could even be used to generate an FM signal for European deployments.

There are at least two primary ways of designing this LMA. The first way uses a

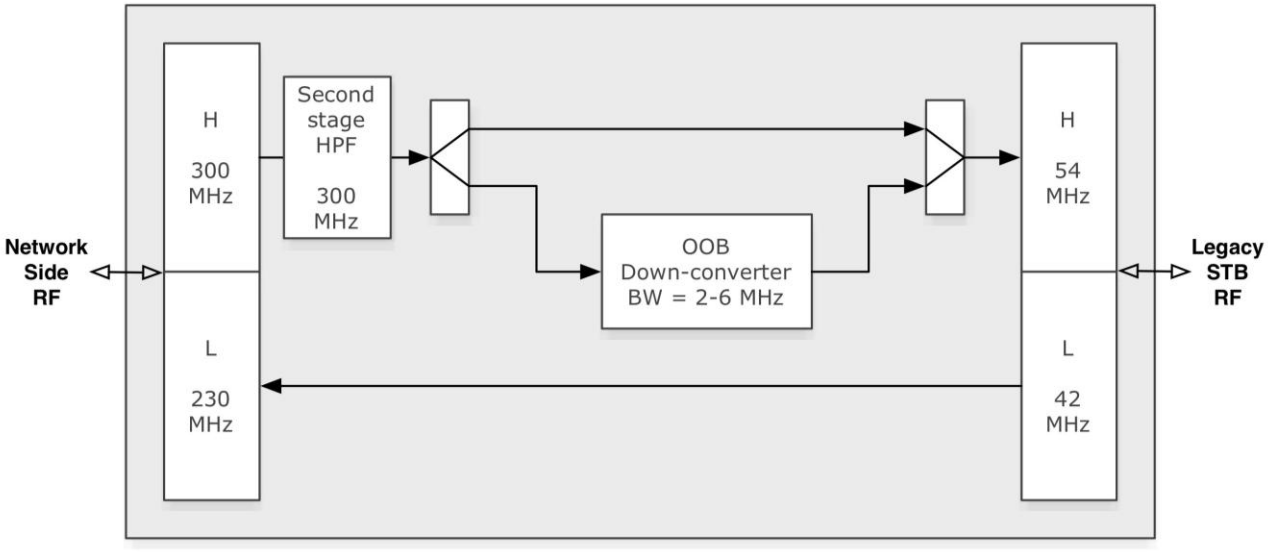


Figure 4 – LMA with Down-Conversion

simple down-conversion method. The second way uses an embedded circuit.

Another interest aspect of the LMA is that it interfaces between the new and old HFC spectrum plans. On the network side of the LMA, it interfaces into the high-split, 200 MHz (for example) plant. On the subscriber side of the LMA, it interfaces into the legacy sub-split 42 MHz plant.

3.4.1 LMA with Down-Conversion

In this approach, the headend would generate two OOB downstream carriers. The first one would be the standard downstream OOB carrier. This first carrier might be at 75 MHz for example.

The headend then generates a second OOB carrier at a frequency that is in the available downstream spectrum that is above the upstream cut-off frequency. This second carrier might be at 750 MHz for example.

This second carrier would fit into a 6 MHz or 8 MHz TV channel slot. This channel would be wide enough that multiple carriers could be fit. That way, any plants that are dual-carry with two STB manufacturers on it could be accommodated.

If necessary, the bandwidth could be expanded to allow for the FM band to be placed at a higher frequency as well.

The first carrier at the lower frequency would be received by legacy STB on areas of the plant that have not been upgraded. The second carrier would be received by the LMA that has been placed behind the legacy equipment.

The use of two carriers at different frequencies presumes a scenario where the LMAs are distributed over a period of time prior to the HFC plant upgrade. Thus, during the transition period, there would be legacy devices on both carriers.

A block diagram of the down-converting LMA is shown in Figure 4. Starting at the network side, the RF signal is separated with a diplexer into downstream and upstream frequency paths. The

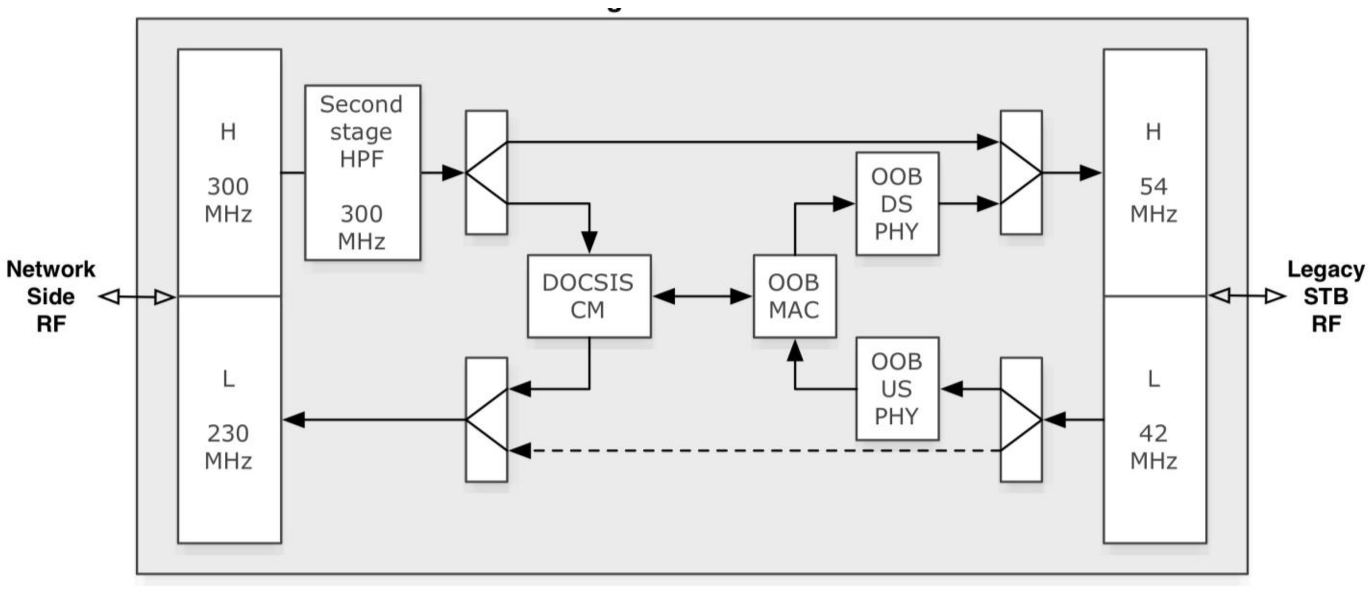


Figure 5 – LMA with CM

downstream path may require further filtering to remove any upstream energy.

The higher frequency OOB carrier is tapped off and passed to a down-converter. In the example used here, the down-converter would down convert from 750 MHz to 75 MHz. This carrier is then combined back into the downstream spectrum and then passed to the legacy STB.

To further reduce the cost of the LMA, the upper frequency that is used for the OOB carrier could be standardized through CableLabs. The LMA would then be a fixed frequency device and would not require any configuration.

The return path is left intact as the legacy STB will need to send an OOB carrier back to the headend.

3.4.2 LMA with DOCSIS CM

This approach achieves similar goals but with a different method. In this method, a DOCSIS CM is used to communicate the OOB information over IP from the headend to a local OOB circuit. This design would be good for operators who are using DSG as a baseline to control their network or for a scenario where the LMA needs to be configured.

DSG can be used on the network side in the downstream. Alternatively, a basic IP tunnel can be used to transport the raw OOB channel. An IP tunnel will have to be defined for the upstream that carries the upstream OOB information to the headend. This can be done at CableLabs.

The LMA has an entire two-way OOB MAC and PHY. This circuit generates a local OOB circuit. A clever implementation could implement both the SCTE 55-1 and SCTE 55-2 OOB standards. Otherwise, there would need to be two separate LMAs.

This design could use a DOCSIS 1.1 CM as part of a reduced cost implementation as only single carrier implementations are needed.

The return path from the home to the network could be disabled so that the LMA would isolate the ingress from the home from getting to the network.

3.5 Downstream Concerns

The downstream frequency band above 1 GHz will have a few challenges as well. In addition to the higher attenuation and micro-reflections, there are some frequency bands to be careful of. Here are two of the more common spectrum usages to be aware of.

3.5.1 MoCA®

MoCA is a technology that allows peer to peer communication across coax in a home environment. It typically is used for communicating between set-top boxes.

The concern would be that new frequencies on the cable plant above 1 GHz could interfere with MOCA in homes that are both upgraded to DOCSIS NG that don't isolate the HFC plant from the home and homes that are legacy.

MoCA 1.1 defines a 100 Mbps data channel that consumes 50 MHz of spectrum that can be located anywhere in between 1125 MHz and 1525 MHz.

MoCA 2.0 defines a 500 Mbps data channel that consumes 100 MHz of spectrum that can be located anywhere in between 500 MHz and 1650 MHz. MOCA 2.0 also has a special 1 Gbps data channel that is bonded across two 100 MHz channels.

A key observation is that MOCA does not occupy the entire operating frequency range. The large frequency range allows multiple MoCA system to coexist.

The most probably solution is to set aside some amount of downstream spectrum, say 200 MHz, for use by MoCA, and let MoCA find it.

3.5.2 GPS

GPS L3 (1381.05 MHz) is an encoded alarm signal broadcast worldwide by the GPS constellation. It is used by part of the US DOD Nuclear Detection System (NDS) package aboard GPS satellites (NDS description [29]). Encoding is robust and is intended for receipt by military ground-based earth stations. These installations are not susceptible to terrestrial signal interference (i.e. skyward-looking antennas).

Despite being so, large scale, wide area leakage into L3 (as from a distributed cable plant) would not be looked upon favorably by either the US or Canadian governments, or by radio astronomy organizations, who already suffer from GPS L3 signals corrupting “their” skyward-looking receive bands near 1381 MHz. [30]

In contrast, L1 (1575.42 MHz) and L2 (1227.60 MHz) are susceptible to terrestrial interference, despite CDMA encoding. This is due to the low-cost nature of the patch antennas and receivers used to detect them in consumer applications. Unlike the military receive systems and precision GPS packages used in commercial navigation (aviation and shipping), which are robust in the presence of terrestrial interference, consumer GPS are not so. Consumer GPS (including auto and trucking) navigation systems rely upon a wide-pattern patch antenna with a low-noise, high-gain preamplifier.

Such a configuration has no discrimination against terrestrial signals. The low level of received signal at the preamp creates a condition ideal for “blanking” of L1 and L2 should a terrestrial signal of sufficient spectral power density –

particularly from overhead cable plant – be present.

Finally, new applications of the latest civilian GPS frequency, L5 (1176.45 MHz), are currently emerging. Despite being CDMA encoded with FEC, it is not possible to predict how consumer receivers for this latest band will perform in the presence of broad-area interference.

It is of some interest to note that the target application for L5 is “life safety”, see [31]. To get a feel for a L1, L2, and L3 receiver architectures, see the following overview paper on civilian GPS receiver parameters, [32].

3.6 Summary

While initially there were many concerns about the logistics of implementing high-split, there are good mediation strategies. Analog video can be removed or remapped. Adjacent device interference should not be a general problem, and a filter LMA or tap plate filter can manage exception cases. Even the OOB channel is quite manageable with DSG or with an LMA.

This LMA can be multi-purpose and include OOB support and downstream high-split filtering. There may be other functions such as FM radio support that may also be interesting to consider.

The LMA has two different implementations. One is a down-conversion. The advantage is low cost, no ASIC needed, and re-use of OOB headend equipment. The second design could be low-cost if done right, requires ASIC integration, and is better suited to a DSG environment.

More research is needed on the impact to the aeronautical band and to the adjacent tuners below 54 MHz.

It is clear, however, that there are no logistical show stoppers in the deployment of a mid-split or high-split system.

4 COAXIAL NETWORK COMPONENTS AND TOPOLOGY ANALYSIS

The goal of any cable operator is a drop in upgrade to add spectrum capacity when needed. This saves time and money in resizing the network such as node and amplifier location and spacing. Adding network elements or changing network element locations will impact cost for electrical powering requirements. [35]

Ideally, the upgrade would touch the minimum number of network elements to reduce cost and time to market. In the section, the technologies, systems and architecture options are explored. The analysis will examine some of the pros and cons of several technologies and architectures, which could be used to provide additional capacity.

The analysis considered the capabilities of a "Drop in Upgrade" to determine the viability and impact for upstream spectrum expansion as a starting point. [35]

- Target starting point is a "Typical" 500 HHP Node Service Group
- Typical number of actives (30) and passives (200)
- Existing spacing, cabling types and distance (see Figure 6)

4.1 Overview of Important Considerations and Assumptions

This report has highlighted some important areas for network planners to consider while making the decisions for the next generation cable access network.

4.1.1 Avoidance of Small Node Service Groups or FTTLA

The analysis and conclusions found in this report indicates that the need for smaller node groups with few actives and passives such as Node +3 or even Fiber to the Last

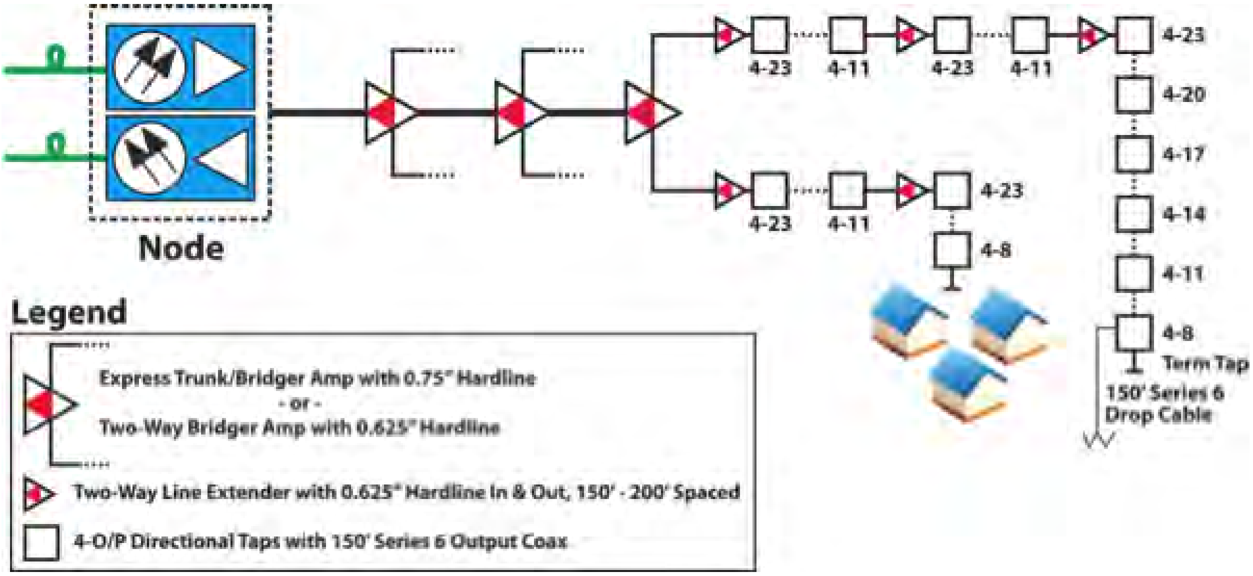

Node
4-23
4-11
4-23
4-11
4-23
4-20
4-17
4-14
4-11
4-8
Term Tap
150' Series 6
Drop Cable
4-23
4-11
4-23
4-8
Legend
Express Trunk/Bridger Amp with 0.75" Hardline
- or -
Two-Way Bridger Amp with 0.625" Hardline
Two-Way Line Extender with 0.625" Hardline In & Out, 150' - 200' Spaced
4-O/P Directional Taps with 150' Series 6 Output Coax

Figure 6 – Coaxial Network Assumptions

Active (FTTLA) is not required to meet capacity, service tier predictions or network architecture requirements for this decade and beyond.

4.1.2 500 HHP Node Long-Term Viability

Our analysis finds that upstream and downstream bandwidth needs may be met while leveraging a 500 HHP node service group for a majority of this decade and even beyond. The maintaining of a 500 HHP service group is of immense value to the MSOs. The ability to solve capacity changes while maintaining the node size and spacing enables an option for a drop-in capacity upgrade.

If the goal is to achieve 1 Gbps capacity upstream this may be achieved using a typical 500 HHP node service group with 30 actives and 200 passives, and over 6 miles of coax plant in the service area as fully described later in this analysis, see Table 5.

The existing 500 HHP node has long-term viability in 750 MHz or higher systems providing enough downstream capacity to last nearly the entire decade. In the upstream a 500 HHP node is predicted to last until mid-decade when the sub-split spectrum may reach capacity and then a choice of node split, node segment or add spectrum like mid-split to maintain the 500 HHP service group are options.

The physical 500 HHP node service group may remain in place with High-split (238) beyond this decade providing 999 Mbps or 1 Gbps of MAC layer capacity. The Top-split 900-1050 with Sub-split has more capacity than Mid-split and will last through the decade.

4.1.3 1 GHz (plus) Passives - A Critical Consideration for the Future

The industry will be considering several spectrum splits and special consideration should be made to the most numerous network elements in the outside plant, the passives. Avoiding or delaying modification to the existing passives will be a significant cost savings to the MSO. Below are key factors about the 1 GHz passives:

1. Introduced in 1990 and were rapidly adopted as the standard
2. This was prior to many major rebuilds of the mid-late 90s and early 2000s
3. Prior even to the entry of 750 MHz optical transport and RF amplifiers/ products in the market place
4. Deployment of 1 GHz passives that would have more capacity than the electronics would have for nearly 15 years
5. Passives are the most numerous network element in the Outside Plant (OSP)
6. Volumes are astounding perhaps as many as 180-220 behind every 500 HHP Node or about 30 per every plant mile (perhaps 40-50 Million in the U.S. alone)
7. 1 GHz Passives may account for 85% of all passives in service today
8. Vendor performance of the 1 GHz Passives will vary and some support less than 1 GHz
9. Our internal measurements indicate that most will support up to 1050 MHz
10. Taps in cascade may affect capacity, thus additional testing is required

4.1.3.1 Assessment of the Passives

The authors believe that special consideration should be given to solutions that leverage the existing passive. This will avoid upgrades that may not be needed until the 2020 era when the MSOs may pursue spectrum above 1 GHz.

If the 1 GHz passives are considered and the desired use is over 1 GHz we believe that 1050 MHz is obtainable. There will be challenges with AC power choke resonances, which may impact the use of passives greater than 1050 MHz with predictably.

4.2 Characterization of RF Components

The network components that most affect signals carried above 1 GHz are the coaxial cable, connectors, and taps. The characteristics of these components are critical, since the major goal in a next generation cable access network is to leverage as much of the existing network as possible.

Before getting into the specifics about the RF characterization and performance requirements, it is worthwhile to establish the quality of signals carried above 1 GHz and below 200 MHz. The bottom line is that while return path signals can be carried above 1 GHz, they cannot be carried with as high order modulation as is possible at lower frequencies.

For example, if the goal is to meet similar return path data capacity the signal carriage above 1 GHz is possible using QPSK for 300 MHz of RF spectrum (47 channels of 6.4 MHz each). Whereas below 200 MHz 256-QAM is possible (due to lower coaxial cable loss) and only 24 channels occupying about 180 MHz spectrum are required, using rough estimates.

Additionally, the over 1.2 GHz solutions will require a 125 HHP service group to support QPSK, where as the High-split 200 solution may use a 500 HHP service group, this is a key contributing factor to the cost deltas of the split options.

4.3 Path Loss and SNR

In a typical HFC Node + N architecture, the return path has many more sources for extraneous inputs, “noise” than the forward path. This includes noise from all the home gateways, in addition to all the return path amplifiers that combine signals onto a single return path (for a non-segmented node).

For now we will ignore the gateway noise, since in principle it could be made zero, or at least negligible, by only having the modem return RF amplifier turned on when the modem is allowed to “talk”.

The RF return path amplifier noise funneling effect is the main noise source that must be confronted; and it cannot be turned off! This analysis is independent of the frequency band chosen for the “New Return Band” (e.g., Mid-split 5-85 MHz; High-split 5-200 MHz; or Top-split with UHF return), although the return path loss that must be overcome is dependent on the highest frequency of signals carried. For a first cut at the analysis, it suffices to calculate the transmitted level from the gateway required to see if the levels are even possible with readily available active devices.

The obvious way to dramatically reduce the funneling noise and increase return path capacity is to segment the Node. That is not considered here to assess how long the network remains viable with a 4x1 configuration, a 500 HHP node service group.

The thermal mean-square noise voltage in 1 Hz bandwidth is kT, where k is the Stefan-Boltzmann constant, 1.38x10^-23

J/deg-K, and T is absolute temperature in degrees Kelvin. From this we have a thermal noise floor limit of -173.83 dBm/Hz. For a bandwidth of 6.4 MHz and 75-ohm system, this gives -57.0 dBmV per 6.4 MHz channel as the thermal noise floor. With one 7 dB noise figure amplifier in the chain, we would have a thermal noise floor of -50 dBmV/6.4 MHz channel.

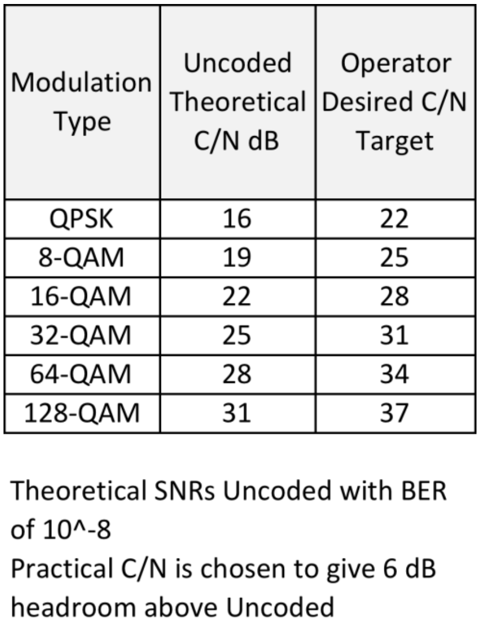
Two amplifiers cascaded would give 3 dB worse; four amplifiers cascaded give 6 dB worse than one. And since the system is balanced to operate with unity gain, any amplifiers that collect to the same point also increase the noise floor by 10*log(N) dB, where N is the total number of amplifiers in the return path segment.

For a typical number of 32 distribution amplifiers serviced by one node, this is five doubles, or 15 dB above the noise from one RF Amplifier, or -35 dBmV/6.4 MHz bandwidth. The funneling effect must be considered in the analysis for the NG Cable Access Network.

If the return path signal level at the node from the Cable Modem (CM) is +15 dBmV, it is clear that the Signal-to-Noise Ratio (SNR) in a 6.4 MHz bandwidth is 50 dB; very adequate for 256-QAM or even higher complexity modulation. But if the Return path level at the node port is 0 dBmV, the SNR is 35 dB; this makes 256-QAM theoretically possible, but usually at least 6 dB of operating margin is desired.

If only -10 dBmV is available at the node return input, the SNR is 25 dB; and so even the use of 16-QAM is uncertain. This illustrates (Table 4) the very high dynamic range of "Pure RF" (about 15 dB higher than when an electrical-to-optical conversion is involved).

Table 4 – Legacy Modulation and C/N Performance Targets



Modulation Type	Uncoded Theoretical C/N dB	Operator Desired C/N Target
QPSK	16	22
8-QAM	19	25
16-QAM	22	28
32-QAM	25	31
64-QAM	28	34
128-QAM	31	37

Theoretical SNRs Uncoded with BER of 10^-8
Practical C/N is chosen to give 6 dB headroom above Uncoded

Table 5 documents many important assumptions and assumed node configuration conditions. An important assumption is the CM maximum power output level of +65 dBmV into 75 ohms.

What this means is that if many channels are bonded (to increase the amount of data transmitted), the level of each carrier must be decreased to conform to the CM maximum power output constraint. Two channels bonded must be 3 dB lower each; four channels must be 6 dB lower than the Pout(max).

Since the channel power levels follow a 10*log(M) rule, where M is the number of channels bonded to form a wider bandwidth group. For 16 channels bonded, each carrier must be 12 dB lower than the Pout(max).

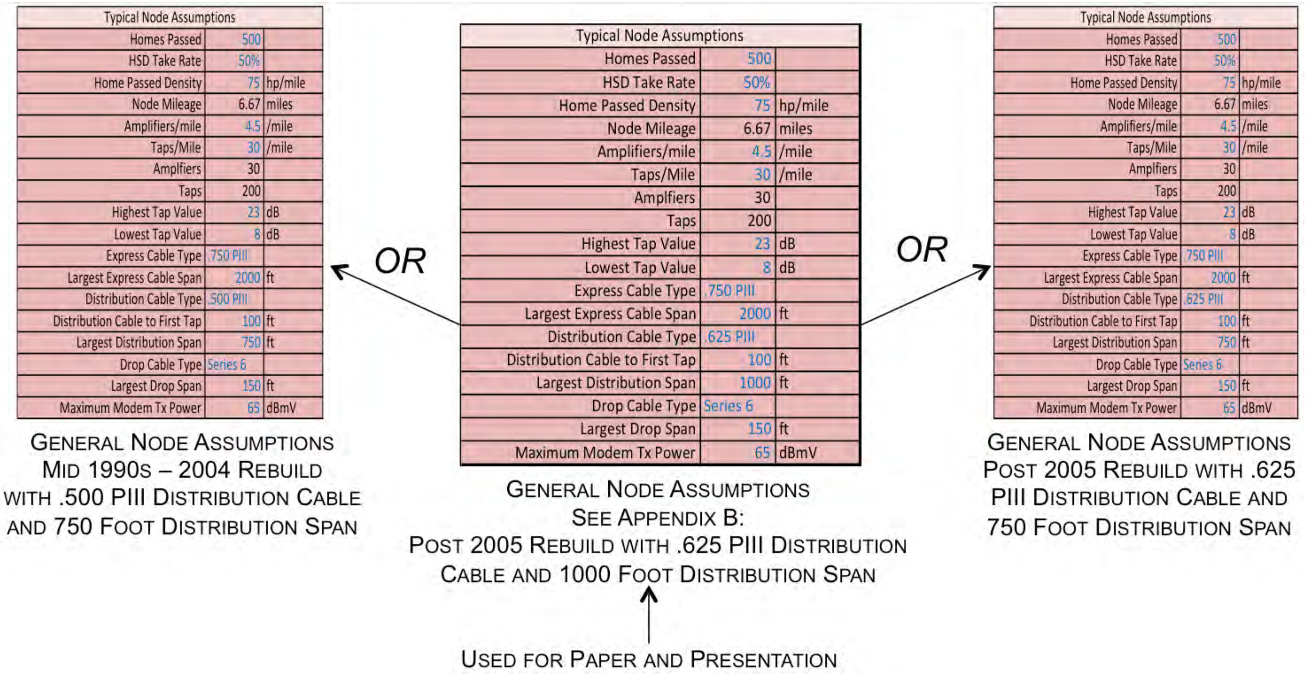
For 48 channels bonded, each must be 16.8 dB lower than the Pout(max). So for 48-bonded channels, the level per channel is at most 65 dBmV -17 dB = +48 dBmV. If there is more than 48 dB of loss in the return path to the node return input, the level is <0 dBmV and 64-QAM or lower modulation is required. The node and system configuration assumptions are as follows.

4.4 Cable Loss Assessment

Two different lengths of 1/2” diameter hardline coax were tested for Insertion Loss and Return Loss (RL). The loss versus frequency in dB varied about as the square root of frequency. But as can be seen below, the loss at 2 GHz is about 5% higher than expected by the simple sq-rt(f) rule. The graph below illustrates a slightly more than twice the loss at 2 GHz compared to 500 MHz, see Figure 7.

In the plot of Figure 8, the coax Return Loss (RL) did not vary as expected above 1200 MHz. This appears due to an internal low-pass matching structure in the hardline-to-75N connectors (apparently for optimizing the 1-1.2 GHz response). The connectors are an important element to return loss with signals above 1 GHz.

Table 5 – Node and Coaxial Network Assumptions Typical of U.S based MSOs



Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.500 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	750	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

General Node Assumptions
Mid 1990s – 2004 Rebuild
with .500 PIII Distribution Cable
and 750 Foot Distribution Span

Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	1000	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

General Node Assumptions
See Appendix B:
Post 2005 Rebuild with .625 PIII Distribution
Cable and 1000 Foot Distribution Span

Used for Paper and Presentation

Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	750	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

General Node Assumptions
Post 2005 Rebuild with .625
PIII Distribution Cable and
750 Foot Distribution Span

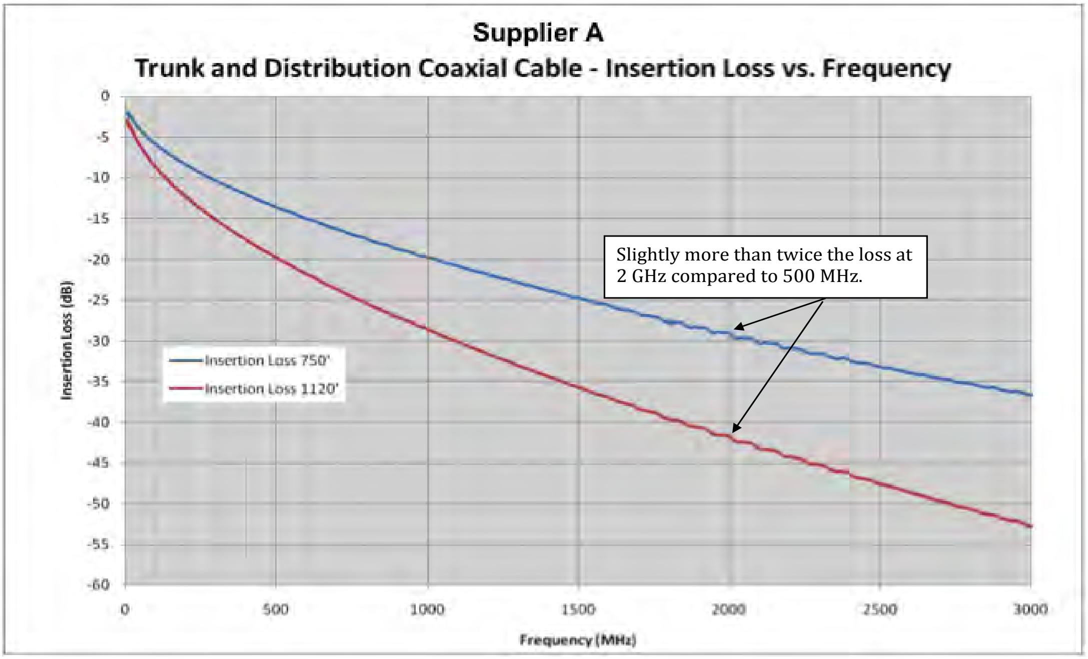


Figure 7 – Distribution Coaxial Cable – Insertion Loss vs. Frequency

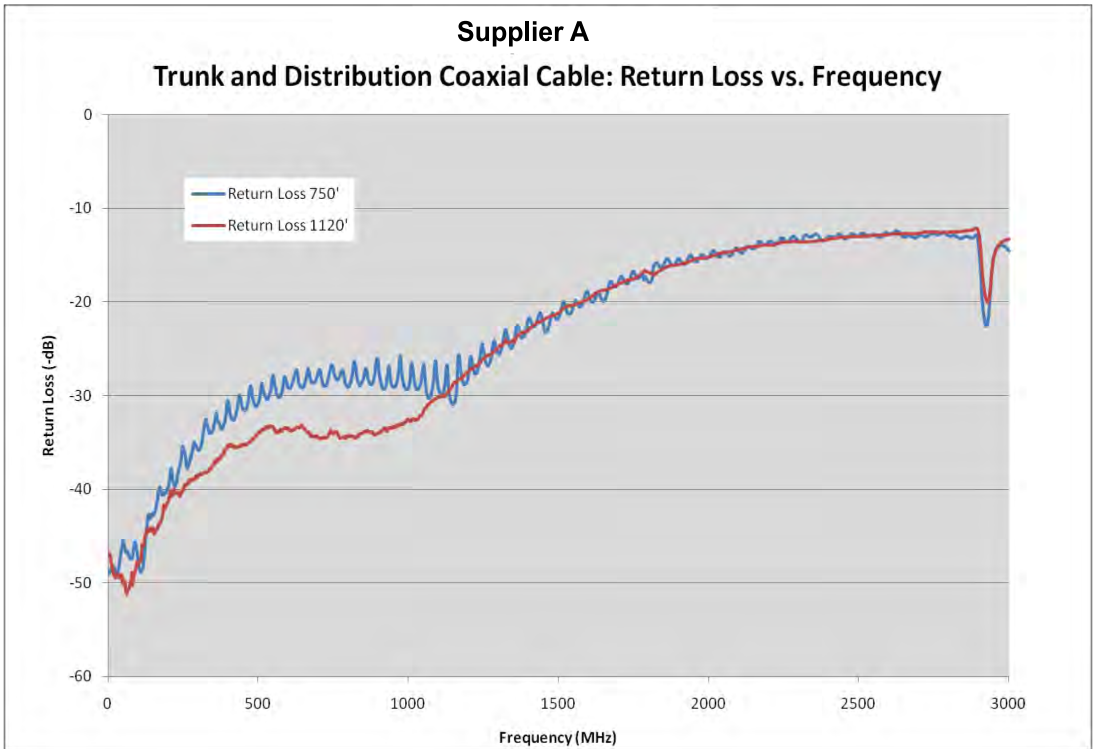


Figure 8 – Distribution Coaxial Cable – Return Loss vs. Frequency

4.5 Tap Component Analysis

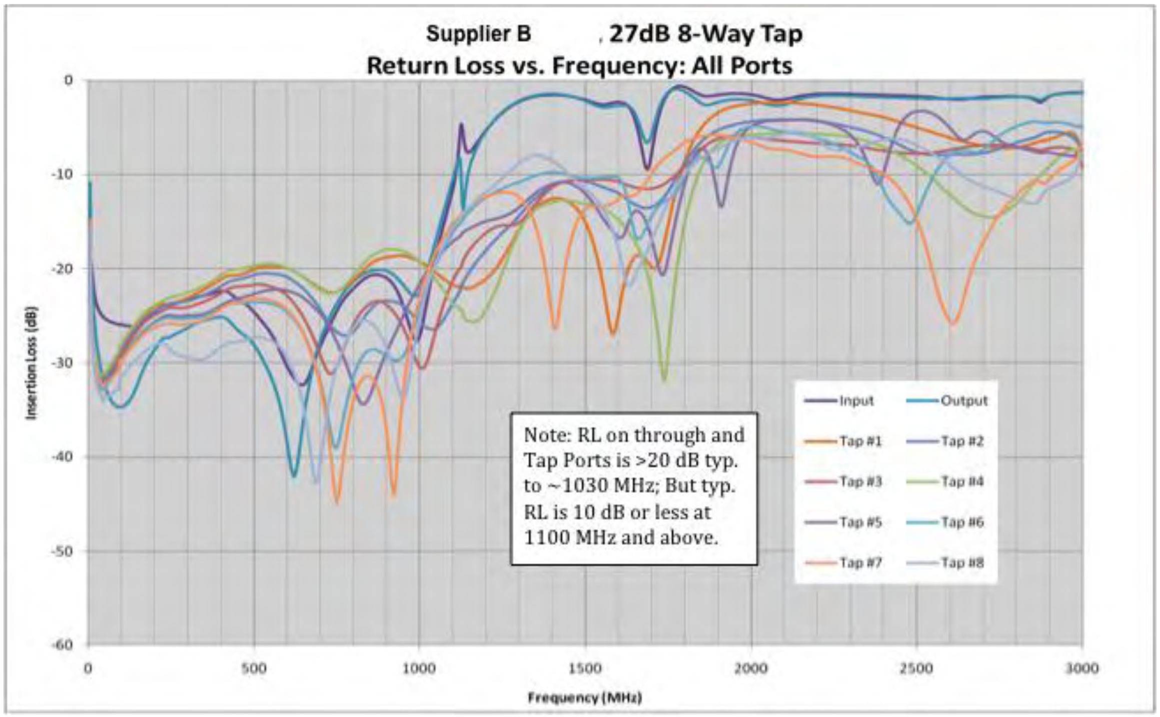
Taps are the components with the most variability in passband characteristics, because there are so many different manufacturers, values, and number of outputs. Most were designed more than ten years ago, well before >1 GHz bandwidth systems were considered. One of the serious limitations of power passing taps is the AC power choke resonance.

This typically is around 1100 MHz, although the “notch” frequency changes with temperature. Tap response resonances are typical from ~1050 to 1400 MHz. A limitation of power passing taps is the AC power choke resonance. This is an important finding when leveraging the existing passives; therefore the use above 1050 MHz may not be predictable or even possible.

Even the newer, extended bandwidth taps, with passband specified 1.8 GHz or 3 GHz, the taps usually have power choke resonances (or other resonances, e.g., inadequate RF cover grounding) resonances in the 1050 MHz to 1300 MHz range. Especially on the tap coupled port. However, most Taps work well to ~1050 MHz.

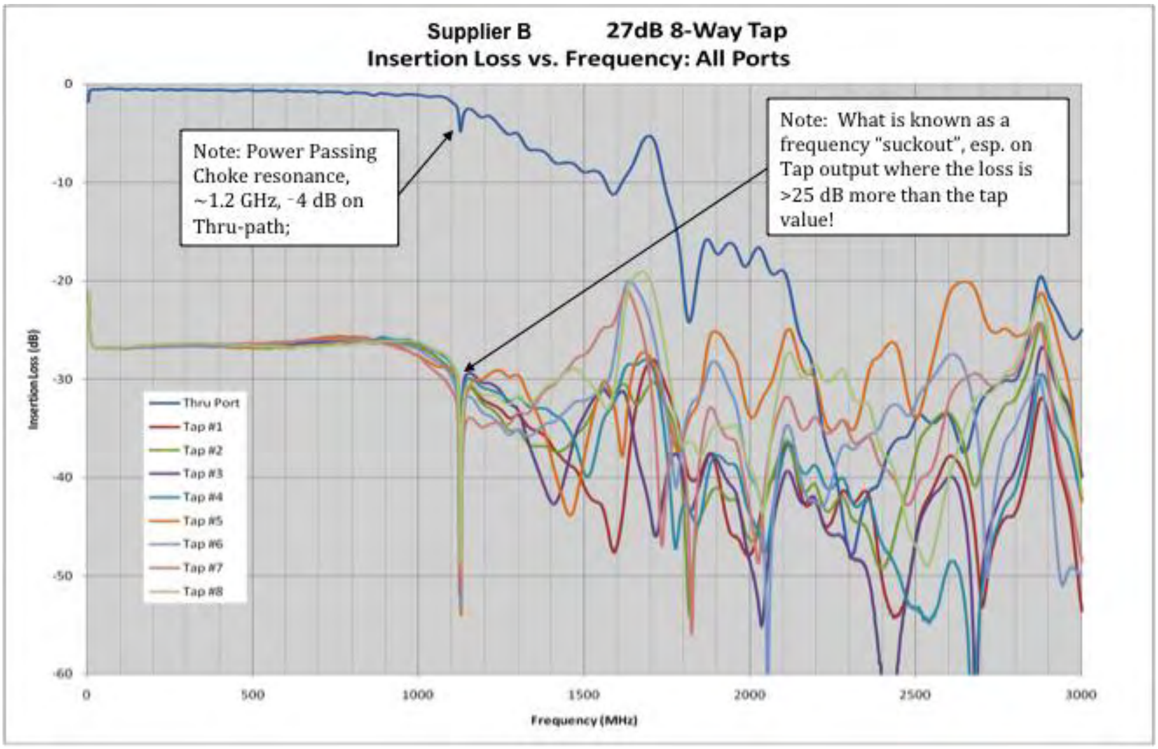
Nearly all taps exhibit poor RL characteristics on all ports above 1400 MHz. Some are marginal for RL (~12 dB), even at 1 GHz. Therefore tap cascades must be tested and over temperature to verify the actual pass band response due to close-by tap reflections.

Figure 9 to Figure 11 show examples of the variability of key RF parameters for an array of Taps evaluated.



Note: RL on through and Tap Ports is >20 dB typ. to ~1030 MHz; But typ. RL is 10 dB or less at 1100 MHz and above.

Figure 9 – 27 dB x 8 Tap - Return Loss vs. Frequency: All Ports



Note: Power Passing Choke resonance, ~1.2 GHz, -4 dB on Thru-path;

Note: What is known as a frequency "suckout", esp. on Tap output where the loss is >25 dB more than the tap value!

Figure 10 – 27 dB x 8 Tap - Insertion Loss vs. Frequency: All Ports

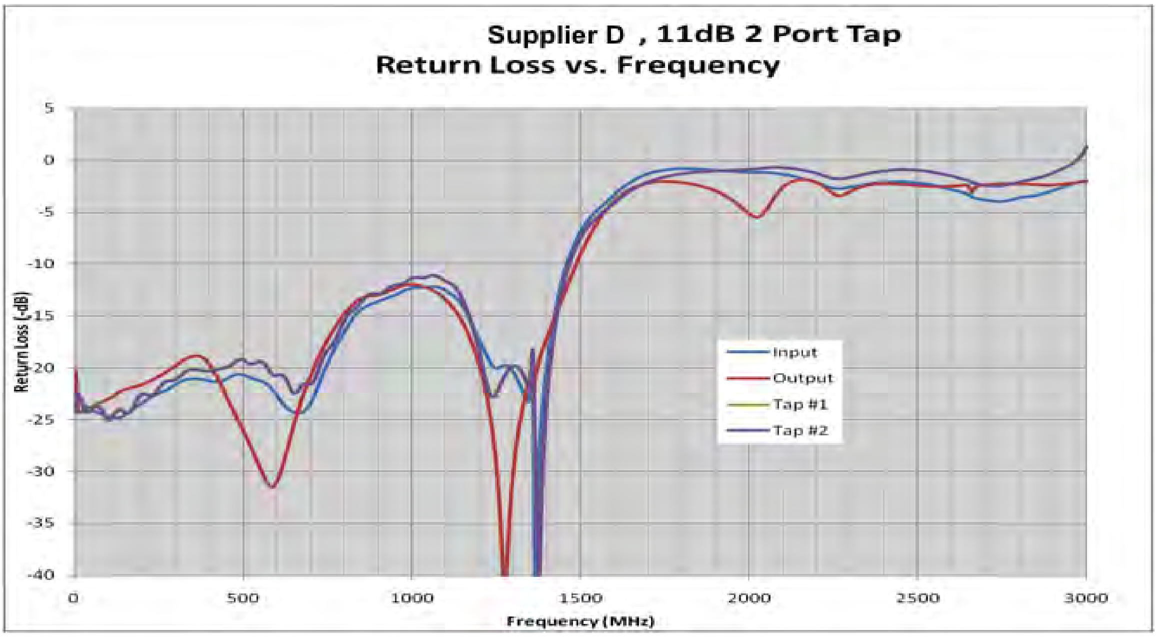


Figure 11 – 11 dB x 2 Tap - Return Loss vs. Frequency

4.6 Field Performance – Passive Coax Above 1 GHz

Let's pull together what we have discussed around taps and passives, the analysis of Section 4.2 and summarize how these components behave together in the context of recent field characterizations performed for the AMP initiative.

As discussed above, coaxial cable and even some current 1 GHz taps are indeed capable of supporting useful bandwidth above 1 GHz [4]. However, the frequency dependence of cable loss (see Figure 7) quickly attenuates signals above 1 GHz when we consider its use relative to attenuation characteristics of a low band upstream. The combination of drop cables, trunk cable, and taps add up to significant losses to the first active.

We can anticipate almost twice the loss (in dB) extending the return band to 200 MHz, such as in the high-split architecture introduced. However, above 1 GHz, the loss may increase by roughly a factor of five (in dB, dependent on Top-Split case chosen) compared to legacy return for such a span. CPE devices must make up for that loss to maintain equivalent performance, all else the same. As we observed in analyzing the case with an increasing amount of channel bonding, they also must generate additional total power associated with the wider bandwidth they would occupy to enable peak rates of a Gbps, relative to today's maximum of 6.4 MHz single or 2-4x bonded channel power.

This is not your father's cable modem – an L-Band, wideband, high power linear transmitter. It is a significantly more complex RF device. It is not a technology challenge, but it will come at a cost premium relative to retail CPE today. Quantifiably, the result is that very high CPE transmit power becomes necessary to close a bandwidth efficient link budget.

Conversely, for a given maximum transmit power, such as 65 dBmV chosen previously, we can favorably assume it is the same transmit power number for low split or for top split frequencies. The additional top-split loss translates to lower SNR at the first active, and every subsequent one if a cascade is in place. This impacts composite SNR formed by the combination of RF funneling and optical link performance.

The end result is that potential bps/Hz of top split is inherently lower for top split, and to achieve an equivalent modulation efficiency, the top split must be deployed over smaller service groups to reduce the noise contributions associated with the lower inherent SNR created by the loss. We will quantify this in further detail in Section 9.6.

However, Motorola performed field measurements as part of the AMP initiative, and the conclusions provide insight into the nature of this issue. We illustrate with a simple, and best case (N+0) example from field characterization done exactly for this purpose. Figure 12 shows field characterized loss [4] [5] of an RF leg of recently-built underground plant, measured from the end of a 300 ft coaxial drop from the final tap of a five-tap string on an otherwise typical suburban architecture.

The five taps, manufactured by Javelin Innovations, where extended bandwidth models, utilizing modified faceplates installed within existing tap housing to extend the RF passband of the network.

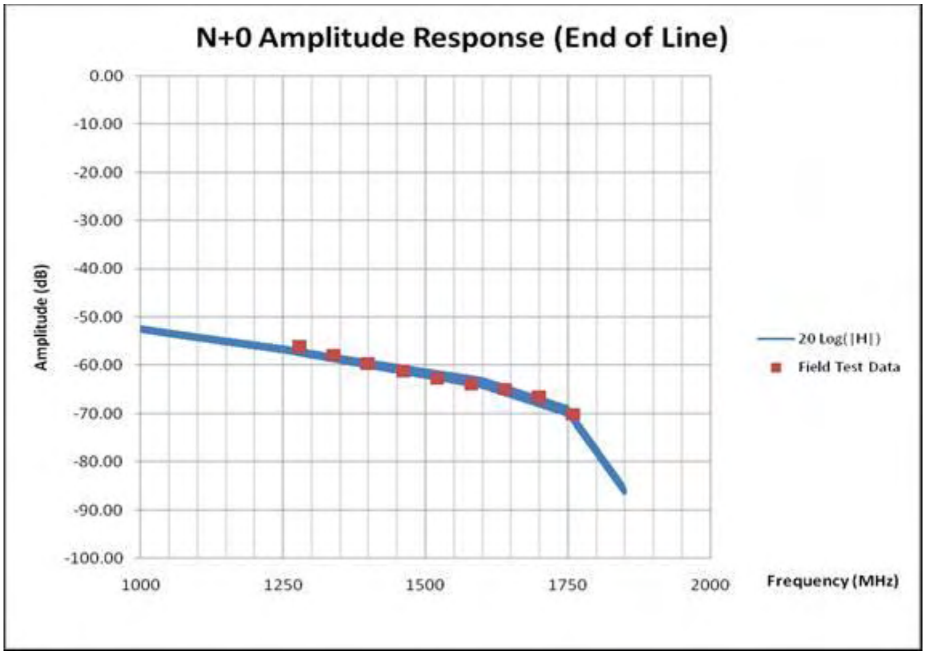


Figure 12 – Top Split Loss Characterization vs Model

Losses from 50-70 dB are observed, with measured data points highlighted in Figure 12. While the drop length represents an extended length scenario, the lack of any home connection removes any effects of additional splitters commonly found inside the home and outside the reach of the MSO until there is a problem in the home.

Let's take a look at the lowest, least attenuation part of the band, 1-1.2 GHz. A reasonable case can be made for a bandwidth efficient link budget for a remote PHY termination, as transmitters that increase the transmit power level over today's requirements to support 65 dBmV will reach the first active with solid SNR.

Mathematically, consider the following:

- Thermal Noise Floor: -65 dBmV/MHz
- Signal BW: 200 MHz
- Total Noise: -42 dBmV/200 MHz
- Active NF+Loss: 8 dB (est)
- Rx Noise Power, Plant Terminated: -34 dBmV

Using the 55 dB of loss observed at the low end of the band for the first 200 MHz, a 58 dBmV transmitter will leave us with an SNR of 37 dB. This is in the neighborhood of the SNR required, with margin, for 1024-QAM if advanced FEC is assumed. In Table 4,1024-QAM is quantified as SNR = 39 dB without FEC using typical HFC upstream optics. Higher orders would become challenging. A 65 dBmV capability would more ably support a higher modulation profile.

Based on the attenuation slope in Figure 12 above 1200 MHz, this gets more challenging as higher bands are considered. Note that the tap performance of the extended band units is very good, but there is simply unavoidable attention associated with deployed coaxial infrastructure that becomes the dominant SNR characteristic of the link.

Now, consider that the above characterization included the following favorable conditions:

- Faceplate tap replacements
- N+0
- Pristine, unused plant
- Extra transmit power assumed in a much higher frequency band
- No connected users
- No home losses

We can easily remove the first of these assumptions for most practical networks. Without the investment in tap faceplate change-outs, typical 1 GHz taps in the band directly above their specified maximum have more loss than these specially designed faceplates.

The additional loss observed is up to 9 dB for the cascade of taps at the end of the usable band, in this case characterized as 1160 MHz [5](worse above that, less below). More loss comes directly off of the SNR as the signal power is dropped into the noise floor.

Thus, in current tap architectures, under N+0 conditions, and constrained to the lowest end of "top-split," in good plant conditions, we are already seeing pressure on SNR for bandwidth efficient modulation profiles as the SNR drops to 30 dB or less. The sensitivity of QAM profile to SNR loss in Table 4 – Legacy Modulation and C/N Performance Targets shows that 2-3 modulation profiles, and the associated capacity, become compromised.

Now, to remove another assumption, if we instead think of the actives as amplifiers, and cascade them on the way to a node with equivalent degradation and potentially combining noise impacts at the node a described in Section 4.3, we find that a bandwidth efficient link budget becomes even more difficult to achieve.

Thus, top-split, while potentially within technology and investment reach, is off to a very difficult start as a viable alternative. The potential bps/Hz efficiency metric is inherently lower, and to achieve an equivalent modulation efficiency, the top-split must be deployed over smaller service groups to reduce the noise contributions associated with the lower inherent SNR created by the loss. This has been shown to be the case analytically as well as in field characterization in a better-than-typical environment.

4.7 Using "Top-Split" Spectrum for New Forward Path Capacity

While the challenges on the upstream above the forward band are significant obstacles to practical deployment, this is not necessarily so on the downstream. This is important, because as the upstream side of the HFC diplex extends, it intrudes on downstream bandwidth and thus removes available downstream capacity. We believe that use of new coaxial spectrum will be required in the evolution of HFC and of DOCSIS, and that both should be part of cable's migration plan. However, in the case of new spectrum above 1 GHz, we believe that is best utilized for new forward capacity.

We have discussed the possibility of a phased architecture. While forward bandwidth loss is relatively modest for an 85 MHz split, if the band extends further, such as to 200-300 MHz, then a significant chunk of downstream capacity is lost. Today, this band may be only carrying analog services, and thus is not reducing the actual deployed downstream capacity, but it is reducing the available capacity for future growth – i.e. it is assumed that at some point analog services will be removed in favor of digital capacity.

With this loss of downstream bandwidth, it then becomes important to uncover new downstream bandwidth, and the logical place to find this is directly above today's forward band. If the architecture is 750 MHz or 870 MHz, then of course there is already technology in place to exploit out to 1 GHz. Beyond 1 GHz, there is very little outdoor gear designed to operate in this band, and no CPE designed to work in this band (just as is the case for upstream).

We can identify at least three compelling advantages to considering use of the band over the end of the defined tap bandwidth for forward services, as opposed to reverse:

1) High Fidelity Forward Path – The fundamental characteristics of the forward path have always been to around a high SNR, low distortion environment to ably support analog video. As we know, the reverse path was not originally architected with high fidelity in mind. Over time, technology has been introduced to enable a high-speed data channel, but the low noise and high linearity architected into the forward path is orders of magnitude above the return path. This difference translates to a much more straightforward exploitation of bandwidth with high performance on the downstream.

2) Broadband RF Power – The forward path levels are designed for RF path losses out to 1 GHz. Because of this, the parasitic losses above 1 GHz of the coax, and the minimal additional attenuation, are not a stretch to achieve when extending the forward path. It is an entirely different case in the return, where the architecture has relied on the low loss end of the band, which increases only modestly as it is extended to 85 MHz or even 200 MHz. This issue was highlighted in Sections 4.3 and 1.1.

3) Cost of New RF BW – Forward path RF systems already extend to the 1 GHz range, so are designed with the expectation of the loss implications. There has therefore been continuing investment in broadband RF hybrids driving higher levels over increasing forward bandwidths, still based on supporting a full analog and digital multiplex. As a result, the output levels of these hybrids and nonlinear characteristics have continued to improve. However, investment in these premium devices for the forward path is spread over the number of homes serviced by the actives. The HFC downstream delivers high linearity and high levels over multiple octaves, and the hybrids are shared, spreading the investment across a subscriber pool. In the reverse path, each home needs a high power, linear transmitter (though less than an octave), and also in a much higher frequency band that would likely require a higher cost technology implementation.

4) The use of spectrum above the forward band implies a new guard band. Since guard bands are a percentage of edge frequency, the lost spectrum is sizable, cost significantly lost capacity. The eliminated spectrum will remove prime forward path digital bandwidth from use, costing on the order of 1 Gbps for DOCSIS NG technology, in order to enable *less* capable upstream bandwidth above 1 GHz.

Without question, HFC will need to mine new bandwidth to enable new capacity for continued traffic growth. Today's coax remains unexploited above 1 GHz in all cases, and above 750 MHz and 870 MHz in other cases in North America. Current forward path technology is already within striking distance and readily capable of being extended to take advantage of latent coaxial capacity above wherever the forward path ends today [6]. And, while this spectrum is non-ideal in the forward path as well, it will benefit from the introduction of OFDM for NG DOCSIS, but without the spectrum loss and RF power implications of use as upstream band.

Based on the above reasoning, our recommendation is to enable additional coaxial capacity above today's forward band, and to exploit this spectrum for downstream purposes exclusively. We will quantify this band for downstream use in subsequent sections derving data capacity, network performance, and lifespan.

In Section 0, we will estimate the available data capacity of the forward path under various implementations of an extended forward band.

Then, in Sections 10.2.1 and 10.2.2, we will quantify available network capacity and discuss the implications to forward path lifespan.

Finally, in Sections 10.2.3 and 10.2.4, we will describe how this bandwidth could be managed within the system engineering of downstream HFC, implemented within linear optics and RF (not an RF overlay).

5 HFC OPTICAL TRANSPORT TECHNOLOGY OPTIONS

The optical layer will be examined in this section. We will look at two technologies of optical transport return, analog return path and digital return, which may commonly be referred to as Broadband Digital Return (BDR), or simply Digital Return. First, we will review the forward path. [36]

5.1 Overview - Analog Forward Path Transport

Analog Forward path is currently the only economical method for the transmission of cable signals downstream. The advances in analog forward laser technologies enable transmission of the 54-1002 MHz of spectrum this is over 150 channels, each 6 MHz wide. This is approximately 6 Gbps of data capacity assuming the PHY layer transmission utilizing 256-QAM (8 bits per Hz BW efficiency, excluding overhead).

The forward path is a layer 1 media-converter style architecture. The optical transmission may be shared with multiple HFC nodes. There are two network architectures for the forward: Full Spectrum as illustrated in Figure 13; and another called QAM Narrowcast Overlay, or simply Narrowcast Overlay, as in Figure 14.

The MSO serving area between headend and node will be in most cases is less than 40 km. Therefore this will be easily

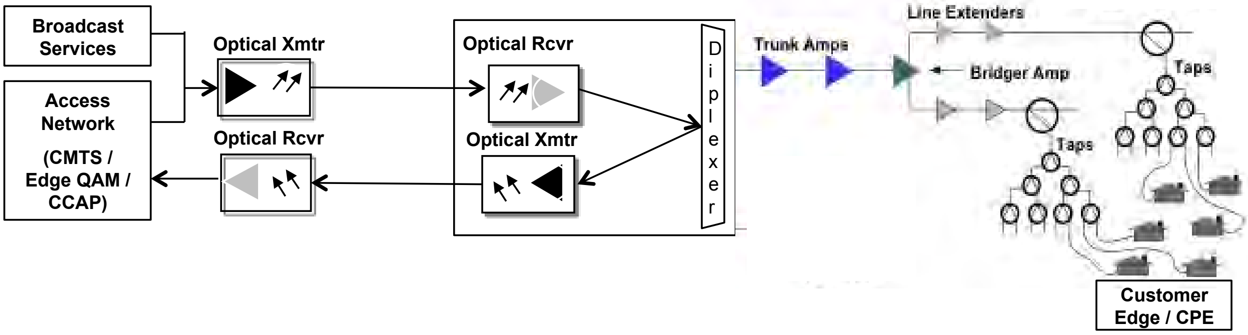


Figure 13 – Hybrid Fiber Coax (HFC) with Full Spectrum and Node +N

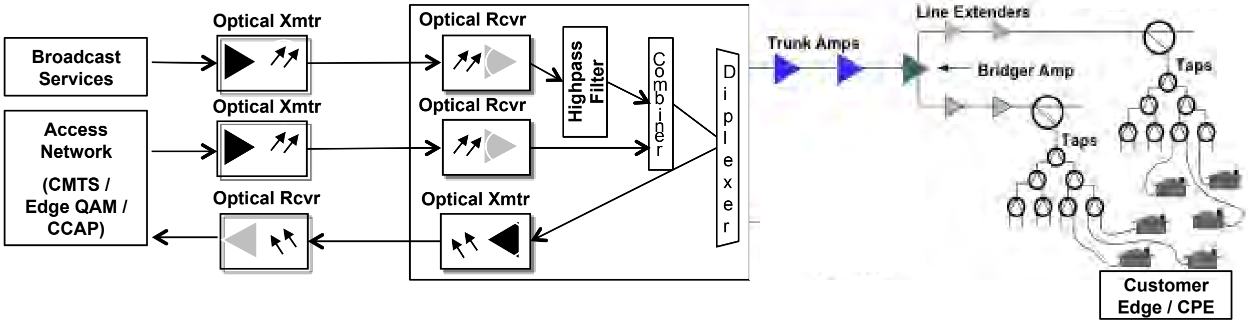


Figure 14 – Hybrid Fiber Coax (HFC) with QAM Narrowcast Overlay and Node +N

supported with an HFC architecture. The support for extremely long distance to and from the node may be a factor for the HFC. The optical capabilities of HFC simply have lots of dependencies, variables, and trade-offs to determine the HFC optical link distance.

We will use round numbers and generalities to discuss some the capabilities of HFC optical transport when considering long distances. So, we will use an example of HFC analog optical transmission of full spectrum, no analog video, and 150 QAM channels, we will assume a 100 km optical reach is achievable in most cases.

In a narrowcast overlay architecture, we assume as many as 40 wavelengths / lambdas per fiber, 80 QAMs of narrowcast spectrum, and a reach of approximately 100 km to the node. HFC optical distance will vary based on many factors, including narrowcast channel loading, the number of analog video channels, and many other factors. We could assume that a greater distance is achievable with an HFC Digital Forward, as well as DFC (Digital Fiber Coax) style optical transport, compared with HFC analog forward optics without the use of EDFAs (erbium-doped fiber amplifier).

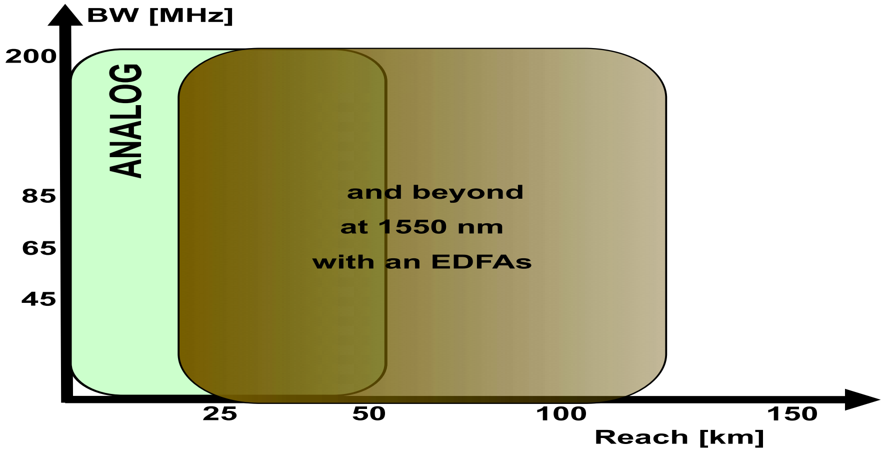


Figure 15 – Return Analog Optical bandwidth and Reach

In some cases, fiber count is insufficient, regardless of the distance. Therefore, to avoid over lashing new fiber to service groups, separate wavelengths are placed on the fiber. The use of HFC analog optics today supports far fewer optical wavelengths than that which is supported using optical Ethernet technology. This may be a challenge for HFC style architectures.

5.2 Overview - Analog Return Path Transport

Analog return path transport is now mostly done with a Distributed Feedback (DFB) laser located in the node housing and an analog receiver located in the headend or hub. Analog return path transport is considered as a viable option for Mid-split, High-split, and Top-split returns. Supporting short to moderate return path distances of 0-50 km with full spectrum High-split is achievable. If the wavelength is changed to 1550 nm with an EDFA, then greater distances are possible. This is shown in Figure 15.

The analog optical return path transport presently supports up to 200 MHz loading; but typically only 5-42 MHz or 5-65 MHz is carried, depending on the distribution diplex filter split. The major benefit with analog optical return is its simplicity and flexibility, when compared with HFC style digital optical transmission. Distance is the chief challenge of analog optical transport. Refer to the Figure 15 and Figure 16.

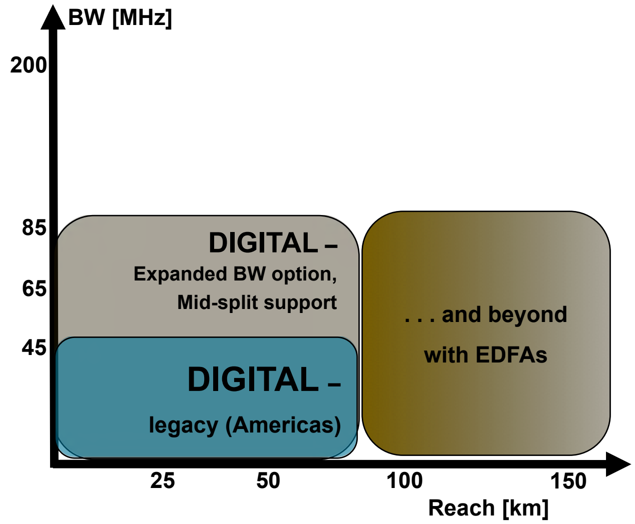


Figure 16 – Return Optical bandwidth and Reach

Pros

The chief advantage of analog return is its cost effectiveness and flexibility. If analog return optics are in use in the field today, there is a good chance that they will perform adequately at 85 MHz; and even 200 MHz loading may be possible, if required in the future. This would allow an operator to fully amortize the investment made in this technology over the decade.

Cons

There are drawbacks to using analog optics. Analog DFB's have demanding setup procedures. RF levels at the optical receiver are dependent on optical modulation index and the received optical power level. This means that each link must be set up carefully to produce the desired RF output at the receiver (when the expected RF level is present at the input of the transmitter). Any change in the optical link budget will have a dramatic impact on the output RF level at the receiver, unless receivers with link gain control are used.

Also, as with any analog technology, the performance of the link is distance dependent. The longer the link, the lower the input to the receiver, which delivers a lower C/N performance. The practical distance over which an operator can expect to deliver 256-QAM payload on analog return optics is limited.

Assessment

The analog return transmitter will work well for the low and high frequency return. Analog return path options should be available for the higher frequency return options at 900-1050 MHz and 1200-1500 MHz. However the cost vs. performance at these frequencies when compared to digital alternatives may make them less attractive. There will be distance limitations and EDFAs will impact the overall system performance noise budgets. The distance of 0-50 km are reasonable and longer distance would be supported with an EDFA.

5.3 Overview – Digital Return Path

Digital return path technology is commonly referred to as broadband digital return (BDR). The digital return approach is "unaware" of the traffic that may be flowing over the spectrum band of interest. It simply samples the entire band and performs an analog to digital conversion continuously, even if no traffic is present. The sampled bits are delivered over a serial digital link to a receiver in the headend or hub, where digital to analog conversion is performed and the sampled analog spectrum is recreated.

The parameters of analog to digital conversion will need to be considered when determining the Digital Return optical transport requirements. There are two important factors in the A-to-D conversion:

1. Sampling Rate and
2. Bit Resolution (number of bits of resolution).

Sampling Rate

- Inverse of the time interval of which samples of the analog signal are taken.
  - Referred to as Samples per Second or Sampling Frequency.
- Nyquist Sampling Theorem governs the minimum sampling rate.
  - Minimum sampling frequency must be at least twice the frequency width of the signal to be digitized.
  - Example: Return band from 5 – 42 MHz must be sampled at 84 MHz (at least). For practical filter realization, the sampling rate should be at least 10-20% greater.

Bit Resolution

- Number of bits to represent the amplitude for each sample taken.
- Each bit can be "1" or "0" only, but multiple bits can be strung together as "words" of "n" number of bits.
- Number of amplitude levels can be calculated as 2^n, where "n" is the number of bits of resolution. Example: 8 bits leads to 2^8 = 256 levels.

Pros

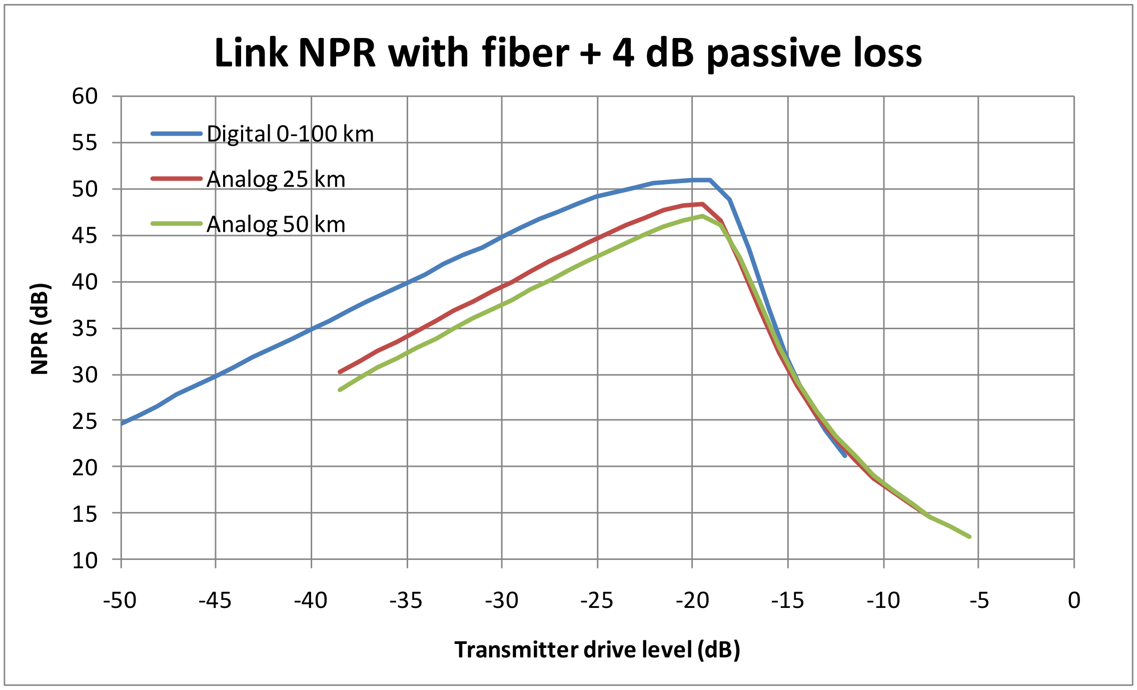


Figure 17 – Analog & Digital Return NPR

There are a number of advantages to the digital return approach. The output of the receiver is no longer dependent on optical input power, which allows the operator to make modifications to the optical multiplexing and de-multiplexing without fear of altering RF levels. The link performance is distance independent – same MER (Modulation Error Ratio) for 0 km as for 100 km, and even beyond as Figure 17 illustrates. The number of wavelengths used is not a factor since on/off keyed digital modulation only requires ~20dB of SNR; thus fiber cross-talk effects do not play a role in limiting performance in access-length links (<160 km)

The RF performance of a digital return link is determined by the quality of the digital sampling, rather than the optical input to the receiver; so consistent link performance is obtained regardless of optical budget. The total optical budget capability is dramatically improved since the optical transport is digital. This type of transport is totally agnostic to the type of traffic that flows over it.

Multiple traffic classes (status monitoring, set top return, DOCSIS, etc) can be carried simultaneously. Figure 17 below is an illustration of performance and distance when examining the analog and digital optical transport methods. With regards to the link noise power ratio (NPR) with fiber and 4 dB optical passives loss, the digital return used 1470 – 1610 nm; analog 25 km used 1310 nm, while the analog 50 km used 1550 nm. The optical output power of each transmitter was 2 mW (+3 dBm).

The Digital Return main drivers are as follow:

- "Set it and forget it" – technician and maintenance friendly
- Signal to noise performance does not degrade with distance
- Supports redundancy over uneven lengths/longer lengths
- Pairs well with "fiber deep" architectures, enables "service group aggregation"
- Pluggable optics for less costly inventory

Cons

The chief drawback to digital return is the fact that nearly all equipment produced to date is designed to work up to 42 MHz. Analog receivers are not useable with digital return transmissions. Further, the analog-to-digital converters and digital return receivers aren't easily converted to new passbands. It requires "forklift upgrades" (remove and replace) of these optics when moving to 85 MHz and 200 MHz return frequencies. There is currently no standardization on the digital return modulation and demodulation schemes, or even transport clock rates.

Another chief drawback to digital return is the Nyquist sampling theorem. It requires a minimum sampling rate, f_s >2B for a uniformly sampled signal of bandwidth, B Hz. For n-bit resolution, this requires a Transport Clock frequency >2nB. It is assumed that the higher the transport clock, the more costly it is. And with higher clock speed, there is more fiber dispersion, which sets an upper limit on transport rate! This causes some practical limitations as to how high the return spectrum can cost effectively reach when considering digital return.

The key points about Nyquist Sampling are captured below. This may be a major driver for the use of analog optics when modest distances are possible and also a major reason to move away from HFC style architectures to a Digital Fiber Coax (DFC) class of architecture when distance is a challenge.

Nyquist Sampling Theorem governs the minimum sampling rate

- Minimum sampling frequency must be at least twice the frequency width of the signal to be digitized

Nyquist Theorem causes some practical limitations

- A 6 MHz baseband signal requires a sampling frequency of 12 MHz minimum
- A 42 MHz return band requires 84 MHz minimum (at least)
- To digitize the entire forward band, we would need to sample at 1.1 GHz (550MHz system) to 2.0 GHz (1GHz system)
- Higher speed A/D converters typically have less Effective Number of Bits (ENOB), translating to decreasing performance at increasing clock speeds for a fixed number of bits.

The total data rate for any given digitized signal can be calculated as follows:

- Determine the minimum sampling rate. As discussed, this is always at least 2X the frequency width of the signal to be digitized (at least). Multiply by the number of resolution bits desired, n, to get the minimum transport clock. And add overhead bits for error correction and framing.

Example: Digital Return

- Typical Return band is 5-42 MHz
- Minimum Sampling frequency is 84 MHz (2*42 MHz) (at least for practical filter realization the sampling rate may be at least 10-20% greater to allow for an anti-aliasing filter.)
- For simple math, we will use 100 MHz or 100 Million samples/second
- Determine the bit resolution will be largely dependent on the SNR required
- For simple math we will use 10-bit resolution or 10 bits/sample
- Multiply bit resolution and sampling rate
  - 100 Million samples/second * 10 bits per sample = 1,000,000,000 bits/second
  - Approximately 1 Gb/s required to digitize the return band

Key Summary:

- >1 Gbps of optical transport was required to transport the 5-42 MHz of spectrum / data capacity
- Estimate of 4 Gbps plus of optical transport was required to transport the 5-250 MHz of spectrum / data capacity at 10 bits per sample (490 Million samples/second * 10 bits per sample = 4,900,000,000 bits/second. This is an estimate only)

Example: Digital Forward

- How about a 550 MHz forward band requiring 52 dB SNR?
- >1.1 Giga samples/second * 10 bits per sample = 11.0 Gb/s!!!

Assessment

It is more difficult and therefore more costly to manufacture digital return products. This may be a driver to use Analog DFB products for the new return applications. The selection of digital return products may be driven by distance and performance requirements. Another driver to move to digital return will be when there is near cost parity with DFB. This may be the case in the future with the new spectrum returns.

5.4 HFC Return Path Analysis and Model

Analog return path transmitters used in HFC applications need to be examined to determine their capability to transmit higher orders of modulation or additional channel loading while maintaining adequate performance. Operating conditions such as the optical link budget, actual channel loading, and desired operational headroom are all contributing factors with respect to performance of these transmitters. Here, operational headroom can be defined as the amount of dynamic range required to provide sufficient margin against the effects of temperature variation, variation from system components (transmitter, receiver, CM/CMTS, etc…), and ingress noise.

In optical networking, the amount of dynamic range for a given modulation format needs to be considered to ensure proper operation of the transmitter under fielded conditions. Typically, 12dB of operational headroom has been recommended for robust operation. However, there may be opportunities in the future to reduce the operational headroom by up to 3dB (perhaps to 9dB). In the future, smaller node sizes and shorter cascades may reduce the amount of ingress noise and the impact of temperature can be lessened with the use of analog DWDM lasers, which are tightly controlled over temperature.

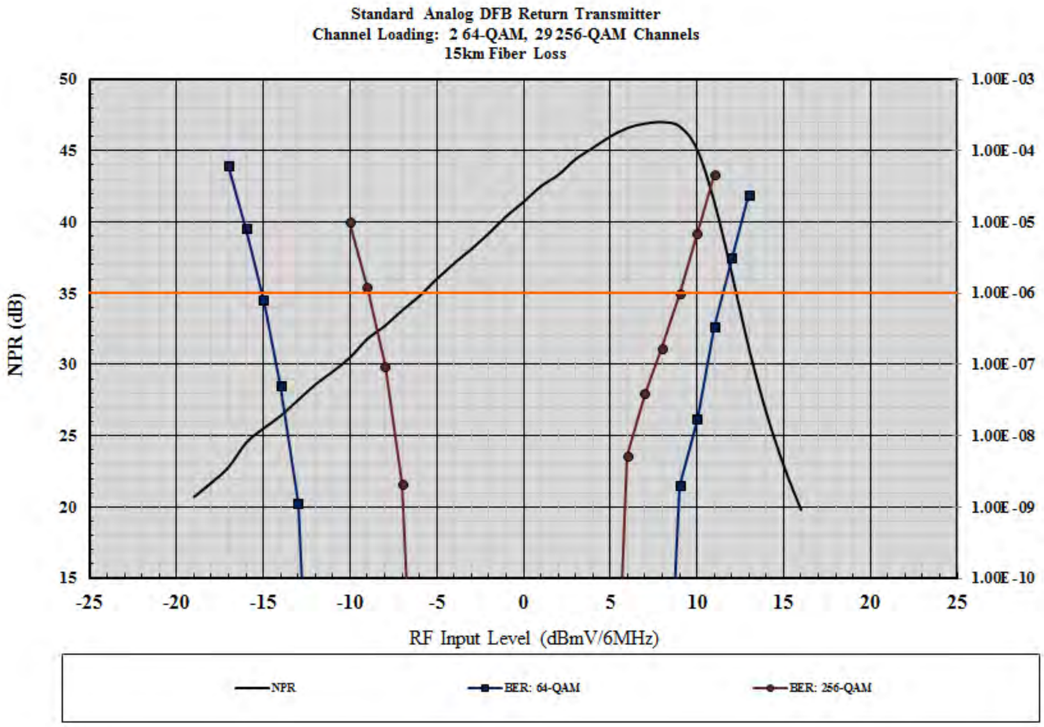


Figure 18 – High-split Standard Analog DFB Return Transmitter

Testing conducted on a standard, analog DFB return transmitter (+3dBm) and an analog DWDM return transmitter, under "high split" loading conditions yielded acceptable dynamic range for 256 QAM operation. Figure 18 provides the results of the +3dBm analog DFB return transmitter. This test was conducted over a 15km link budget with a received power of -3dBm. The RF channel loading consisted of 31 QAM channels upstream containing two 64 QAM channels and twenty-nine 256 QAM channels. The measured dynamic range for a BER< 1E-06 for the 256 QAM channels is 18dB, which provides adequate operational headroom.

Figure 19 and Figure 20 provide data, taken at three frequency splits (low, mid, and high) using 64 QAM and 256 QAM channel loading, for an analog DWDM return transmitter, operating at +8dBm output power over a 16dB optical link (40km of fiber plus 8dB of passive loss). In the "high split" case, this transmitter provides 13dB of dynamic range (1E-06) for 256 QAM, adequate both for present day scenarios where 12dB of operational headroom may be required and for future scenarios where reduced operational headroom is sufficient.

.

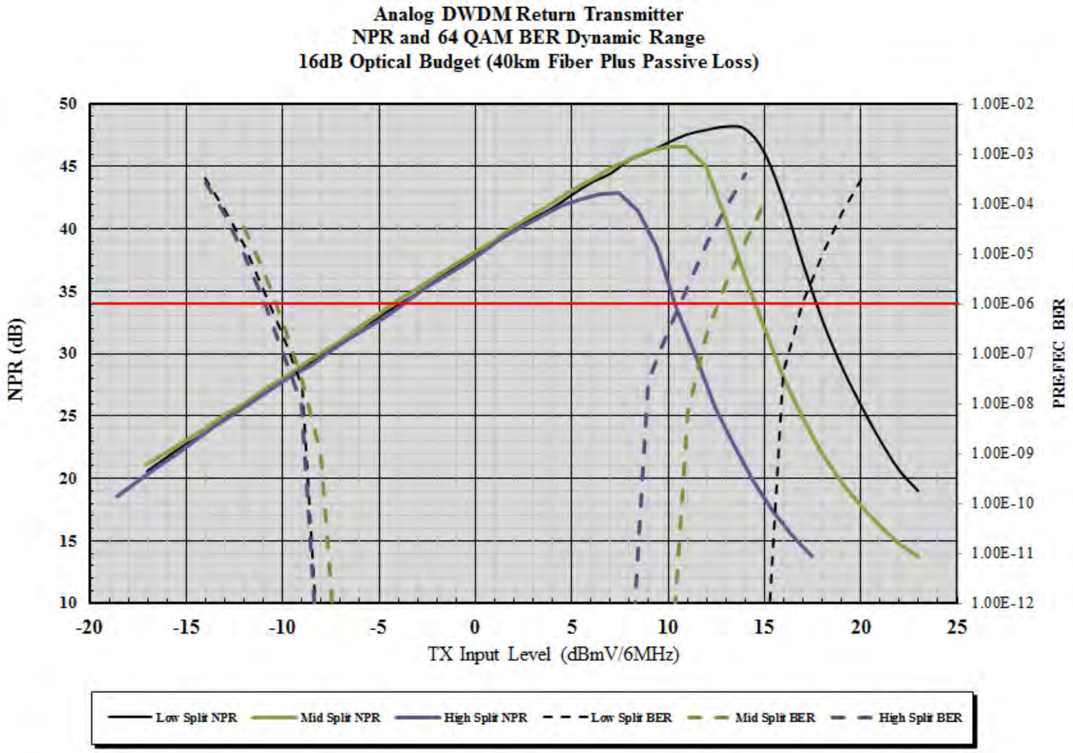


Figure 19 – Analog DWDM Transmitter: 64 QAM (Low/Mid/High Split)

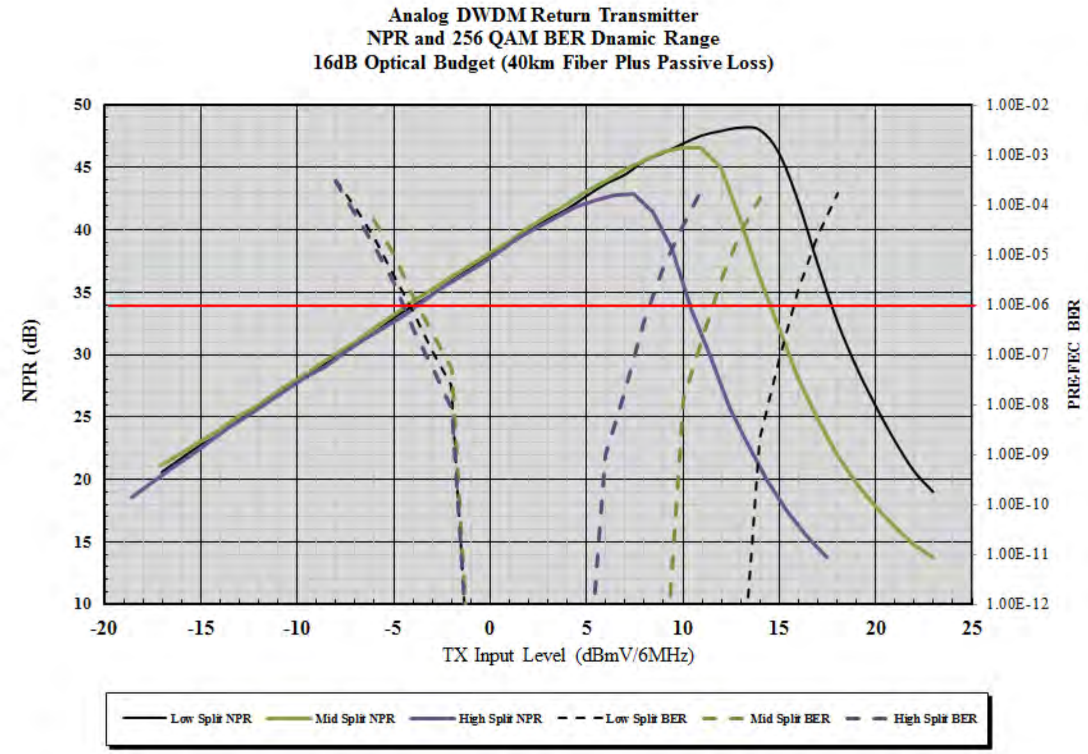


Figure 20 – Analog DWDM Transmitter: 256 QAM (Low/Mid/High Split)

6 SUMMARIES FOR HFC NETWORK COMPONENTS AND TOPOLOGY ANALYSIS

The analyses of the coaxial and optical network, the Hybrid Fiber Cox (HFC) network and the issues that need to be considered that may impact performance are summarized in Table 6. The spectrum selection will play a major role in terms of data capacity and network architecture.

6.1 Major Considerations for Coaxial Network Performance

- **First Major Consideration:** Spectrum Selection
- **Second Major Consideration:** Path Loss or Attenuation
  - Overall System loss progressively increases as frequency increases, thus a major factor when considering higher frequency return.
  - Path Loss from the Last Tap including: Tap Insertion, Tap Port, Cable Loss Hardline, Cable Loss Drop, In Home Passive Loss to Modem/Gateway (these impact Top-splits)
- **Third Major Consideration:** Transmit Power Constraints
  - Modem maximum power output composite not to exceed +65 dBmV (to minimize power and cost, and maintain acceptable distortion)
- **Fourth Major Consideration:** Noise Funneling Effect
  - The effects of large number of return path amplifiers. This is not a factor at low frequency because the cable loss is low enough that a cable modem can provide adequate power level to maintain high C/N.
- **Fifth Major Consideration:** Optical CNR Contribution
- **Sixth Major Consideration:** Error Correction Technology

6.2 Analysis

An analysis will be performed on the network in Figure 21 and described by Table 6

Table 6 – Node Service Group and Coaxial Network Assumptions

Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	1000	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

For this analysis, 0.75” PIII class cable was assumed for express amplifier spans and 0.625” PIII class cable was assumed for tapped feeder spans. Table 7 shows what the gain requirements would be for an upstream express amplifier at the ranges of Figure 21.

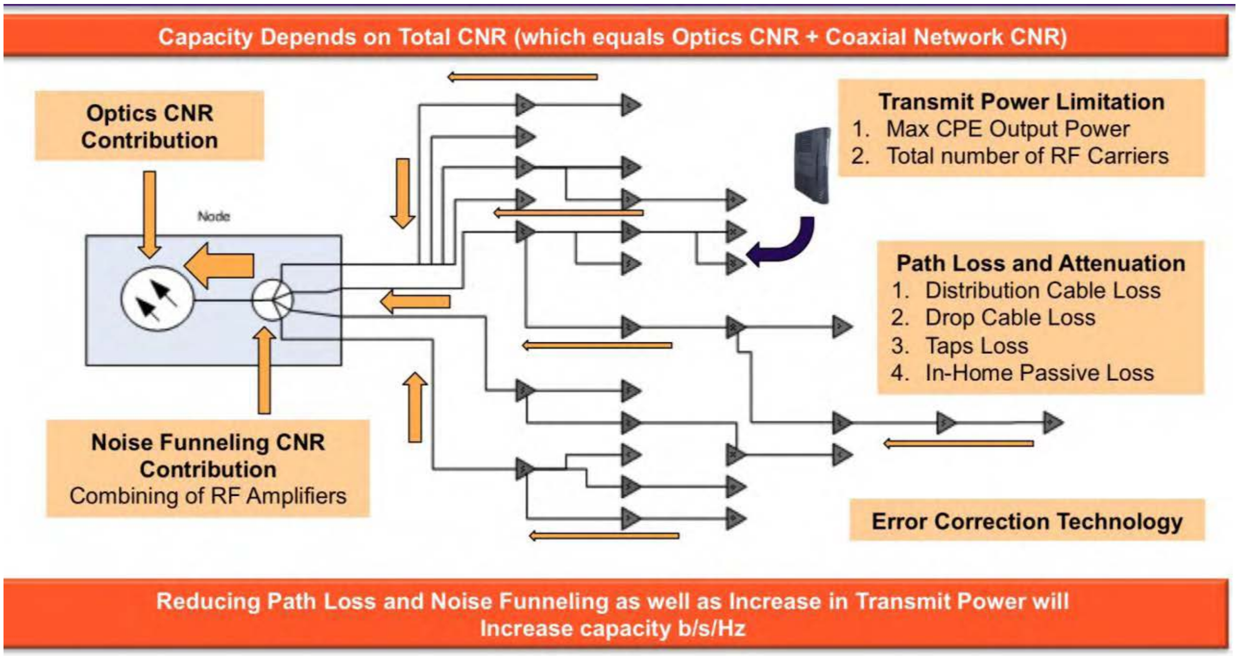

Capacity Depends on Total CNR (which equals Optics CNR + Coaxial Network CNR)
Optics CNR Contribution
Node
Transmit Power Limitation
1. Max CPE Output Power
2. Total number of RF Carriers
Path Loss and Attenuation
1. Distribution Cable Loss
2. Drop Cable Loss
3. Taps Loss
4. In-Home Passive Loss
Noise Funneling CNR Contribution
Combining of RF Amplifiers
Error Correction Technology
Reducing Path Loss and Noise Funneling as well as Increase in Transmit Power will Increase capacity b/s/Hz

Figure 21 – Major Considerations for Coaxial Network Performance

It is worth noting that the Sub-split, Mid-split and High-split gain requirements can be satisfied with commonly available components that are currently used in amplifier designs today and would likely involve no cost premium. However, the Top-Split options would likely require multistage high gain amplifiers to overcome predicted losses, which would be more costly.

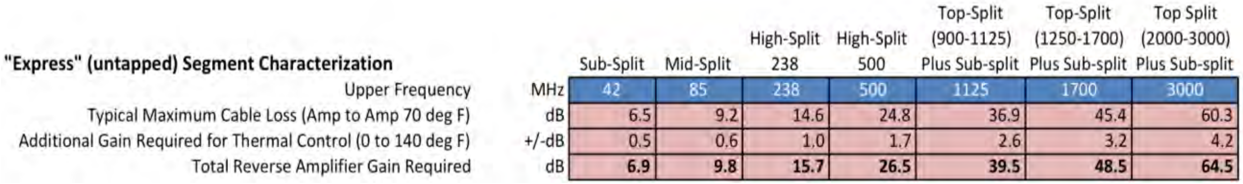
It is also important to note that thermal control would likely become a major issue in the Top-split designs. Table 7 shows seasonal temperature swings of 5 to 6 dB loss change per amplifier span would be likely in the top-split solutions.

Reverse RF AGC systems do not exist today, and could be complex and problematic to design. Thermal equalization would be sufficient to control the expected level changes at 200 MHz and below, but it is not certain that thermal equalization alone will provide the required control above 750MHz. This needs more study.

Table 8 is a summary of path loss comparisons from home to the input of the first amplifier, which will ultimately determine the system operation point. It is interesting to note that as soon as the upper frequency is moved beyond the Sub-split limit, the maximum loss path tends toward the last tap in cascade as opposed to the first tap. There is a moderate increase in expected loss from 42 to 200 MHz, and a very large loss profile at 1000 MHz and above. The expected system performance can be calculated for each scenario.

Table 7 shows the compared performance calculations for the 500 home passed node outlined in Figure 21 and Table 6. The desired performance target is 256-QAM for each scenario; if it can be achieved, the throughput per subscriber will be maximized.

Table 7 – Express" (untapped) Segment Characterization



"Express" (untapped) Segment Characterization		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Typical Maximum Cable Loss (Amp to Amp 70 deg F)	dB	6.5	9.2	14.6	24.8	36.9	45.4	60.3
Additional Gain Required for Thermal Control (0 to 140 deg F)	+/-dB	0.5	0.6	1.0	1.7	2.6	3.2	4.2
Total Reverse Amplifier Gain Required	dB	**6.9**	**9.8**	**15.7**	**26.5**	**39.5**	**48.5**	**64.5**

For each approach, it is assumed that a CPE device is available with upstream bonding capability that can use the entire spectrum available at a reasonable cost. The number of bonded carriers transmitting must not exceed the maximum allowable modem transmit level, so the maximum power per carrier is calculated not to exceed 65 dBmV total transmitted power.

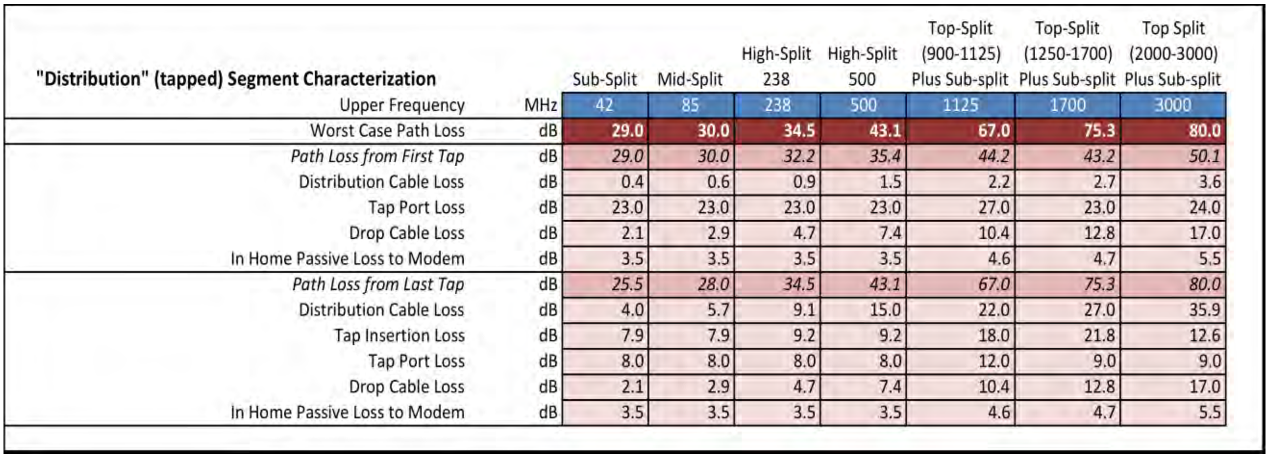
The maximum power, along with the worst-case path loss, yields the input level to the reverse amplifiers in the HFC Network. If the return level was greater than 15 dBmV, it was assumed that it would be attenuated to 15 dBmV.

Armed with the input level and station noise figure, the single station amplifier C/N is calculated and then funneled through the total number of distribution amplifiers serving the node to yield the C/N performance expected at the input of the node.

The HFC return optical links considered in the model are the analog DFB lasers or broadband digital return (BDR) systems. The selection DFB option was selected for the low frequency returns up to the High-split of 238 MHz. However, High-split 500 was modeled with Digital HFC Return. All the Top-split spectrum options used the Digital HFC Return optics as well.

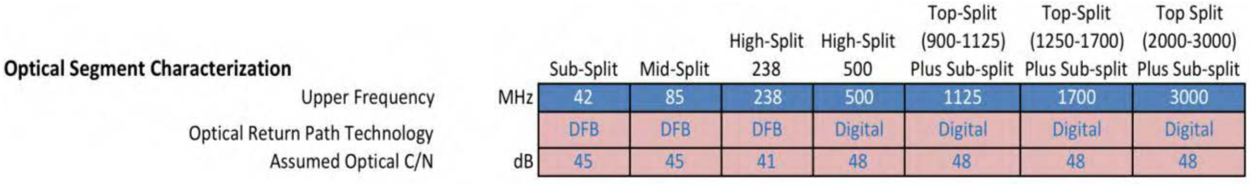
In the model used to determine the performance of the optical link at several we used the following inputs for the various spectrum options and as well as optical link types, see the Table 9 below.

Table 8 – "Distribution" (tapped) Segment Characterization



"Distribution" (tapped) Segment Characterization		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Worst Case Path Loss	dB	**29.0**	**30.0**	**34.5**	**43.1**	**67.0**	**75.3**	**80.0**
Path Loss from First Tap	dB	*29.0*	*30.0*	*32.2*	*35.4*	*44.2*	*43.2*	*50.1*
Distribution Cable Loss	dB	0.4	0.6	0.9	1.5	2.2	2.7	3.6
Tap Port Loss	dB	23.0	23.0	23.0	23.0	27.0	23.0	24.0
Drop Cable Loss	dB	2.1	2.9	4.7	7.4	10.4	12.8	17.0
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5
Path Loss from Last Tap	dB	*25.5*	*28.0*	*34.5*	*43.1*	*67.0*	*75.3*	*80.0*
Distribution Cable Loss	dB	4.0	5.7	9.1	15.0	22.0	27.0	35.9
Tap Insertion Loss	dB	7.9	7.9	9.2	9.2	18.0	21.8	12.6
Tap Port Loss	dB	8.0	8.0	8.0	8.0	12.0	9.0	9.0
Drop Cable Loss	dB	2.1	2.9	4.7	7.4	10.4	12.8	17.0
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5

Table 9 – Optical Segment Characterization Assumed per Spectrum Split



Optical Segment Characterization		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	48	48

The inputs and results in Table 9 show following:

- 5 - 238 MHz have sufficient performance to support 256-QAM modulation at a 500 HHP node.
- 5 - 500 MHz have sufficient performance to support 128QAM modulation at a 500 HHP node.
- The top-split options suffer from cable loss, not to exceed +65 dBmV, and noise funneling.
  - The Top-split (900-1125) may operate at QPSK modulation with only 24 carriers at 6.4 widths.
  - The Top-split (1250-1700) may operate at QPSK modulation with only 3 carriers at 6.4 widths.
  - The Top-split (2000-3000) may operate at QPSK modulation with only 1 carrier at 6.4 widths. .

Further analysis of the Top-split options as shown in Table 10 through Table 13 concludes that reducing the node size, and thereby the funneled noise in the serving group could yield higher modulation capability. In these tables are red arrows, which highlight the key service group size and performance.

The comparison of low spectrum return options like that of Sub-split, Mid-split, and High-split versus the Top-split spectrum choices are measured in the following tables.

These table show that spectrum selection is one of the most important choices the cable operators could make for expanding the upstream. The spectrum options have vastly different performance capabilities when compared in the same cable topology. The Top-split option "MUST" reduce the noise funneling level, which requires smaller service group to increasing loading. Top-split allows only low order modulation and few carries will operate.

All of these assumptions are based on the use of single carrier QAM based systems using Reed-Solomon codes. Section 7 "DOCSIS PHY Technologies" describes the use of different error correction technologies and improvement that may be achieved in operating conditions and use of higher order modulation.

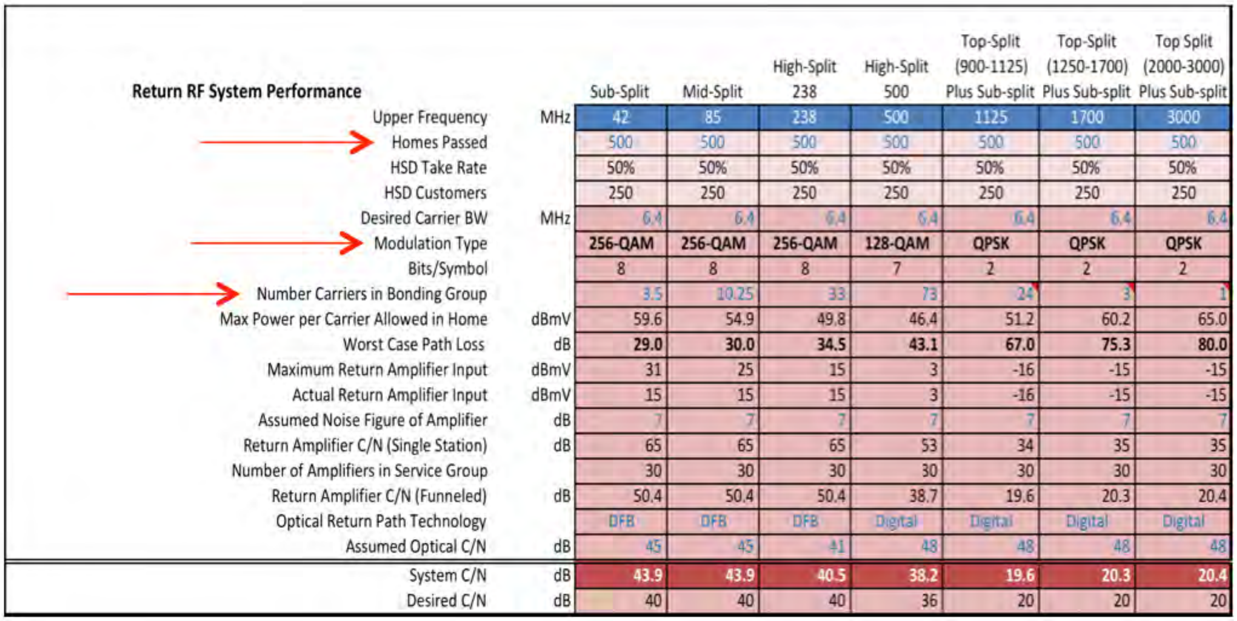
The use of Top-split frequencies will drive higher costs for additional node segmentation, nodes splits, and even running fiber deeper in the network.

The existing passive have an AC power choke resonances, which varies between 1050 - 1400 MHz making portions unusable or predictable. The recommendation on the low side is not to exceed 1050 MHz and high side 1125 MHz. Some passives may not even reach 1 GHz in cascade, so test your passives.

Plan to use low frequency return (Mid-split and High-split) and allow the downstream to use 1 GHz plus, like 1125 MHz or as high as the cascade of existing taps will allow.

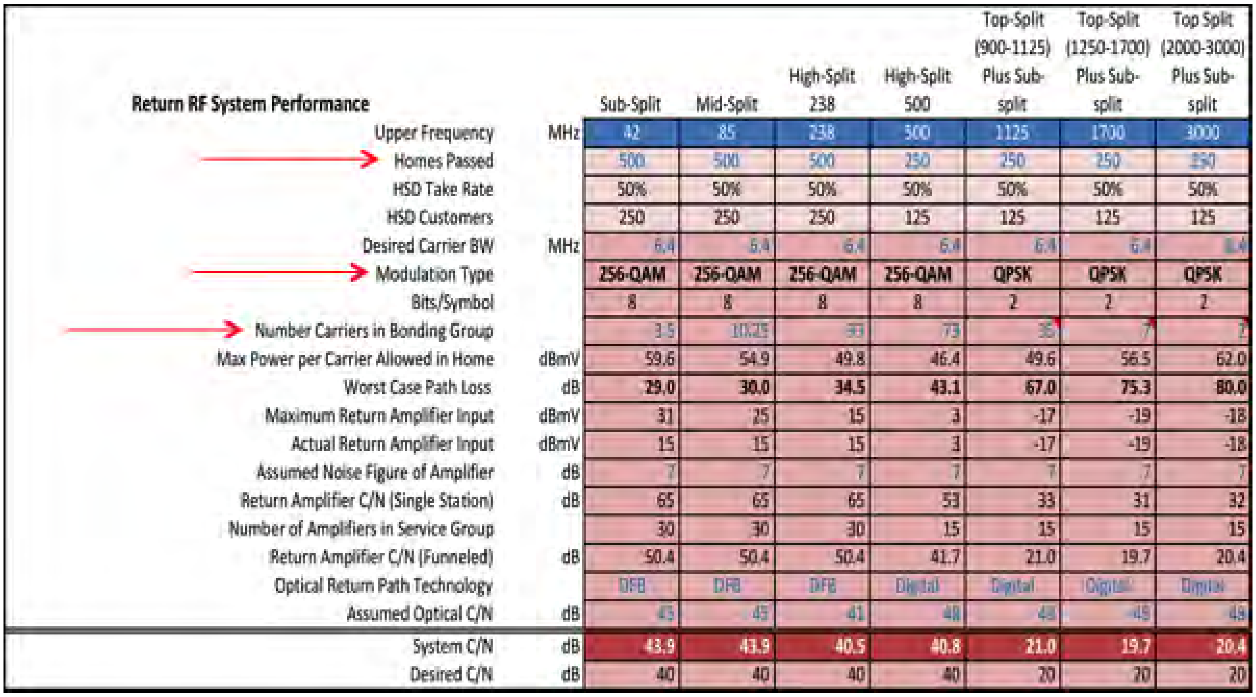
Consider touching the taps as a last resort.

Table 10 – Network Performance of a 500 HHP Optical Service Group



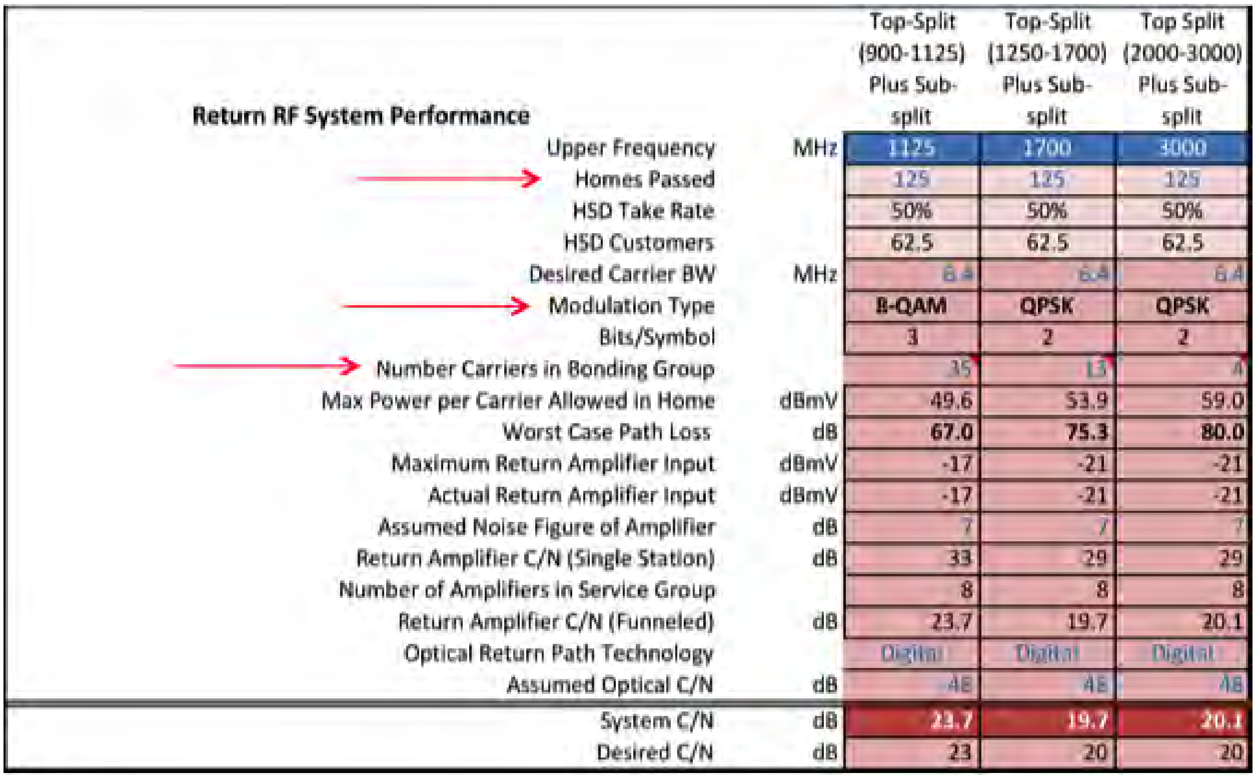
Return RF System Performance		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Homes Passed		500	500	500	500	500	500	500
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	250	250	250	250
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4
Modulation Type		256-QAM	256-QAM	256-QAM	128-QAM	QPSK	QPSK	QPSK
Bits/Symbol		8	8	8	7	2	2	2
Number Carriers in Bonding Group		3.5	10.25	33	73	24	3	1
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	49.8	46.4	51.2	60.2	65.0
Worst Case Path Loss	dB	**29.0**	**30.0**	**34.5**	**43.1**	**67.0**	**75.3**	**80.0**
Maximum Return Amplifier Input	dBmV	31	25	15	3	-16	-15	-15
Actual Return Amplifier Input	dBmV	15	15	15	3	-16	-15	-15
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7
Return Amplifier C/N (Single Station)	dB	65	65	65	53	34	35	35
Number of Amplifiers in Service Group		30	30	30	30	30	30	30
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	38.7	19.6	20.3	20.4
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	48	48
System C/N	dB	**43.9**	**43.9**	**40.5**	**38.2**	**19.6**	**20.3**	**20.4**
Desired C/N	dB	40	40	40	36	20	20	20

Table 11 – 250 HHP Optical SG High-split 500 & Top-split Options



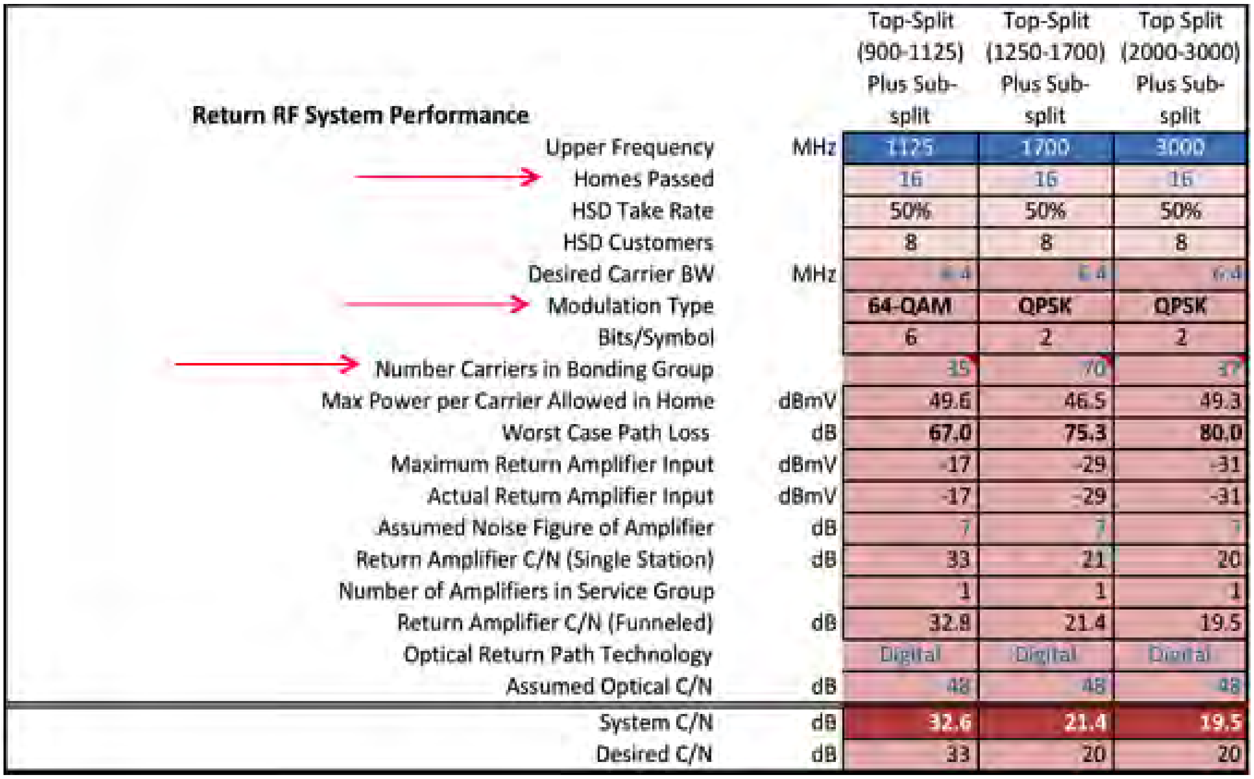
Return RF System Performance		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Homes Passed		500	500	500	250	250	250	250
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	125	125	125	125
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4
Modulation Type		256-QAM	256-QAM	256-QAM	256-QAM	QPSK	QPSK	QPSK
Bits/Symbol		8	8	8	8	2	2	2
Number Carriers in Bonding Group		3.5	10.25	33	73	36	7	7
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	49.8	46.4	49.6	56.5	62.0
Worst Case Path Loss	dB	**29.0**	**30.0**	**34.5**	**43.1**	**67.0**	**75.3**	**80.0**
Maximum Return Amplifier Input	dBmV	31	25	15	3	-17	-19	-18
Actual Return Amplifier Input	dBmV	15	15	15	3	-17	-19	-18
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7
Return Amplifier C/N (Single Station)	dB	65	65	65	53	33	31	32
Number of Amplifiers in Service Group		30	30	30	15	15	15	15
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	41.7	21.0	19.7	20.4
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	45	48
System C/N	dB	**43.9**	**43.9**	**40.5**	**40.8**	**21.0**	**19.7**	**20.4**
Desired C/N	dB	40	40	40	40	20	20	20

Table 12 – 125 HHP Optical SG Top-split Options



Return RF System Performance		Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	1125	1700	3000
Homes Passed		125	125	125
HSD Take Rate		50%	50%	50%
HSD Customers		62.5	62.5	62.5
Desired Carrier BW	MHz	6.4	6.4	6.4
Modulation Type		8-QAM	QPSK	QPSK
Bits/Symbol		3	2	2
Number Carriers in Bonding Group		35	13	4
Max Power per Carrier Allowed in Home	dBmV	49.6	53.9	59.0
Worst Case Path Loss	dB	67.0	75.3	80.0
Maximum Return Amplifier Input	dBmV	-17	-21	-21
Actual Return Amplifier Input	dBmV	-17	-21	-21
Assumed Noise Figure of Amplifier	dB	7	7	7
Return Amplifier C/N (Single Station)	dB	33	29	29
Number of Amplifiers in Service Group		8	8	8
Return Amplifier C/N (Funneled)	dB	23.7	19.7	20.1
Optical Return Path Technology		Digital	Digital	Digital
Assumed Optical C/N	dB	48	48	48
System C/N	dB	23.7	19.7	20.1
Desired C/N	dB	23	20	20

Table 13 – 16 HHP Optical SG Top-split Options



Return RF System Performance		Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	1125	1700	3000
Homes Passed		16	16	16
HSD Take Rate		50%	50%	50%
HSD Customers		8	8	8
Desired Carrier BW	MHz	6.4	6.4	6.4
Modulation Type		64-QAM	QPSK	QPSK
Bits/Symbol		6	2	2
Number Carriers in Bonding Group		35	70	37
Max Power per Carrier Allowed in Home	dBmV	49.6	46.5	49.3
Worst Case Path Loss	dB	67.0	75.3	80.0
Maximum Return Amplifier Input	dBmV	-17	-29	-31
Actual Return Amplifier Input	dBmV	-17	-29	-31
Assumed Noise Figure of Amplifier	dB	7	7	7
Return Amplifier C/N (Single Station)	dB	33	21	20
Number of Amplifiers in Service Group		1	1	1
Return Amplifier C/N (Funneled)	dB	32.8	21.4	19.5
Optical Return Path Technology		Digital	Digital	Digital
Assumed Optical C/N	dB	48	48	48
System C/N	dB	32.6	21.4	19.5
Desired C/N	dB	33	20	20

7 DOCSIS PHY TECHNOLOGIES

7.1 ATDMA & J.83 (Single Carrier QAM)

7.1.1 Potential for Higher Symbol Rate A-TDMA

With the increasing deployment of wideband (6.4 MHz) 64-QAM upstream channels and in some cases bonding of upstream channels, operators are beginning to take advantage of the most powerful set of DOCSIS 2.0 and DOCSIS 3.0 tools available for maximizing capacity of a given channel and delivering higher peak service rates.

Nonetheless, as these advancements have matured – they are 11 years and 6 years since initial release, respectively – the pace of bandwidth consumption and market demand for higher rate service has continued. While it has slowed in the upstream relative to the downstream, it has nonetheless marched forward such that we speak of 10 Mbps and 20 Mbps upstream service tiers today, with an eye towards 100 Mbps in the near future.

The nature of reasonable traffic asymmetry ratios for efficient operation of DOCSIS may pull 100 Mbps along as well as the downstream heads towards a 1 Gbps. Certainly, for DOCSIS-based business subscribers – already outfitted with CMs, for example, or without convenient access to a fiber strand – 100 Mbps is often not just an objective but a requirement.

It is also likely one that operators can derive increased revenue from and consider SLA management options to deliver higher-end services.

7.1.1.1 100 Mbps Residential Upstream

For residential services, while a need for a 1 Gbps service appears far off into the next decade, a 100 Mbps offering is a reasonable target for the near term, and projects as the CAGR-based requirement in 4-6 years for 20 Mbps services today using traffic doubling periods of every two years (approximately 40%) or every three years (approximately 25%).

Unfortunately, today, only through bonding four 64-QAM carriers can 100 Mbps service rate, accounting for overhead loss to net throughput, be provided. The addition of 256-QAM as a modulation profile, to be described in the next section, helps to alleviate this somewhat by enabling a 100 Mbps rate to be offered over three bonded upstreams.

In either case, however, the added complexity of latency of bonding is required to achieve what is expected to be a fundamental service rate target to likely be implemented in bulk. Latency in particular has become a topic generating much interest because of the impact packet processing delay can have on gaming.

While relatively low average bandwidth, high quality gaming demands instantaneous treatment for the fairness and QoE of the gaming audience. Performance has been quantified against latency and packet loss by game type [1], and the variations in performance have led to solution variation exploiting the video architecture, managing server locations, and using potential QoS or priority mapping schemes. While bonding is not the dominant network constraint, elimination of

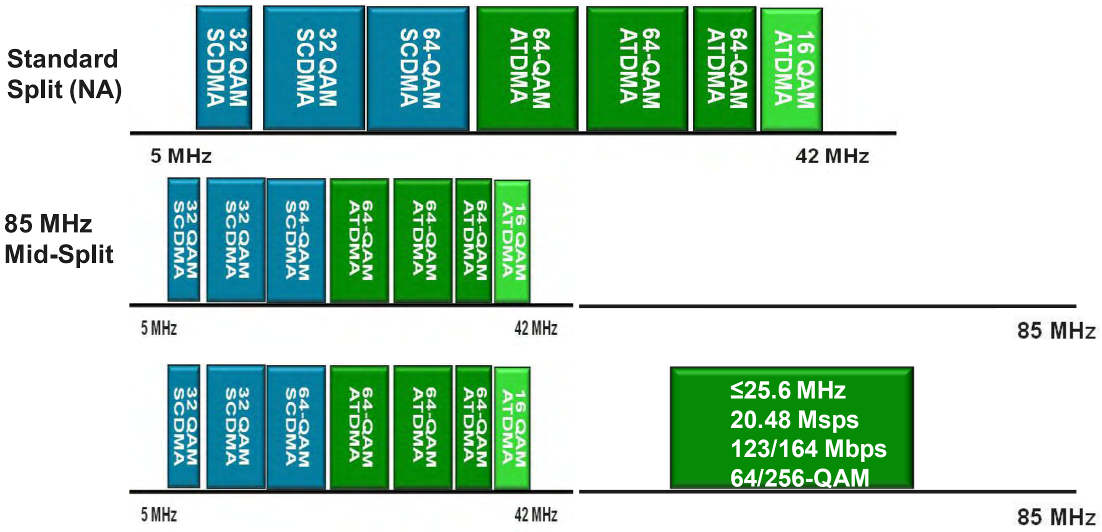


Figure 22 – Higher Symbol Rates Applied Over an 85-MHz Mid-Split Architecture

bonding is favorable for improving processing latency for gaming and other latency-sensitive applications that may arise in the future.

There is also a concern that upstream bonding capability will be limited to a maximum of 8 carriers, due to the increasing complexity associated with the tracking of packets and scheduling operation to process the payload across PHY channels. While operators are not ready to bond even four channels today, if this eight-channel limit were indeed the case, then peak upstream speeds could never exceed 240 Mbps at the PHY transport rate, or 320 Mbps under a 256-QAM assumption.

So, while 1 Gbps of capacity or service rate is likely not a near-term concern, a path to achieve that within the HSD infrastructure should be made available for the long-term health and competitiveness of the network.

Both concerns – 1 Gbps and the bonding implementation for 100 Mbps services – are addressed by a straightforward, integer-scale widening of the symbol rate of today's robust, single-carrier architecture. This approach is shown in Figure 22, where it is displayed as it might be implemented with an 85 MHz Mid-Split architecture. While not obvious from Figure 23, because of the full legacy band, two wider symbol rate channels could be operated within an 85 MHz architecture.

With an excess bandwidth (α) of 15%, there would be a reduced relative bandwidth overhead over today's α = .25. This represents a savings of over 2 MHz of excess bandwidth at 20.48 Msps symbol rates, and two channels would consume less than 48 MHz of spectrum. This leaves plenty of additional spectrum for legacy carriers in a clean part of the lower half of the upstream.

By increasing the maximum symbol rate by a factor of four, from 5.12 Msps to 20.48 Msps, a basic unit of single-carrier operation now is capable of being a 100 Mbps net throughput channel, and simple delivery of this key peak speed service rate is achieved.

7.1.1.2 Achieving 1 Gbps

By bonding eight such carriers together, coupled with the introduction of 256-QAM,

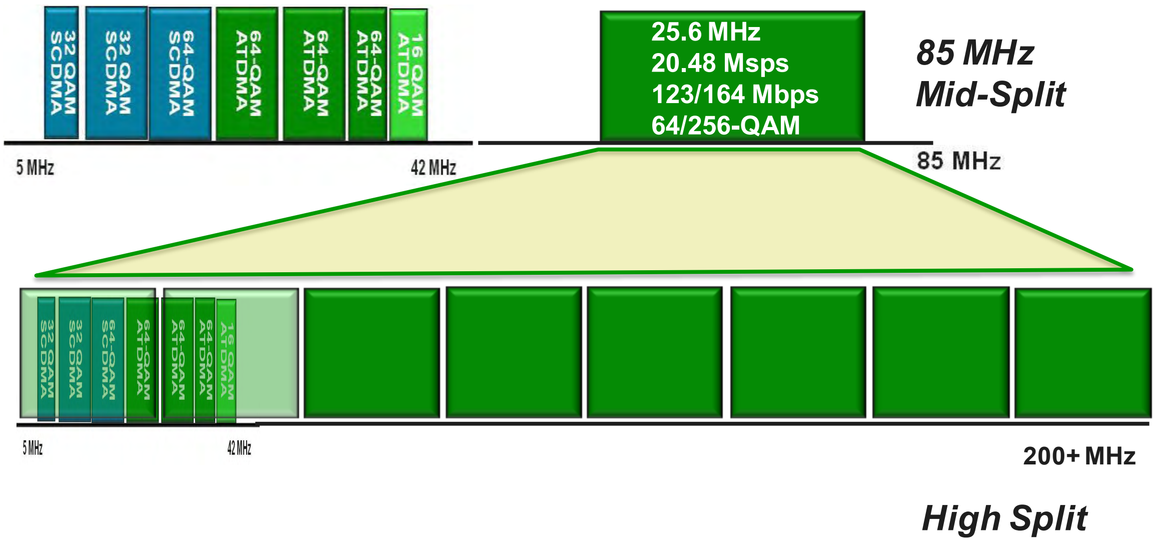


Figure 23 – 8x Bonded Higher Symbol Rates Over a "High-Split" Architecture

an aggregate throughput of over 1 Gbps can also now be enabled with a 4x symbol rate approach, when required. While it is not clear yet if there is an 8-bonded upstream limit, this technique takes that potential risk off of the table. This scenario is shown in Figure 23. In principle, these eight carriers can fit within 200 MHz of spectrum, making the approach comfortably compatible, even with the minimum bandwidth "high-split" spectrum architecture.

In practice, given that legacy services already populate the return path and will only grow between now and any new evolution of the channel or architecture, a high-split based upon a 250 MHz or 300 MHz upstream band is the more likely deployment scenario, with the possibility that it could increase further over time. A flexible FDD implementation would allow the traffic asymmetry to be managed as an operator sees fit based upon need.

7.1.1.3 Wider Band Channel Implications

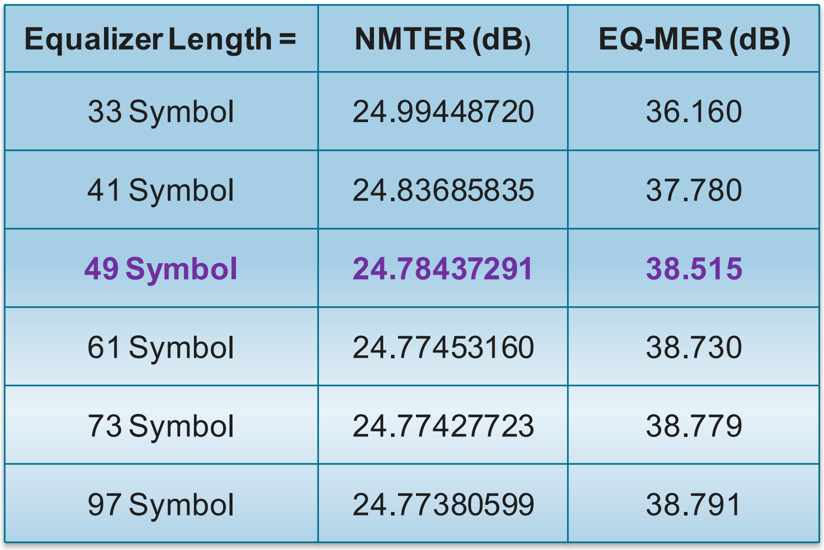
The complexity of DOCSIS 2.0's wideband 64-QAM is largely around the ability to equalize the signal under frequency response distortions. The 24-Tap architecture evolved from the 8-Tap structure of DOCSIS 1.0, providing a very powerful tool for both ISI mediation as well as for plant characterization and diagnostics through the use of the pre-equalization (pre-EQ) functionality.

Every individual CM has its RF channel effectively characterized for reflection content and frequency response distortions, such as roll-off and group delay distortion. Use of pre-EQ has become an immensely powerful tool for MSOs in optimizing their return and efficiently diagnosing and zeroing in on problem locations. Optimization of use has matured and MSOs have learned how best to make use of this powerful tool as wideband 64-QAM has become a critical component of the upstream strategy.

Today's equalizer architecture is also, therefore, quite mature, and the ability to provide real-time processing of burst upstream signals has advanced considerably in the intervening years per Moore's Law as it pertains to processing power. This is important to consider as we ponder higher symbol rates.

Higher symbol rates translate directly to wider channel bandwidths, and thus the equalizer is impacted by this technique. For the T-Spaced implementation of DOCSIS 3.0, if the symbol rate increases by a factor of four, then time span of an equalizer using the same number of taps has *shrunk* by a factor of one-quarter. In other words, the equalizer length must be increased by a factor of four to provide the same span of compensation for micro-reflections, for example.

Table 14 – Post-EQ MER as a Function of Tap Span



Equalizer Length =	NMTER (dB)	EQ-MER (dB)
33 Symbol	24.99448720	36.160
41 Symbol	24.83685835	37.780
49 Symbol	**24.78437291**	**38.515**
61 Symbol	24.77453160	38.730
73 Symbol	24.77427723	38.779
97 Symbol	24.77380599	38.791

Since equalizer taps are a complex multiply operation, it means 16x as many calculations take place in the equivalent algorithm. While this sounds imposing, considering that the 24-Tap structure is over ten years old, a 16x increase in processing is actually well below the “Moore’s Law” rate of compute power capability growth.

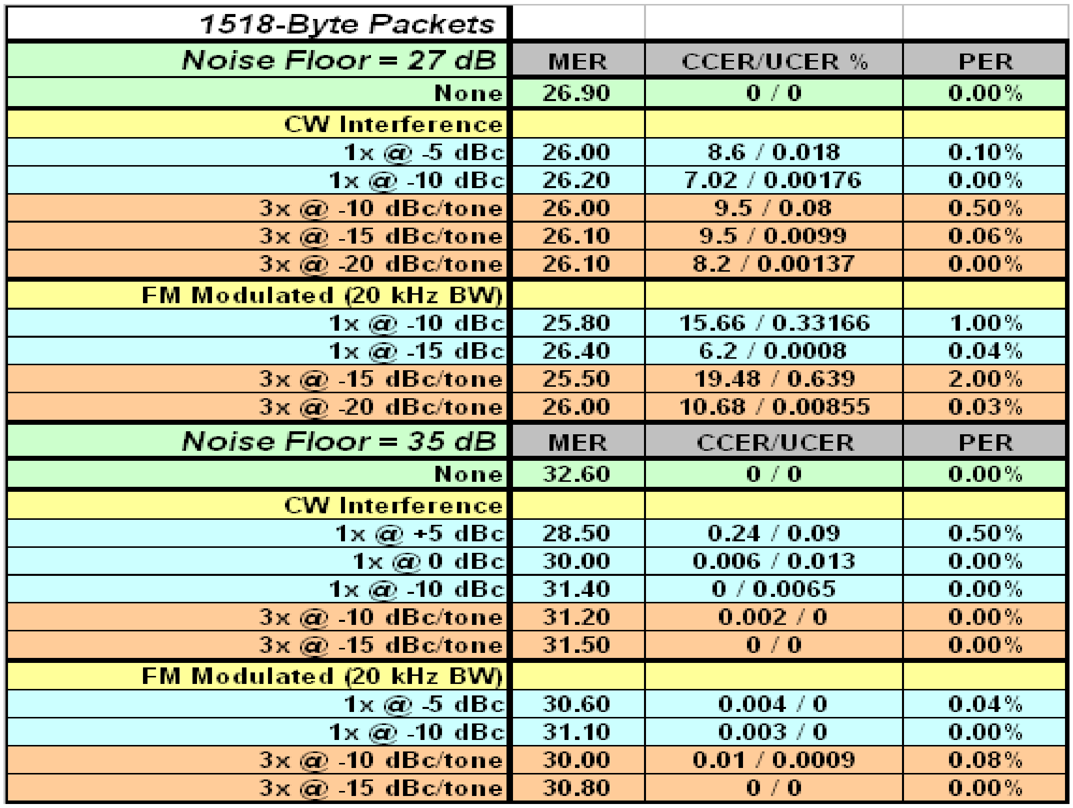
For example, at a doubling of capability even every two years, this would project out to more than 32x the processing power available today than was available when the current equalizer was *deployed*, much less designed. The technology capability to achieve a 96-Tap structure does not appear to be an obstacle, although its fit within modest variations to existing silicon is an important consideration.

There is some evidence that the 4x symbol rate may be a reasonable extension for today’s equalizer architecture to handle. Recent characterization of wideband channels in the > 1 GHz band has shown that the dithering on the last few taps in the equalizer may be minimal for short cascades.

In these environments, spectral roll-off caused by many filters in cascade is limited, as is the group delay impact of this roll-off. Also, fewer connected homes means fewer opportunities for poor RF terminations and the micro-reflections they cause.

Table 14 quantifies test results for a 4x symbol width in an unspecified part of the coaxial band at 1.5 GHz through a cascade of taps in the passive leg of the plant. The frequency response above 1 GHz is generally not specified today. However, this characterization was done with taps with faceplates installed to extend their bandwidth to about 1.7 GHz.

Table 15 – A-TDMA Narrowband Interference Suppression Capability



1518-Byte Packets			
Noise Floor = 27 dB	MER	CCER/UCER %	PER
None	26.90	0 / 0	0.00%
CW Interference			
1x @ -5 dBc	26.00	8.6 / 0.018	0.10%
1x @ -10 dBc	26.20	7.02 / 0.00176	0.00%
3x @ -10 dBc/tone	26.00	9.5 / 0.08	0.50%
3x @ -15 dBc/tone	26.10	9.5 / 0.0099	0.06%
3x @ -20 dBc/tone	26.10	8.2 / 0.00137	0.00%
FM Modulated (20 kHz BW)			
1x @ -10 dBc	25.80	15.66 / 0.33166	1.00%
1x @ -15 dBc	26.40	6.2 / 0.0008	0.04%
3x @ -15 dBc/tone	25.50	19.48 / 0.639	2.00%
3x @ -20 dBc/tone	26.00	10.68 / 0.00855	0.03%
Noise Floor = 35 dB	MER	CCER/UCER	PER
None	32.60	0 / 0	0.00%
CW Interference			
1x @ +5 dBc	28.50	0.24 / 0.09	0.50%
1x @ 0 dBc	30.00	0.006 / 0.013	0.00%
1x @ -10 dBc	31.40	0 / 0.0065	0.00%
3x @ -10 dBc/tone	31.20	0.002 / 0	0.00%
3x @ -15 dBc/tone	31.50	0 / 0	0.00%
FM Modulated (20 kHz BW)			
1x @ -5 dBc	30.60	0.004 / 0	0.04%
1x @ -10 dBc	31.10	0.003 / 0	0.00%
3x @ -10 dBc/tone	30.00	0.01 / 0.0009	0.08%
3x @ -15 dBc/tone	30.80	0 / 0	0.00%

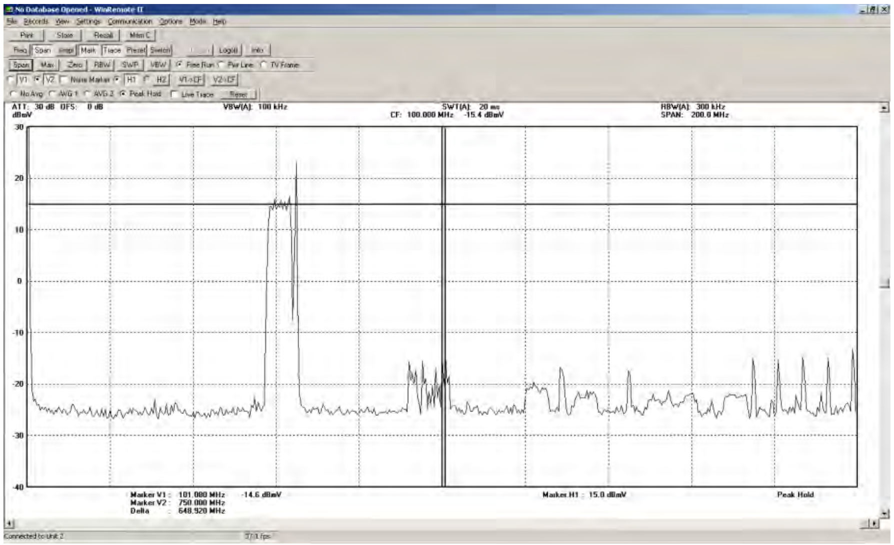
Evident in this essentially “N+0” segment is that the MER after equalization improves only incrementally as we include more taps up to about T = 49 symbols. The T=48 symbols would, of course, mean a doubling of the Tap span for a quadrupling of the symbol rate.

As cascades reduce and new, cleaner upstream bands are used to exploit more capacity, favorable channel condition with respect to frequency response are likely to result. This data certainly is favorable to the thought that even above 1 GHz, where little has been defined for CATV, a 4x symbol rate can be accommodated for the downstream.

Now, switching to the upstream, the spectrum expected to be exploited is in fact well-defined – return loss requirements and all – and will benefit from the same architectural migration shifts to shorter cascades and passive coax architectures. Because of this, the potential complexity increase of a 96-Tap equalizer and the corresponding time span that it supports may not be necessary to effectively use an extended upstream with 4x symbol rate transport. This may be valuable news to silicon implementers who may then be able to allocate silicon real estate and MIPS to other receiver processing functions.

7.1.1.4 Narrowband Interference

Another concern associated with increased symbol rates is the increased likelihood by a factor of four on average (slightly less with less excess bandwidth, of course) that narrowband interference will fall in-band and degrade the transmission. Unlike multi-carrier techniques, which can drop sub-channels out that can become



No Database Opened - WinRemote II
File Records View Settings Communication Options Mode Help
Print Store Recall Mem C
Freq Span Ampl Mark Trace Preset Switch Logoff Info
Span Max Zero RBW SWP VBW
Free Run Pwr Line TV Frame
V1 V2 Noise Marker H1 H2 V1->CF V2->CF
No Avg AVG 1 AVG 2 Peak Hold Live Trace Reset
dBmV
30
20
10
0
-10
-20
-30
-40
Connected to Unit 2
37.1 fps

ATT: 30 dB OFS: 0 dB
VBW(A): 100 kHz
SWT(A): 20 ms
CF: 100.000 MHz -15.4 dBmV
RBW(A): 300 kHz
SPAN: 200.0 MHz

Marker V1 : 101.080 MHz -14.6 dBmV
Marker V2 : 750.000 MHz
Delta : 648.920 MHz
Marker H1 : 15.0 dBmV
Peak Hold

Figure 24 – Observed FM Band Interference on Deliberately Poor CM RF Interface

impaired by such interference (at the expense of throughput), a single carrier system must find a way to suppress the interference and reconstruct the symbol without it.

Fortunately, such techniques have matured, and today's ingress cancellation technology is very powerful in delivering full throughput performance in the face of strong narrowband interference. These processing algorithms sense ingress and adapt the rejection to the location and level of detected interference.

Table 15 quantifies the measured robustness under controlled testing of the DOCSIS 3.0 narrowband interference mechanism in suppressing interference [8]

It is readily apparent that today's DOCSIS 3.0 narrowband incision capability handles in-band interference very effectively over a range of much-worse-than-typical SNR, impulse, and interference conditions.

For example, at an SNR of 27 dB, which represents the return path quality of very old Fabry-Perot return paths long since replaced in most cases (DOCSIS minimum being 25 dB), it takes three tones of 20 kHz bandwidth a piece and adding up to about a 10 dB C/I to register a PER that might be considered objectionable (2%) from a user QoE perspective.

A borderline 1% PER occurs at C/I = 10 dB for a single interferer. These C/I values represent very high levels of plant interference in practice, although not completely uncommon, especially at the low end, shortwave area of the return band.

At SNRs closer to what is expected today (35 dB), no static interference case has PER of any consequence, even with C/I taken to 5 dB (modulated) and -5 dB (unmodulated) tones. This data suggests that wider symbols in the ever-cleaner part of the spectrum are likely to comfortably operate, quite robustly.

As the high-split architecture is deployed, interference levels over the air bands – particularly FM radio in North America, as discussed in Section 3.3.2– become important to understand. Figure 24 shows a field test with a diplex split extended above the 85 MHz Mid-Split for purposes of quantifying the potential for such interference.

In what was a very harsh metropolitan environment, with older plant cabling and nearby FM towers, a deliberately loose fitted CM resulted in relatively modest. However, because it is a wideband spectrum of channels, it would not be able to be compensated for by receiver ingress suppression. The roughly 30 dB of SNR would still yield high throughput, though because the interference effect may have non-Gaussian qualities, the uncorrected error rates may be higher.

However, it is expected this would be well within FEC capability to yield error-free output. Similar C/I's resulted with various arrangements of splitters, modems and deliberately radially and longitudinally damaged cables. While only one example, given the ground conditions, this trial was highly encouraging with respect to the high split running well in the region of spectrum occupied by FM radio over the air.

Note that the ingress-only performance shown in Table 16 in fact identifies a potential *advantage* of the single carrier approach to interference suppression relative to OFDM – there is no loss of available data rate; there is instead an overhead increase for channel knowledge. In OFDM, the C/I on a single sub-channel and closest neighbors, must be removed or have their modulation profile decreased at the cost of available data rate. If the C/I environment worsens however, OFDM can gracefully degrade where SC has threshold behavior.

7.1.1.5 Joint Impairment Thresholds

When impulse noise is added as a joint impairment, we can then begin to count more cases of potentially objectionable PER from a user QoE perspective. However, it is quite clear from the comparison that the error rate is being dictated by the very impulse noise component. This is indeed an area where OFDM would have benefits, much like will be seen with S-CDMA, through the use of longer symbol times to outlast the impulse events.

Of course, impulse noise tends to be restricted to the low end of the return band. Above about 20 MHz, there is little evidence that the joint impairment scenario occurs in a meaningful way to degrade A-TDMA performance. Indeed, where A-TDMA is the most vulnerable is relative to impulse noise. It is left to defend itself only with FEC today, and this has been proven to be sufficient in the vast majority of 64-QAM deployments implemented in the middle to high end of the 42 MHz upstream spectrum.

7.1.1.6 Summary

DOCSIS is currently a predominantly A-TDMA system, and exclusively so in the vast majority of deployment worldwide. As such, a natural and simple extension, with perhaps only minor impact on silicon development, is the increase the symbol rate of the already existing protocol to be better aligned with service on the near-term horizon, but also compatible with the direction of data services requirements for the long term.

Table 16 – A-TDMA Performance with Interference and Impulse Noise

	None - Narrowband Interference Only		Impulse Noise: 4 usec @ 100 Hz			
			-10		-5	
SNR = 35 dB	**MER**	**PER**	**MER**	**PER**	**MER**	**PER**
None	32.60	0.00%	32.30	0.00%	32.30	0.30%
CW Interference						
1x @ -10 dBc	31.40	0.00%	31.30	1.40%	31.20	2.50%
3x @ -15 dBc/tone	31.50	0.00%	31.40	1.50%	31.50	2.80%
3x @ -20 dBc/tone	31.60	0.00%	31.60	1.00%	31.40	2.20%
3x @ -25 dBc/tone			31.70	0.40%	31.60	1.70%
3x @ -30 dBc/tone						
FM Modulated (20 kHz BW)						
1x @ -10 dBc	31.10	0.00%	31.00	0.10%	30.60	3.70%
3x @ -15 dBc/tone	30.80	0.00%	30.60	2.80%	29.90	3.70%
3x @ -20 dBc/tone	31.20	0.00%	31.10	1.70%	31.00	3.50%
3x @ -25 dBc/tone			31.50	0.70%	31.40	2.10%
3x @ -30 dBc/tone						
SNR = 27 dB	**MER**	**PER**	**MER**	**PER**	**MER**	**PER**
None	26.90	0.00%	26.70	0.01%	26.70	0.50%
CW Interference						
1x @ -10 dBc	26.20	0.00%	26.30	0.50%	26.10	1.60%
3x @ -15 dBc/tone	26.10	0.06%	25.90	0.90%	26.10	2.50%
3x @ -20 dBc/tone	26.10	0.00%	26.10	0.50%	26.10	2.50%
3x @ -25 dBc/tone			26.20	0.10%	26.20	1.50%
3x @ -30 dBc/tone						
FM Modulated (20 kHz BW)						
1x @ -10 dBc	25.80	1.00%	25.60	6.00%	25.60	5.00%
3x @ -15 dBc/tone	25.50	2.00%	25.40	5.00%	25.40	6.00%
3x @ -20 dBc/tone	26.00	0.03%	25.90	1.00%	25.80	0.60%
3x @ -25 dBc/tone			26.20	0.20%	26.20	1.70%
3x @ -30 dBc/tone						

While many advances in PHY technology have occurred, the existing signal flow, knowledge base, silicon maturity, and understanding of management of the single carrier approach all favorably weigh in towards working to tweak something that doesn't need outright fixing. Couple this maturity with the ability of single carrier tools to handle the upstream channel environment across the vast majority of the spectrum, creating a higher symbol rate of 4x, as described here, represents a logical, incremental, low-risk step for the transmission system portion of the PHY.

7.1.2 256-QAM Upstream

With the introduction of DOCSIS, cable operators created a specification for high speed data services that was built around the architecture and technology realities of the time – large serving groups of subscribers funneled through deep cascades of amplifiers and onto into a single laser transmitter – typically of the low-cost, low quality, Fabry-Perot variety – and with the anticipation of a lot of unwanted interference coming along for the ride.

The resulting requirements spelled out ensured robust operation under the condition of a 25 dB SNR assumption, among other impairments defined. Robust performance was assured through the use of relatively narrowband, robust modulation formats (QPSK and 16-QAM), a limited number of channels competing for spectrum power , and the ability to use powerful forward error correction.

Now, of course, many of the characteristics that defined the return have changed significantly, and DOCSIS 2.0 took advantage of many of them by calling for support of a 64-QAM modulation profile of up to twice the bandwidth if conditions allowed it.

It was not the case everywhere that it could be supported, but all phases of evolution were trending towards the ability to squeeze more and more capacity out of the return. Better, Distributed Feedback (DFB), analog optics became cost effective, digital return optics came on the scene, cascades shortened as serving groups shrunk during node splitting operations, and lessons learned over the years brought improvements in return path alignment and maintenance practices.

These same lessons brought about the introduction of S-CDMA, based on a better understanding of the characteristics of the low end of the return spectrum.

DOCSIS 2.0 itself is now over ten years old. DOCSIS 3.0 subsequently added channel bonding for higher peak speeds, as well as calling our support for return path extension in frequency up to 85 MHz.

Fortunately, the HFC architecture and supporting technology has continued to evolve favorably towards more upstream bandwidth, used more efficiently. In Section 2, the case was made for the use of the 85 MHz mid-split as an excellent first step for cable operators looking to add essential new bandwidth for upstream services. In this section, we will show how today's return paths, extended to 85 MHz, are now capable of exploiting this band while also increasing the modulation profile to 256-QAM. It is within the capability of the upstream and demonstrably proven in the field that a 256-QAM modulation profile can be supported, and over a wider band than the legacy 42 MHz bandwidth in North America and the 65 MHz Euro split.

7.1.2.1 Upstream Link Analysis

While early generation CMTS equipment was designed to support 16-QAM as the maximum modulation profile, vendors generally provided enough margin in their systems to enable 64-QAM once networks evolved towards better HFC optics. 64-QAM was subsequently embraced in DOCSIS 2.0.

In Figure 17 through Figure 20 in Section 5, we introduced noise power ratio (NPR) curves to characterize return path optical technologies. NPR curves have the desirable feature of representing a worst-case (no TDMA operating) fully loaded return link from a signal stimulus standpoint while simultaneously quantifying the SNR and S/(N+D) on a single curve.

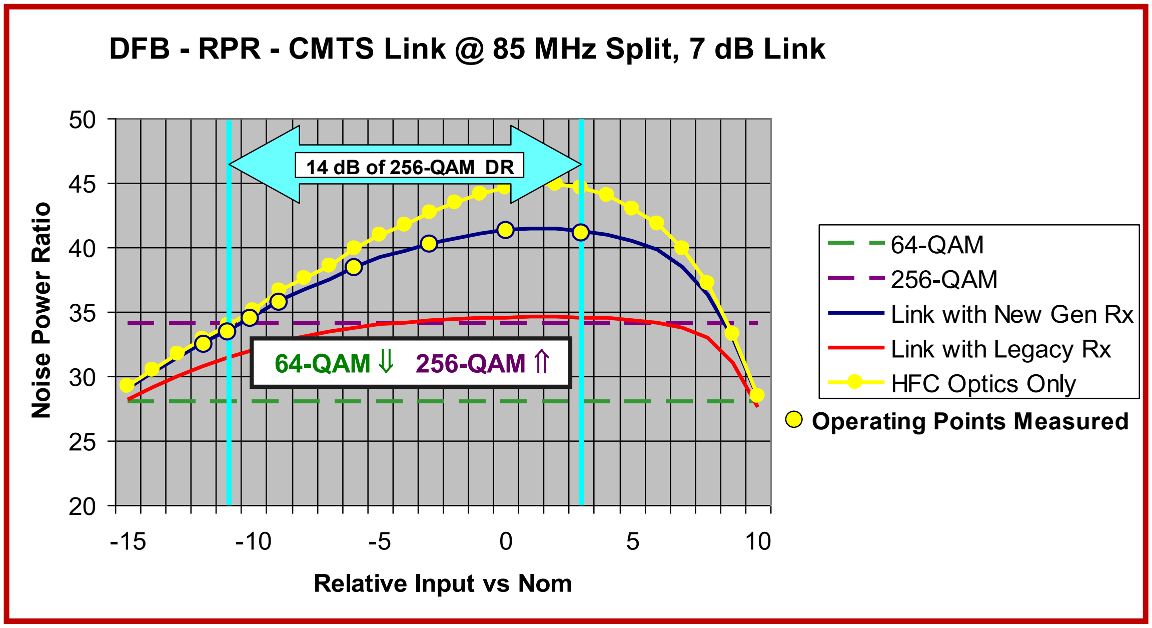


Figure 25 – HFC DOCSIS System Performance

In the NPR curves shown in this section, the optical performance will be augmented with other contributors to the link SNR – in particular RF contributions in the form of noise funneling previously discussed, and receiver noise figures associated with receivers, such as DOCSIS CMTS front ends. We will consider "legacy" DOCSIS receiver – designed originally for 16-QAM maximum profiles, and modern receivers aimed at higher sensitivity for better modulation efficiency.

Consider Figure 25. The red curve marks the performance characteristics of and HFC+CMTS link for legacy-type receivers optimized for 16-QAM and a DFB-RPR link of nominal length under an assumption of 85 MHz of spectrum loading. Clearly, it shows margin over and above the (green) 64-QAM threshold (chosen at 28 dB – an uncorrected 1e-8 error rate objective).

DFB HFC optics plus most of today's CMTS receivers comfortably support 64-QAM with sufficient, practical, operating dynamic range. This lesson is being proven everywhere DOCSIS 3.0 is being deployed. In some cases newer, high quality FP lasers can support 64-QAM as well. While DFBs are recommended for upstream as new channels are added and profiles enabled, it is comforting to realize that newer FPs can get 64-QAM started while the large task of exchanging lasers methodically takes place.

Though legacy receiver exceeded their original design requirements in being extended to 64-QAM (with the help of plant upgrades), enabling 256-QAM design margin – an additional 12 dB of performance over 16-QAM – was not cost effective to consider in early stages of DOCSIS.

As a result, there is zero margin to run 256-QAM (purple), as shown in Figure 25, or otherwise insufficient margin if we aid the factor in more power-per-Hz by limiting the bandwidth to the 65 MHz Euro split by comparison (about 1 dB higher peak) or the 42 MHz split (about 3 dB higher peak).

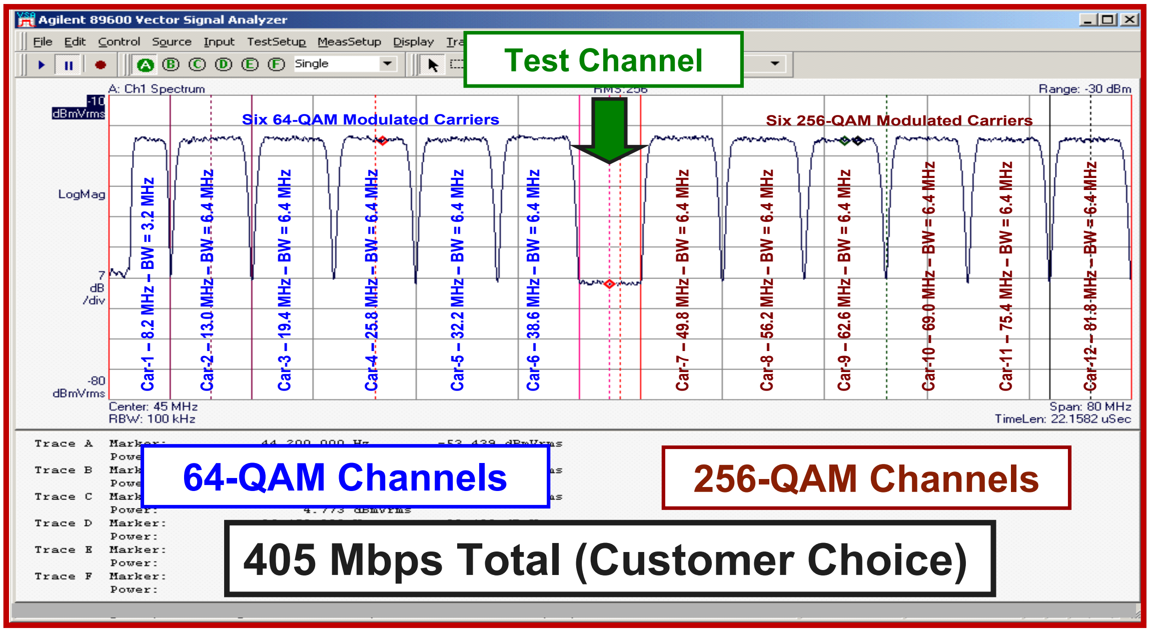

Agilent 89600 Vector Signal Analyzer
File Edit Control Source Input TestSetup MeasSetup Display Tra
A B C D E F Single
Test Channel
A: Ch1 Spectrum
RMS:256
Range: -30 dBm
-10 dBmVrms
LogMag
7 dB /div
-80 dBmVrms
Six 64-QAM Modulated Carriers
Six 256-QAM Modulated Carriers
Car-1 – 8.2 MHz – BW = 3.2 MHz
Car-2 – 13.0 MHz – BW = 6.4 MHz
Car-3 – 19.4 MHz – BW = 6.4 MHz
Car-4 – 25.8 MHz – BW = 6.4 MHz
Car-5 – 32.2 MHz – BW = 6.4 MHz
Car-6 – 38.6 MHz – BW = 6.4 MHz
Car-7 – 49.8 MHz – BW = 6.4 MHz
Car-8 – 56.2 MHz – BW = 6.4 MHz
Car-9 – 62.6 MHz – BW = 6.4 MHz
Car-10 – 69.0 MHz – BW = 6.4 MHz
Car-11 – 75.4 MHz – BW = 6.4 MHz
Car-12 – 81.8 MHz – BW = 6.4 MHz
Center: 45 MHz
RBW: 100 kHz
Span: 80 MHz
TimeLen: 22.1582 uSec
Trace A Marker: 44.200 000 Hz -53.439 dBmVrms
Trace B
Trace C Power: 4.773 dBmVrms
Trace D Marker: Power:
Trace E Marker: Power:
Trace F Marker: Power:
64-QAM Channels
256-QAM Channels
405 Mbps Total (Customer Choice)

Figure 26 – Mid-Split Channel Loading

New receivers, however, provide a higher fidelity upstream termination in order to support 64-QAM with margin and S-CDMA synchronization. Because of these requirements and the continued advances in performance of DFB return optics (higher power laser transmitters), 256-QAM can now be comfortably supported.

The performance of the combined HFC+CMTS link for modern receivers is shown in the blue curve of Figure 25. DOCSIS does not yet call out 256-QAM, although this is a change currently in process.

However, much of the existing silicon base already supports this mode. Note that the yellow points on the blue curve represent points measured in the field that achieved low end-of-line packet error rate performance, as a way of verifying the predicted dynamic range on a real HFC link (NPR would be an intrusive measurement).

Note also that the dynamic range supported for 256-QAM is nearly the same dynamic range that existing receivers provide for 64-QAM – an indication of the robustness potential for 256-QAM links.

Finally, comparing the HFC (yellow) NPR trace to the HFC+CMTS (blue) trace, it is apparent also how little loss of NPR is incurred by new high fidelity CMTS receivers.

Figure 26 shows a snapshot of a recent trial of an Mid-Split architecture, where the upper half of the band was used to support 256-QAM channels, but with all signals at the same power level except for the lowest frequency (narrower) channel. A mid-band test channel was left unoccupied for monitoring the most probable location of maximum distortion build-up as dynamic range was exercised.

Evident from Figure 26 is the high available SNR delivered by the HFC link using existing analog DFB return optics at nominal input drive. The available SNR as measured at the input to the CMTS receiver is about 45 dB. In this case, the tested link was an N+3 architecture.

Table 17 shows a full 85 MHz optimization, using 12 carriers of both S-CDMA and A-TDMA, employing modulations from 32-QAM to 256-QAM across the band. The results indicate a maximum of nearly 400 Mbps of Ethernet throughput under the packetized traffic conditions used.

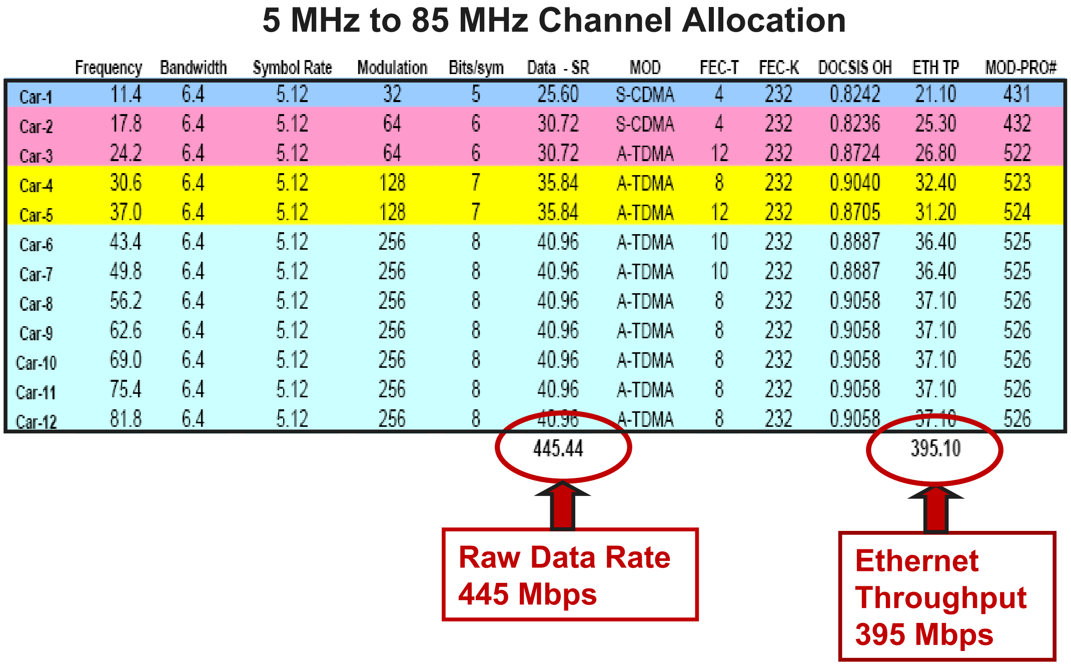
7.1.2.2 *Extended HFC Performance*

To show the robustness potential of 256-QAM upstream, we can extend the performance calculations in Figure 25 to include longer HFC links and the contribution of potentially long RF cascades summed together, resulting in the "noise funnel" aggregation of amplifier noise figures.

The cases shown in Figure 27 assumes a deep cascade (N+6) in a 4-port node, and thus 24 amplifiers summed, and optical links of 7 dB and 10 dB. While the yellow curve still represents 7 dB optics only, both 7 dB and 10 dB links are shown with the RF cascade included (dashed), and then each of the same with the CMTS receiver contribution included (solid).

The loss due to an analog optical link length is very predictable, as the optical receiver SNR drops as input light level drops. The RF cascade can be shown to create the effect of pushing the performance peak down, reflecting the SNR contribution of amplifier noise to the optical link. However, its effect on the dynamic range for supporting 256-QAM is negligible.

The stronger dynamic range effect is the extended optical link of 10 dB, which ultimately reduces 256-QAM dynamic range by about 2 dB, but with the dynamic range still showing a healthy 11 dB of robust wiggle room.



5 MHz to 85 MHz Channel Allocation

	Frequency	Bandwidth	Symbol Rate	Modulation	Bits/sym	Data - SR	MOD	FEC-T	FEC-K	DOCSIS OH	ETH TP	MOD-PRO#
Car-1	11.4	6.4	5.12	32	5	25.60	S-CDMA	4	232	0.8242	21.10	431
Car-2	17.8	6.4	5.12	64	6	30.72	S-CDMA	4	232	0.8236	25.30	432
Car-3	24.2	6.4	5.12	64	6	30.72	A-TDMA	12	232	0.8724	26.80	522
Car-4	30.6	6.4	5.12	128	7	35.84	A-TDMA	8	232	0.9040	32.40	523
Car-5	37.0	6.4	5.12	128	7	35.84	A-TDMA	12	232	0.8705	31.20	524
Car-6	43.4	6.4	5.12	256	8	40.96	A-TDMA	10	232	0.8887	36.40	525
Car-7	49.8	6.4	5.12	256	8	40.96	A-TDMA	10	232	0.8887	36.40	525
Car-8	56.2	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-9	62.6	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-10	69.0	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-11	75.4	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-12	81.8	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
						445.44					395.10	

Raw Data Rate 445 Mbps

Ethernet Throughput 395 Mbps

Table 17 – Optimized 85 MHz Mid-Split Channel Loading

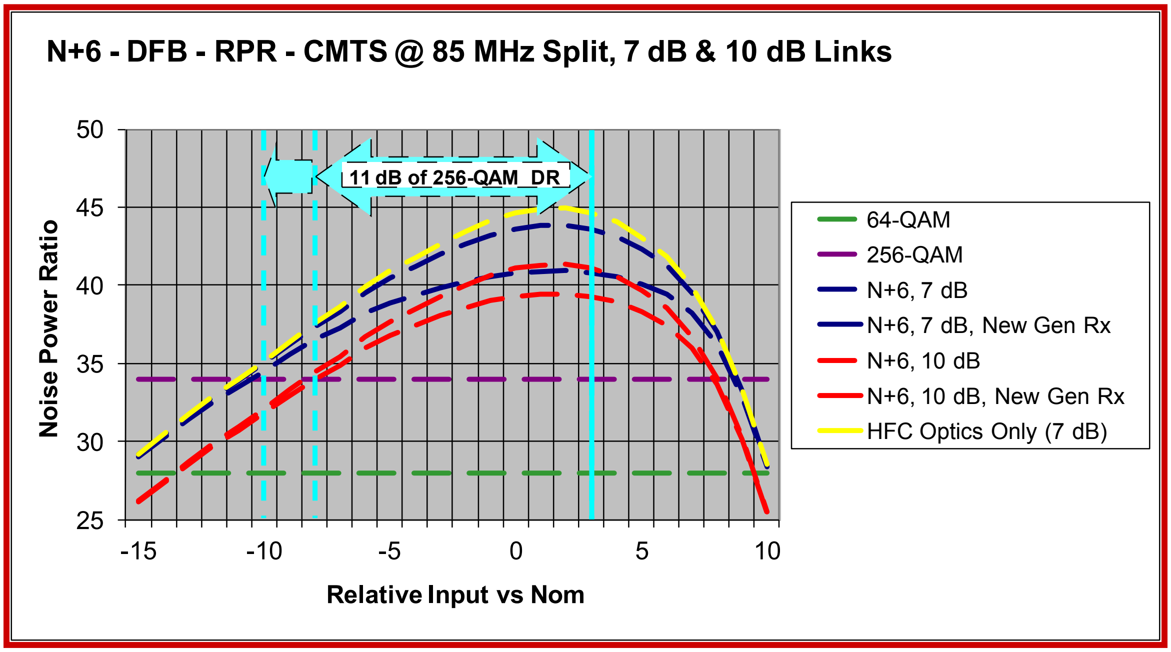


Figure 27 – HFC DOCSIS System Performance for Longer RF Cascades

7.1.2.3 Extended "High-Split" Bandwidth Projection

A 1 Gbps capacity threshold upstream requires the split to move to 200 MHz or higher. The 5-200 MHz bandwidth itself supports well over 1 Gbps of theoretical capacity, but legacy use may not make the full spectrum available for higher efficiency, and overhead loss will decrease transport capacity to a lower net throughput.

A higher spectrum diplex will likely therefore be required. However, we quantify the 200 MHz case because of its potential compatibility with current equipment outfitted with 200 MHz RF hybrids, or with minor modifications thereof.

Figure 28 is the analogous figure to Figure 25 for 85 MHz Mid-Split, showing, in this case, projected performance on a 200 MHz "high" split when factoring in an "equivalently performing" CMTS receiver (DOCSIS does not extend to 200 MHz) and DFB optics performing at today's noise density (adjusted only for power loading).

As would be expected, with the receiver performance equivalent to legacy CMTS receivers, inherently not equipped for 256-QAM, performance does not even breach the threshold. However, with a new generation of high fidelity receivers, system analysis projects that there exists 10 dB of dynamic range to 256-QAM performance over a fully loaded 200 MHz return path.

This would see degradation when RF amplifiers are included, but again to minor effect on dynamic range. Conversely, it is anticipated that by the time the need for high split is required, very small serving groups have already been established, leading to a much less significant noise funnel.

While dynamic range (10 dB) is still relatively high, there is observable loss of peak above the 256-QAM threshold, meaning much of the dynamic range exists over a relatively low steady-state operating margin. This could make the link more

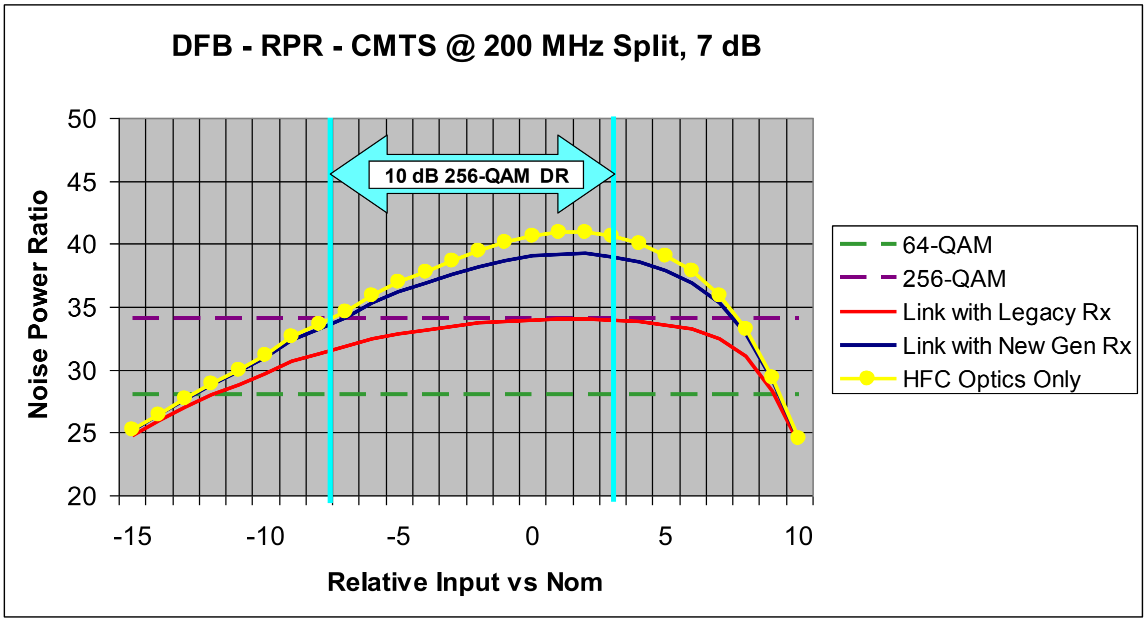


Figure 28 – HFC-DOCSIS System Performance using 200MHz "High Split"

susceptible to moderate transients, drift, temperature extremes, or misalignment, and thus require more regular maintenance.

As such, Figure 28 points out the near term potential for high split operation over HFC optics, but also indicates that performance improvements over time will be welcome to ensure robust operations. Also, note that measured performance for a high split return to 185 MHz, shown in Figure 20, is similar to the analysis in Figure 28. In fact, measured performance of the 1550 nm DWDM return in Figure 20 is slightly better (by about 1.5 dB) than the extrapolated performance in Figure 28 using a standard 1310 nm DFB, pointing out additional margin for the high split case already existing today.

7.1.2.4 Modem Performance Characterization Findings

Recent results [17] have evaluated 256-QAM transmission in the presence of narrowband interference to assess the capability of the ingress suppression capability for the higher order of modulation. Table 18 quantifies these results in terms of Codeword Errors (CCER, UCER) and Packet Errors (PER) as are calculated and made available in the DOCSIS MIB.

Results for 64-QAM were shared, along with results for 256-QAM, in [16]. However, Table 18 updates the results for 256-QAM with a more robust performance assessment using higher performance recovers for the proper SNR baseline. This is simply mirroring what was already described and identified in Figure 25 – legacy DOCSIS receivers do not have acceptable margin to run a robust 256-QAM profile.

Nonetheless, it is difficult to make apples-to-apples ingress suppression comparisons, as the SNR margin for 64-QAM offers inherently 6 dB more room for the ingress cancellation to operate under than 256-QAM.

The DFB-RPR link in Table 18 was setup to provide higher SNR than the 64-QAM case in [16] in order than a very low

Table 18 – 256-QAM Interference Performance Low PER Thresholds

	256-QAM				
	Level (dB, dBc)	**UNCORR%**	**CORR%**	**PER%**	**MER (dB)**
Baseline - AWGN	**36**	0.000%	0.000%	0.000%	37
Single Ingressor Case					
QPSK 12kHz 0.5%	3	0.254%	0.435%	1.060%	34
QPSK 12kHz 1.0%	1	0.447%	0.944%	2.300%	34
FSK 320ksym/s 0.5%	29	0.278%	0.032%	0.110%	35
FSK 320ksym/s 1.0%	27	0.633%	0.230%	0.810%	35
FM 20kHz 0.5%	2	0.128%	0.295%	0.750%	34
FM 20kHz 1.0%	1	0.187%	0.554%	1.260%	34
Three Ingressor Case					
CPD 0.5%	28	0.297%	0.041%	0.190%	34
CPD 1.0%	27	0.698%	0.144%	0.750%	33

BER threshold in each was a baseline. However, it was not the same absolute margin of the M-QAM to the SNR of the link (5 dB vs 2 dB). It did lead to a very important conclusion, however.

With this low BER steady state case in [8] for 256-QAM, for nearly equivalent relative performance (6 dB difference) for nominal single-interference cases was observed. However, for multiple interferers and for wideband (100's of kHz) there was still substantially more robustness in the case of 64-QAM. Refer to [8] for full details.

Overall, proof of the functionality of ingress cancellation was achieved for 256-QAM, but with degraded performance when the channel is at its noisiest. Of course, the strategy for deploying 256-QAM is to place in the clean part of the upstream, where it can be supported – above 25 MHz. And, certainly consider it to extract capacity in the 85 MHz Mid-Split case above 42 MHz.

This is the approach used to "optimize" the 85 MHz band and shown in Table 18 – a mixture of 256-QAM, 128-QAM, 64-QAM, and S-CDMA based 64-QAM and 32-QAM.

This is the upstream line-up that led to the 445 Mbps transport rate proof of concept reported in [12].

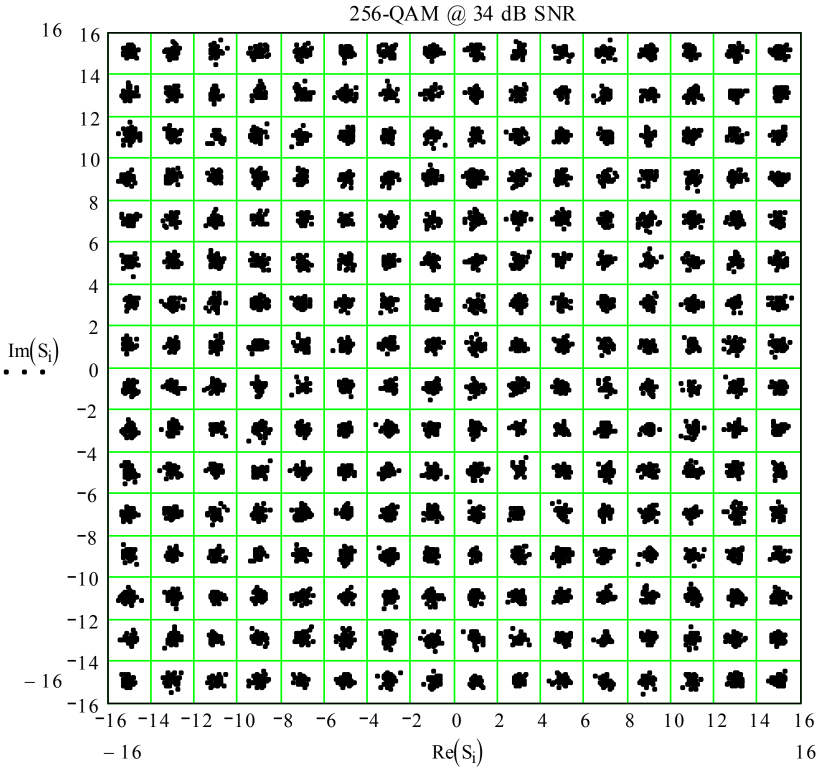


Figure 29 – 256-QAM @ 34 dB SNR

7.1.3 1024-QAM Downstream

In Section 9.5 "Downstream Capacity", we will calculate the downstream capacity for a fully digitized forward band, multiplying the number of 6 MHz slots by the modulation profile allowed by DOCSIS (256-QAM) to arrive at data capacities for 750 MHz, 870 MHz, and 1 GHz networks. We then calculated the case for a Next Generation PHY using LDPC and OFDM, making the reasonable assumption that by updating the FEC, we can achieve two QAM orders of modulation higher in bandwidth efficiency, which effectively suggests 6 dB can be gained.

However, not all of this may be in the FEC (depending on code rate). Some incremental link budget dB may be obtained through some of the business-as-usual operations of fiber deeper and cascade reduction, which reduces noise and distortion accumulation, and through the conversion of analog carriers to digital, which reduced (2x analog + digital) composite carrier-to-noise (CCN) distortion effects. Lastly, newer STBs in the field tend to higher sensitivity (lower noise figure).

Because of this, the FEC is not left to make up all of the dB between 256-QAM and 1024-QAM. And, in fact, it is now possible to make a case based only on these HFC changes that 1024-QAM may be possible in evolved architectures today, even without the addition of new FEC on silicon that can support this QAM mode. This offers the potential for 25% more bandwidth efficiency. This section quantifies this potential.

Let's begin the discussion with the use of QAM over HFC for downstream video as it has evolved to date.

The cable plant has kept up with the bandwidth consumption by adding RF bandwidth and using efficient digital modulations to mine the capacity effectively and with robustness. What started as 64-QAM digital signals became yet more bandwidth efficient with the deployment of 256-QAM downstream, the dominant QAM approach today. The ability to successfully deploy such schemes is due to the very high SNR and very low distortion downstream.

This was to ensure proper conditions for supporting much less robust analog video. In addition to high linearity and low noise, the downstream channel has a flat frequency response on a per-channel basis, minimizing both amplitude and phase distortion, although it can be prone to reflection energy.

As a simple example of the possibilities, the theoretical capacity of a 6 MHz channel with a 40 dB SNR is approximately 80 Mbps. Yet, for J.83-based 256-QAM, the transmission rate is only about 40 Mbps. When accounting for overhead, there is even less throughput.

The next higher order, square-constellation, modulation is 1024-QAM. This technique achieves an efficiency of 10 bits/symbol, or another 25% efficiency over 256-QAM, and an impressive 67% improvement relative to 64-QAM. To support 1024-QAM, a more stringent set of specifications must be met.

Analysis was performed to identify implications to the plant and its performance requirements for robust downstream transmission [1]. The analysis quantified SNR, beat distortion interference, and phase noise, and interpreted the results. We summarize the problem statement here and describe the conclusions.

7.1.3.1 SNR

Let's consider the implications of 1024-

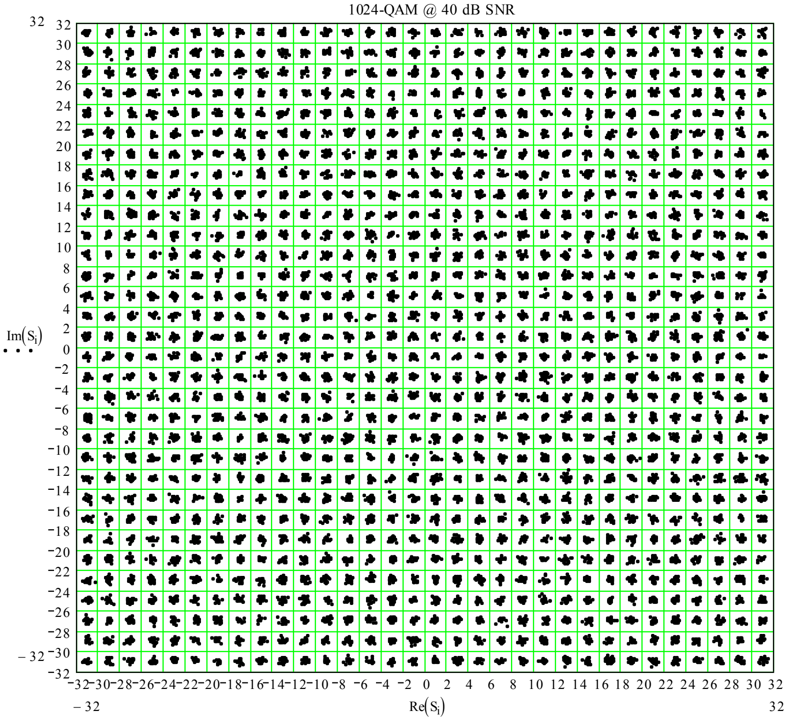


Figure 30 – 1024-QAM @ 40 dB SNR

Table 19 – Power Loading Effects of Analog Reclamation - 870 MHz

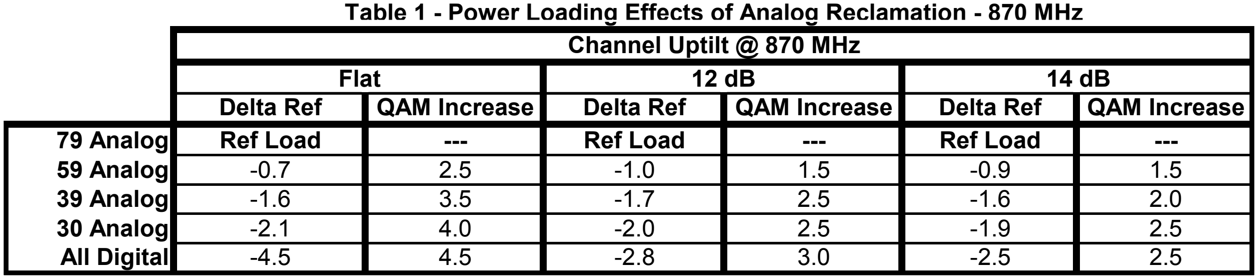


Table 1 - Power Loading Effects of Analog Reclamation - 870 MHz

	Channel Uptilt @ 870 MHz					
	Flat		12 dB		14 dB	
	Delta Ref	QAM Increase	Delta Ref	QAM Increase	Delta Ref	QAM Increase
79 Analog	Ref Load	---	Ref Load	---	Ref Load	---
59 Analog	-0.7	2.5	-1.0	1.5	-0.9	1.5
39 Analog	-1.6	3.5	-1.7	2.5	-1.6	2.0
30 Analog	-2.1	4.0	-2.0	2.5	-1.9	2.5
All Digital	-4.5	4.5	-2.8	3.0	-2.5	2.5

QAM. Figure 29 and Figure 30 show constellation diagrams of 256-QAM @ 34 dB SNR and 1024-QAM @ 40 dB SNR. Being 6 dB apart, these are equivalent uncorrected error rate cases (@1E-8). The congested look of the 1024-QAM diagram, emphasized by the small symbol decision regions, signals the sensitivity this scheme has to disturbances.

Now consider what 40 dB means in terms of use on the plant. For an end-of-line 46 dB of plant (analog) CNR, QAM SNR becomes 40 dB when backed off by 6 dB. We've thus removed virtually all link available margin under an objective of 1E-8, and are now into a region of measurable errors, relying on FEC to finish the job under even the most benign circumstance of thermal noise only.

On the STB side, there is similar margin-challenged mathematics. For a STB noise figure of 10 dB, and for QAM signals arriving at the STB at the low end of the power range, some simple math shows the following:

- Residual Thermal Noise Floor: -58 dBmV/5 MHz
- STB Noise Figure, NF = 10 dB: -48 dBmV/5 MHz
- Analog Level into STB: 0 dBmV
- Digital Level into STB: - 6 dBmV
- STB SNR contribution: -6 -(-48) = 42 dB

Note that NF = 10 dB is not a technically difficult performance requirement. However, in practice, given the cost sensitivity of CPE equipment and without a historical need to have better RF sensitivity, 10 dB and higher is quite normal.

The combined link delivers an SNR of about 38 dB. This simple example leads to the conclusion that existing conditions and existing deployment scenarios create concerns for a seamless 1024-QAM roll-out under a "J.83"-type PHY situation. It reveals the necessity of at least 2 dB of coding gain to ensure robust link closure.

Improving the noise performance of CPE is of course one option to enable more bandwidth efficient link budgets, particularly as yet more advanced modulation profiles beyond 1024-QAM are considered. The sensitivity of CPE cost and the existing deployment of 1024-QAM capable receivers and current noise performance, however, leads to a desire to remain conservative in the expectation of CPE performance assumptions.

7.1.3.2 Favorable Evolution Trends

A couple of favorable trends are occurring in HFC migration that potentially free up some dB towards higher SNR of the

Table 20 – Noise and Distortion @ 550 MHz vs Analog Channel Count

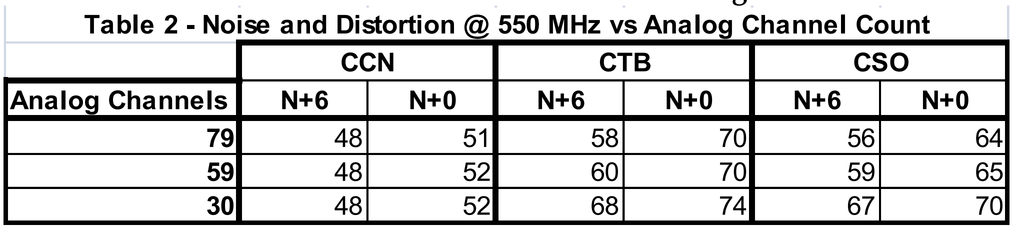


Table 2 - Noise and Distortion @ 550 MHz vs Analog Channel Count

	CCN		CTB		CSO	
Analog Channels	N+6	N+0	N+6	N+0	N+6	N+0
79	48	51	58	70	56	64
59	48	52	60	70	59	65
30	48	52	68	74	67	70

QAM channels – analog reclamation and cascade shortening.

Table 19 shows the potential for higher SNR by taking advantage of the RF power load when compared to a reference of 79 analog channels for 870 MHz of forward bandwidth. In the table, the left hand column for each case – Flat, 12 dB tilt, 14 dB tilt – represents the decrease in total RF load compared to the 79-analog channel reference. The right column for each case represents how much more power could be allocated to each digital carrier in order to maintain the same total RF power load. This is the potential available theoretical SNR gain.

The flat case represents the effect on the optical loading of the analog reclamation process. There is headroom that can be exploited in the optical link and RF cascade by increasing the total power of the analog plus digital multiplex, gaining SNR for all channels and offering potential mediation against the 6 dB increased SNR requirement.

The SNR discussion above refers only to the improvement relative to the thermal noise floor. The additional distortion component (composite inter-modulation noise or CIN) and practical RF frequency response means not all of the theoretical dB will be realized (refer to [1] for details).

Now consider Table 20 quantifying modeled performance for a sample HFC link under different assumptions of line-up and cascade. The data underscores the impact on noise and distortion of decreasing analog channel loads and shorter RF cascades. CCN represents Composite Carrier-to-Noise – a combination of the CNR or SNR and digital distortion products.

Moving across rows, noise and distortion improvements associated with the elimination of the RF cascade (N+6 to N+0) is clear. Moving down columns, the benefits of doing analog reclamation also becomes clear. Both activities enable the network to more ably support higher order modulation SNR performance requirements.

From the perspective of noise (CCN), shortening of the cascade reduces the accumulation of amplifier noise, freeing up 3-4 dB additional SNR available relative to a typical line-up and cascade depth of today. When coupled with possible loading adjustments with the larger digital tier and new headroom available – a few dB here and a few dB there approach – we can come close to 6 dB of new SNR as we evolve the network and use the gains to our benefit. This is, of course, the amount of increased SNR sensitivity of 1024-QAM compared to 256-QAM.

Table 21 – Inner (5/6) LDPC Coded M-QAM Throughput and Comparison to J.83 [2]

Mode	Efficiency (bits/symbol)	Representative Symbol Rate (Msps)	Representative Inner Code Bit Rate (Mbps)	TOV Es/No (dB)	Delta from Capacity* (dB)
Proposed 64QAM	5.333	5.056	26.96	18.02	0.52
Proposed 256QAM	7.333	5.361	39.31	24.27	0.46
Proposed 1024QAM	9.333	5.361	50.03	30.42	0.50
J83.B 64QAM	5.337	5.056	26.97	20.75	3.25
J83.B 256QAM	7.244	5.361	38.84	26.90	3.44
“J83.B” 1024QAM	9.150	5.361	49.05	33.03	3.80

*Note that “Capacity” in this case is an abbreviation for *Constrained* Capacity, as opposed to Shannon Capacity. For this example, the constraint is a symbol set of uniformly distributed QAM symbols. Please refer to above text and [2] for details.

7.1.3.3 Modern FEC

So far, we have considered only existing FEC with 1024-QAM, relying on HFC migration phases to extract additional dB from the plant to create sufficient operational margin. Fortunately, we are not limited to legacy error corrections schemes. While powerful in its day, concatenated Reed-Solomon FEC used in J.83 is now roughly 15 years old – an eternity in information theory technology development. While J.83 leaves us several dB from theoretical PHY performance, modern FEC, typically built around Low Density Parity Check (LDPC) codes – also concatenated to avoid error flooring – achieves performance within fractions of dB of theoretical.

A proposal made during DOCSIS 3.0 discussions [2] quantified additional gains available using LDPC for current 64-QAM and 256-QAM systems, as well as for potential 1024-QAM use. Table 21 summarizes some of the core findings of that system design. The analysis references a common Threshold of Visibility (TOV) threshold for video of 3e-6 and compares constrained capacity (limited to QAM signal sets) of the various profiles. This constraint has an inherent offset from Shannon capacity that grows as a function of SNR.

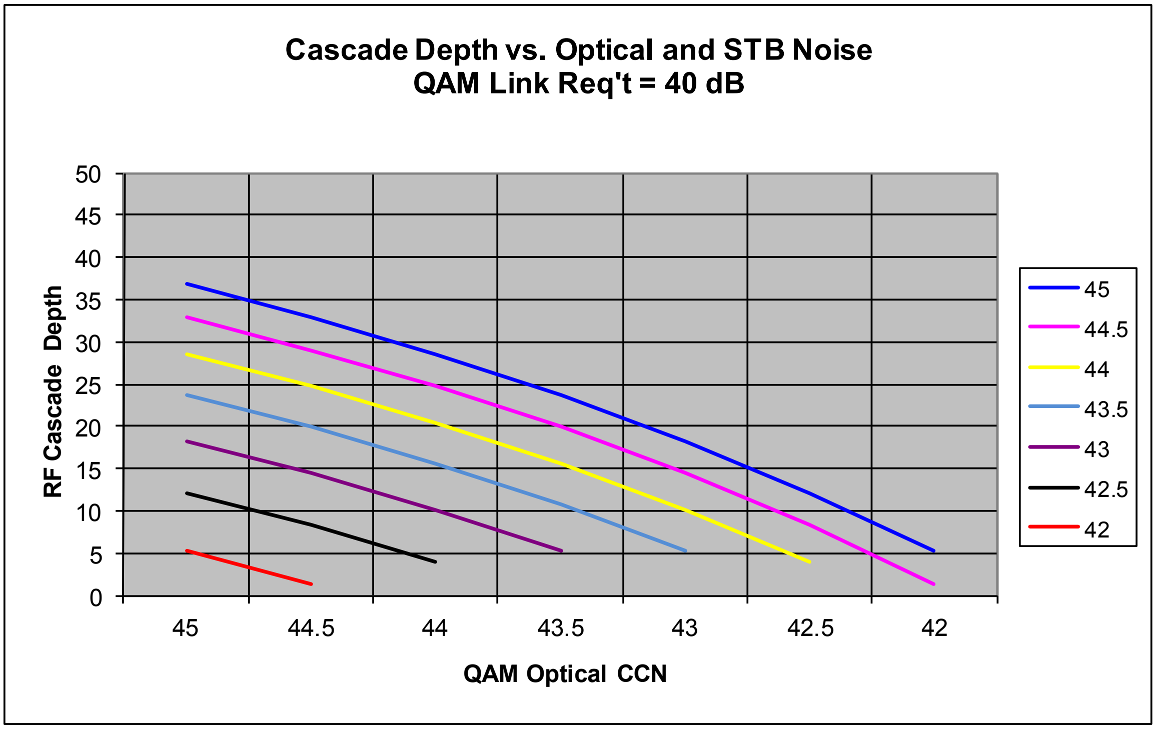


Figure 31 – 1024-QAM, Noise, and Cascade Depth – 40 dB Link Requirement

With the recognition of another 3.3 dB of coding gain, the proposal pointed out the accessibility of 1024-QAM for the downstream channel in a legacy 6 MHz format. This constraint (6 MHz) can also be removed for wider band channels, leading to more flexibility in code design and thus more available coding gain. However, we will see below that even just assuming a modest 3 dB more coding gain provides very meaningful SNR margin for robust 1024-QAM.

We can now execute architecture trade-offs of noise contributions and the depth of the RF cascade to evaluate support for 1024-QAM. HFC cascade thresholds are shown in Figure 31 and Figure 32, as a function of STB noise figure and optical link CCN, as a function of a pre-defined overall SNR link objective (40 dB or 37 dB). Each curve represents a different value of SNR as set by the STB alone, associated with the noise figure and digital level (de-rated from analog) at its input.

Note from the figures that there is a wide range of SNR combinations that essentially offer no practical limit to RF cascade depth as it relates to noise degradation. Clearly, tolerating a 37 dB link requirement is exactly this scenario, and this is quite a reasonable requirement under the capability of new FEC. It provides a very comfortable range of operation, even for poor performing optical links with respect to noise.

However, the 40 dB range includes conditions that could lead to a sharp reduction in the cascade acceptable. From a sensitivity analysis standpoint, such conditions hinge on small dBs and even fractions thereof. This makes it more valuable to be able to earn back, for example, just 1-2 dB SNR in the analog reclamation process.

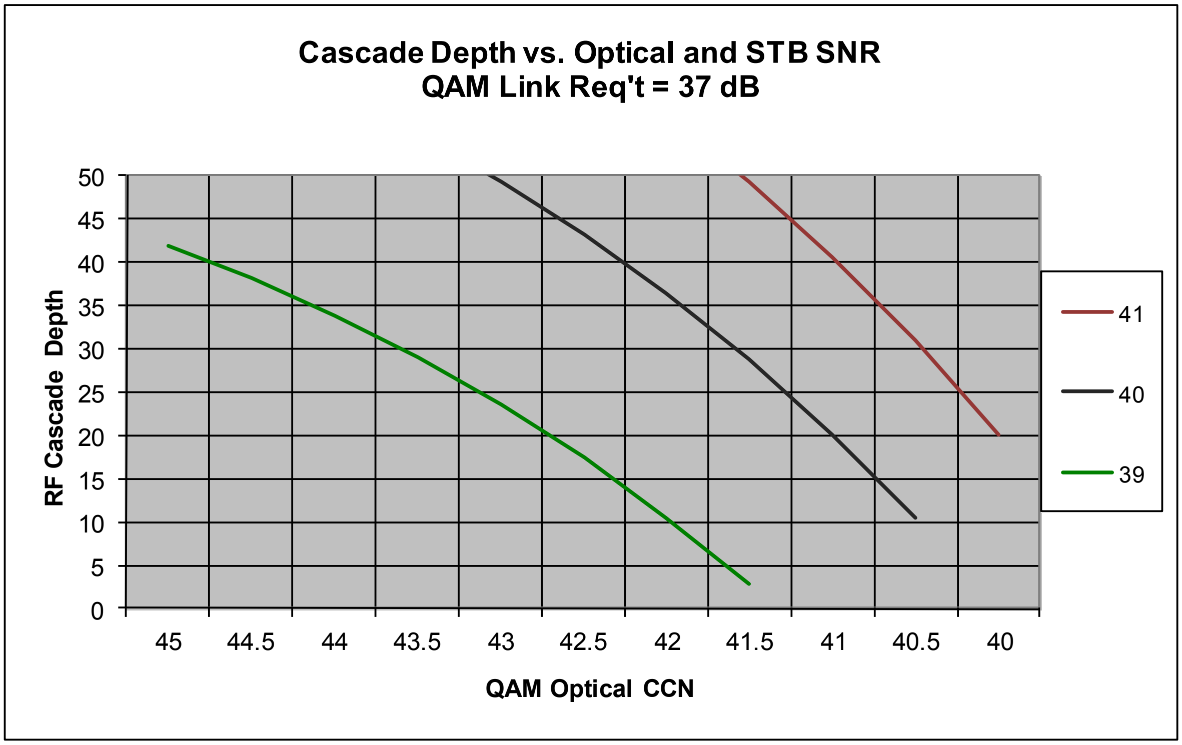


Figure 32 – 1024-QAM, Noise, Cascade Depth – 37 dB Link Requirement (Improved FEC)

Finally, note specifically the SNR = 42 dB at Optical CCN = 45 point on the bottom left of Figure 31. For a quite typical 51 dB Optical CNR requirement, a digital CCN of 45 dB would occur under 6 dB back-off. These conditions yield a cascade depth of five (N+5) as tolerable. Note, however, that 42 dB was a NF = 10 CPE, and, as previously identified, higher NF's (10-14 dB) may be the case.

This points out simply that STB clients of higher NF than 10 dB, under nominal optical link performance and deeper cascades may struggle to achieve the 40 dB requirement for 1024-QAM. FEC may save the link from a QoE perspective, but this example points out how relatively nominal conditions of legacy plant add up to make 1024-QAM a challenge. It also emphasizes the value of the dB available in migration, and especially the value of new FEC, most readily observable in Figure 32.

7.1.3.4 Distortion

As observed in Table 20, in addition to its positive effects on digital SNR, analog reclamation offers benefits in the distortion domain as well. Table 20 results are arrived at through tools such as shown in Figure 33 – a sample of a distortion beat map for 79 analog channels on a 12 dB tilt to 870 MHz. Such analysis is used to calculate the impact of varying channel line-ups on relative distortion level. Coupled with the sensitivity of 1024-QAM under CTB/CSO impairment, we can then evaluate the ability of an HFC cascade to support 1024-QAM.

The performance thresholds for CTB were taken from laboratory evaluation of error-free or nearly error-free 1024-QAM with actual live-video CTB generated as the impairment source [1]. It is interesting to note in that testing how pre-FEC and post FEC results are related, indicative of CTB as a "slow" disturbance relative to the symbol rate, and thus a burst error mechanism that challenges FEC decoding.

A result of the use of these CTB thresholds to find HFC architecture limitations is shown in Figure 34. It plots cascade depth thresholds over a range of given RF amplifier CTBs, specified at typical RF output levels, and varying analog channel counts used using a CTB threshold of 58 dBc [1].

It is clear to see that analog reclamation to 30 channels enables virtually any practical RF cascade depth. However, it also becomes clear how for 79-channel systems and 59-channel systems, some limitations may appear.

Prior analysis had investigated the effects of analog beat distortions on 256-QAM, developing relationships for the comparative performance of 64-QAM and 256-QAM [3]. It was observed that 10-12 dB difference existed in susceptibility to a single, static, in-band narrowband interferer at the main CTB offset frequency. Under the assumption that ingress mediation performance can achieve equivalent rejection relative to the M-QAM SNR (potentially an aggressive assumption), this relationship might be assumed hold between 256-QAM and 1024-QAM for narrowband interference.

7.1.3.5 Phase Noise

Untracked phase error leads to angular symbol spreading of the constellation diagram as shown in Figure 35 for 1024-QAM with .25° rms of Gaussian-distributed untracked phase error imposed. This non-uniform impact on symbols is critical to understand to explain phase noise sensitivities for increasing M in M-QAM. It was observed in [1] that .25° rms represents a loss due to phase noise of about 1 dB, assuming low error rate conditions, and with no practical phase noise-induced BER floor.

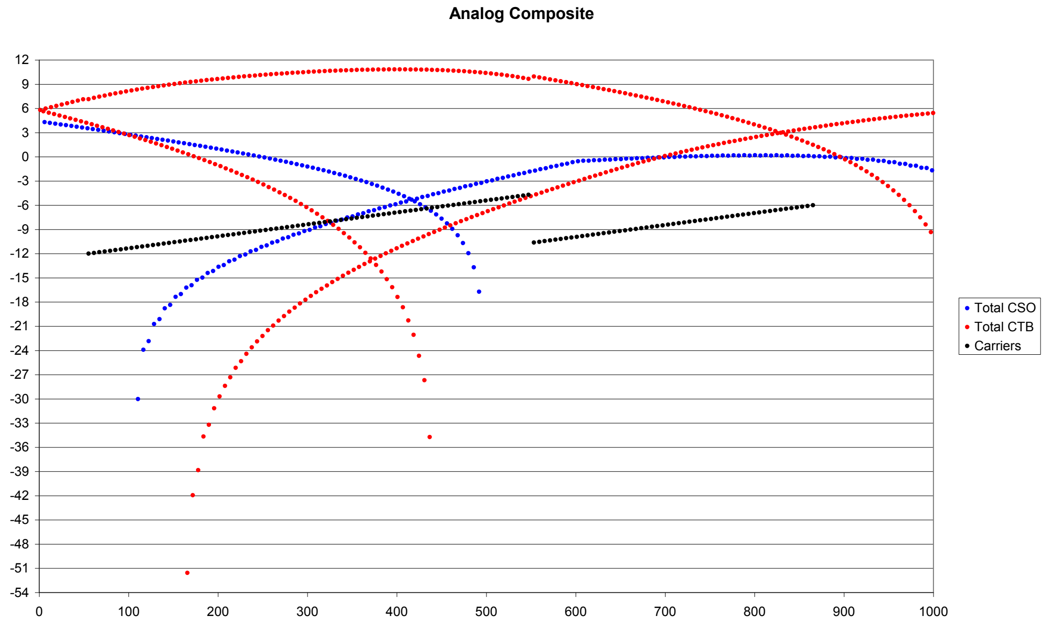


Figure 33 – Distortion Map - 79 Analog Channels, 12 dB Tilt

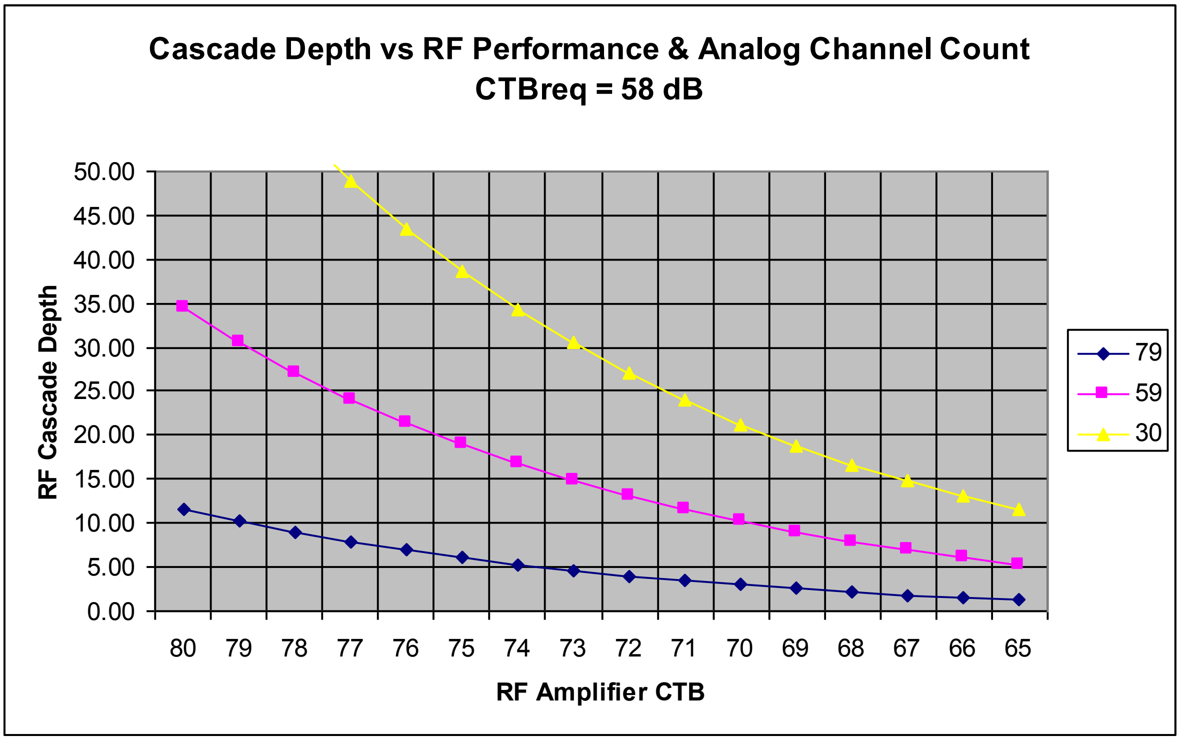


Figure 34 – 1024-QAM, CTB, and Cascade Depth, Thresh = 58 dB

A floor in the 1E-8 or 1E-9 region will be induced at roughly 50% more jitter, or .375 deg rms. Measurements of phase noise showed that for high RF carrier frequencies, typically associated with higher total phase noise, wideband carrier tracking still left about .33 deg rms of untracked error, enough to cause a BER floor to emerge at very high SNR.

The use of degrees rms is more easily understood when expressed as signal-to-phase noise in dB. Note that 1° rms is equivalent to 35 dBc signal-to-phase noise. Doubling or halving entails 6 dB relationships. Thus, we have the following conversions:

1 deg rms = 35 dBc SNRφ

.5 deg rms = 41 dBc SNRφ

.25 deg rms = 47 dBc SNRφ

The values .33 deg rms and .375 deg rms represent 44.6 dBc and 43.5 dBc, respectively. This is instructive to compare to the SNR under AWGN only (40 dB used above), as it illustrates the nature of the phase noise impairment on M-QAM with high M.

Error rate measurements [1] show that error flooring appears to be occurring as measured by pre-FEC errors, suggesting that there have not been significant enough tuning (historically analog, now full-band capture) noise improvements or carrier recovery system changes to mitigate this effect.

However, although phase noise is a slow random process that challenges burst correcting FEC, the combination of the interleaver, Reed-Solomon, and the relatively low floor, has been seen to result in zero post-FEC errors. Note that the phase noise alone is requiring the FEC to work to clean up the output data, and is thus consuming some FEC "budget" in the process.

Phase noise can be improved through design as well, almost without limit, but as strong function of cost for broadband performance. Current performance appears sufficient, although perhaps coming at the expense of increased sensitivity to other impairments that may also require FEC help.

These observations are likely a harbinger of issues to come as M increases further in search of higher bandwidth efficiency, such as 4096-QAM.

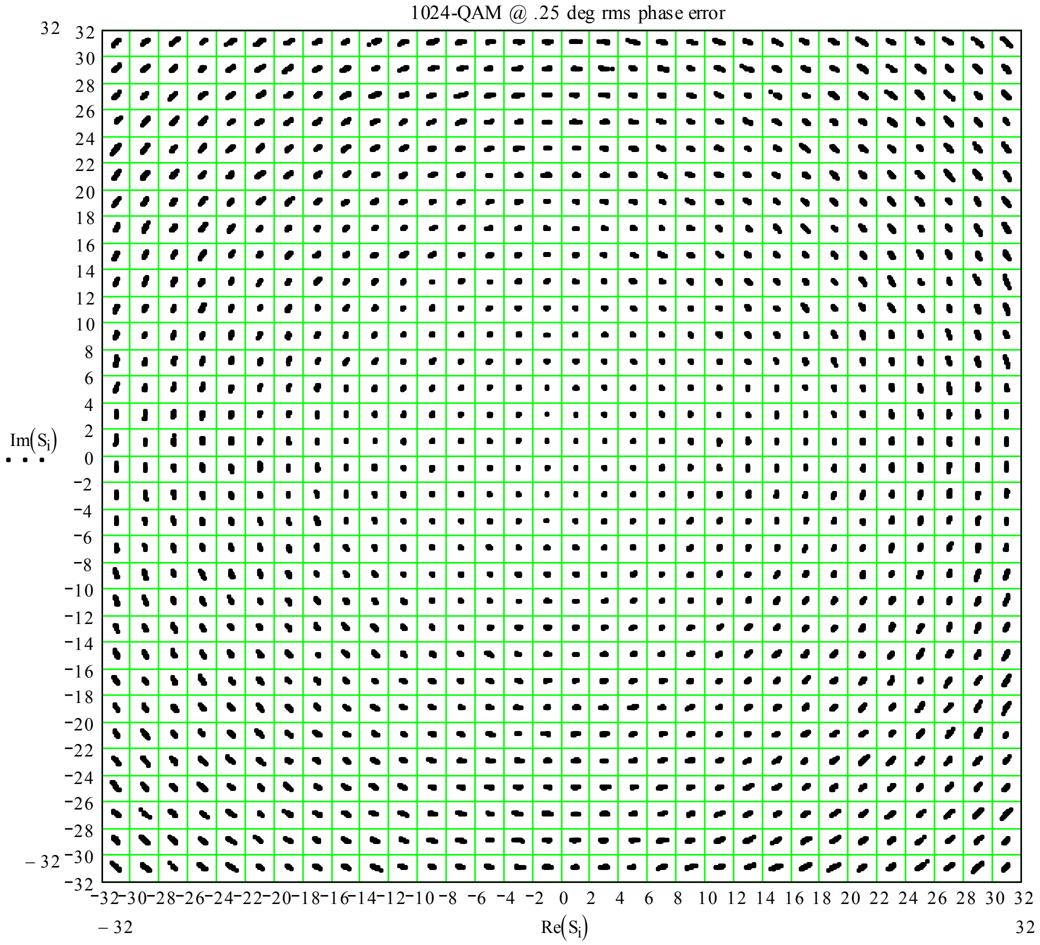


Figure 35 – 1024-QAM with .25° RMS Phase Noise

7.2 S-CDMA

Leveraging S-CDMA has many benefits, including reclamation of regions of upstream spectrum considered previously unusable with TDMA, lower overhead for FEC, and even feasibility of higher-order constellations. Some frequency regions are, of course, readily accessed leveraging Advanced Time Division Multiple Access (A-TDMA).

A-TDMA can be made very robust to a broad set of impairments including noise, distortion, and interference when it's coupled powerful tools such as Forward Error Correction (FEC), Equalization, and Ingress Cancellation. Problems arise when impairments exceed the performance limits of what A-TDMA can mitigate, resulting in objectionable codeword errors and packet loss.

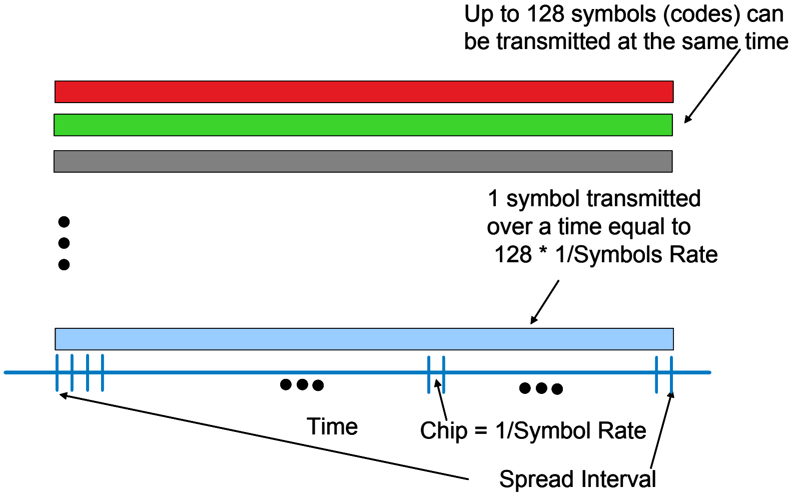

Up to 128 symbols (codes) can be transmitted at the same time
1 symbol transmitted over a time equal to 128 * 1/Symbols Rate
Time
Chip = 1/Symbol Rate
Spread Interval

Figure 36 – S-CDMA Parallel Symbol Transmission

Fortunately, DOCSIS 2.0 and later includes Synchronous Code Division Multiple Access (S-CDMA), which offers additional robustness against impairments, and in particular against impulse noise. This robustness against impulse noise exceeds that of A-TDMA by a factor of 100 times or more [14].

As powerful as DOCSIS 2.0 S-CDMA has been proven to be in field trials, DOCSIS 3.0 has S-CDMA features that further enhance robustness against impairments. These techniques were standardized to create a very high-performance, sophisticated PHY for cable, capable of supporting high data rates in the most difficult of environment.

The latest features include Selectable Active Codes (SAC) Mode 2, Trellis Coded Modulation (TCM), Code Hopping, and Maximum Scheduled Codes (MSC). Despite these advances aimed at adding more capability to the upstream, most of the DOCSIS 3.0 features remain largely unused, and DOCSIS 2.0 deployments are minor in scale in North America.

Let's take a look at what is available in DOCSIS to maximize the throughput of the upstream band, and discuss how today's PHY toolsets complement one another. First, let's understand what S-CDMA does best – high throughput performance under difficult channel conditions.

7.2.1 Impulse Noise Benefits of S-CDMA

There are several benefits to S-CDMA, but the most important by far is its burst protection capability. The ingredient that makes the robustness to impulse noise possible is the spreading out of the symbols by as much as 128 times in the time domain, which directly translates to stronger protection against impulse noise.

This spreading operation is pictured in Figure 36. Noise bursts that may wipe out many QAM symbols of an A-TDMA carrier must be two orders of magnitude longer in duration to have the same effect on S-CDMA, which is very unlikely. It is the spread signaling approach itself, without even considering FEC settings, that enables S CDMA to withstand much longer impulsive events.

There is no reduction in throughput as a result of this spreading, of course, because the slower symbols are transmitted simultaneously. S-CDMA has similarities conceptually to OFDM in this manner, with the difference being S-CDMA's use of the orthogonality in the code domain versus OFDM's use of orthogonality in the frequency domain.

Now consider Figure 37, which illustrates how S-CDMA's primary benefit translates to return path bandwidth access. Through its effectiveness against impulse noise, S-CDMA facilitates efficient use of what is otherwise very challenging spectrum for A-TDMA. It is a critical tool for squeezing every last bit-per-second possible out of return spectrum.

Additionally, the lower the diplex split used in the system, the more important S-CDMA becomes. It has become well-understood that the most consistently troublesome spectrum is at the low end of the band, typically 5-20 MHz.

This region is where S CDMA shines in comparison to A-TDMA. As such, S-CDMA matters more for maximizing use of 42 MHz than it does to 65 MHz (Euro Split) or 85 MHz (Mid-Split) because of the percentage of questionable spectrum.

Purely in terms of spectrum availability then, S-CDMA is most valuable to the North American market, where upstream spectrum is the scarcest and use of DOCSIS services is high. Depending on the upstream conditions, about 35-50% of extra capacity can be made available using S-CDMA.

Nonetheless, S-CDMA's benefits have been largely unused in practice by operators, despite its availability in DOCSIS 2.0 and DOCSIS 3.0 certified equipment.

7.2.2 Quantifying Performance

Again, by far S-CDMA's most compelling advantage is its ability to perform in harsh impulse noise environments. Impulse noise is, by definition, a transient event – interference of finite duration and often periodic or with

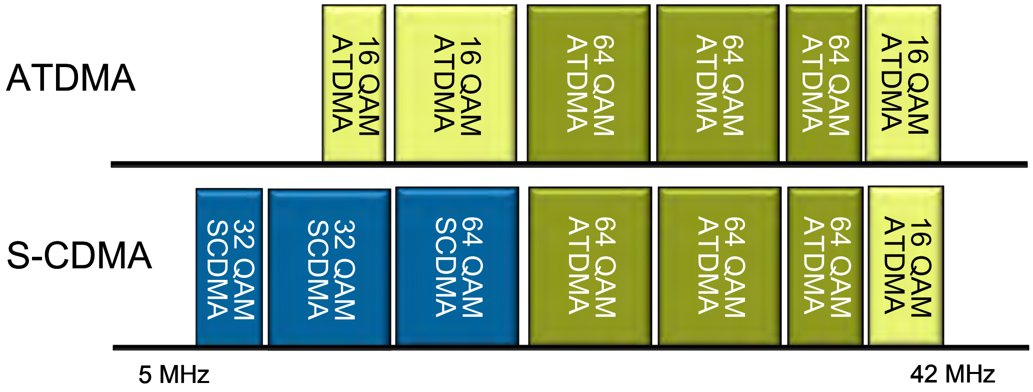


Figure 37 – Maximizing 5-42 MHz Throughput Using S-CDMA

repetitive frequency of occurrence.

Characterization of impulse noise includes duration, rate, and amplitude. It is generated in a variety of ways. When noisy devices such as dimmer switches, hair dryers, garage door openers, power tools, automobile ignition circuits – the list goes on – are in close proximity to the cable network, impulse noise may enter into upstream. The majority of impulse noise originates in and around the home.

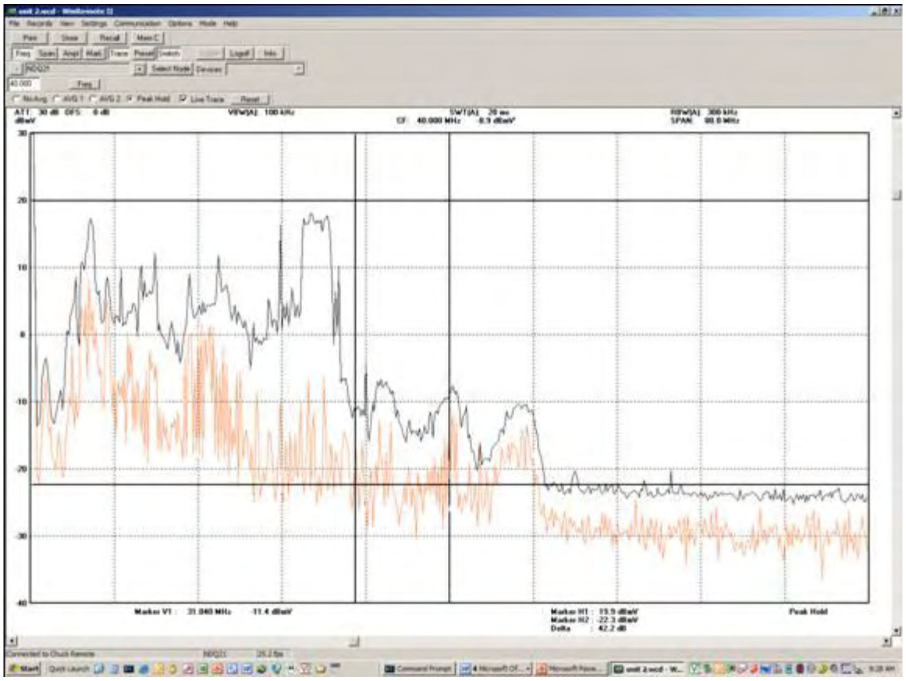
Figure 38 is a spectral snapshot of impulse noise, where a noticeable wideband burst above the noise average (red) is very likely interfering with DOCSIS signaling by creating a temporary condition whereby the SNR is only about 18 dB.

The impact of such a burst on a discrete set of QAM symbols is to cause the symbols to jump decision boundaries, or increase the probability that they will do so, resulting in codeword errors, as shown in Figure 39. Note the wideband nature of the degradation in the frequency domain of short duration impulse noise.

Consider just the DOCSIS-described scenario of duration 10µs and rate 1 kHz. A 10 usec burst will corrupt 52 symbol at 5.12Msps, which translates to 39 bytes of data for 64-QAM. This is beyond the capability of the Reed-Solomon FEC, with a maximum burst protection of t = 16 bytes.

For this scenario, the FEC cannot be effective without assistance of interleaving. An interleaver, in theory, could be used to break-up clusters of impacted bytes so that they span multiple codewords, allowing FEC to be more effective. However, byte interleaving requires longer packets for adequate shuffling of the bytes. Minimum packet lengths of 2x the designated codeword length are necessary, and the longer the better.

Unfortunately, of course, most upstream packets tend to be short and not suited to effective interleaving.



ATT: 30 dB OFS: 0 dB
VBW(A): 100 kHz
SWT(A): 20 ms
CF: 40.000 MHz 8.9 dBmV
RBW(A): 300 kHz
SPAN: 80.0 MHz
Marker V1: 31.040 MHz -11.4 dBmV
Marker H1: 13.9 dBmV
Marker H2: -22.3 dBmV
Delta: 42.2 dB
Peak Hold
Connected to Chuck Remote

Figure 38 – Impulse Noise Illustration, -18dBc

Such situations are where S-CDMA is the best choice for achieving high throughput. S-CDMA has greater ability to recover transmissions through long noise bursts, and is not sensitive to packet size the way interleaving is in a burst environment.

A most recent head-to-head comparison under simultaneous RF impairments of impulse noise and interference is shown in Table 22.

Three impulse noise sources were used:

1. Duration = 10µs, Rate = 1kHz (per DOCSIS specification)
2. Duration = 20µs, Rate = 4kHz
3. Duration = 40µs, Rate = 4kHz

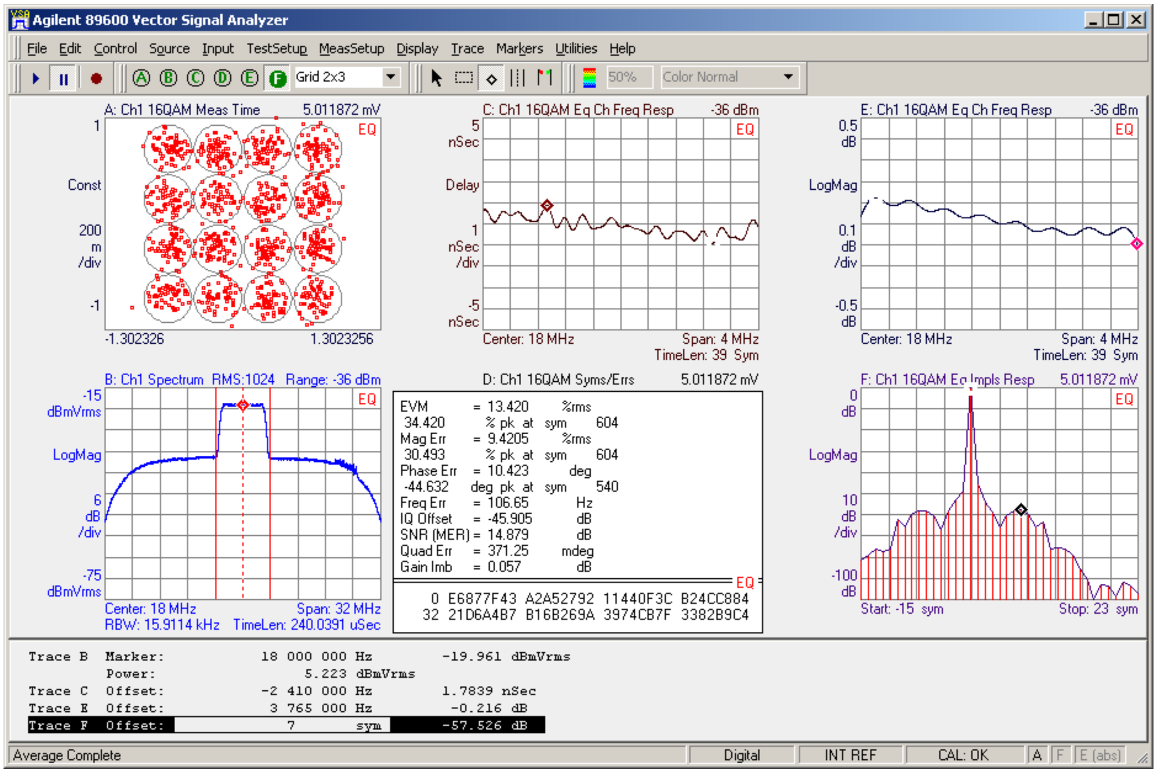
Three interference patterns used, centered around the signal center frequency:

- A. 4x π/4-DQPSK Carriers @16ksym/s, Spacing = 400kHz
- B. 2x π/4-DQPSK Carriers @16ksym/s, Spacing = 1600kHz
- C. 1 π/4-DQPSK Carriers @16ksym/s

The interference was modulated in order to randomize it and give it some spectral width, which makes ingress cancellation more challenging.

Table 22 shows the comparative results, with S-CDMA clearly and significantly outperforming A-TDMA under the dual impairment conditions. A-TDMA FEC is working much harder in each of the cases evaluated, primarily because of the impulse noise.

Uncorrected Codeword Error Rate (UCER) and packet error rate (PER) for A-TDMA under each of the impairment conditions shows performance that would



```
EVM        = 13.420      %rms
 34.420      % pk  at  sym     604
Mag Err    = 9.4205      %rms
 30.493      % pk  at  sym     604
Phase Err  = 10.423       deg
 -44.632   deg  pk  at  sym     540
Freq Err   = 106.65        Hz
IQ Offset  = -45.905       dB
SNR (MER)  = 14.879        dB
Quad Err   = 371.25      mdeg
Gain Imb   = 0.057         dB

 0 E6877F43 A2A52792 11440F3C B24CC884
32 21D6A4B7 B16B269A 3974CB7F 3382B9C4
```

```
Trace B  Marker:        18 000 000 Hz     -19.961 dBmVrms
         Power:               5.223 dBmVrms
Trace C  Offset:        -2 410 000 Hz     1.7839 nSec
Trace E  Offset:         3 765 000 Hz     -0.216 dB
Trace F  Offset:                 7   sym  -57.526 dB
```

Figure 39 – Impulse Noise Impaired 16-QAM

likely noticeably degrade the customer experience.

Not only is the S-CDMA FEC not working as hard as A-TDMA FEC, there is also less S-CDMA FEC applied. FEC for A-TDMA was at its maximum setting of t=16, and k=219, whereas field trial results previously published [1] resulted in lower FEC for S-CDMA of t=6, and k=239.

As previously discussed, FEC operating requirements can be lowered for S-CDMA because the robustness of the spreading function itself.

Clearly, for equal or even more strenuous impairment scenarios than the A-TDMA cases, S-CDMA offers error-free UCER and PER with no impact to the customer experience.

Additionally, proactive monitoring of Corrected Codeword Error Rate (CCER) with S-CDMA could better facilitate impulse noise problem diagnostics, whereas A-TDMA links would not.

Additional testing in the field on live plants has confirmed the advantage that S-CDMA delivers in the poorer part of the upstream spectrum. A result from a comparison of S-CDMA and A-TDMA on the same return path channel using logical channel operation, centered in a noisy portion of the upstream (about 13 MHz), is shown in Figure 40.

Apparent from Figure 40 is that A-TDMA is taking errors in transmission at a nearly 20% clip, while S-CDMA is taking

Table 22 – S-CDMA & TDMA Performance against Impulse Noise + Interference

16-QAM. 6.4MHz						
1518-Byte Packets	S-CDMA			ATDMA		
Noise Floor = 35dB	MER	CCER/UCER	PER	MER	CCER/UCER	PER
Interference Characteristics	Impulse Noise Characteristics: Duration = 10us, Rate = 1kHz, Level = -11dBc					
Pattern A @ -20dBc	33.1	3.2653%/0.0000%	0.00%	32.2	9.8190%/0.3643%	1.82%
Pattern B @ -18dBc	33.3	2.2164%/0.0004%	0.00%	30.4	9.4996%/0.4362%	1.84%
Pattern C @ -16dBc	33.6	6.0938%/0.0000%	0.00%	30.5	9.1357%/0.9920%	4.86%
Interference Characteristics	Impulse Noise Characteristics: Duration = 20us, Rate = 4kHz, Level = -13dBc					
Pattern A @ -22dBc	29.0	6.2512%/0.0000%	0.00%	29.6	39.7214%/0.2657%	1.46%
Pattern B @ -22dBc	23.0	6.4386%/0.0000%	0.00%	28.2	36.8949%/0.0730%	0.39%
Pattern C @ -20dBc	33.5	5.3450%/0.0000%	0.00%	25.6	36.5901%/1.1087%	4.61%
Interference Characteristics	Impulse Noise Characteristics: Duration = 40us, Rate = 4kHz, Level = -14dBc					
Pattern A @ -22dBc	17.3	13.1082%/0.0000%	0.00%	26.6	39.7623%/0.0639%	0.40%
Pattern B @ -22dBc	26.1	13.8848%/0.0000%	0.00%	20.3	35.1569%/0.0079%	0.05%
Pattern C @ -13dBc	34.2	7.6259%/0.0000%	0.00%	28.0	38.3802%/1.7060%	6.91%

16-QAM. 3.2MHz						
1518-Byte Packets	S-CDMA			ATDMA		
Noise Floor = 35dB	MER	CCER/UCER	PER	MER	CCER/UCER	PER
Interference Characteristics	Impulse Noise Characteristics: Duration = 10us, Rate = 1kHz, Level = -7dBc					
Pattern A @ -22dBc	32.2	6.9036%/0.0000%	0.00%	33.5	18.1515%/2.7396%	14.87%
Pattern B @ -26dBc	21.1	4.0558%/0.0000%	0.00%	28.9	19.2957%/0.7367%	3.99%
Pattern C @ -11dBc	33.1	3.6618%/0.0000%	0.00%	34.0	16.8403%/5.2196%	22.86%
Interference Characteristics	Impulse Noise Characteristics: Duration = 20us, Rate = 4kHz, Level = -10dBc					
Pattern A @ -23dBc	25.6	8.1255%/0.0005%	0.00%	26.2	79.9084%/4.3388%	22.07%
Pattern B @ -24dBc	19.5	17.1071%/0.0000%	0.00%	24.8	81.1037%/0.1378%	0.85%
Pattern C @ -12dBc	32.6	13.3983%/0.0000%	0.00%	18.0	65.1727%/20.9625%	65.44%
Interference Characteristics	Impulse Noise Characteristics: Duration = 40us, Rate = 4kHz, Level = -12dBc					
Pattern A @ -20dBc	22.9	15.8017%/0.0000%	0.00%	18.6	85.0225%/2.8658%	13.41%
Pattern B @ -23dBc	31.3	16.5487%/0.0000%	0.00%	20.0	83.8348%/0.4118%	2.01%
Pattern C @ -13dBc	31.6	24.5632%/0.0000%	0.00%	23.0	71.7126%/17.3259%	56.71%

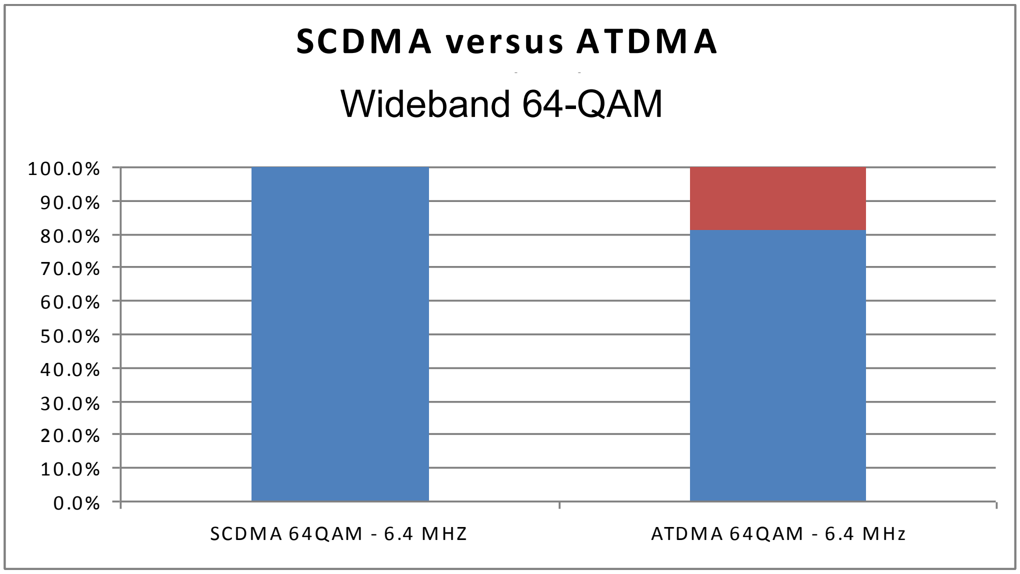


Figure 40 – Corrected Error Statistics

none. In this case, FEC settings for A TDMA are again t=16, while for S-CDMA, they are set to just t=2. S-CDMA inherently takes advantage of its impulse immunity properties rather than relying on FEC.

It is worth noting that, for A-TDMA, impulse noise can also wreak havoc on adaptive processes such as equalization and ingress cancellation, resulting in appreciable variation in cancellation estimates. For example, Figure 41 shows Non-Main Tap to Total Energy Ratio (NMTER) for a population of eight cable modems where impulse noise caused significant variation in equalizer correction.

NMTER is useful as a Figure of Merit to describe the linear distortion level of the upstream path. Here, it is indicating that the frequency response correction process is being significantly disturbed, resulting in a period of increased ISI until the impulse noise subsides and the taps updated.

Even should FEC be able to handle the impulse duration, this increase in ISI can degrade performance because of the increased susceptibility to detection errors at the slicer. The FEC budget may be required to deal with both ISI and burst correction, and is therefore more likely to be overwhelmed until the next tap update can be processed.

7.2.3 More Capability Remains

S-CDMA's impulse noise robustness has been demonstrated, but there is still more that can be leveraged to take advantage of all of the DOCSIS 3.0 features of S-CDMA.

Additional features include Selectable Active Codes (SAC) Mode 2, Trellis Coded Modulation (TCM), Code Hopping, and Maximum Scheduled Codes (MSC). These features provide more flexibility and capability for extracting bandwidth from noisy, limited spectrum, and yet remained largely unused despite more being standardized for many years.

Briefly, these features provide the following:

SAC Mode 2 – Allows for customization of the active codes. Instead of fixed active codes (SAC Mode 1) codes

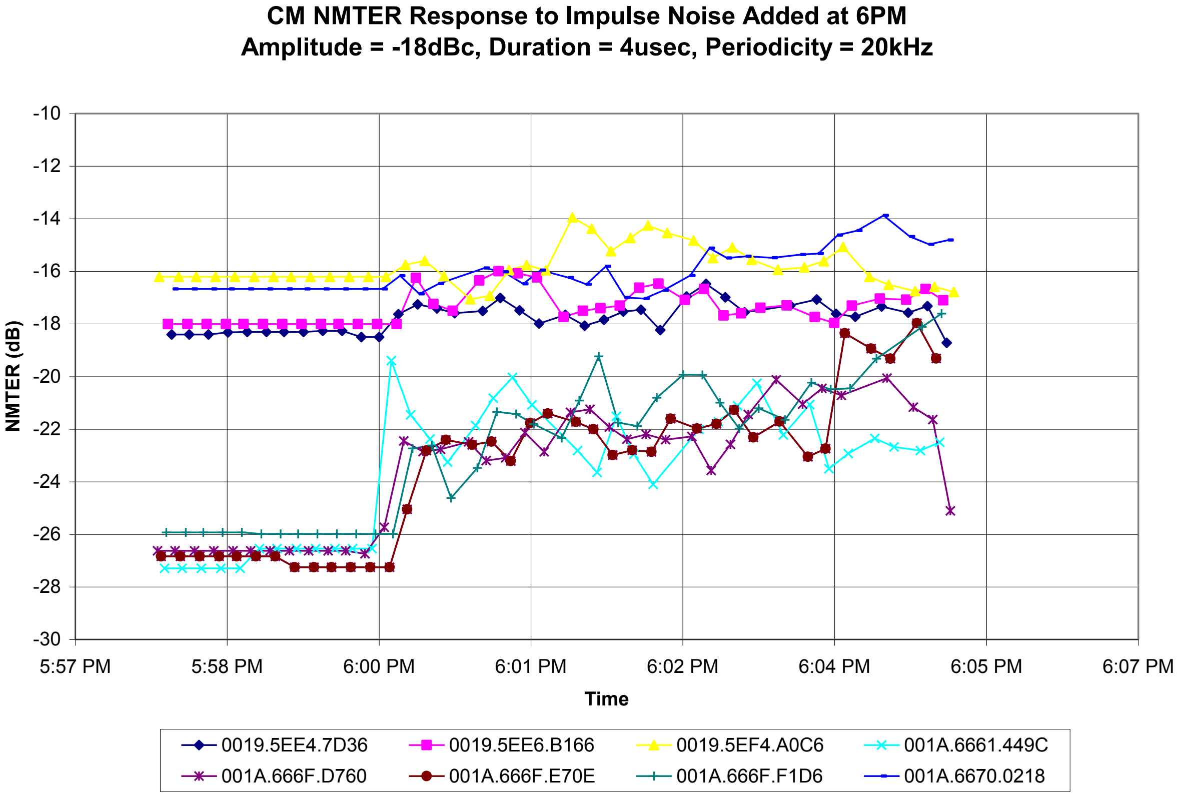


Figure 41 – NMTER vs. Time Impaired by Impulse Noise

may now be optimally allocated between spreading and ingress cancellation.

Trellis-Coded Modulation (TCM) – The well-known technique for optimizing coding structure through integration with symbol mapping, adding gain without adding bandwidth overhead to do so.

Code Hopping – Provides cyclic shifts of the active code set at each spreading interval, further randomizing code allocation to achieve a uniformity of robustness of performance

Maximum Scheduled Codes (MSC) – Offers the flexibility to trade-off between the power allocated per-code and the number of codes turned on. For example, if 128 codes are on transmitting at Pmax, each code is allocated Pmax/128. If only 64 codes are used, each code is allocated Pmax/64, or 3 dB more power per code. This comes at the expense of throughput, but offers some choices to the operator that may be better than an equivalent A-TDMA alternative.

7.2.4 Summary

S-CDMA delivers proven, substantial gains in impulse noise robustness – performance verified in detailed lab testing and in the field, around the world.

It clearly outperforms A-TDMA on difficult channels, enables high-throughput access to the otherwise abandoned lower portion of the return spectrum, and has been shown to operate robustly on channels where A-TDMA will not operate at all.

Many available, but as yet unused, features of S-CDMA, including SAC Mode

2, MSC, Code Hopping, and TCM, provide further capability against upstream impairments. Nonetheless, while a long-standardized tool in DOCSIS, operators have not widely deployed S-CDMA.

In low-diplex architectures, where DOCSIS extensions may be the most straightforward, low-complexity way to light up new spectrum, S-CDMA already exists to support the delivery of high throughput on difficult low-end spectrum. It is capable of providing the same benefits as in any new spectrum deployed for upstream that becomes prone to high interference and noise levels.

The combination of updated A-TDMA with the full features of S-CDMA may, in fact, be a sufficient PHY toolset for upstream growth and lifespan extension, eliminating the need to develop a third upstream PHY, such as an OFDM-based system.

7.3 OFDMA, OFDM & LDPC (A Proposal for a New PHY)

7.3.1 Problem Statement

Once it is acknowledged that current DOCSIS 3.0 MAC provides all the necessary capabilities to extend DOCSIS service to future gigabit rates, the challenge becomes optimizing the PHY layer.

Before choosing the technology for that new PHY, key selection criteria need to be established. These criteria apply to both upstream and downstream.

1. Bandwidth capacity maximization
2. Transparency toward the existing D3.0 MAC
3. Robustness to interference
4. Robustness to unknown plant conditions
5. Throughput scalability with plant condition (SNR)
6. Implementation complexity and silicon cost
7. Time to market
8. PAPR considerations
9. Frequency agility

7.3.1.1 Bandwidth Capacity Maximization

According to Shannon theorem the maximum achievable throughput capacity for a communication system is a function of signal to noise ratio and bandwidth. Both of these resources, the signal power relative to an unavoidable noise and the useful bandwidth of the coaxial part, are limited in an HFC plant.

An upgrade of the HFC plant is costly, and therefore before (or in parallel with) this upgrade, the available SNR and bandwidth utilization can, and must be maximized using state-of-the-art modulation and coding techniques.

7.3.1.2 Transparency Toward the Existing D3.0 MAC

One of the extremely useful features of the D3.0 MAC is the physical channel bonding. This feature allows trafficking of logical flows on information through multiple and different physical channels. Apart from the lower level convergence layer features, the DOCSIS 3.0 MAC is not aware what type of Physical channel(s) the information is flowing through, be it 256-QAM or 64-QAM in downstream, or ATDMA or SCDMA in upstream.

Allowing the new PHY to follow the same transparency will allow the products introduced to the market migrate gradually from using the old PHY to using the new PHY by utilizing (rather than giving up) throughput from existing legacy channels, until these are gradually replaced with new ones. For example, there are CMs deployed in the field with eight downstream channels. Until all these CMs are replaced, those eight channels will continue to occupy the shared spectrum. A transition period product will be able to make use of both the legacy PHY and the new PHY through channel bonding; and hence will maximize the data throughput as illustrated in Figure 42 and Figure 49.

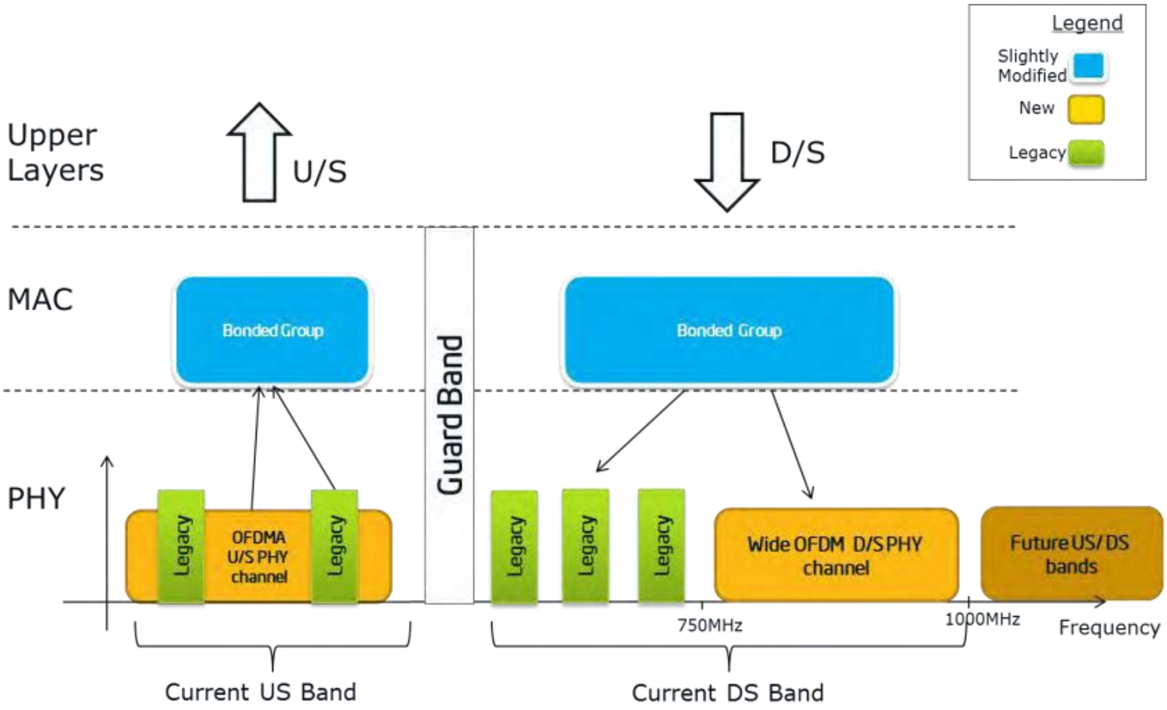


Figure 42 – Illustration of bonding the legacy and the new PHY channels

As a comparison, a non-DOCSIS technology will not be able to benefit from the bandwidth occupied by legacy.

7.3.1.3 Robustness to Interference

As the home and business environment becomes flooded with electronic equipment, the level of interference becomes a significant limiting factor of bandwidth usage in some regions of the HFC spectrum, particularly in the upstream. A modulation scheme of choice should be designed to minimize the effect of interference on the achievable throughput.

7.3.1.4 Robustness to Unknown Plant Conditions

The new PHY should be well equipped to be deployed in spectrum that is currently unused for cable systems, such as spectrum beyond 1GHz. Also, it should be equipped to maximize throughput given unknown parameters in the existing installation, as these differ significantly by region, type of installation, countries, etc. Planning for the worst case adds inefficiency and cost, hence agility to optimize capacity per given condition is required.

7.3.1.5 Throughput Scalability with Plant Condition (SNR)

As mentioned above, SNR sets the maximum achievable capacity over a given bandwidth. Ability to scale the throughput accordingly with the SNR available to the modem will allow squeezing the maximum throughput possible per given installation condition. Simply put, more bits/sec/Hz configurations are needed with finer granularity, spanning a wide SNR scale.

7.3.1.6 Implementation Complexity and Silicon Cost

Adding more throughput capability to the modem will result in more silicon complexity that translates to silicon cost. It is essential that the new PHY technology chosen is able to offer cheaper implementation in terms of dollars per bits/sec/Hz over other alternatives. As a side note, one thing worth noting is that process technology scaling (Moore's law) allows increasing the PHY complexity without breaking the cost limits.

7.3.1.7 Time to Market

It is important to isolate the proposed changes to specific system elements without affecting system concepts. Changing only the PHY channel, without any significant changes to the MAC minimizes the scope of impact of the change and allows quicker standardization and implementation of the change. Utilizing existing, proven, and well-studied technologies helps accelerate the standardization and the productization.

7.3.1.8 PAPR Considerations

Good (low) peak to average ratio properties of the modulation technique may help in squeezing more power out of the amplifiers in the system by moving deeper into the non-linear region. Hence, good PAPR properties are desirable, as these have system impact beyond the end equipment.

7.3.1.9 Frequency Agility

The ability of the new PHY channel to be deployed in any portion of the spectrum is a great advantage. This is especially useful during the transition period where various legacy services occupy specific frequencies and bands and cannot be moved.

Next we consider the alternatives of the PHY channels in light of the above-mentioned criteria, focusing on the parameters of the suggested proposal.

7.3.2 Solution Analysis

7.3.2.1 Channel Coding – Optimizing Spectral Efficiency

FEC has the most significant impact on spectral efficiency. Traditional error control codes such as J.83 Annex B are concatenations of Trellis and Reed-Solomon block codes. Modern coding techniques such as LDPC and Turbo use iterative message passing algorithms for decoding, thereby yielding significant coding gains over traditional techniques. LDPC has been shown to out-perform Turbo codes at relatively large block sizes. LDPC also has the parallelism needed to achieve high throughputs.

Figure 43 shows a comparison of different coding schemes used in Cable technologies[1]. 256-QAM modulation is taken as baseline for comparison. The horizontal axis is the code rate and the vertical axis shows the SNR required to achieve a BER of 1e-8. The two DVB-C2 LDPC codes are shown, the long code with a block size of 64800 bits and the short code with a block size of 16200 bits. [28]

As expected, the code with the longer block size does provide better performance

[1] Although code rate 0.8 is not present in current J83 specification, the system was simulated with RS codes (204, 164) for J.83 Annex A and (128, 108) for J.83 Annex B to get the effective performance of these codes at a rate of 0.8.

although the difference is very small (0.2 dB) for high code rates needed for cable applications. The two DVB-C2 LDPC codes do include a weak BCH code to assist with the removal of the error floor.

The graph in Figure 43 shows that the DVB-C2 LDPC offers about 3 dB more coding gain over J.83 Annex B code for a code rate of 0.9 implying an increase of capacity of 1 bit/s/Hz, i.e. a 12.5% increase with respect to 256-QAM. The increase in coding gain and hence the capacity is much higher (about 5 dB) with respect to just the RS code used in J.83 Annex A, i.e. DVB-C.

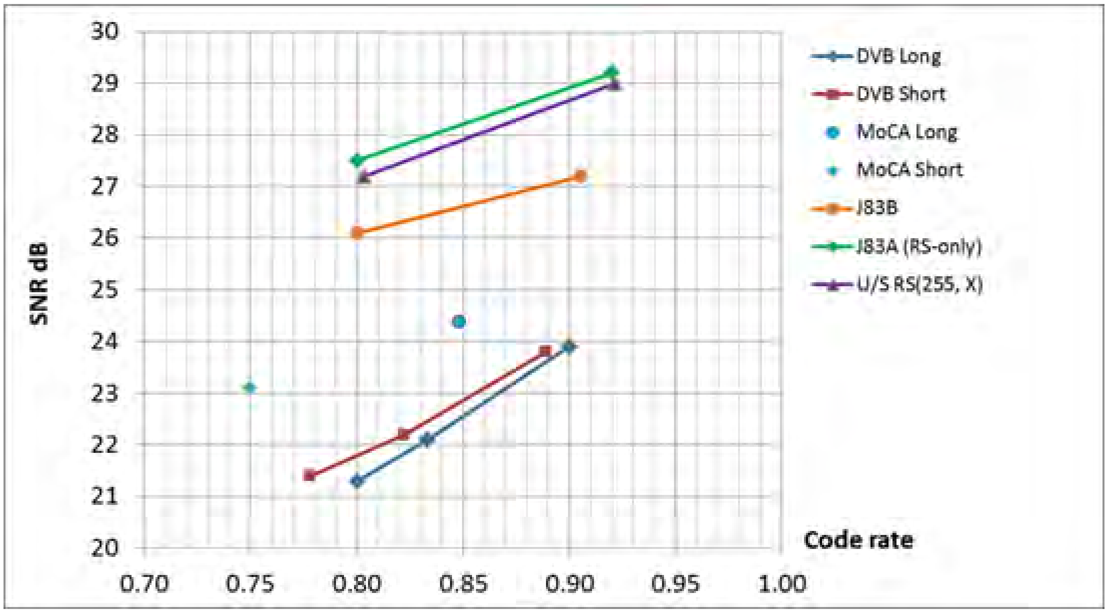


Figure 43 – FEC Comparison for 256-QAM Modulation

Note that since existing coding schemes are compared, the code word lengths are not the same, implying an advantage to longer code words. Theoretically, if the J.83 Annex B FEC is extended to a longer code word, the difference will be less than 3 dB, but the DVB-C2 code will still give the better performance.

To enable efficient stuffing of upstream bursts with code words, two types of codes with different code word length are necessary. A short code word for short bursts, and a long code word for long bursts are recommended. Since the ambitious throughput requirements are usually on the long bursts (streaming data, rather than maintenance messages), no system throughput loss is expected due to usage of shorter code word.

7.3.2.2 Modulation Scheme

The options considered for the modulation scheme of the next gen PHY are as follows:

1. Legacy modulation, narrow Single Carrier QAM, 6/6.4 MHz channels
2. A new, wide Single-Carrier channel modulation, e.g. Single Carrier QAM 24 MHz channel
3. A new, wide Multi-Carrier OFDM channel modulation

A comparison of these options is discussed next against the established criteria.

7.3.2.3 Implementation Complexity

To contain the total complexity increase due to scaling to gigabit throughputs, both PHY layer implementation itself, and its effect on the MAC layer need to be considered.

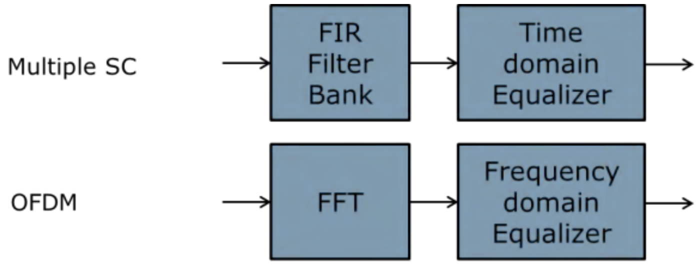


Figure 44 – Signal Processing Block for Computational Complexity analysis

If narrow channels are used to attain the high throughputs, a large number of such channels will be required, which may lead to a non-linear increase in the MAC complexity. Hence, there is a benefit of using wide channels to reduce the total number of bonded channels.

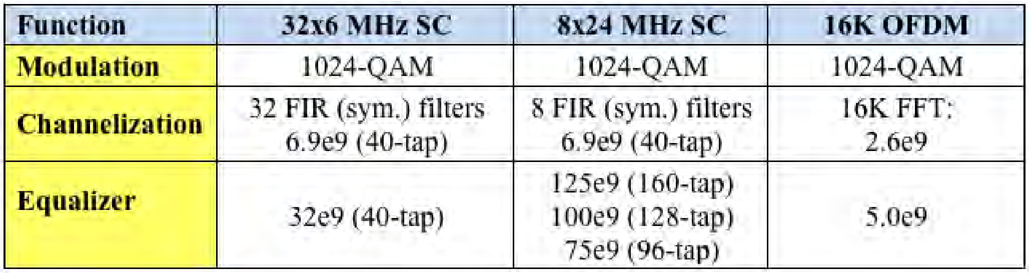
However, only OFDM out of the three options considered can give a computational differences in channel processing and equalization. The channelization for OFDM is based on FFT, which is computationally more efficient than the multiple sharp channel filters required for single carrier. Also the frequency domain equalization in OFDM is much lighter computationally than time domain equalization required with SC-QAM. Figure 44 and Table 23 show the processing power analysis of the options based on the number of real multiplication per second required. A clear advantage of OFDM is observed.

Another thing worth noting in favor of wide channels is that since today's analog front end technology for cable is based on direct digital-to-analog and analog-to-digital conversion, having wide channels does not pose a new implementation challenge for the analog front end design. All the channelization and up/down frequency conversion can be done digitally.

7.3.2.4 Channel Equalization

A common assumption for OFDM modulation is that the guard interval (GI) needs to be of a length equal to or higher than the longest reflection in the channel. However, this does not have to be the case. The reflection that is not completely covered by the GI affects only small part of the benefit given a wide channel, due to

Table 23 – Number of Multiplications per sec (real*real) for different modulations schemes



Function	32x6 MHz SC	8x24 MHz SC	16K OFDM
Modulation	1024-QAM	1024-QAM	1024-QAM
Channelization	32 FIR (sym.) filters 6.9e9 (40-tap)	8 FIR (sym.) filters 6.9e9 (40-tap)	16K FFT: 2.6e9
Equalizer	32e9 (40-tap)	125e9 (160-tap) 100e9 (128-tap) 75e9 (96-tap)	5.0e9

symbol, reducing the power of the inter-symbol-interference (ISI) on the entire symbol accordingly (approximately by 10log(T_interference_overlap/T_symbol) on top of the already weak power of the long echoes).

The result is extra gain in throughput of OFDM symbol, due to the GI being shorter than the longest anticipated reflection. To illustrate this, a simulation result of a 16K FFT OFDM system with 200 MHz channel bandwidth and DVB-C2 LDPC code with rate 8/9 is depicted in Figure 45. SCTE-40 reflection profile (SCTE-40) is simulated, as well as AWGN.

The 4.5 us SCTE-40 echo (-30 dB) is outside the 3.33 us cyclic prefix guard interval. However, the loss with respect to the 5 us guard interval is only 0.15 dB because the ICI/ISI noise floor due to echo outside guard is at -42 dB.

An OFDM scheme can have multiple options for guard intervals without any silicon cost penalty, whilst the SC time equalizer approach needs to be designed for the worst case. As DOCSIS moves into new spectrum, this additional flexibility gives OFDM an advantage over SC.

7.3.2.5 Robustness to Interference

In OFDM, narrow interference typically affects only a small number of carriers, causing only a minor loss in capacity. If the locations of the interferences are known, it is possible not to transmit at those carriers or reduce the modulation order of transmission for those carriers only. Also, since the LDPC decoding is done based on SNR estimation per carrier, the error contribution of the noisy carriers will be minimized by the LDPC decoder even if the location of the interference is not known.

Robustness to interference of wide single carrier channels would be based on the same ingress cancellation techniques

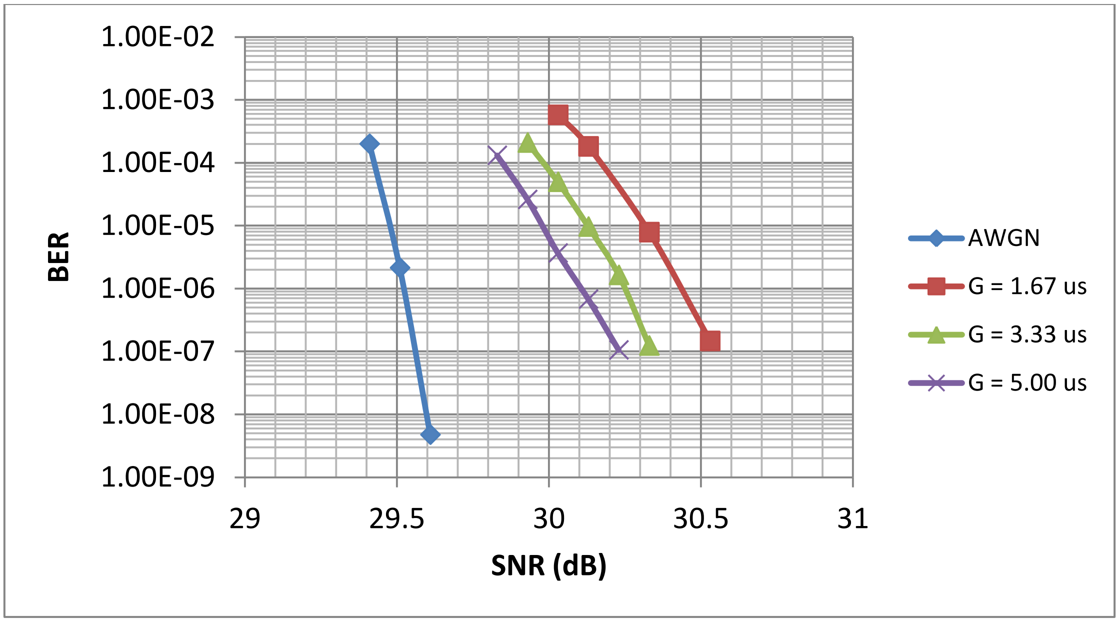


Figure 45 – OFDM/LDPC system performance in presence of SCTE-40 channel echoes

currently used in downstream and upstream receivers. However, for wider channels, these functions could become more challenging because of the increased probability of multiple interferers. This would result in inferior performance compared to today's single carrier in spectral regions beset by interference, or an increase in complexity to achieve the same performance.

In general, OFDM offers particular, understood simplicity and flexibility advantages for dealing with the narrowband interference environment. These could benefit DOCSIS, particularly as previously unused, unpredictable bands become used.

7.3.2.6 Throughput Scalability with SNR

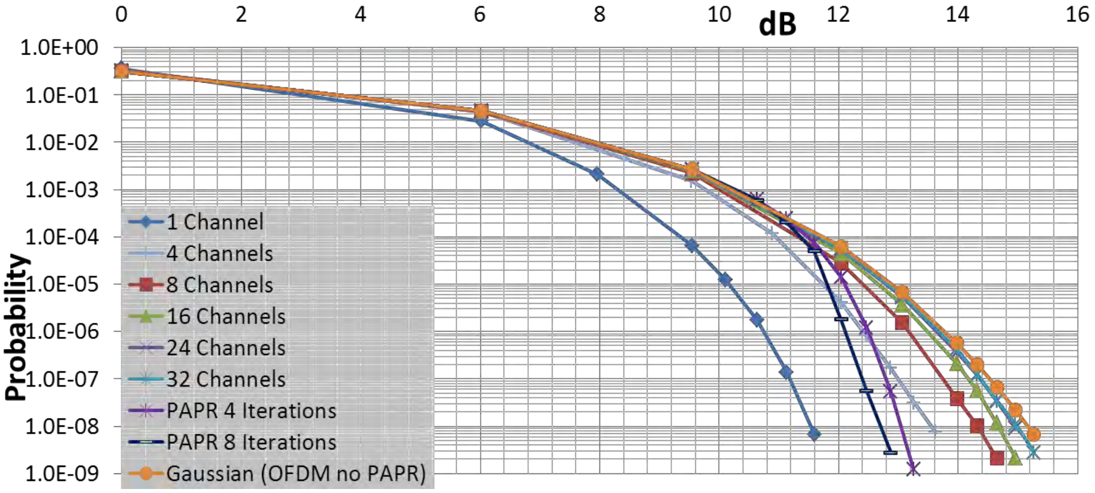
Another useful feature of OFDM modulation is that it enables use of different QAM constellations per carrier (also known as "bit loading"). This allows keeping all the benefits of a wide channel, while having the ability to fit modulation per the existing SNR at a narrow portion of spectrum. This enables maximizing throughput when the SNR is not constant within the channel band.

The non-flat SNR case is especially relevant for spectrum beyond 1 GHz, where signal attenuation falls sharply with frequency, or above the forward band of sub-1 GHz systems. Using a wide single carrier channel in this case would mean a compromise on throughput, and using a narrow singe carrier channel would require a myriad of channels.

7.3.2.7 Peak to Average Power Ratio (PAPR)

Peak to Average Ratio of OFDM modulation is frequently considered as its disadvantage due to the fact that OFDM symbol has Gaussian amplitude distribution (that's because of its multicarrier nature). It is true, but mainly in comparison to a single channel or a small number of channels.

DOCSIS 3.0 systems have at least 4 upstream channels, and this number will continue going up as long as single carrier channels are used to reach higher rates.



Notes: 6 MHz channels with 0.15 alpha and wide OFDM with Peak to Average Reduction Algorithm.

Figure 46 – Probability of Clipping as a Function of Peak to RMS Ratio

Figure 46 shows the PAPR profiles for OFDM and different numbers of single-carrier channels. The vertical axis is the clipping probability for the clipping threshold given in the horizontal axis.

The Gaussian profile is for OFDM with no PAPR reduction. Graphs for different numbers (1, 4, 8, 16, 24 and 32) of single-carrier channels are also shown (each with 0.15 RRC roll-off). It is seen that even when the number of single carrier channels is as low as four, the PAPR is not too different from Gaussian.

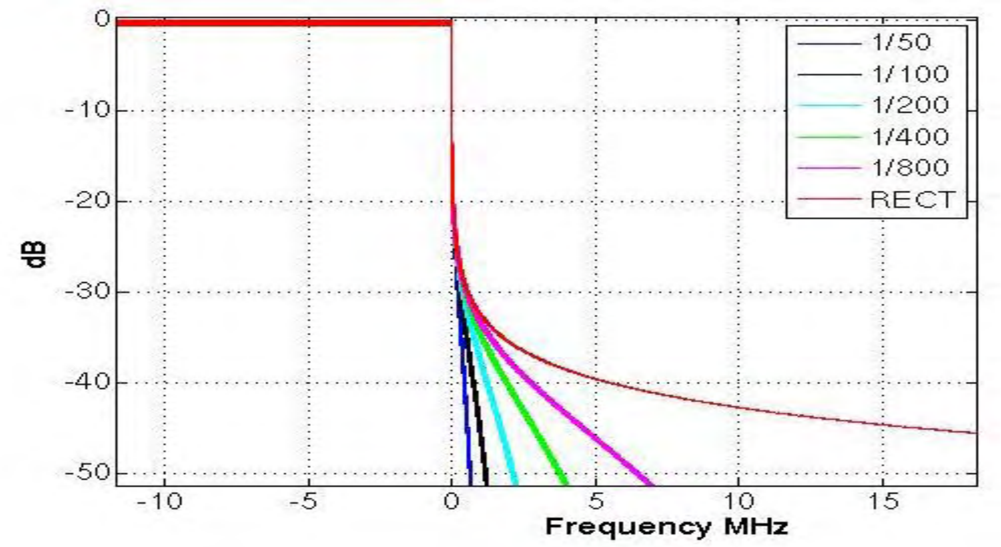
However, unlike single-carrier, OFDM offers ways of reducing peak-to-average power. One such method illustrated using this graph is called tone reservation. In this method a few (< 1%) of the tones are reserved to reduce the high amplitudes in an OFDM FFT. The results shown have been obtained by simulating the specific method given in the DVB-T2 specification. It is seen that the peak power of OFDM can be made to be less than four single-carrier channels at clipping probabilities of interest to cable applications.

Hence, as far as next gen DOCSIS PHY is concerned, OFDM actually has an advantage over bonded single carrier modulation of four channels or greater in terms of PAPR.

7.3.2.8 Frequency Agility

All options considered for downstream have width of multiples of 6 MHz or 8 MHz, for compatibility with the existing downstream grid.

A wide OFDM channel allows creating a frequency "hole" in its spectrum to enable legacy channels inside it, should there be a frequency planning constraint (as graphically shown in Figure 42. With this feature, OFDM retains the frequency agility of a narrow channel, while keeping all the benefits of a wide channel. A wide single carrier channel will be at a disadvantage in that respect.



Notes: lengths (ratio to total number of carriers)

Figure 47 – 16K symbol frequency response with different pulse shaping

To reduce the interference of OFDM

channel to the QAM channel inside it, an OFDM symbol shaping (windowing) can be employed as shown on Figure 47. This windowing makes the OFDM symbol length longer which implies a reduction in the bit rate. Nevertheless, as seen from the figure, windowing significantly sharpens the edge of the OFDM spectrum. This allows data carriers to be inserted until very close to the edge of the available bandwidth. So we have a capacity loss seen from the time domain representation and a capacity gain seen from the frequency domain representation. The net effect is a significant capacity gain and the optimum excess time for windowing has been found (for 12.5 KHz carrier separation) to be 1% of the useful OFDM symbol period (black line in Figure 47).

7.3.2.9 Upstream Multiple Access Considerations

Allowing simultaneous access of multiple CMs is essential for containing latency and for ease of CM management. OFDM modulation can be extended into an OFDMA (Orthogonal Frequency Division Multiple Access) modulation where several modems can transmit on different carriers at the same time.

The good news is that the DOCSIS 3.0 MAC convergence layer already supports that type of access for a case of SCDMA modulation in DOCSIS 3.0. The same concepts can be adopted with minor adjustments for OFDMA convergence layer. The concept of minislots that serves as an access sharing grid for the upstream transmission opportunities can be kept. The two dimensional minislot numbering used in SCDMA can also be kept for OFDMA. The contention, ranging and station maintenance arrangements can be kept.

In order to allow different bit loading per carrier, the minislots, if chosen as constant in time, may be different in size. That would be a change from constant size minislots in legacy DOCSIS, but this is an isolated change. Figure 48 shows an example of such access.

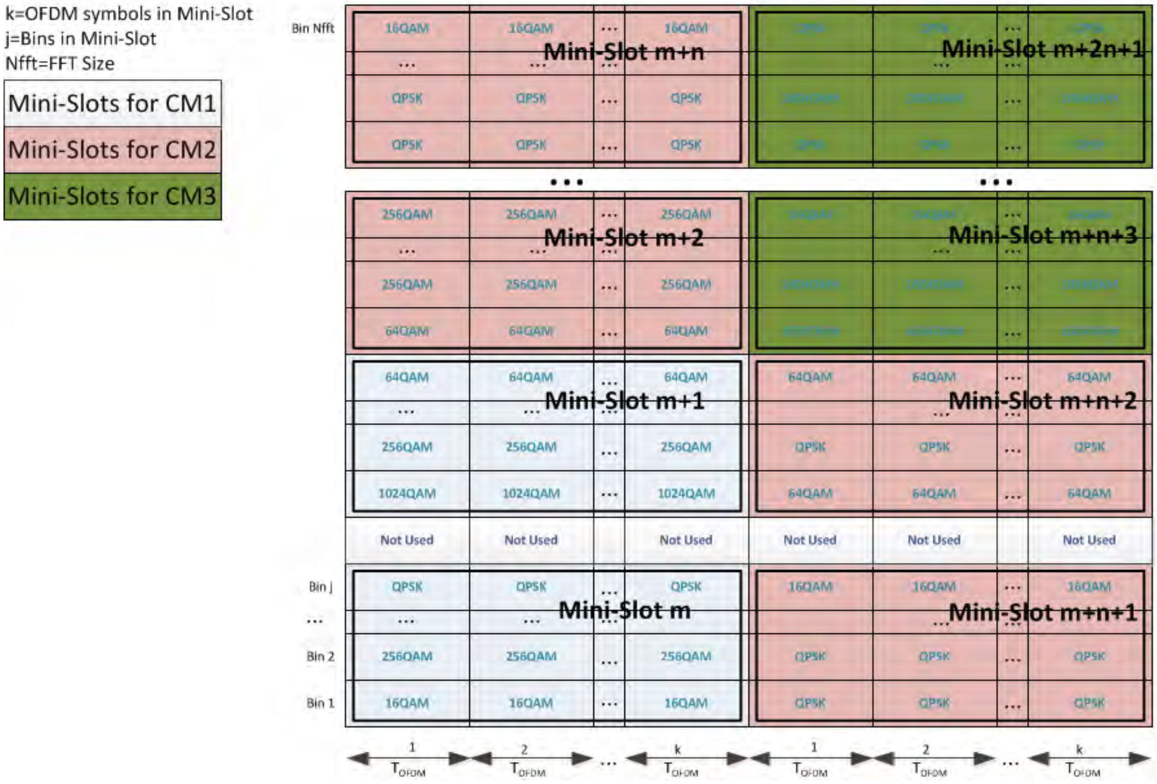


Figure 48 – Mini-slot based scheduling for OFDMA

7.3.3 OFDM Channel Parameter Examples

Table 24 – OFDM Channel Parameters for 192 MHz Wide Channel

Parameter	Value
Channel bandwidth	192 MHz
Useful bandwidth	190 MHz (-95 MHz to +95 MHz) -44 dB attenuation at 96 MHz band-edge
FFT size	16384
FFT sample rate	204.8 MHz (multiple of 10.24 MHz)
Useful symbol time	80 us
Carriers within 190 MHz	15200
Guard interval samples	683 (ratio=1/24; 3.33 us)
Symbol shaping samples	164 (ratio=1/100; 0.8 us)
Total symbol time	84.13us
Continuous pilots	128 (for synchronisation)
Scattered pilots	128 (for channel estimation)
PAPR pilots	128 (for PAPR reduction)
Useful data carriers per symbol	14816
QAM Constellations	4096-QAM, 1024, 256, 64, 16
Bit rate for 4096-QAM w/o FEC	2.11 Gbit/s (11.0 bits/s/Hz)
Bit rate for 1024-QAM w/o FEC	1.76 Gbit/s (9.17 bits/s/Hz)

Table 25 – OFDM Channel Parameters for 96 MHz Wide Channel

Parameter	Value
Channel bandwidth	96 MHz
Useful bandwidth	94 MHz (-47 MHz to +47 MHz) -44 dB attenuation at 48 MHz band-edge
FFT size	8192
FFT sample rate	102.4 MHz (multiple of 10.24 MHz)
Useful symbol time	80 us
Carriers within 94 MHz	7520
Guard interval samples	341 (ratio=1/24; 3.33us)
Symbol shaping samples	82 (ratio=1/100; 0.8 us)
Total symbol time	84.13us
Continuous pilots	64 (for synchronisation)
Scattered pilots	64 (for channel estimation)
PAPR pilots	64 (for PAPR reduction)
Useful data carriers per symbol	7328
QAM Constellations	4096-QAM, 1024, 256, 64, 16
Bit rate for 4096-QAM w/o FEC	1.05 Gbit/s (10.9 bits/s/Hz)
Bit rate for 1024-QAM w/o FEC	0.87 Gbit/s (9.07 bits/s/Hz)

Table 26 – OFDM Channel Parameters for 48 MHz Wide Channel

Parameter	Value
Channel bandwidth	48 MHz
Useful bandwidth	46 MHz (-23 MHz to +23 MHz) -44 dB attenuation at 24 MHz band-edge
FFT size	4096
FFT sample rate	51.2 MHz (multiple of 10.24 MHz)
Useful symbol time	80 us
Carriers within 46 MHz	3680
Guard interval samples	171 (ratio=1/24; 3.33us)
Symbol shaping samples	41 (ratio=1/100; 0.8 us)
Total symbol time	84.13us
Continuous pilots	32 (for synchronisation)
Scattered pilots	32 (for channel estimation)
PAPR pilots	32 (for PAPR reduction)
Useful data carriers per symbol	3584
QAM Constellations	4096-QAM, 1024, 256, 64, 16
Bit rate for 4096-QAM w/o FEC	0.51 Gbit/s (10.65 bits/s/Hz)
Bit rate for 1024-QAM w/o FEC	0.43 Gbit/s (8.88 bits/s/Hz)

Table 27 – OFDM Channel Parameter for 37 MHZ Wide Channel, Upstream NA Band

Parameter	Value
Channel bandwidth	37 MHz
Useful bandwidth	36 MHz (-18 MHz to +18 MHz) -40 dB attenuation at 18.5 MHz (TBC)
FFT size	2048
FFT sample rate	51.2 MHz
Sub-carrier spacing	25 KHz
Useful symbol time	40 us
Carriers within 36 MHz	1440
Guard interval samples	192(ratio=3/32; 3.75 us)
Symbol shaping samples	41 (ratio=1/50; 0.80 us)
Total symbol time	44.55us
Continuous pilots	16 (for synchronisation)
Scattered pilots	none (Channel est. via preamble)
PAPR pilots	16 (for PAPR reduction)
Useful data carriers per symbol	1408
QAM Constellations	1024-QAM, 256, 64, 16, QPSK
Bit rate (for 1024-QAM)	0.32 Gbit/s (8.56 bits/s/Hz)

7.3.3.1 Modulation Summary

DOCSIS 3.0 equipment, completed in 2006, is now seeing increasing field deployment. While deployed CM percentages are still modest, CMTS capabilities are being installed and spectrum plans have been put into place. It has been proven to be rugged and capable, and it is now timely to consider the next phase of DOCSIS evolution. And, as powerful as DOCSIS 3.0 may be, it most certainly can be enhanced by taking advantage of modern tools and the continued advancement in cost-effective, real-time processing power.

Two such approaches have been identified here – adding new symbol rates, similar to the DOCSIS 2.0 extension in 2002 that introduced 5.12 Msps, or introducing multi-carrier modulation, which has been embraced in standards bodies across industries. Table 28 summarizes various attributes of these PHY modulation alternatives relative to today’s available DOCSIS 3.0 baseline for the scaling of services to Gbps rates.

Table 28 – Relative Impact of Extensions to DOCSIS 3.0 for Gigabit Services

Attribute	Wide SC	Wide OFDM	Comments
Silicon Complexity (cost per bit)	-	+	Based on # of real-time multiplication operations
Transparency to existing D3.0 MAC	Same		OFDM: Minor mods to convergence layer
Field Technician Familiarity	+	-	
Robustness to interference	-	+	SC-QAM improved with SCDMA (upstream only)
Robustness to unknown plant (e.g. > 1 GHz operation)	-	+	
Throughput scalability per plant condition (SNR)	-	+	
Peak-to-Avg Power Ratio (PAPR)	Same		OFDM: better with PAPR reduction algorithms
Spectrum Allocation Flexibility	-	+	
New Requirements Definition	+	-	

__Notes__: Wide SC-QAM refers to 8x24 MHz. Wide OFDM refers to 16k IFFT 192 MHz.
"+" and "-" compare wide SC and wide OFDM to a 6.4 Mhz channel-bonded DOCSIS 3.0 baseline.

7.3.4 In Summary

By first stating the criteria, and then analyzing the available options against the criteria, it is suggested that the OFDM/OFDMA/LDPC wide channel is the best candidate for next generation gigabits capable DOCSIS PHY layer. This scheme is based on well-studied, widely adopted methods, allowing quick standardization turn around.

It enables to maximize the throughput with the available and future bandwidth and SNR resources. It is flexible enough to cope with new, less studied spectrum portions and interferences. It is more cost efficient than other alternatives for same throughputs (cost per bit). All these traits suggest that this PHY can optimally serve the DOCSIS evolution going into the gigabit rates, minimizing the investment needed by doing it “once and for all”.

8 DOCSIS MAC TECHNOLOGIES

8.1 DOCSIS Channel Bonding

DOCSIS Channel bonding may support full spectrum downstream. Additional DOCSIS channel bonding upstream may support higher upstream capabilities with targets to 1 Gbps. Achieving larger bonding group will require software, hardware and perhaps specification changes.

A future release of DOCSIS should enable bonding across legacy DOCSIS 3.0 and the new DOCSIS NG, even if they use dissimilar PHY technologies. The MAC layer and IP bonding will stitch the PHY systems together.

8.2 DOCSIS Scheduler Benefits

The DOCSIS protocol allows multiple users to "talk" or transmit at same moment in time and on the same channel, this was part of DOCSIS 2.0 introduction of SCDMA. The introduction of channel bonding allowed ATDMA based system to transmit at the same moment in time on differ frequencies while part of a channel bonding group.

Unlike DOCSIS, the EPON MAC allows "only one" subscriber to "talk" or transmit at any given moment in time. If we consider a single Home Gateway with multiple services and devices behind it, these will contend with each other and neighbors for time slots for transport of voice service, video conferencing, real-time data services, and even normal data and IPTV TCP acknowledgments.

Now, we must consider all the Home Gateways in a serving area domain competing for time slots allocated only on a "per home" basis, if the MSOs move to this style of architecture.

In many ways the EPON and EPOC MAC is most equivalent to a DOCSIS 1.1 MAC, of the 2000 era, because this supports multiple service flows, however allows only "one" user to talk or transmit at a time. The DOCSIS 2.0 and 3.0 specifications changed this limitation to accommodate for more devices, bandwidth, services, and concurrency of users and latency sensitivity; this is a powerful difference between the MAC standards.

The DOCSIS MAC designers knew that shared access meant contention for both bandwidth resources "and" time, this is why DOCSIS 2.0 and 3.0 support simultaneous transmission upstream enabling Quality of Service (QoS) and Quality of Experience (QoE).

There is another major factor with the DOCIS MAC, the development and feature set is controlled by the Cable Industry and not a third party standards organization, like the IEEE or ITU. This allows the MSO to make design request directly to systems vendors for continue innovation and support for new features that come along over time.

The DOCSIS MAC continues to change as the MSOs think of new service differentiation features and the flexible DOCISS MAC enable this support and creating a best in breed and cost effective MAC for the cable industry.

8.3 Services Enabled by DOCSIS

The DOCSIS technology can support virtually any service. DOCSIS technology may enable support for the full range of IP and Ethernet based services. The challenges for support for advanced layer 2 and layer 3 VPN services are not found in the DOCSIS access layer technology, but rather the network elements.

The DOCSIS CMTS will need to add support for desired layer 2 and layer 3 VPN services. The DOCSIS protocol with the use of the advanced MAC should support Ethernet Services types and Bandwidth Profiles defined by the Metro Ethernet Forum (MEF).

8.4 Importance of Backward Compatibility with DOCSIS 3.0 and Any Successor

The authors of this analysis believe that DOCSIS and any successor should consider the value of backwards compatibility especially across channel bonding groups. This assures previous and future investment may be applied to create a large IP based bandwidth network while not stranding previous capital investment and spectrum.

The use of channel bonding leverages every MHz, which is finite and not free, this is all towards an effort to create one large IP pipe to and from the home. The use of backwards compatibility has benefitted the cable industry as well as other industries which use technologies like IEEE Ethernet, WiFi, and EPON creating consumer investment protection, savings, and a smooth migration strategy.

The adoption of backward compatibility simply allows the MSOs to delay and perhaps avoid major investment to the network such as adding more data equipment, spectrum, node splits, and running fiber deeper.

The Data over Cable System Interface Specification (DOCSIS) began development in the late 1990's and has since had four versions released. DOCSIS standards include DOCSIS 1.0, 1.1, 2.0 and 3.0. The standards allowed for backwards compatibility and coexistence with previous versions of the standard.

As the needs of subscribers and providers continued to evolve, the DOCSIS standard was progressively upgraded to accommodate the change in services. The DOCSIS 2.0 standards increased upstream speeds and the DOCSIS 3.0 standard dramatically increased upstream and downstream bandwidth to accommodate higher speed data services.

These transitions capitalized on the availability of new technologies (ex: SCDMA) and the processing power of new silicon families (ex: Channel Bonding).

The authors of this analysis believe that DOCSIS and any successor should consider the value of backwards compatibility especially across channel bonding groups. This assures previous and future investment may be applied to create a large IP based bandwidth network while not stranding previous capital investment and spectrum.

The use of channel bonding leverages every MHz, which are finite and not free, this is all towards an effort to create one large IP pipe, to and from the home. The use of backwards compatibility has benefitted the cable industry as well as other industries which use technologies like IEEE Ethernet, WiFi, and EPON creating consumer investment protection, savings, and a smooth migration strategy.

The adoption of backward compatibility simply allows the MSOs to delay and

perhaps avoid major investment to the network such as adding more data equipment, spectrum, node splits, or running fiber deeper.

1. DOCSIS 3.0 QAM based and any successor should consider that every MHz should all share the same channel bonding group, this maximizes the use of existing spectrum and delays investment

2. Sharing channel bonding groups with DOCSIS 3.0 and Any Successor creates "one" IP Network (cap and grow networks hang around awhile)

3. Sharing the same bonding group assures previous and future investment may be applied in creating larger IP based bandwidth and not stranding previous capital investment

4. Backward Compatibility has benefitted industries like the IEEE Ethernet, WiFi, and EPON saving the entire eco-system money

5. Backward Compatibility simply allows the MSOs to delay and perhaps avoid major investment to the network such as adding more spectrum or running fiber deeper.

6. Avoids the MSO having a RF Data Simulcasting Tax (as discussed in this report)

7. All of our analysis in this report assumes backward compatibility with DOCSIS 3.0 QAM and any successor, like DOCSIS OFDM; thus creating a larger and larger IP bonding group with each year's investment. If this is not the case the investment in HFC upgrades will pull forward. It is uncertain of the exact level of financial impact but the total cost of ownership may be higher when deploying two separate IP based network technologies.

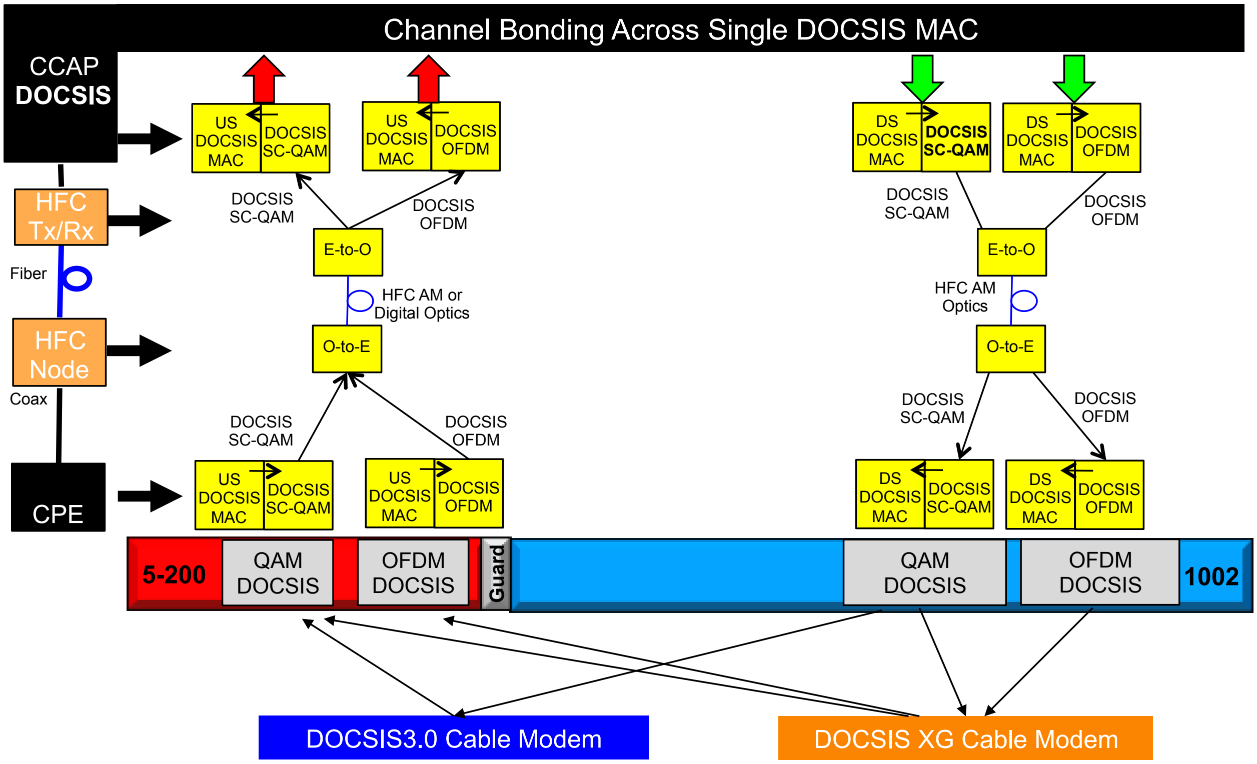


Figure 49 – Channel Bonded DOCSIS 3.0 and DOCSIS NG System

This is an illustration of channel bonding across a DOCSIS 3.0 and potential DOCSIS NG system. Figure 49 shows a DOCSIS 3.0 system coexisting with a DOCSIS NG system, then adding a DOCSIS NG system this platform could support legacy DOCSIS 3.0 SC-QAM, modulation and perhaps add 256-QAM upstream and 1024-QAM downstream, and RS and also supporting the new DOCSIS NG PHY. This will allow backward compatibility for the DOCSIS 3.0 cable modems and CMTS, while supporting the new PHY and likely in new spectrum.

Figure 50 is an illustration of the possible integration of HFC optics in the CCAP that will support DOCSIS 3.0 and DOCSIS NG. DOCSIS 3.0 and DOCSIS NG will likely be supported on the same card in the future without requiring HFC optical integration to the CCAP.

8.5 RF Data Simulcasting Tax

We would recommend strongly examining the history and impact of simulcasting services. If an alternative to DOCSIS is considered this will require new spectrum. The existing DOCSIS service and spectrum allocation may actually continue to grow during the initial introduction of the new data MAC/PHY technology, such as EPOC.

New spectrum that likely mirrors the size of DOCSIS would have to be found, so that at least the same services may be offered using an EPOC technology. The amount of new spectrum allocated by the MSO for DOCSIS and EPOC would begin the RF Data Simulcasting Tax Period.

It is true, that legacy networks tend to hang around for a long time. For example,

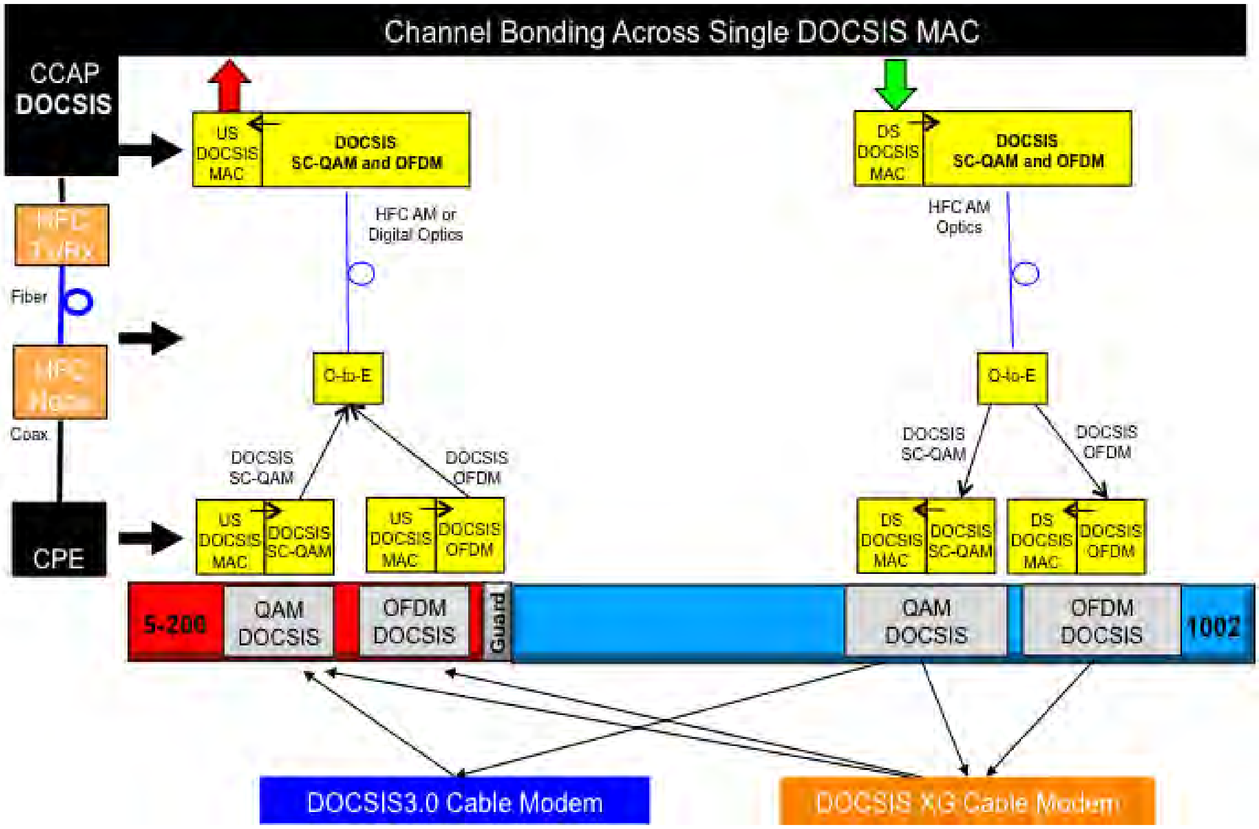


Figure 50 – Channel Bonded DOCSIS 3.0 and 4.0 System with CCAP Integrated HFC Optics

MSOs that deployed constant bit rate voice services, known as CBR voice, may still have these technologies occupying spectrum, even though they also have voice services using DOCSIS in the same network. The challenge is cost; the cost to reclaim spectrum is substantial, it requires new CPE and Headend systems, for no additional revenue.

The additional impact is finding new spectrum to offer what is a duplicate service using a different technology. It is fair to say that the cost for supporting parallel RF data networking technologies will have a capital and operational impact that will likely be more than expanding the current technology over the existing HFC network.

DOCSIS has the ability with each passing year investment to create larger and larger IP bonding groups, to enable higher speed service tiers and support traffic growth. Additionally, the DOCSIS CPEs may be channel bonded with legacy PHY and/or new PHY technologies, while all sharing the same MAC layer-bonding group.

Also, not a single DOCSIS CPE would be required to change to reclaim spectrum, because of backward compatibility or to eliminate the RF data simulcasting tax, as this network tax could be avoided with DOCSIS current and future systems.

This is a compelling feature of continuing to leverage DOCSIS 3.0 and why next generation DOCSIS needs to be backward compatible at the MAC layer, with different PHYs.

1. The amount of new spectrum allocated by the MSO for DOCSIS and EPOC would begin the RF Data Simulcasting Tax Period.
2. The existing DOCSIS service and spectrum allocation may actually continue to grow during an initial introduction of a new data MAC/PHY technology, such as EPOC.
3. Legacy networks tend to hang around for a long time, CBR Voice.
4. A challenge is the cost to reclaim spectrum is substantial; it requires new CPE and Headend systems, for likely no additional revenue.
5. The additional impact is finding new spectrum to offer what is a duplicate service offering using a different technology, to find capacity node splits, new node placement in the field, and/or spectrum expansion, new powering for the OSP equipment, and more are all impacts.
6. It is fair to say that the cost for supporting a parallel RF data networking technology will have a capital and operational impact.
7. The ability that DOCSIS has is that with each passing year spectrum is allocated creating larger and larger IP bonding groups, to enable higher speed service tiers and support traffic growth.
8. This is a compelling feature of continuing to leverage DOCSIS 3.0 and why next generation DOCSIS needs to be backward compatible.

9 NETWORK CAPACITY ANALYSIS

9.1 Intro

The network capacity of the cable access network is determined by the amount of spectrum available and the data rate possible within the spectrum. The modern cable network is incredibly flexible allowing the MSO to make targeted investments where and when needed to either incrementally or in some cases substantially increase network capacity depending on the capacity expansion method selected.

The use of capacity expansion methods may be applied across an entire network footprint or with laser beam focus to address capacity challenges. Figure 51 is an attempt to capture the various methods available to increase or improve capacity of the network. The diagram brings together methods and techniques used by various disciplines within the MSO, such as outside/inside plant, IP/Data, SDV, and Video Processing. The techniques will allow the MSO to transform their network from broadcast to unicast and from analog/digital to IP.

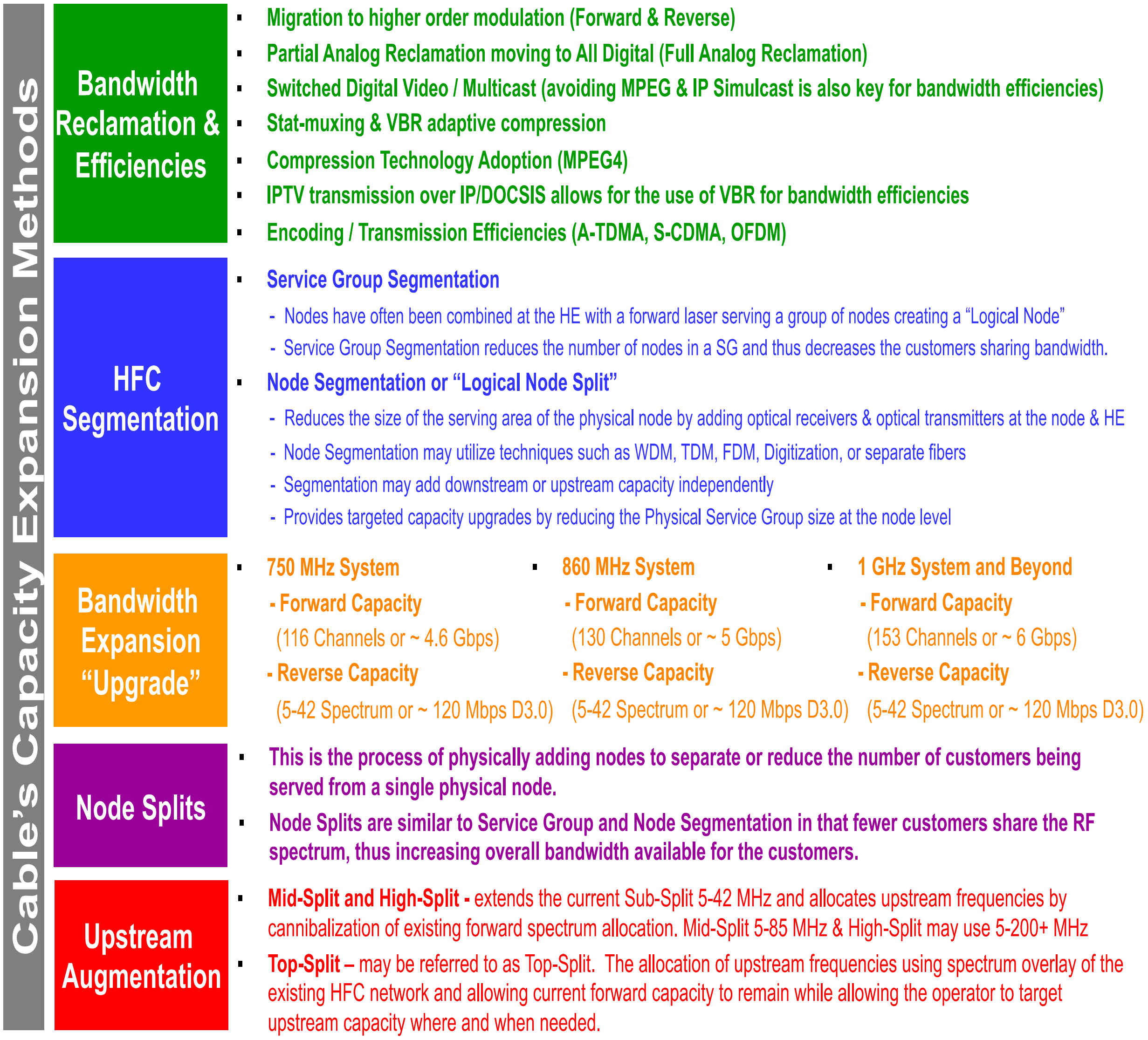
Today, in fact MSOs may use techniques to increase capacity without touching the outside plant; this is dramatically different than the approaches that were used for decades. The technique referred to as Bandwidth Reclamation and Efficiencies, as illustrated in the top of Figure 51 is becoming the primary method to address system wide capacity challenges. In most cases this technique may be implemented with equipment in the headend and home, thus not requiring conditioning of the outside plant or headend optics.

A technique recently put into practice by some cable operators is partial or even full analog reclamation. This enables the operator to transition the channels currently transmitted in analog and to transmit them only in digital format allowing greater bandwidth efficiencies by requiring the use of a digital terminal adapter (DTA) alongside televisions that may have only had analog services.

Another technique for Bandwidth Reclamation and Efficiencies is the use of Switch Digital Video (SDV). The use of SDV allows the cable operator to transmit in the network only the video streams that are being viewed by consumers. This allows the operator to increase the number of channels offered to consumers, in fact the actual channels offered to the consumers may exceed the throughput capabilities of the network but through careful traffic engineering and capacity planning this approach is an excellent way of adding additional capacity to the network.

This technique is a form of over-subscription and has been in practice for decades by the telecommunication industry. The items captured in Bandwidth Reclamation and Efficiencies are the modern methods to expand capacity. In many respects the Bandwidth Expansion “upgrade” approach as illustrated in Figure 51 whereby the entire network was upgraded to increase capacity, may be seldom used in the future. If used, this may be part of a joint plan to increase the spectrum allocation of the return path.

In the future, the use of IP for video delivery will provide even greater bandwidth efficiencies. IP used for digital video transmission and will also provide functionality similar to the techniques used in SDV. Another key advantage is that IP allows for the use of variable bitrate (VBR)



Cable's Capacity Expansion Methods

Bandwidth Reclamation & Efficiencies

- Migration to higher order modulation (Forward & Reverse)
- Partial Analog Reclamation moving to All Digital (Full Analog Reclamation)
- Switched Digital Video / Multicast (avoiding MPEG & IP Simulcast is also key for bandwidth efficiencies)
- Stat-muxing & VBR adaptive compression
- Compression Technology Adoption (MPEG4)
- IPTV transmission over IP/DOCSIS allows for the use of VBR for bandwidth efficiencies
- Encoding / Transmission Efficiencies (A-TDMA, S-CDMA, OFDM)

HFC Segmentation

- Service Group Segmentation
  - Nodes have often been combined at the HE with a forward laser serving a group of nodes creating a "Logical Node"
  - Service Group Segmentation reduces the number of nodes in a SG and thus decreases the customers sharing bandwidth.
- Node Segmentation or "Logical Node Split"
  - Reduces the size of the serving area of the physical node by adding optical receivers & optical transmitters at the node & HE
  - Node Segmentation may utilize techniques such as WDM, TDM, FDM, Digitization, or separate fibers
  - Segmentation may add downstream or upstream capacity independently
  - Provides targeted capacity upgrades by reducing the Physical Service Group size at the node level

Bandwidth Expansion "Upgrade"

- 750 MHz System
  - Forward Capacity (116 Channels or ~ 4.6 Gbps)
  - Reverse Capacity (5-42 Spectrum or ~ 120 Mbps D3.0)
- 860 MHz System
  - Forward Capacity (130 Channels or ~ 5 Gbps)
  - Reverse Capacity (5-42 Spectrum or ~ 120 Mbps D3.0)
- 1 GHz System and Beyond
  - Forward Capacity (153 Channels or ~ 6 Gbps)
  - Reverse Capacity (5-42 Spectrum or ~ 120 Mbps D3.0)

Node Splits

- This is the process of physically adding nodes to separate or reduce the number of customers being served from a single physical node.
- Node Splits are similar to Service Group and Node Segmentation in that fewer customers share the RF spectrum, thus increasing overall bandwidth available for the customers.

Upstream Augmentation

- **Mid-Split and High-Split -** extends the current Sub-Split 5-42 MHz and allocates upstream frequencies by cannibalization of existing forward spectrum allocation. Mid-Split 5-85 MHz & High-Split may use 5-200+ MHz
- **Top-Split –** may be referred to as Top-Split. The allocation of upstream frequencies using spectrum overlay of the existing HFC network and allowing current forward capacity to remain while allowing the operator to target upstream capacity where and when needed.

Figure 51 – Cable's Capacity Expansion Methods

encoding increasing the capacity of the network and the utilization of higher order compression techniques.

Cable operator's selection priority of the capacity expansion methods has and will continue to vary. The cable operators will eventually use all or nearly all of the capacity expansion methods in Figure 51

9.2 Importance of Error Correction Technologies

The paper by David J.C. MacKay and Edward A. Ratzer, titled "Gallager Codes for High Rate Applications", published January 7, 2003 [27], examines the improvements obtained by switching from Reed-Solomon codes to Gallager codes or Low Density

Parity-Check (LDPC) code. It is the opinion of this author, that the MacKay paper is one of the best comparisons of illustrating the benefits of switching to LDPC from Reed-Solomon. The paper initially released in 2003, suggests some modifications to Gallager codes to improve performance. The paper suggest about a 5 dB gain. The paper lists further ideas worth investigating that may improve performance.

The use of LDPC has expanded recently with the adoption by the IEEE WiMAX 802.16e, ITU-T G.hn. and the cable industry use for downstream transmission in DVB-C2. The use of LDPC may be used in any carrier modulation method, such as SC-QAM, OFDM, or Wavelet, and the expectation is the use of higher order modulation is achievable compared with Reed-Solomon based systems. It is reasonable to suggest a 6 dB gain is possible by switching from Reed-Solomon to LDPC and this will allow an increase in modulation by perhaps two orders, in other words perhaps one could move from 64-QAM to perhaps 256-QAM. In Table 29, the R-S using approximately 86-87% coding and LDPC using the inner code of 5/6 or 83% yields a 6 dB difference and will allow an increase of two orders of the modulation.

The key takeaway is the use of LDPC will improve network capacity or actual bit per second per Hertz over Reed-Solomon based systems, and this is achieved by enabling the use of higher order modulation with the same signal-to-noise ratio (SNR) condition. This allows operators to allocate less spectrum compared to Reed-Solomon based systems or have more network capacity in occupied spectrum.

The benefits of the cable industry's use can be seen in DVB-C2 systems. However, the use of LDPC for upstream cable data use is still under study as seen in this report. There are also other error correction technologies to consider that have been adopted by other standards groups.

This section will state the major differences and reasons why the use of modern error correction technology is key to increasing network capacity. The new error correction technology and the assumed two-order increase in modulation while operating in the same Signal to Noise Ratio (SNR) environment is the major reason there is an improvement in capacity.

Refer to Table 29 to Table 31 for the DOCSIS Single Carrier-QAM with Reed-Solomon system verse the performance estimates of a DOCSIS Multi-carrier OFDM with LDPC system and also refer to Table 32 to Table 34 for the analysis of these competing PHY layer technologies.

This section compares DOCSIS Single Carrier QAM and the current error correction technology with the proposed DOCSIS NG use of OFDM and the modern LDPC error correction technology.

9.3 DOCSIS 3.0 Single Carrier-QAM with Reed-Solomon

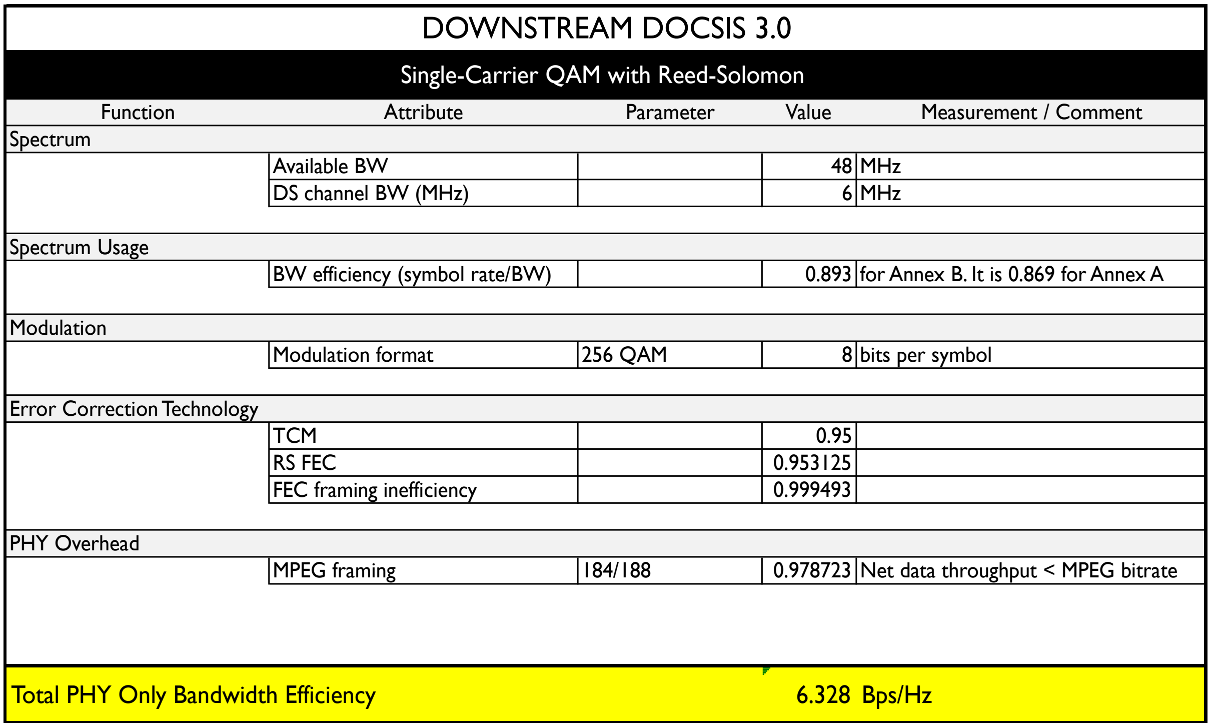
The DOCSIS SC-QAM 256-QAM downstream, as shown in Table 29 and the following two tables models the upstream using DOCSIS SC-QAM 64-QAM and DOCSIS 256-QAM. Each scenario assumes ATDMA.

These tables measure the PHY layer spectral efficiency of DOCSIS QAM based solutions. The channel coding for controlling errors in data transmission for the DOCSIS examples use Reed-Solomon forward error correction (RS-FEC) and Trellis Modulation or also known as Trellis Coded Modulation (TCM).

These are used to calculate the network capacity of the cable network considering several spectrum options found in the Network Capacity section.

A key take away is performance gap between 256-QAM PHY and 64-QAM layer efficiencies. The assumptions for 64-QAM at 4.1 bps/Hz would require 33% more spectrum and DOCSIS channels to maintain the equivalent PHY layer throughput. The use of DOCSIS 256-QAM for the upstream is not part of the DOCSIS standards. However some CMTS and CM products support this modulation profile in hardware.

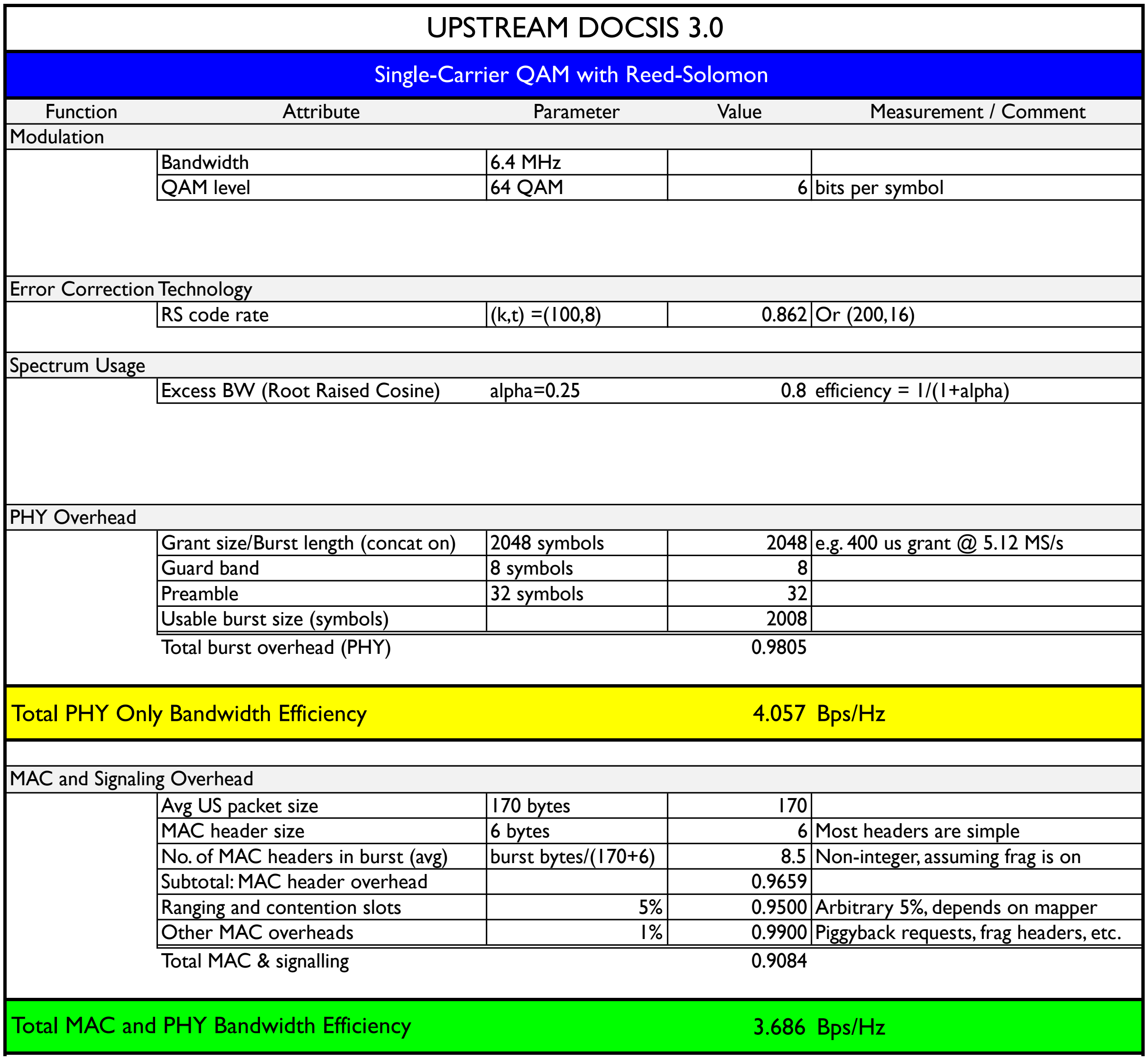
Table 29 – Downstream DOCSIS 3.0 256-QAM with Reed-Solomon & TCM



DOWNSTREAM DOCSIS 3.0				
Single-Carrier QAM with Reed-Solomon				
Function	Attribute	Parameter	Value	Measurement / Comment
Spectrum				
	Available BW		48	MHz
	DS channel BW (MHz)		6	MHz
Spectrum Usage				
	BW efficiency (symbol rate/BW)		0.893	for Annex B. It is 0.869 for Annex A
Modulation				
	Modulation format	256 QAM	8	bits per symbol
Error Correction Technology				
	TCM		0.95	
	RS FEC		0.953125	
	FEC framing inefficiency		0.999493	
PHY Overhead				
	MPEG framing	184/188	0.978723	Net data throughput < MPEG bitrate
Total PHY Only Bandwidth Efficiency			6.328	Bps/Hz

The DOCSIS specifications could be modified to include 256-QAM upstream as well as 1024-QAM in the upstream and downstream. However, the real major gains would be achieved by changing the error correction technology.

Table 30 – Upstream DOCSIS 3.0 64-QAM with Reed Solomon

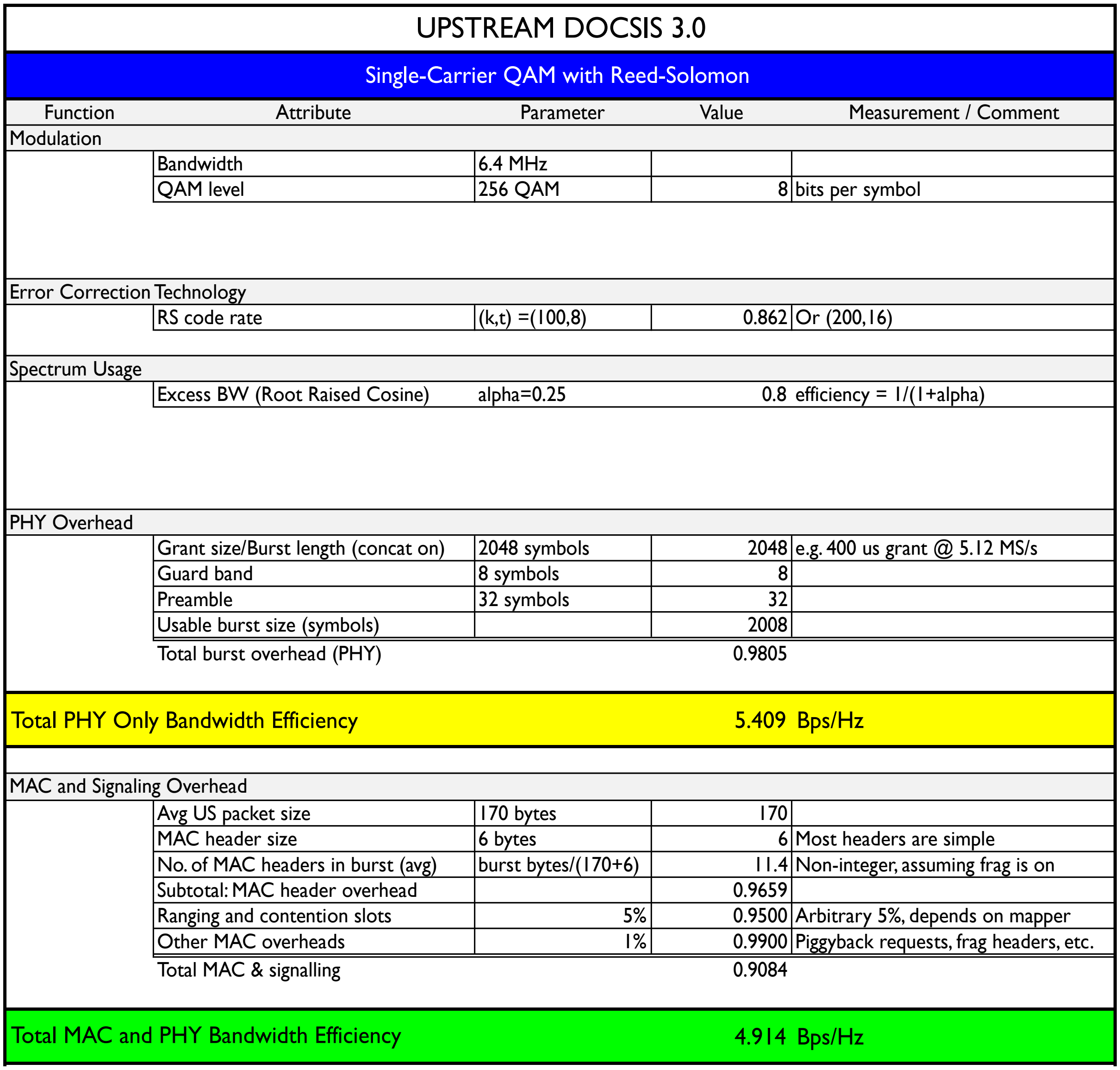


UPSTREAM DOCSIS 3.0

Single-Carrier QAM with Reed-Solomon

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	6.4 MHz		
	QAM level	64 QAM	6	bits per symbol
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Excess BW (Root Raised Cosine)	alpha=0.25	0.8	efficiency = 1/(1+alpha)
PHY Overhead				
	Grant size/Burst length (concat on)	2048 symbols	2048	e.g. 400 us grant @ 5.12 MS/s
	Guard band	8 symbols	8	
	Preamble	32 symbols	32	
	Usable burst size (symbols)		2008	
	Total burst overhead (PHY)		0.9805	
Total PHY Only Bandwidth Efficiency			4.057	Bps/Hz
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	MAC header size	6 bytes	6	Most headers are simple
	No. of MAC headers in burst (avg)	burst bytes/(170+6)	8.5	Non-integer, assuming frag is on
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Piggyback requests, frag headers, etc.
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			3.686	Bps/Hz

Table 31 – Upstream DOCSIS 3.0 256-QAM with Reed Solomon



UPSTREAM DOCSIS 3.0

Single-Carrier QAM with Reed-Solomon

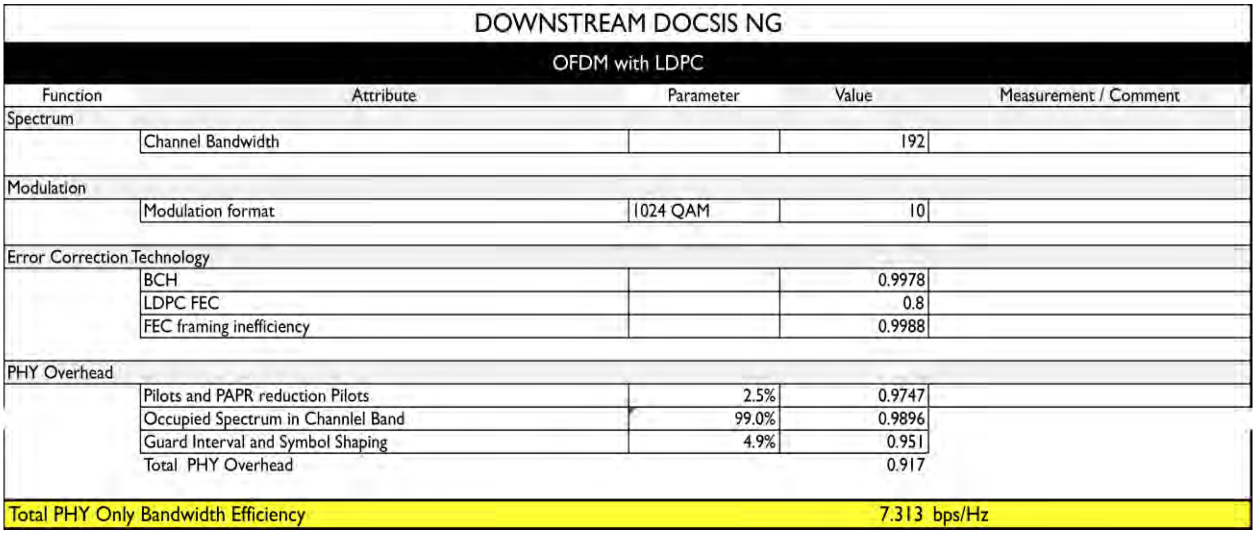
Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	6.4 MHz		
	QAM level	256 QAM	8	bits per symbol
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Excess BW (Root Raised Cosine)	alpha=0.25	0.8	efficiency = 1/(1+alpha)
PHY Overhead				
	Grant size/Burst length (concat on)	2048 symbols	2048	e.g. 400 us grant @ 5.12 MS/s
	Guard band	8 symbols	8	
	Preamble	32 symbols	32	
	Usable burst size (symbols)		2008	
	Total burst overhead (PHY)		0.9805	
Total PHY Only Bandwidth Efficiency			5.409	Bps/Hz
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	MAC header size	6 bytes	6	Most headers are simple
	No. of MAC headers in burst (avg)	burst bytes/(170+6)	11.4	Non-integer, assuming frag is on
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Piggyback requests, frag headers, etc.
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			4.914	Bps/Hz

9.4 DOCSIS NG Multi-carrier OFDM with Low Density Parity-Check (LDPC) code

The analysis in this section provides measurements using OFDM/OFDMA. Again OFDM is not part of the DOCSIS 3.0 standard. The channel coding for controlling errors in data transmission is assumed to use Low Density Parity-Check (LDPC) code also referred to as Gallager codes.

The analysis also uses values as described in Section 7.3.3 OFDM Channel Parameter Examples discuss in this paper. The target for these DOCSIS NG OFDM and LDPC estimates is to use an error correction amount referred to as 5/6 inner code rates or .833. The strong error correction used for the LDPC is modeled to achieve the Carrier to Noise target of 6 dB below Reed Solomon code rate of 86%. This will mean for the same modulation format R-S will yield greater b/s/Hz than LDPC using a stronger FEC in this effort to achieve a 6 dB decrease in C/N.

Table 32 – Downstream DOCSIS OFDM 1024-QAM with LDPC



DOWNSTREAM DOCSIS NG				
OFDM with LDPC				
Function	Attribute	Parameter	Value	Measurement / Comment
Spectrum				
	Channel Bandwidth		192	
Modulation				
	Modulation format	1024 QAM	10	
Error Correction Technology				
	BCH		0.9978	
	LDPC FEC		0.8	
	FEC framing inefficiency		0.9988	
PHY Overhead				
	Pilots and PAPR reduction Pilots	2.5%	0.9747	
	Occupied Spectrum in Channlel Band	99.0%	0.9896	
	Guard Interval and Symbol Shaping	4.9%	0.951	
	Total PHY Overhead		0.917	
Total PHY Only Bandwidth Efficiency			7.313 bps/Hz	

The downstream DOCSIS OFDM 1024-QAM with LDPC system has about a 20% performance improvement of DOCSIS SC-QAM 256-QAM with Reed-Solomon. This is attributed primary to the FEC and not to the change in multi-carrier OFDM. The modern FEC will support greater Modulation QAM Format in the same SNR.

In the previous figures, 256-QAM was analyzed using estimates for PHY and MAC layer efficiency comparing DOCSIS single carrier 256-QAM and DOCSIS OFDM 256-QAM. The use of LDPC may allow higher upstream modulation schemes to be used compared with Reed-Solomon based approaches.

This could mean that 64-QAM Reed-Solomon system may actually be compared with an OFDM 256-QAM LDPC based system in the same Signal to Noise Ratio environment. Moreover, a 256-QAM Reed-Solomon system may actually be compared with a OFDM 1024-QAM LDPC based system in the same SNR environment.

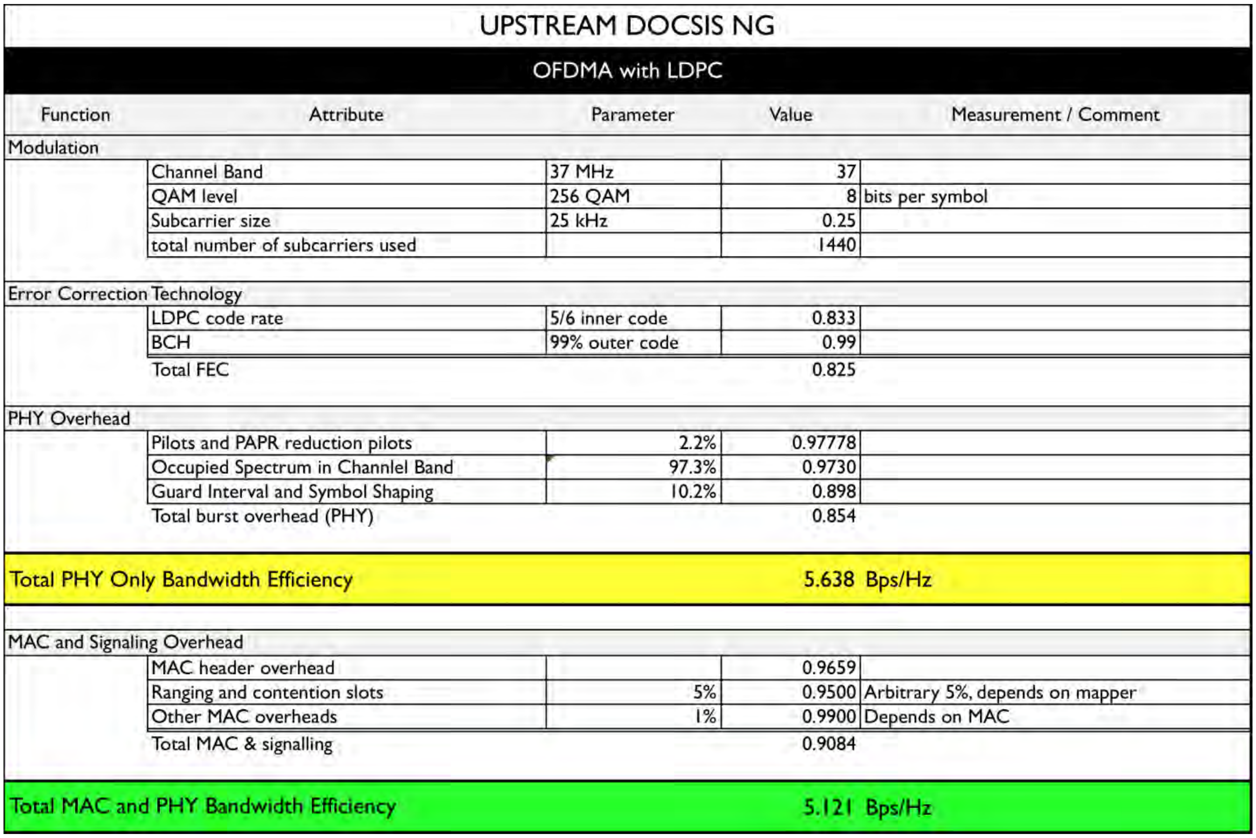
The goal to target the OFDM and LDPC system to operated in the same SNR environment and with two orders increase in

QAM level, required us to apply more error correction codes to LDPC.

Again, because we are assuming that LDPC will be capable of operating in the same SNR environment while using 2 orders higher modulation than a Reed Solomon system. This accounts for the added FEC overhead and lower performance when using the same QAM level.

The actual performance of either system in real-world HFC deployments is unknown. There are many attributes and assumptions than can be modified. We used an estimate that we considered to be fair for single carrier QAM and OFDM. These are subject to debate until systems are tested in a cable system.

Table 33 – Upstream DOCSIS OFDM 256-QAM with LDPC

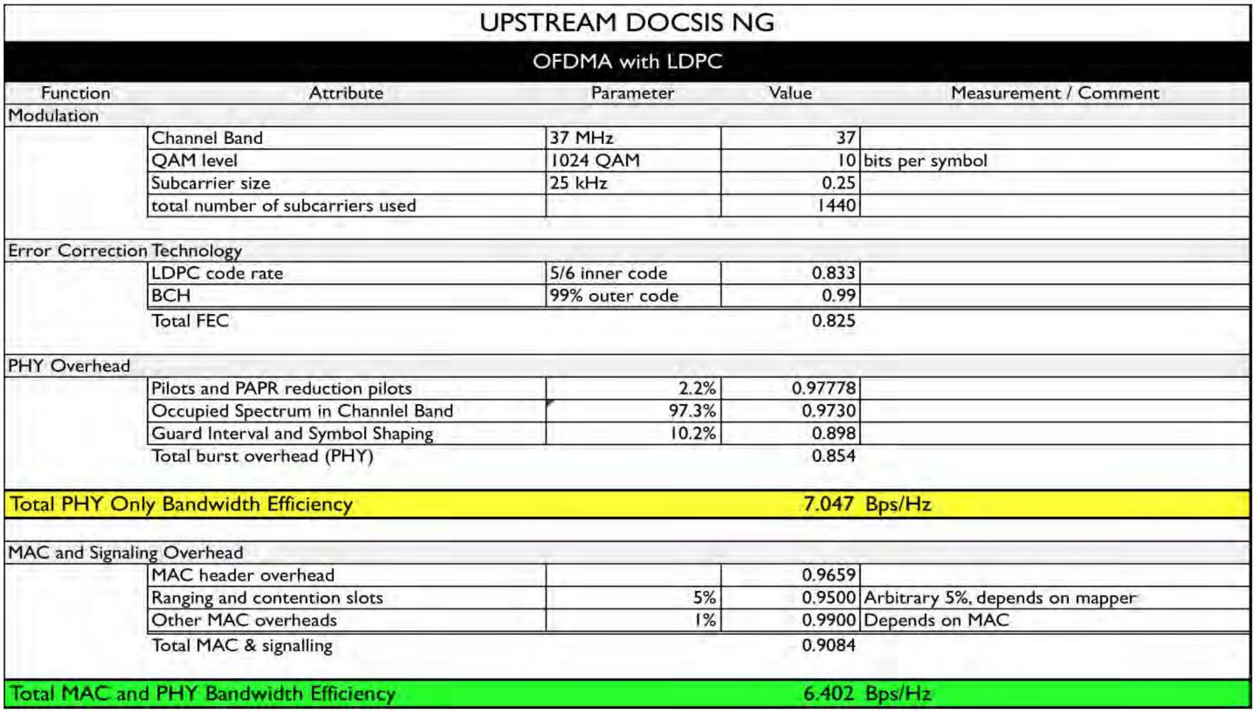


UPSTREAM DOCSIS NG

OFDMA with LDPC

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Channel Band	37 MHz	37	
	QAM level	256 QAM	8	bits per symbol
	Subcarrier size	25 kHz	0.25	
	total number of subcarriers used		1440	
Error Correction Technology				
	LDPC code rate	5/6 inner code	0.833	
	BCH	99% outer code	0.99	
	Total FEC		0.825	
PHY Overhead				
	Pilots and PAPR reduction pilots	2.2%	0.97778	
	Occupied Spectrum in Channlel Band	97.3%	0.9730	
	Guard Interval and Symbol Shaping	10.2%	0.898	
	Total burst overhead (PHY)		0.854	
Total PHY Only Bandwidth Efficiency			5.638	Bps/Hz
MAC and Signaling Overhead				
	MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Depends on MAC
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			5.121	Bps/Hz

Table 34 – Upstream DOCSIS OFDM 1024-QAM with LDPC



UPSTREAM DOCSIS NG

OFDMA with LDPC

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Channel Band	37 MHz	37	
	QAM level	1024 QAM	10	bits per symbol
	Subcarrier size	25 kHz	0.25	
	total number of subcarriers used		1440	
Error Correction Technology				
	LDPC code rate	5/6 inner code	0.833	
	BCH	99% outer code	0.99	
	Total FEC		0.825	
PHY Overhead				
	Pilots and PAPR reduction pilots	2.2%	0.97778	
	Occupied Spectrum in Channlel Band	97.3%	0.9730	
	Guard Interval and Symbol Shaping	10.2%	0.898	
	Total burst overhead (PHY)		0.854	
Total PHY Only Bandwidth Efficiency			7.047	Bps/Hz
MAC and Signaling Overhead				
	MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Depends on MAC
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			6.402	Bps/Hz

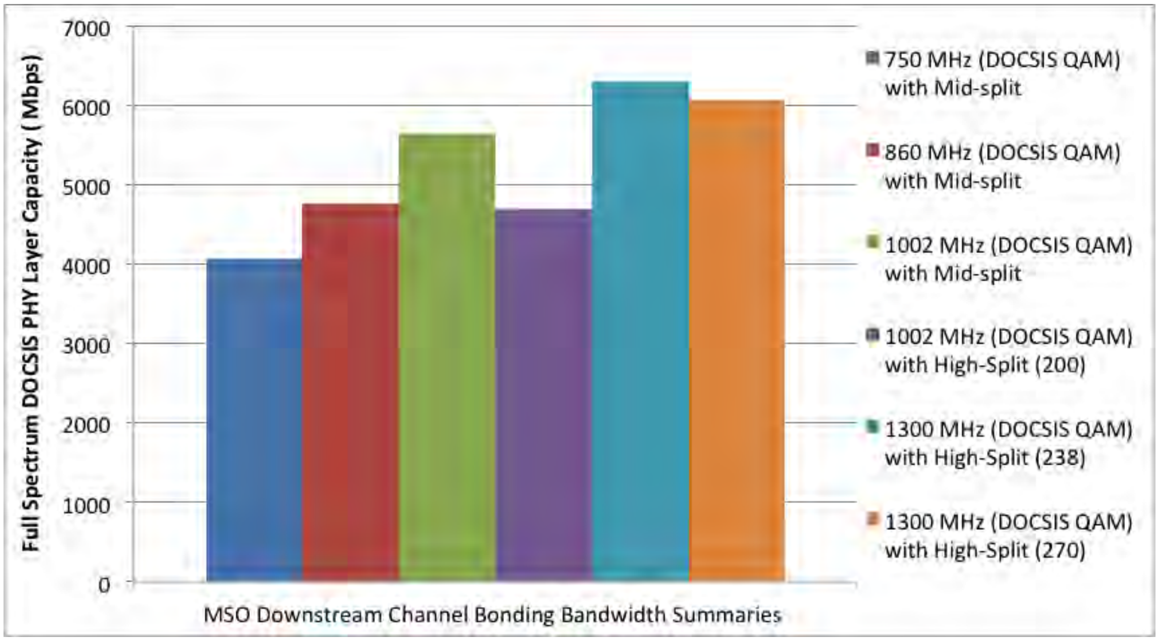


Figure 52 – 256 SC-QAM RS Codes PHY

9.5 Downstream Capacity

The most critical determination for the capacity of the network is the amount of spectrum available. The determination of the downstream capacity will assume the eventual migrations to an all IP based technology. The migration to all IP on the downstream which will optimize the capacity of the spectrum providing the versatility to use the network for any service type and provide the means to compete with PON and the flexibility to meet the needs of the future.

Table 35 provides capacity projections considering the upstream spectrum split and the use of DOCSIS Single Carrier QAM using several downstream spectrum allocations from 750 MHz to 1002 MHz. Certainly there are other spectrum options that could be considered such as moving the downstream above 1 GHz such as 1300 MHz as well as other spectrum options for the upstream. This table will calculate the estimated downstream PHY layer capacity using several spectrum options with limits of 256-QAM though higher modulations are possible.

Figure 52 shows different downstream spectrum allocations as well as the removal of upstream spectrum from the downstream. The downstream network capacity is illustrated using DOCSIS 256 SC-QAM Reed-Solomon Codes PHY or DOCSIS 1024-QAM OFDM LDPC capacity assuming full spectrum.

9.6 Upstream Capacity

The upstream capacity measurements are more complicated and not as straightforward as the downstream capacity projections. In the Figure 53, many of the spectrum split options were evaluated considering several PHY layer options and modulation schemes within each spectrum split.

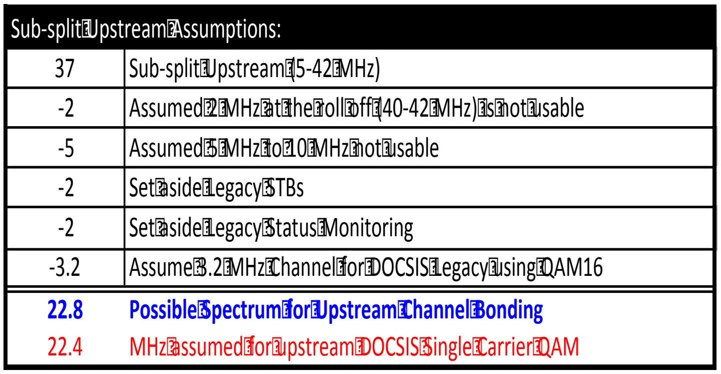
These are some key assumptions about the upstream capacity estimates:

Table 35 – 256 SC-QAM RS PHY or 1024-QAM OFDM LDPC Full Spectrum Capacity

Split Type	MSO Downstream Channel Bonding Bandwidth Summaries	Total Downstream Spectrum Available	DOCSIS QAM Usable Data Rate Per MHz 256 QAM)	DOCSIS OFDM Usable Data Rate Per MHz (Assuming 1024 QAM OFDM w/ LDPC)	Total Capacity Data Rate Usable (Mbps)
Downstream Capacity with Sub-split (5-42 MHz)	750 MHz (DOCSIS QAM) with Sub-split	696	6.328	7.313	4404
	750 MHz DOCSIS OFDM OFDM w/ LDPC with Sub-split	696	6.328	7.313	5090
	860 MHz (DOCSIS QAM) with Sub-split	806	6.328	7.313	5100
	860 MHz DOCSIS OFDM OFDM w/ LDPC with Sub-split	806	6.328	7.313	5894
	1002 MHz (DOCSIS QAM) with Sub-split	948	6.328	7.313	5999
	1002 MHz DOCSIS OFDM OFDM w/ LDPC with Sub-split	948	6.328	7.313	6933
Downstream Capacity with Mid-split	1002 MHz (DOCSIS QAM) with Mid-split	897	6.328	7.313	5676
	1002 MHz DOCSIS OFDM OFDM w/ LDPC with Mid-split	897	6.328	7.313	6560
Downstream Capacity with High-Split (238)	1050 MHz (DOCSIS QAM) with High-Split (238)	750	6.328	7.313	4746
	1050 MHz DOCSIS OFDM OFDM w/ LDPC with High-Split (238)	750	6.328	7.313	5485
	1300 MHz (DOCSIS QAM) with High-Split (238)	1000	6.328	7.313	6328
	1300 MHz DOCSIS OFDM OFDM w/ LDPC with High-Split (238)	1000	6.328	7.313	7313
Downstream Capacity with Top-split (900-1125)	750 MHz (DOCSIS QAM) with Top-split (900-1050)	696	6.328	7.313	4404
	750 MHz DOCSIS OFDM OFDM w/ LDPC with Top-split (900-1050)	696	6.328	7.313	5090
Downstream Capacity with Top-split (1250-1750)	1002 MHz (DOCSIS QAM) with Top-split (1250-1750)	948	6.328	7.313	5999
	1002 MHz DOCSIS OFDM OFDM w/ LDPC with Top-split (1250-1750)	948	6.328	7.313	6933

- Sub-split and/or Mid-split channel bonding spectrum was counted in capacity summaries with any new spectrum split (Figure 54 does illustrate Top-split spectrum options and the capacity. Note that Sub-split and Mid-split are add to these options)

- Included in the analysis are PHY layer efficiency estimates as well as MAC layer efficiency estimates. This will be labeled in each model



Sub-split Upstream Assumptions:	
37	Sub-split Upstream (5-42 MHz)
-2	Assumed 2 MHz at the roll off (40-42 MHz) is not usable
-5	Assumed 5 MHz to 10 MHz not usable
-2	Set aside Legacy STBs
-2	Set aside Legacy Status Monitoring
-3.2	Assume 3.2 MHz Channel for DOCSIS Legacy using QAM16
22.8	Possible Spectrum for Upstream Channel Bonding
22.4	MHz assumed for upstream DOCSIS Single Carrier QAM

Figure 53 – Sub-split Assumptions

An important assumption is that the upstream capacity measurements assume that spectrum blocks from the sub-split region and any new spectrum split will all share a common channeling bonding domain. This is essentially assuming that backwards compatibility is part of the upstream capacity projections.

The upstream capacity projections for each split will assume DOCSIS QAM – and if adopted in the future – DOCSIS OFDM based systems will all share the same channel-bonding group. This will allow for previous, current, and future investments made by the cable operator to be applied to a larger and larger bandwidth pipe or overall upstream capacity.

If backwards compatibility were not assumed, the spectrum options would have to allocate spectrum for

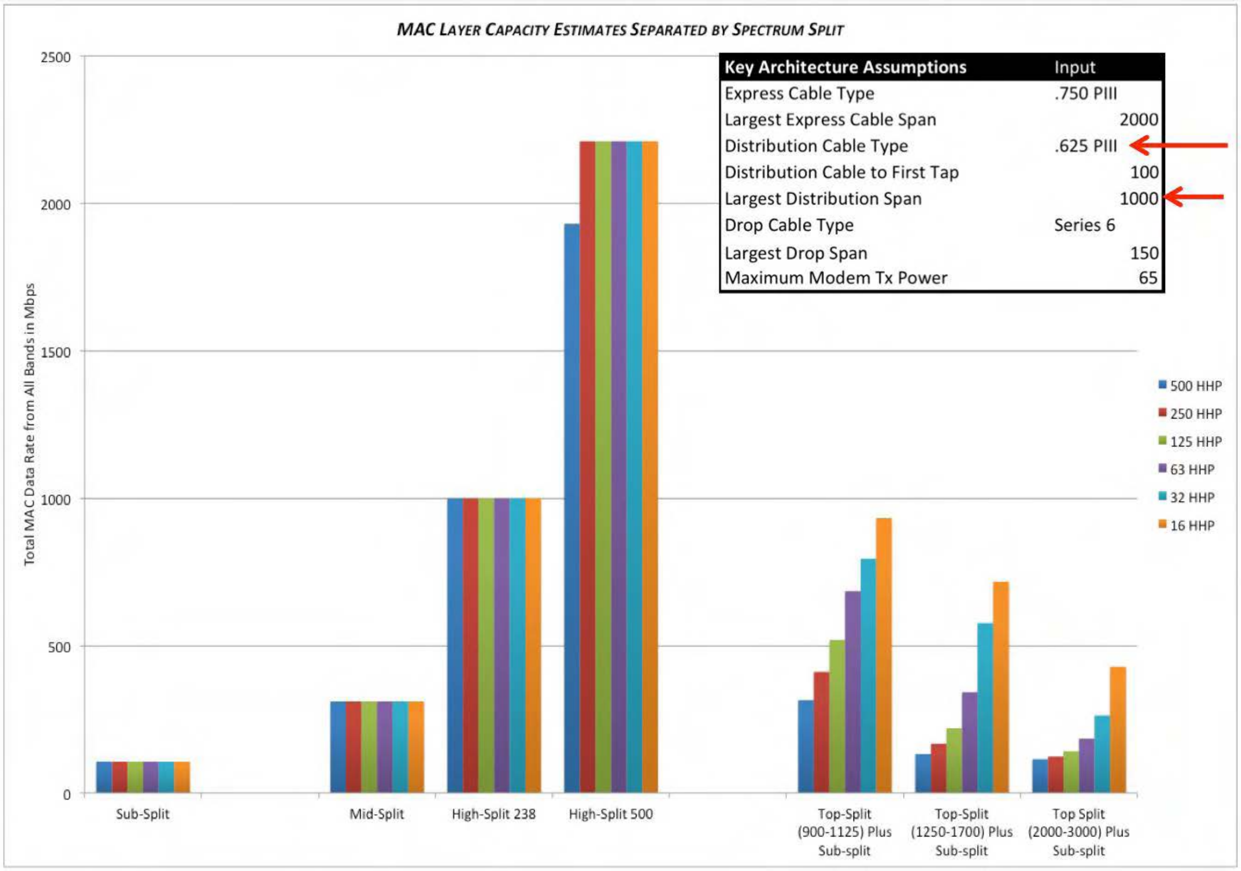
DOCSIS QAM and separate capacity for any successor technology, resulting in a lower capacity throughput for the same spectrum allocation. This would compress the duration of time that the same spectrum may be viable to meet the needs of the MSO.

9.6.1 Achieving 1 Gbps Symmetrical Services and Beyond with DOCSIS 3.0

A major interest of the cable operators is the understanding of the architecture requirements for each spectrum split option to achieve 1 Gbps MAC layer performance. The migration strategy to reach 1 Gbps may be of interest as well, so that an operator can make incremental investment if desired to meet the capacity needs over time, this is sort of a pay as you grow approach.

We have modeled the MAC layer capacity estimates for each node service group size starting at 500 HHP and splitting the service group size in half until reaching 16 HHP, equivalent of fiber to the last active (FTTLA). The model assumes .625 PIII distribution cable with the largest span of 1000 feet in the architecture calculations as shown in Figure 54.

The upstream capacity measurements found in Figure 54 compares various spectrum splits using DOCSIS single carrier QAM with Reed Solomon with a maximum of 256-QAM. The spectrum splits found in the table include Sub-split, Mid-split, High-split (238), High-split (500), Top-split (900-11125) with Sub-split, Top-split (1250-1700) with Sub-split, Top-split (2000-3000)



Key Architecture Assumptions	Input
Express Cable Type	.750 PIII
Largest Express Cable Span	2000
Distribution Cable Type	.625 PIII
Distribution Cable to First Tap	100
Largest Distribution Span	1000
Drop Cable Type	Series 6
Largest Drop Span	150
Maximum Modem Tx Power	65

Figure 54 – Upstream D3.0 MAC Layer Capacity Estimates over Dist. Cable .625 PIII at 1000′

with Sub-split.

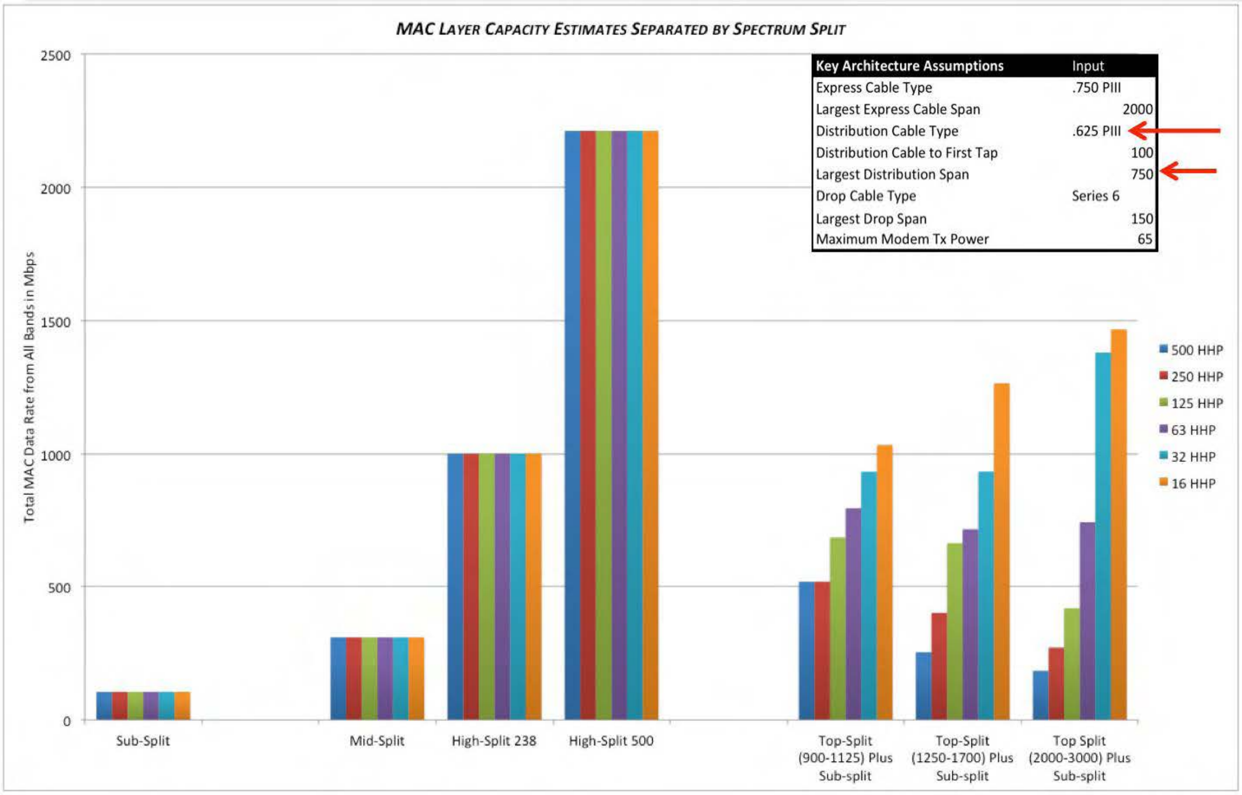
The various spectrum splits, along with the overhead contributed from the current DOCSIS PHY, the MAC, the use of SC-QAM and the highest possible modulation type, are examined in Figure 54 to determine the Total MAC Channel Bond Capacity Usable. Traffic engineering and capacity planning should consider the headroom needed for peak periods.

Similar to the examination of the downstream capacity projections above, the upstream use of a new error correction technology such as LDPC will allow high order modulations to be used, thus increasing capacity compared to Reed-Solomon based systems. Higher order modulations will also mean less spectrum required for a desired data rate.

The actual gain for the upstream across an HFC network will need to be determined in the real-world deployments. All upstream capacity is limited to 256-QAM, all though higher order modulation may be possible under certain conditions. Figure 54 through Figure 56 are meant to show the vast difference in capacity and network architecture with upstream spectrum just for having different distribution cable and span of this section of the network. It is this layer of the cable network that is vastly different among MSOs and even within MSOs.

Figure 55 represents cable rebuilds or new builds after the year 2005. Figure 56 represents the Mid 1990s – 2004 Rebuild. Again maximum 256-QAM limitations are assumed as well other assumptions defined in the paper.

A ***major*** finding is that Top-split



Key Architecture Assumptions	Input
Express Cable Type	.750 PIII
Largest Express Cable Span	2000
Distribution Cable Type	.625 PIII
Distribution Cable to First Tap	100
Largest Distribution Span	750
Drop Cable Type	Series 6
Largest Drop Span	150
Maximum Modem Tx Power	65

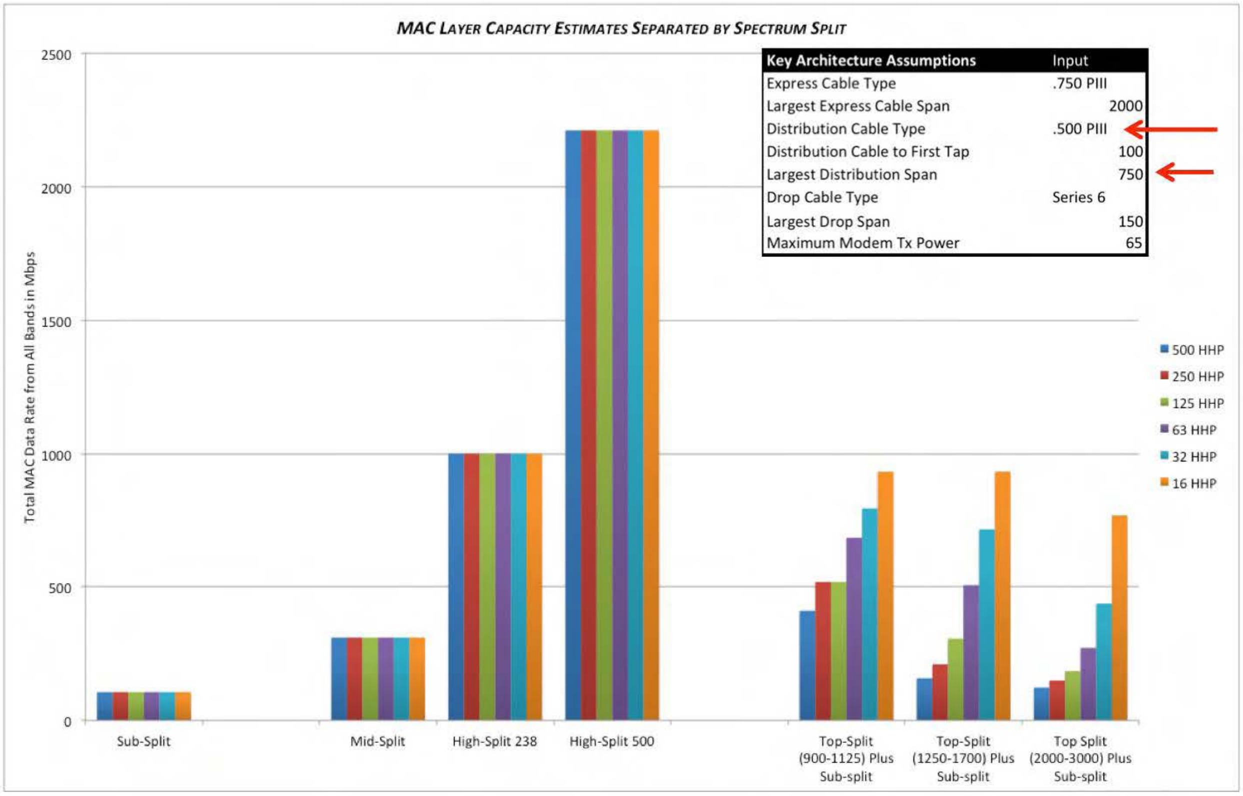
Figure 55 – Upstream D3.0 MAC Layer Capacity Estimates over Dist. Cable .625 PIII at 750′

options require Fiber to the Last Active (~16 HHP) and the placement of a node at each location to maximize the spectrum capacity. However, all Top-split options even if combined with the existing Sub-split will not reach the capacity any of the High-split option. If these two Top-split options are not combined with Sub and Mid-split achieving 1 Gbps MAC Layer performance is not possible, given the assumptions described in this analysis, .625 PIII at 1000 foot spans to last tap and other assumptions.

Another ***major*** finding is that even, given the assumption of the widely deployed cable architecture using .500 PIII distribution cable with 750 foot spans to the last tap, none of the Top-split with Sub-split reaches 1 Gbps with current DOCSIS PHY as shown in Figure 56. Only Top-split with .625 PIII at 750 foot spans to last tap will meet or exceed the 1 Gbps capacity.

Another very important point is that the network architecture and performance characteristics of the plant in the real world will determine the spectrum capacity to be used. The determination of the network architectures that may work at various spectrum splits, modulations, and number of carriers in different cable types and distance to the subscriber was a critical finding.

We have modeled the network architecture and performance assumptions to estimate the modulation and capacity possible for each spectrum split. This allowed us to determine the overall requirements and impacts to cost of the various split options and the ability for the spectrum split to meet the business needs of the MSO.



Key Architecture Assumptions	Input
Express Cable Type	.750 PIII
Largest Express Cable Span	2000
Distribution Cable Type	.500 PIII
Distribution Cable to First Tap	100
Largest Distribution Span	750
Drop Cable Type	Series 6
Largest Drop Span	150
Maximum Modem Tx Power	65

Figure 56 – Upstream D3.0 MAC Layer Capacity Estimates over Dist Cable .500 PIII at 750'

9.6.2 DOCSIS NG Network Capacity Estimates Upstream

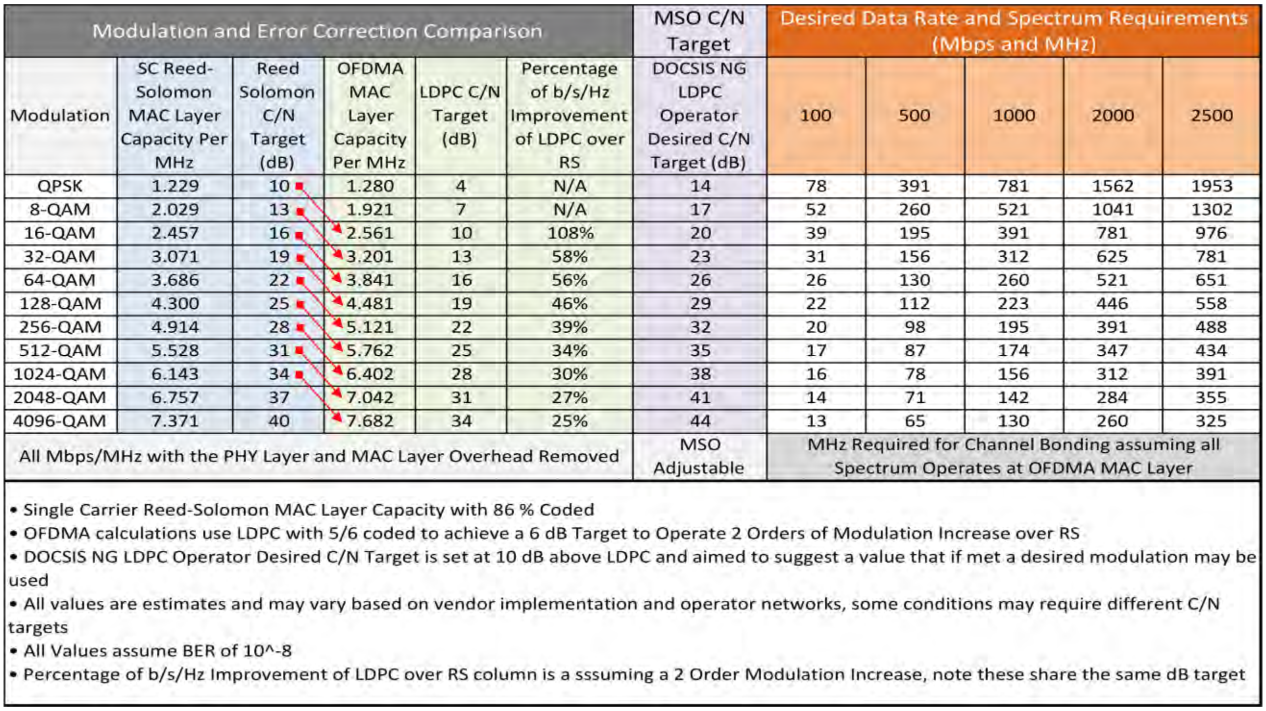
We have modeled the network architecture using several HFC coaxial network topologies using DOCSIS 3.0, however in this section DOCSIS NG will be compared. This section will provide a summary of the key methods and measurements to estimate sizing for DOCSIS NG.

The adoption of higher modulation formats in DOCSIS NG will increase b/s/Hz. A key finding is the use of DOCSIS 3.0 Single Carrier Reed Solomon verse OFDM using LDPC may allow two (2) orders of modulation increase. In Figure 57, DOCSIS 3.0 verse DOCSIS NG Modulation C/N and Capacity Estimates this summarize the major benefits of moving to DOCSIS NG.

Figure 57 illustrates that the use of Reed Solomon and LDPC with different code rates will have different b/s/Hz using the same modulation format. The major takeaway from the table is the use of a stronger error correction code will allow LDPC to operate in the same carrier to noise environment as Reed Solomon but LDPC may use two orders of modulation higher.

The table uses red arrows to illustrate the corresponding Reed Solomon modulation and C/N to the OFDMA LDPC modulation format, which shares the same C/N dB. The table will show that in the same modulation format Reed Solomon will have more b/s/Hz than LDPC and this is due to a higher code rate percentage applied to LDPC. The percentage of gain is measured using the SC Reed Solomon data rate for a given modulation and the used of two order of modulation increase using LDPC.

For example, in the table SC Reed Solomon b/s/Hz of QPSK is measured against OFDMA LDPC using 16-QAM, the percentage of gain in b/s/Hz 89%. As



Modulation and Error Correction Comparison						MSO C/N Target	Desired Data Rate and Spectrum Requirements (Mbps and MHz)				
Modulation	SC Reed-Solomon MAC Layer Capacity Per MHz	Reed Solomon C/N Target (dB)	OFDMA MAC Layer Capacity Per MHz	LDPC C/N Target (dB)	Percentage of b/s/Hz Improvement of LDPC over RS	DOCSIS NG LDPC Operator Desired C/N Target (dB)	100	500	1000	2000	2500
QPSK	1.229	10	1.280	4	N/A	14	78	391	781	1562	1953
8-QAM	2.029	13	1.921	7	N/A	17	52	260	521	1041	1302
16-QAM	2.457	16	2.561	10	108%	20	39	195	391	781	976
32-QAM	3.071	19	3.201	13	58%	23	31	156	312	625	781
64-QAM	3.686	22	3.841	16	56%	26	26	130	260	521	651
128-QAM	4.300	25	4.481	19	46%	29	22	112	223	446	558
256-QAM	4.914	28	5.121	22	39%	32	20	98	195	391	488
512-QAM	5.528	31	5.762	25	34%	35	17	87	174	347	434
1024-QAM	6.143	34	6.402	28	30%	38	16	78	156	312	391
2048-QAM	6.757	37	7.042	31	27%	41	14	71	142	284	355
4096-QAM	7.371	40	7.682	34	25%	44	13	65	130	260	325
All Mbps/MHz with the PHY Layer and MAC Layer Overhead Removed						MSO Adjustable	MHz Required for Channel Bonding assuming all Spectrum Operates at OFDMA MAC Layer				

- Single Carrier Reed-Solomon MAC Layer Capacity with 86 % Coded
- OFDMA calculations use LDPC with 5/6 coded to achieve a 6 dB Target to Operate 2 Orders of Modulation Increase over RS
- DOCSIS NG LDPC Operator Desired C/N Target is set at 10 dB above LDPC and aimed to suggest a value that if met a desired modulation may be used
- All values are estimates and may vary based on vendor implementation and operator networks, some conditions may require different C/N targets
- All Values assume BER of 10^-8
- Percentage of b/s/Hz Improvement of LDPC over RS column is a sssuming a 2 Order Modulation Increase, note these share the same dB target

Figure 57 – DOCSIS 3.0 versus DOCSIS NG Modulation C/N and Capacity Estimates

expected the percentage of gain will decrease as modulation increases, for example moving from 256-QAM to 1024-QAM is a smaller gain, than moving than the doubling of QPSK to 16-QAM.

The table estimates the use of OFDMA and the MAC layer bit rate in a given modulation as explained in the paper. The table calculated several desired MAC layer throughput capacities from 100 Mbps, 500 Mbps, 1,000 Mbps, 2,000 Mbps, and 2,500 Mbps and using the OFDMA estimated MAC layer data rate a required spectrum calculation and corresponding modulation format are aligned.

Table 36 – DOCSIS NG Modulation and C/N Performance Targets

Modulation Type	Uncoded Theoretical C/N dB	LDPC 5/6 Coded C/N dB	Operator Margin is Desired C/N Target
QPSK	16	4	14
8-QAM	19	7	17
16-QAM	22	10	20
32-QAM	25	13	23
64-QAM	28	16	26
128-QAM	31	19	29
256-QAM	34	22	32
512-QAM	37	25	35
1024-QAM	40	28	38
2048-QAM	43	31	41
4096-QAM	46	34	44

Theoretical SNRs Uncoded with BER of 10^-8
Practical C/N is chosen to give 10 dB headroom
Operator Margin above LDPC 5/6 coded

The MSO may require less upstream spectrum if a high modulation format may be used. The table illustrates a proposed Operator Desired C/N target for each Modulation format using LDPC, please note that the higher the modulation form the higher the C/N requirements but the lower percentage of gain in b/s/Hz.

In the past, our industry may have used The "Operating Margin" (OM) or Operator Desired carrier to noise target to be 6 dB above the theoretical uncoded C/N for a given BER, usually between 10E-6 or 10E-8, without any Forward Error Correction (FEC). The 6 dB of margin typically assumed a 500 HHP case; that is, for "Node +5" (or so), involving up to 30 return path RF amplifiers.

In the future perhaps we need to change the method by which we estimates the "Operating Margin" (OM) and perhaps we need to estimate the operating margin from the coded rate used for a given system and then add the Operating Margin, for the analysis below we used 10 dB above the LDPC dB value.

About the "Operating Margin" (OM) parameter, this is a variable (in dB) to account for the performance changes in the HFC return path system due to temperature variation and setup accuracy of the outside plant. This mainly involves RF level changes due to hardline and drop cable loss changes, Tap loss change, and RF Amplifier/Node Return RF drive path (Hybrid) gain changes, and Node passive loss changes with temperature. It also includes setup level tolerances (due to RF Testpoint accuracy and flatness over frequency) and laser optical power output changes over temperature.

Some of these changes are small or only occur in one place, while others are more significant as they occur at many places and in cascade (e.g., cable segments, RF Amplifiers, and Taps). With many amplifiers in a 500 HHP distribution sector (up to 30 for Node +5 sector), the number of cascaded Amplifiers is typically a maximum of 6. There typically will be 6 or more Taps used between each amplifier, so these elements contribute significantly.

About 2/3 of the 6 dB OM assumed in the calculation matrix is due to the cable part of the plant. The other 2 dB is due to the "optics" part; mainly for the Return laser. The laser is assumed a high quality uncooled CWDM analog laser, with 2 mW or higher optical output. The OM is added to the "Theoretical C/N" at 10E-6 BER (without encoding) to obtain a "Desired C/N" for determining the highest order modulation type allowed.

In the model that will estimate the use of DOCSIS NG and LDPC, we will use a 10 dB Operating Margin, on top of the coded value, please see Table 36 for the allocation.

In order to estimate the capacity of the different spectrum splits using DOCSIS NG we placed the values of the Operator Margin desired C/N target and the b/s/Hz estimates for DOCSIS NG. The model estimates the system C/N and in this case the model used .500 PIII distribution cable at 750 feet.

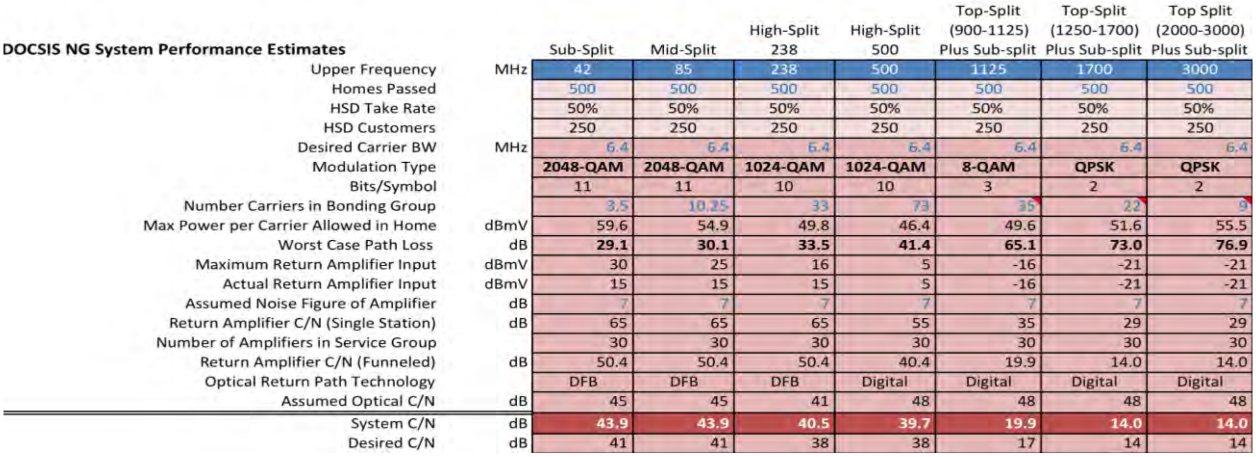
Please note the that model estimates that very high modulation format may be used in a 500 HHP node for the low frequency return while the Top-split spectrum selection is only capable of using substantially lower order modulation formats.

As seen in Table 37, 2048 QAM and 1024 QAM are possible in the upstream in a 500 HHP node with assumption defined in this table. This is an illustration of the modern DOCSIS PHY and the ability to maximize spectrum for the operator.

DOCSIS NG capacity is examined in Figure 58 considering several spectrum-split options. Please note the capacity of Sub-split, Mid-split, and the pair of High-split options. The MSOs may choose any of this spectrum split or others depending on the desired capacity. The estimates assume that the entire spectrum uses the highest modulation rate possible for a given spectrum selection.

9.6.3 DOCSIS 3.0 versus DOCSIS NG Side-by-Side Upstream Capacity Estimate

Table 37 – Upstream DOCSIS NG MAC Layer Capacity Estimates over Distribution Cable .500 PIII at 750 Feet



DOCSIS NG System Performance Estimates		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Homes Passed		500	500	500	500	500	500	500
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	250	250	250	250
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4
Modulation Type		**2048-QAM**	**2048-QAM**	**1024-QAM**	**1024-QAM**	**8-QAM**	**QPSK**	**QPSK**
Bits/Symbol		11	11	10	10	3	2	2
Number Carriers in Bonding Group		3.5	10.25	33	73	35	22	9
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	49.8	46.4	49.6	51.6	55.5
Worst Case Path Loss	dB	**29.1**	**30.1**	**33.5**	**41.4**	**65.1**	**73.0**	**76.9**
Maximum Return Amplifier Input	dBmV	30	25	16	5	-16	-21	-21
Actual Return Amplifier Input	dBmV	15	15	15	5	-16	-21	-21
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7
Return Amplifier C/N (Single Station)	dB	65	65	65	55	35	29	29
Number of Amplifiers in Service Group		30	30	30	30	30	30	30
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	40.4	19.9	14.0	14.0
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	48	48
System C/N	dB	**43.9**	**43.9**	**40.5**	**39.7**	**19.9**	**14.0**	**14.0**
Desired C/N	dB	41	41	38	38	17	14	14

The paper has examines the downstream and upstream features of DOCSIS NG. The analysis has examined modulation profiles such as using LDPC with increased FEC to obtain a 6 dB gain over Reed Solomon in the same modulation format. Figure 59 examines the low frequency return spectrum options using DOCSIS 3.0 using 64 QAM against DOCSIS NG using the maximum modulation format possible given the assumptions and spectrum selection. Please note the much higher aggregate capacity of the DOCSIS NG system over current DOCSIS.

9.6.4 Summaries for Network Capacity

DOCSIS NG will greatly expand the capacity of the cable network and coupled with backward compatibility utilize spectrum efficiently

Downstream Capacity Expansion

1. DTA's & SDV will provide long term downstream plant capacity expansion
2. Reduced service group size enabling fewer customers to share bandwidth
3. Node segmentation and node splits will continue to be used in a targeted basis
4. Use of highest order modulation and channel bonding to increase throughput
5. Consider DOCSIS NG changes with modern error correction technology that allow the modulation rate to increased, given the same SNR, perhaps as much as two orders. For example, 256-QAM could be increased to 1024-QAM
6. Possible downstream bandwidth expansion along with upstream augmentation

Upstream Capacity Expansion

1. Use of highest order modulation and Channel Bonding to increase throughput

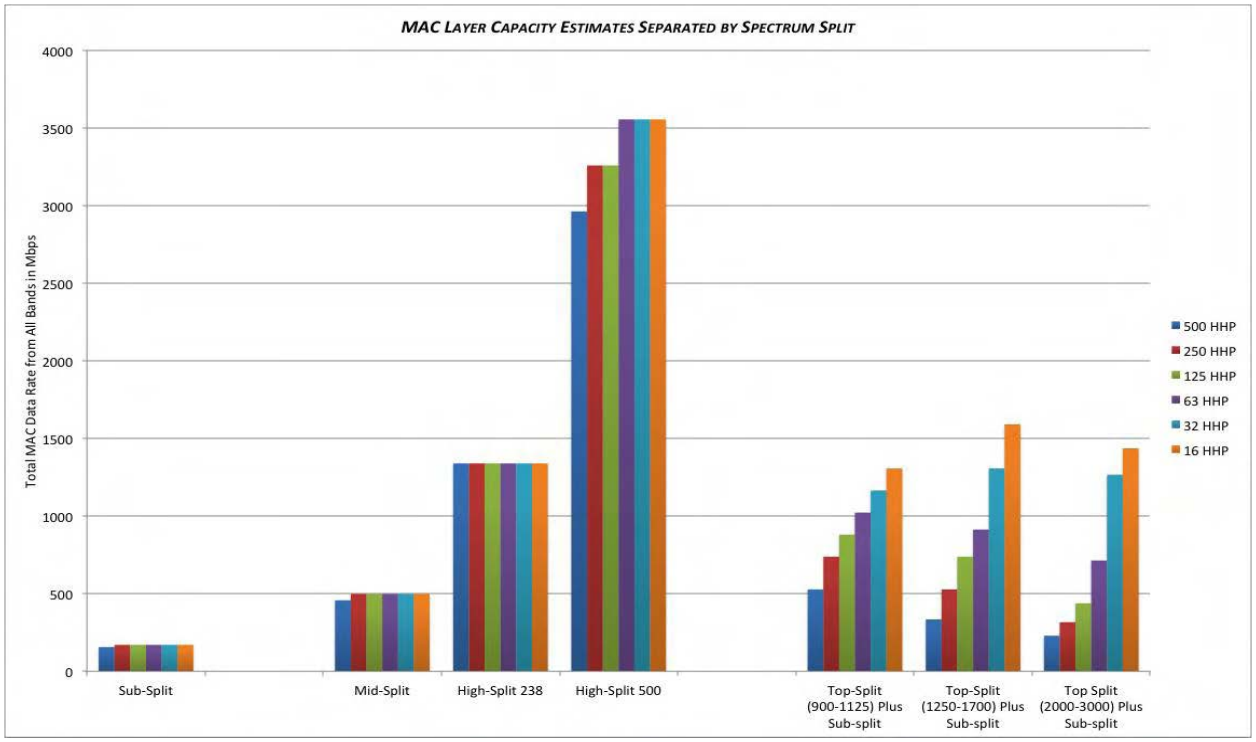


Figure 58 – Upstream DOCSIS NG MAC Layer Capacity Estimates over Dist Cable .500 PIII at 750'

2. Consider DOCSIS NG changes with modern error correction technology that allow the modulation rate to be increased, given the same SNR, perhaps as much as two orders. For example, 64-QAM to 256-QAM and perhaps 256-QAM to 1024-QAM
3. Progressively smaller upstream service groups
4. Ongoing node splits / segmentation
5. These incremental steps should last for a majority of the decade

Upstream augmentation expands upstream spectrum and bandwidth such as conversion to mid-split, high-split, or top-split options.

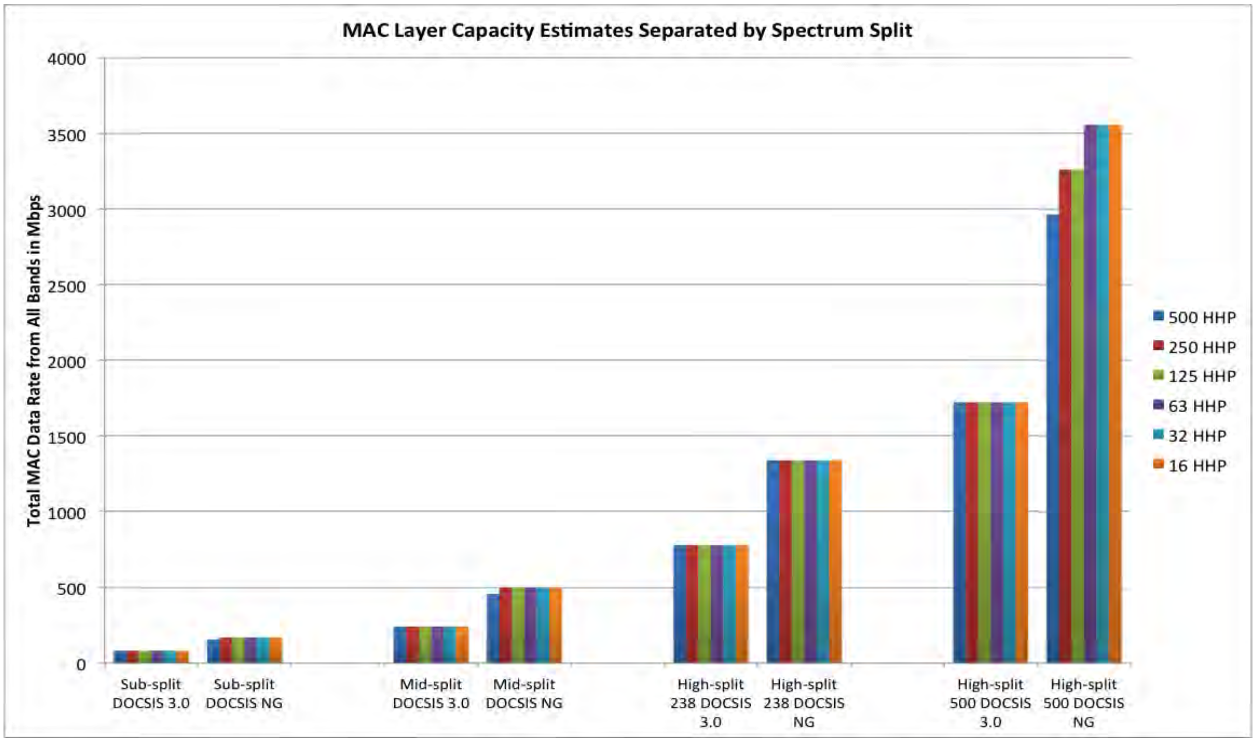


Figure 59 – DOCSIS 3.0 verse DOCSIS NG

10 NETWORK CAPACITY PROJECTION AND MIGRATION STRATEGIES

10.1 Upstream Migration Strategy

10.1.1 Phase 0: Sub-Split and Business as Usual

10.1.1.1 Sub-split Legacy Return Lifespan

Let's put our understanding of upstream data capacities to work in evaluating time-based migration strategies for the HFC upstream. Note that not every capacity number calculated in the paper to this point is represented on a chart in this section. We expect that the reader may have to extrapolate between displayed values in some case to draw conclusions from curves shown for some cases not explicitly plotted.

We introduced a version of an upstream lifespan analysis in Figure 2 of Section 2.6. A more traditional version is shown in Figure 60. Traffic models based on a compound annual growth (CAGR) methodology have been shown to represent historical traffic trends well. However, because of short-term fluctuations, particularly in the upstream, there is a need to engineer ahead of the curve to avoid being unprepared in the case of an unexpected step function in growth (a "Napster" moment).

We will use CAGR analysis such as this and Figure 2 as a guideline to understand the most fundamental of drivers for upstream evolution – the need to find more capacity,

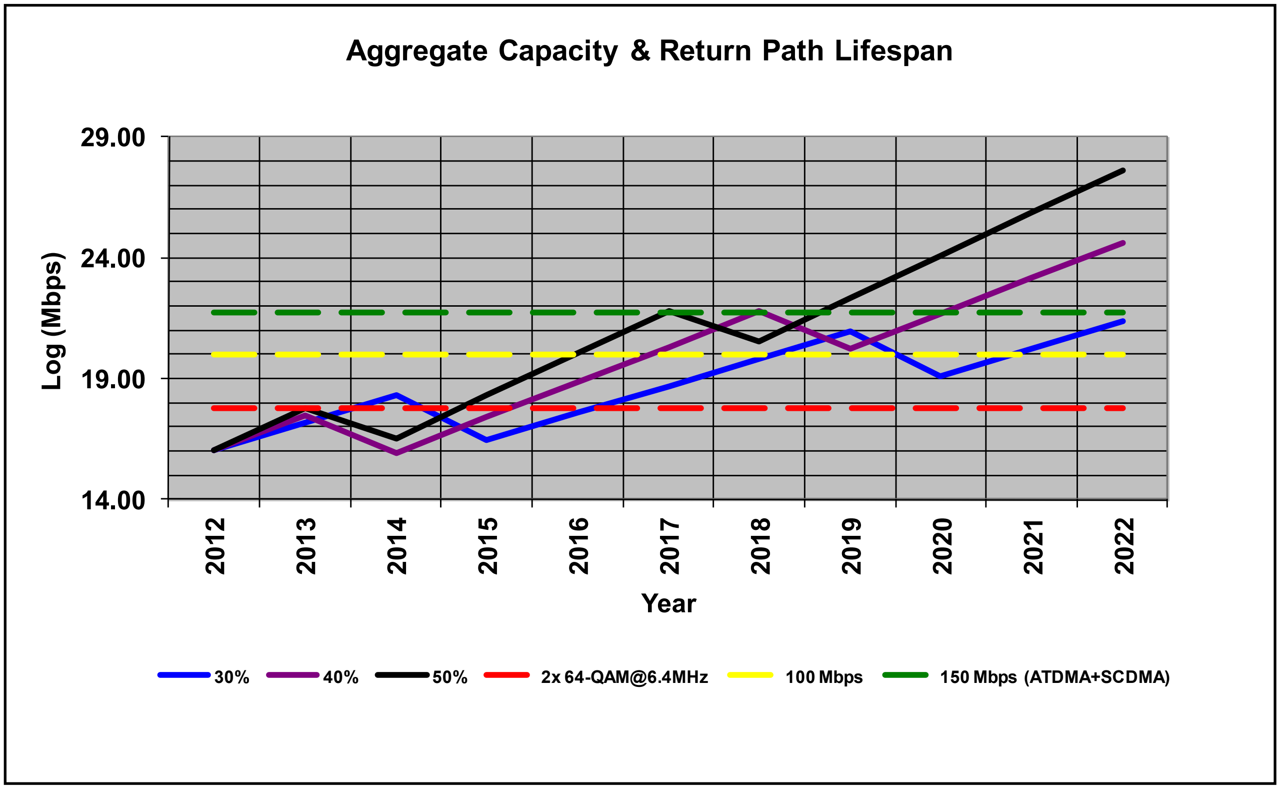


Figure 60 – Upstream CAGR vs. Available Capacity

coupled with a need to deliver competitive service rates, so that the upstream achieves a long and healthy lifespan.

Figure 60 shows this a CAGR approach for the upstream using three different assumptions – 30%, 40% and 50%. The three trajectories, representing a single aggregate service group, are interrupted by two breakpoints over the next ten years.

These represent node and/or service group splits – 3 dB (best case) offsets, or a doubling of average bandwidth per home. Note that the 3 dB is a step straight downward by 3 dB at implementation, so that by the time the next year comes around, some of that has been consumed.

These trajectories are plotted against three different HFC upstream capacity thresholds, using raw physical layer transport rate for simplicity and to remove the ambiguity around overhead of different configurations, packet sizes, and net throughputs. We will use raw transport rate for trajectories and thresholds throughout to simplify apples-to-apples comparisons.

- 60 Mbps – Approximately two 64-QAM DOCSIS channels at 5.2 Msps
- 100 Mbps – Approximate available bit rate in 5-42 MHz with only A-TDMA
- 150 Mbps – Approximately a fully utilized 5-42 MHz using both A-TDMA and S-CDMA

Using these, we can now estimate when various CAGRs exhaust the available upstream. Let’s assume 40 Mbps of upstream consumption at peak busy hour – 50% of 80 Mbps of deployed capacity, for example (2x 64-QAM + 16-QAM, all at 6.4 MHz).

Some key conclusions can be drawn from Figure 60. Clearly, a couple of 64-QAM DOCSIS channels get exhausted within a few years without a service group split. While node splits are costly and intrusive, they are well-understood business-as-usual (BAU) activities.

Most important to craft an evolution strategy is to estimate when 5-42 MHz itself gets exhausted, and when a more significant change must be considered. Referring again to Figure 60, note that a single split supports 4-6 years of growth considering 100 Mbps as the 5-42 MHz throughput boundary.

While further node splitting will provide more average bandwidth, the maximum service rate limit also come into play, where 100 Mbps upstream service rates require more total capacity to be achieved. Aside from merely keeping pace with upstream service rate growth, the service rate upstream should be somewhat aligned with 1 Gbps downstream rates from a timing perspective.

Finally, note that with S-CDMA the upstream could last through the decade for a very robust CAGR (40%).

Figure 60 is a useful guide for visualizing growth versus time. In Figure 61, as in Figure 2, we have displayed the same information differently, allowing us to understand the sensitivity of the exhaustion of the 5-42 MHz return path relative to the CAGR assumptions. Note that service group splits are instead represented by dashed traces for the 100 Mbps and 150 Mbps cases.

The three crosshairs on Figure 61 are positioned to help interpret between Figure 60 and Figure 61. For example, note the point at which a 50% CAGR exhausts a 150 Mbps maximum throughput threshold after one split in Figure 60. This occurs 5 years into the future. We can see this same point represented by the leftmost crosshair in Figure 61. Similarly, we can correlate between the crosshairs at 40% and 30% CAGR on Figure 61 and the corresponding breach of threshold in Figure 60.

We will use the format of Figure 61 in subsequent discussion because of the granularity and clarity it brings in an environment where CAGR tends to have more variation. This variation of CAGR points out why, for network planning decisions, upstream CAGR needs to be considered in the context of an average, long-term CAGR, rather than based on very high or very low periods of growth.

This is particularly true upstream, where there is not a set of knobs and levers at the operator's disposal to manage a spectrum congestion issue as there is in the downstream. In the downstream, while CAGR is consistent and generally higher, but there is more control over service delivery choices to manage spectrum. In the upstream, there is a hard bandwidth cap at 42 MHz in North America, for example, little control over the growth of Internet usage, and limited ability or authority to more actively manage traffic by type. As such, there are not any "easy" answers to creating more upstream capacity in the 42 MHz spectrum.

One area where there is some room to grow is in the low end of the return. A key problem for A-TDMA is its ability to operate efficiently or at all in this region. Some 30-40% of the 5 to 42 MHz return band is polluted by a combination of impulse noise emanating from homes and often times various narrowband interferes managing to get onto the cable in the short wave band.

However, it is the impulse noise that

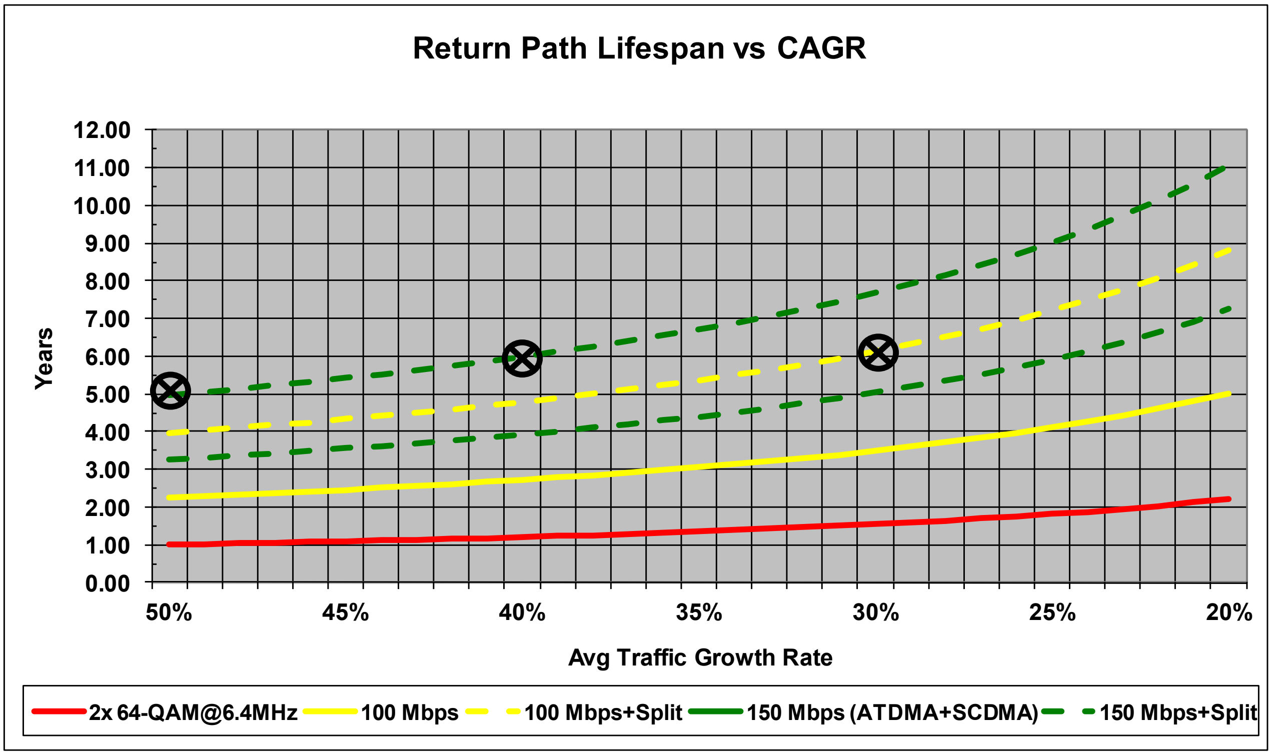


Figure 61 – Lifespan of 5-42 MHz vs CAGR

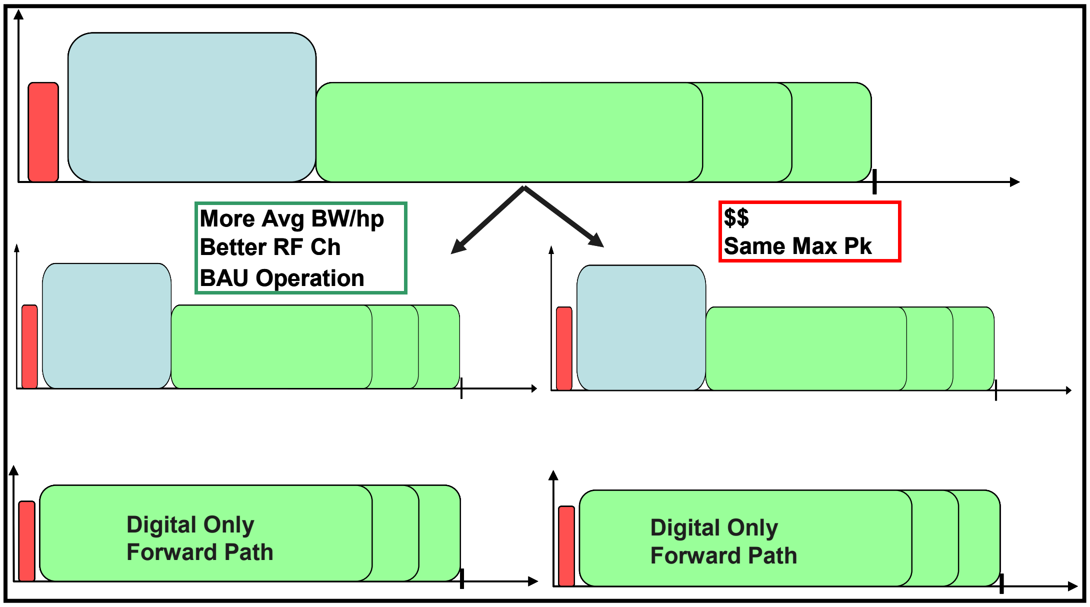


Figure 62 – Serving Group Segmentation

gives A-TDMA the most difficulty, even with powerful Reed-Solomon burst correction employed. To combat this, DOCSIS 2.0 introduced S-CDMA to the standard. By enabling use of the lower portion of the upstream spectrum, the total 5-42 MHz band improves in its total capacity by almost 50%, to about 150 Mbps. We will discussed S-CDMA in Section 7.2, and will use some of the results observed to add to the available capacity in 5-42 MHz to calculate the lifespan of a fully optimized 5-42 MHz.

10.1.1.2 Legacy Relief: Business-As-Usual Node Splitting

The classically deployed tool for improving average bandwidth per user is service group or node splitting. However, this does not enable service rate increases, and splitting nodes in the field runs into diminishing return because of the unbalanced nature of physical architectures.

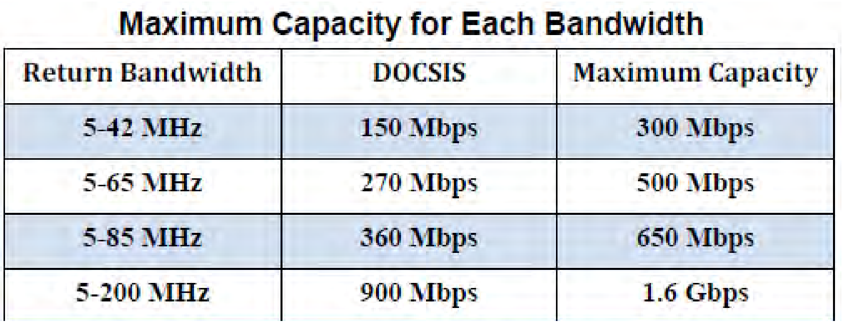
We observed in Figure 60 and Figure 61 how this lead to a longer lifespan for 5-42 MHz by simply sharing the fixed bandwidth among fewer users. The average bandwidth per user, often a good reflection of user QoE, will increase.

The most natural HFC methods to decreasing the service group size are the removal of combiners at the output of the return optical receivers that combine upstreams into a single port, or the splitting of nodes, either through a segmentable node or pulling fiber deeper.

Figure 62 illustrates this approach from a spectral allocation perspective, identifying also the pros and cons commonly associated with this well-understood tool.

The increased BW/user is an obvious benefit. Another key benefit of this straightforward approach is that, while heavy touch, it is a well-understood "business as usual" operation. In addition, reducing the serving group size can improve the RF channel in two ways.

Table 38 – Bandwidth, DOCSIS, and Theory @25 dB SNR



Maximum Capacity for Each Bandwidth

Return Bandwidth	DOCSIS	Maximum Capacity
5-42 MHz	150 Mbps	300 Mbps
5-65 MHz	270 Mbps	500 Mbps
5-85 MHz	360 Mbps	650 Mbps
5-200 MHz	900 Mbps	1.6 Gbps

First, fewer users means a lesser probability of interference and impulse from a troublesome subscriber. While the troublemaker has not gone away, he is now only inflicting his pain on half the number of users. Second, from a system engineering standpoint, the same funneling reduction that increases the probability of not having a troublemaker also reduces any amplifier noise aggregation effect, noticeable when deep RF cascades combine in multiport nodes, for example. All of this can lead to more efficient use of the existing spectrum than had existed prior to the split.

The primary performance disadvantage of only a segmentation strategy is that 5 to 42 MHz ultimately limits the maximum total bandwidth to around 100 Mbps. Under good conditions, a single 100 Mbps serving group may be all that can be obtained in an A-TDMA only system.

This limits the flexibility of this architecture to provide other services, such as mid-size business service tiers, and to support Nielsen's Law-based peak rate growth. And, peak rate offerings generally are topped out at some scale factor of the total available capacity for practical reasons.

Note that in Figure 62 we have added the "digital only" forward example. As we consume forward band for return applications, techniques that make more efficient use of the forward path also draw more focus. Digital only carriage (DTA deployments) is one of the key tools for extracting more from the downstream as upstream imposes on it, and for adding flexibility to the diplex split used in the architecture.

10.1.1.3 Delivering New DOCSIS Capacity

Because of the known limitations of return spectrum, the expectation that traffic growth in the upstream will continue to compound, and the anticipation that peak service rates will do the same, options to find new capacity are required.

There is consensus that new spectrum must eventually be mined for upstream use. The questions that remain are where do we find it and how much do we need. And, of course, at the core of the discussion, how much new capacity, for how long, and what are the practical implications of implementing such a change.

We will focus on the recommended evolution approach whereby cable maintains a diplex-only architecture for optimum bandwidth efficiency. We view a migration that has as a primary objective the most efficient long-term use of the cable spectrum to ensure the longest lifespan of the architecture, and preferably with the simplicity of implementation that cable enjoys today.

A diplex architecture achieves this. We view the selection of the actual frequency split as something that evolves with time, in an efficient way, and based on the traffic mix and projected services.

We note that it is possible that extracting the most bandwidth efficiency with flexibility theoretically involves a TDD implementation. However, the obstacles in place to enable TDD in the HFC environment are so great and will be so for so long, that it does not appear to be a sensible plan for typical HFC architectures.

However, with the very long observation window enabled by fiber deep migration and the recommendations made herein, it may at some point become a more practical consideration for cable if the need for increased flexibility of traffic allocation justifies the increase in complexity.

Table 38 illustrates the available DOCSIS transport rate for various low diplex-based frequency split architectures, and the theoretically available channel capacity at the DOCSIS-specified minimum of 25 dB.

While it is impractical to achieve theoretical capacity, the gap has indeed closed over time between practice and theory. This not a negative reflection on DOCSIS 1.0, only a reflection that its PHY basis is 15 years old – a very long time in technology evolution, and a period of extensive advances in communications theory and practice. For DOCSIS NG, we have already introduced the fact that a new FEC added to the PHY mix will enable a major step closer to capacity by enabling higher order profiles over the same SNR.

One simple conclusion of Table 38 is simply the power of the Shannon-defined proportional relationship between capacity and bandwidth for a fixed SNR. Indeed, for high SNR assumptions, capacity is directly proportional to both bandwidth available and SNR expressed in dB – the assumption being very relevant to the cable architecture. This leads to the inescapable conclusion that when discussing new actual upstream capacity, it is first about architecture and bandwidth, and not waveform.

As previously introduced, a straightforward and surprisingly powerful way to exploit new bandwidth and remain compatible with DOCSIS is use of the 85 MHz Mid-Split.

This band edge was wisely chosen to maximize clean low band return without overlapping the FM radio band and potential harmful effects of proximity to that band. Its advantages are numerous. First, however, let's understand what new spectrum means in terms of that fundamental upstream problem – lifespan – that has us so concerned in the first place.

10.1.2 Phase 1: Deploy 85 MHz Mid-Split

10.1.2.1 Capacity and Lifespan

It was shown in Figure 2 how the 85 MHz Mid-Split delivers long-term new capacity to the HFC upstream. Consider **Error! Reference source not found.**, which adds the Mid-Split case to cases observed in Figure 61 for 42 MHz. The gap between the set of 5-42 MHz options and the maximized Mid-Split is readily apparent at 3.5-5.5 years at 30% CAGR, depending on whether S-CDMA is utilized or not.

The transition to Mid-Split pushes the lifespan of the return path to nearly a decade under a 256-QAM maximum assumption – a very comfortable chunk of next generation network planning time. This lifespan time frame is pushed beyond a decade for CAGRs of 35% and below if the Mid-Split is combined with one service group split, as shown in **Error! Reference source not found.**.

Though not apparent in an upstream analysis, it is straightforward to show that a ten-year lifecycle of growth aligns the upstream with what is also achievable in the downstream under similar assumptions about plant segmentation. Aligning these two in terms of physical plant segmentation has operational benefits.

Because of this result observed in **Error! Reference source not found.**, when combined with a service group split, Mid-Split (440 Mbps), in fact, represents a *long-term* solution, not merely an incremental one.

This is a very important, fundamental conclusion to recognize about the 85 MHz Mid-Split architecture, that is often not fully understood. The amount of lifespan afforded by 85 MHz with just a single split is nearly a decade – a technology eternity. If today's observed, low, CAGRs persist, it is even

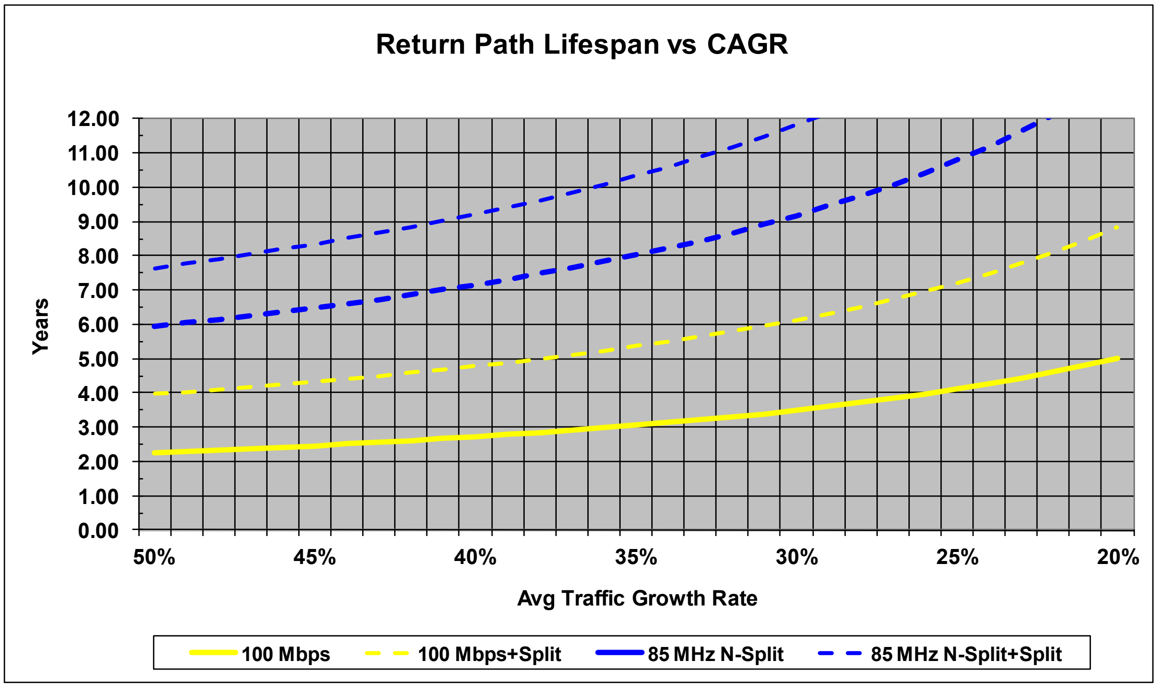


Figure 63 – 85 MHz Mid-Split vs. 42 MHz A-TDMA-Only with Segmentation

longer, and longer still if we assume that modulation profiles extend beyond the 256-QAM examples used for the Mid-Split analysis here. For example, 25% is a three year doubling period, so it offers 50% more lifespan than 40%. Similarly, 1024-QAM, which may become available with LDPC FEC, offers 25% more data capacity, pushing 400 Mbps of 85 MHz throughput to 500 Mbps available for growth.

The window of time to observe trends in traffic, applications, services, and technology, coupled with the runway for managing down legacy in an all-IP transition, is a very meaningful strategy component considering the low risk associated with implementation.

Even under an acceleration of CAGR, the architecture supports 100 Mbps services and an attractive long-term lifespan. A common traffic engineering assumption is to evaluate an increased CAGR resulting from the exploding number of devices looking for access to the upstream, using similar models for average application bandwidth of the access. The net effect for equivalent QoE is the potential requirement to adjust the oversubscription model.

In **Error! Reference source not found.**, we adjust this traffic engineering parameter by a factor of two to account for the increasing number of simultaneous users (devices) looking to access the upstream. Despite this acceleration, the Mid-Split architecture still achieves a decade of lifespan under two segmentations for common CAGR ranges.

Considering that a downstream CAGR analysis typically requires two splits over this same time period, there is the added opportunity to take advantage of this added lifespan to the upstream as well if necessary.

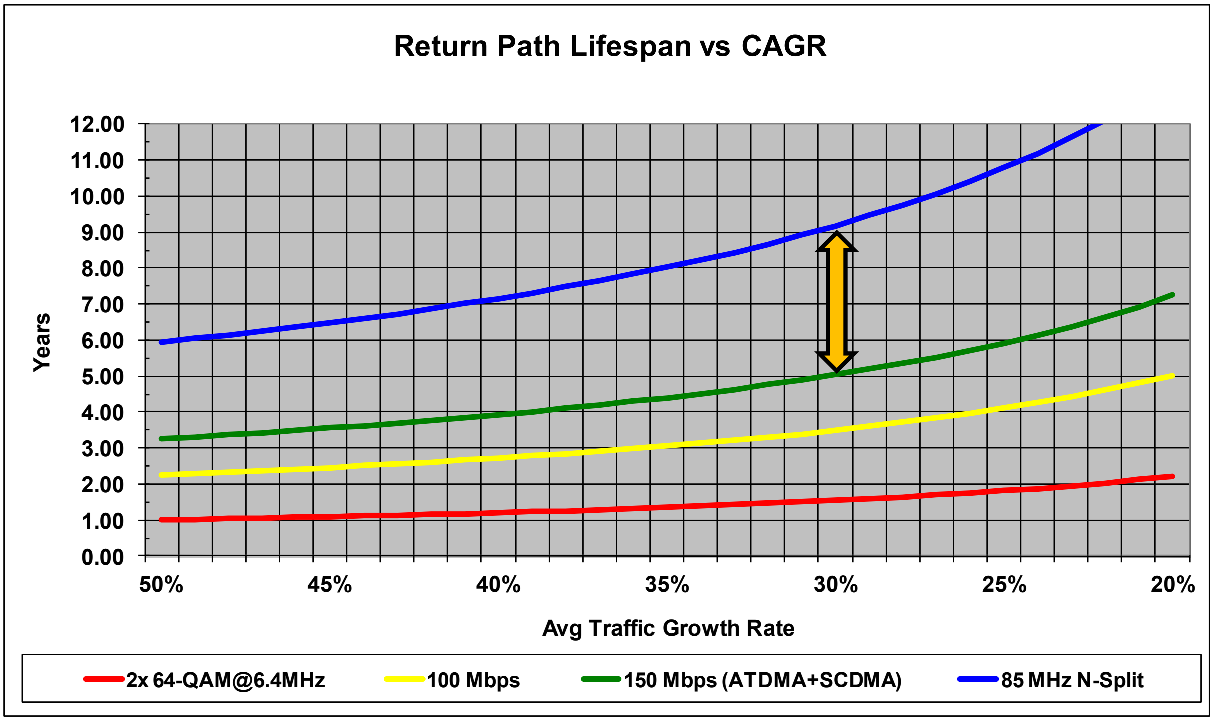


Figure 64 – 85 MHz Mid-Split Years of Growth vs. 5-42 MHz Use

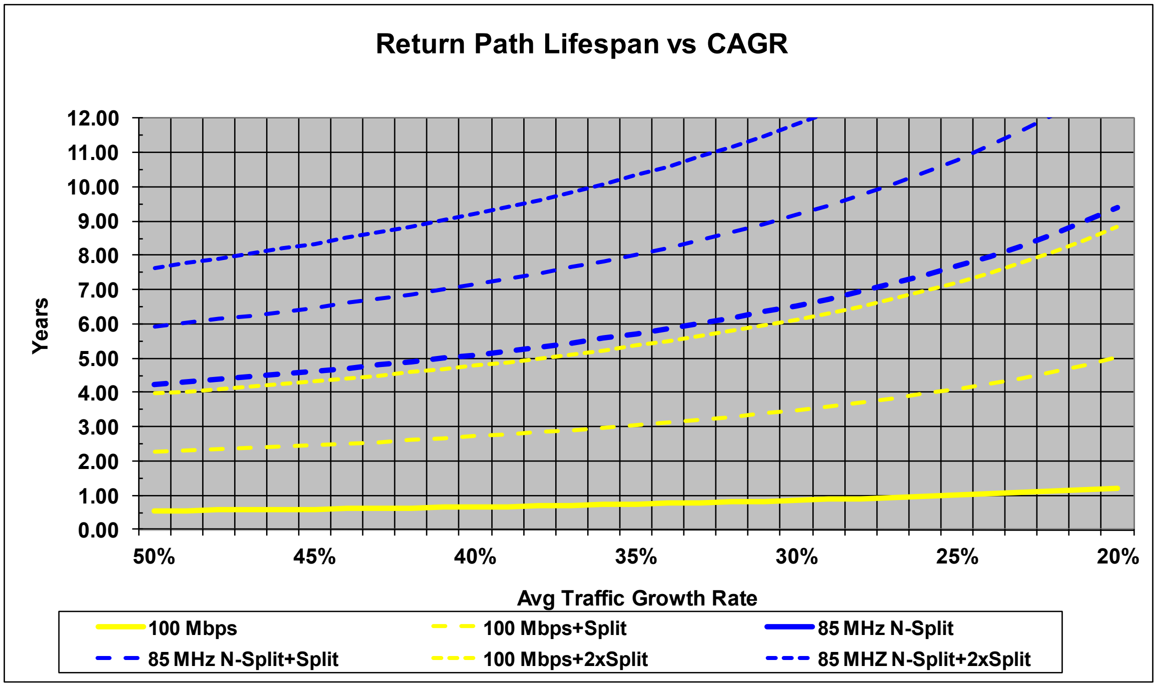


Figure 65 – Upstream Lifespan for Accelerated Usage Patterns

10.1.2.2 Architecture

We observed the clear relationship between available bandwidth and upstream capacity in Table 38. Unfortunately, there simply are no "easy" answers to adding new, real upstream capacity (as opposed to virtual, node splitting).

However, the 85 MHz Mid-Split looks to be the most compelling option in the near term in terms of implementation ease, availability, risk, compatibility, lifespan, and the strength of the value proposition, additional components of which are described in Section 2.1. We have seen in **Error! Reference source not found.** and **Error! Reference source not found.** and **Error! Reference source not found.**, that it also has perhaps unexpectedly powerful benefits.

The 85 MHz spectrum approach is diagrammed in Figure 66. Also shown is the combined case of the Mid-Split and a node split – clearly these are complementary tools.

This architecture has many very valuable and compelling advantages including the most important one of enabling a long upstream lifespan, while supporting key service expectations around data rate.

We summarize the 85 MHz Mid-Split benefits below:

- More than doubles the spectrum available, and more triple the available capacity compared to the use of 5-42 MHz today

- A decade of life OR MORE of upstream growth under aggressive assumptions for traffic growth using only an assumption of 256-QAM

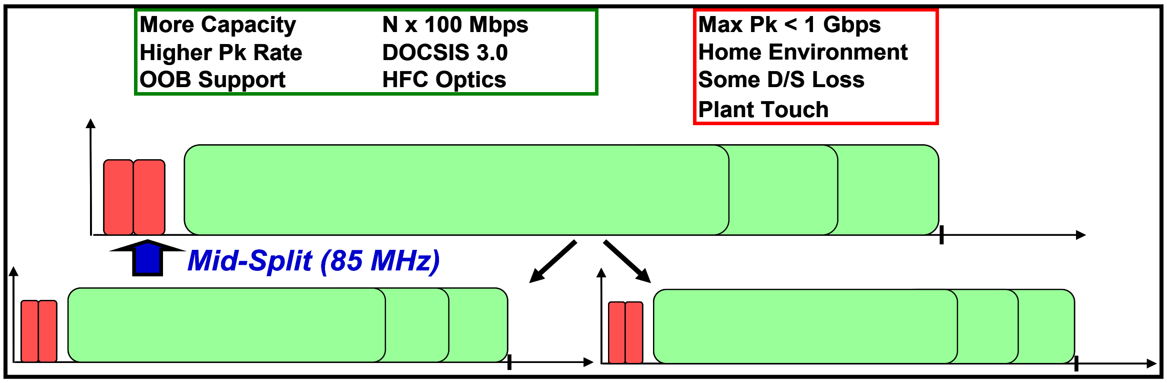


Figure 66 – Step 1: New Return Above the Old Return

- Accommodates multiple 100 Mbps peak rates. Accommodates higher peak rates if desired such as 150 Mbps or 200 Mbps. These may be important to run an effective 1 Gbps DOCSIS downstream service.

- Compatibility with DOCSIS 3.0. Current specification call out support of this extended spectrum. Equipment exists and has been proven for this band.

- Compatibility with standard downstream OOB carriers (70-130 MHz). Thus, no STB CPE using standard OOB is stranded (or at least the vast, vast majority, will not). Over time, as this older population of CPE is removed as part of an all-IP transition, even more flexibility for how to manage return spectrum become available.

- Can be implemented over standard HFC RF and linear optical returns, as well as digital returns. Products exist today for both.

- The new spectrum from 42-85 MHz tends to be cleaner, with less interference and impulse noise, and overall well behaved. This follows the characteristic of the current return that gets cleaner towards the higher end of the band.

- The Mid-Split architecture remains in the low-loss end of the HFC band. Combined with clean spectrum, the DOCSIS 3.0 implementation should have little if any differences, and any updated PHY approaches have the opportunity for even more bandwidth efficient modulation profiles.

- Entails minimal encroachment into the downstream bandwidth as a matter of capacity, and is even less significant when considered in the context of reclaiming the analog spectrum. In this case, it is basically the loss of one 6 MHz slot from a program count perspective – nine lost slots to cover the guard band

- Has similar cable loss versus frequency properties as legacy band – important for understanding CPE implications

- Very low risk, Proven in the field on a fully loaded upstream carrying 64-QAM and 256-QAM. Field trials using standard DFB lasers over typical link length and optical receivers have proven performance.

Note that the proven performance and link characterization for the Mid-Split architecture was discussed in detail in Section 7.1.2, where 256-QAM deployments for upstream were described

A few drawbacks are often cited for the Mid-Split, typically around cost and deployment obstacles. The primary concern is the need to touch actives throughout the plant. It is thus an imperative an upgrade activity be coupled with a segmentation operation and preferably with the ability to enable a Phase 2 of the evolution without requiring the same heavy touch.

Many potential solutions are available to ensure that an elegant transition from 85 MHz to a wider bandwidth in the future can be achieved. Unfortunately, as was originally stated for the upstream, there is no simple solution to more return spectrum.

Recognizing the intrusiveness of the work at hand to modify the frequency split, is commonly observed that the level of touch to the plant means that the "big" step to the 200+ MHz approach should be made.

However, in consulting with operators and suppliers, it is clear that the legacy CPE still requiring the downstream OOB channel for communications must be accommodated. The dynamics associated with this obstacle were detailed in Section 3.3.5. Also, the ability to absorb that amount of loss in the downstream is not tolerable at this phase of the IP migration, which currently might best be described as the "IP Simulcast Bubble" phase of evolution. Therefore, we recommend a phased approach.

Two key items must be recognized in implementing the change. First, it is intrusive, but it is also very low tech, very low risk, available and standardized today. Indeed, it has been proven in existing equipment. Second there is a perception that "just" going to 85 MHz with the effort involved is not enough.

In fact, as shown in the analysis of 85 MHz Mid-Split capacity and lifespan, this is not a band-aid, incremental upgrade, but one that delivers a powerful value proposition in the long term runway it enables, all the while maintaining the fundamental diplex architecture and simplicity of using the low-loss end of the spectrum for the return path.

The deployment challenge often arises out of concern for the home environment when an 85 MHz CM is installed. We described these dynamics in Section 3 and discussed strategies to deal with the challenge. For example, an installation may need to include a blocking filter for some STB CPE. Obviously, the risk here drops considerably if analog channels are removed, or if a Home Gateway architecture is adopted as part of an IP video transition. This is important to characterize and develop a sound operational model for, but is certainly not a technology challenge.

And, in Sections 2.6 we outlined the argument around the limitation often stated that that 85 MHz cannot achieve 1 Gbps of upstream. As was observed in Figure 2, with the time window made available by an Mid-Split upgrade, an extension of the Mid-Split is poised to deliver this capability when necessary and after legacy obstacles have had an opportunity to be addressed. The capacity requirements for residential 1 Gbps of capacity or service rate project well into the next decade on a CAGR basis.

10.1.2.3 Summary – Mid-split Migration Strategy

We recommend an 85 MHz Mid-Split upgrade for a near-term phase of spectrum expansion. Given the lifespan it will be shown to support over CAGRs much more aggressive than are observed today, the 85 MHz Mid-Split should be viewed as a long-term solution and not a temporary fix.

Key benefits are summarized as follows:

1. More than doubles the spectrum and triple the available capacity, providing a path to a decade of life OR MORE of upstream growth
2. Accommodates multiple 100 Mbps peak rates and higher.
3. Compatible with DOCSIS 3.0
4. Compatible with standard downstream OOB carriers (70-130 MHz)
5. Can be implemented over HFC RF and linear optical returns, as well as digital returns.
6. Cleaner spectrum from 42-85 MHz tends to be cleaner
7. Maintains use of the low-loss end of the HFC band. Any updated PHY approaches have the opportunity more bandwidth efficient modulation profiles, and CPE Tx power remains manageable.
8. Entails minimal encroachment into the downstream bandwidth as a matter of capacity
9. Very low risk, proven in the field on a fully loaded upstream carrying 64-QAM and 256-QAM using standard DFB lasers.

While we refer to Mid-Split as "Phase 1", it is a possibility that such a step becomes essentially a "forever" step from a business planning standpoint, on the way to some other long-term approach as greater than ten years of HFC migration is traversed.

Nonetheless, given the projected objectives for the upstream as we see them today, ensuring a path to 1 Gbps in the upstream within the context of HFC tools and technologies is a good long-term objective and a necessary part of long term planning.

Thus, a smooth transition plan beyond Mid-Split requires thinking through the aspects of the Phase 1 implementation that clears the way for this point in the distant future when 1 Gbps becomes a requirement. In this way, the best of multiple key objectives is achieved – many comforting years of immediately available lifespan, support for a long transition window of legacy services, and a strategy for effectively dealing with the continuous traffic growth to come with new bandwidth on-demand.

10.1.3 Phase 2: Deploy High-split – Enabling Gigabit Plus

10.1.3.1 High-Split Extension

Though there are many benefits to an 85 MHz extension, one aspect that cannot be accomplished is support of the 1 Gbps capacity or service rate. This is the case within the parameters of DOCSIS use of the band (360 Mbps), and also the case considering theoretical capacity under DOCSIS SNR assumptions of 25 dB (650 Mbps).

Interestingly, a theoretical 1 Gbps within the 85 MHz Mid-Split architecture would require a 38 dB return path SNR. While well above the DOCSIS requirement, this is, in fact, a relatively easily achievable optical link SNR today using modern DFB transmitters or digital returns. In addition, we can expect higher order modulation profiles enabled at lower SNRs because of the new FEC anticipated – such as 1024-QAM. This would increase data capacity by 25% over 256-QAM and 67% over 64-QAM.

In practice, a manageable operating dynamic range must be considered, as must the other factors that contribute to SNR degradation – RF cascade, user interference, CMTS receivers, and upstream combining, for example. And, though this may be

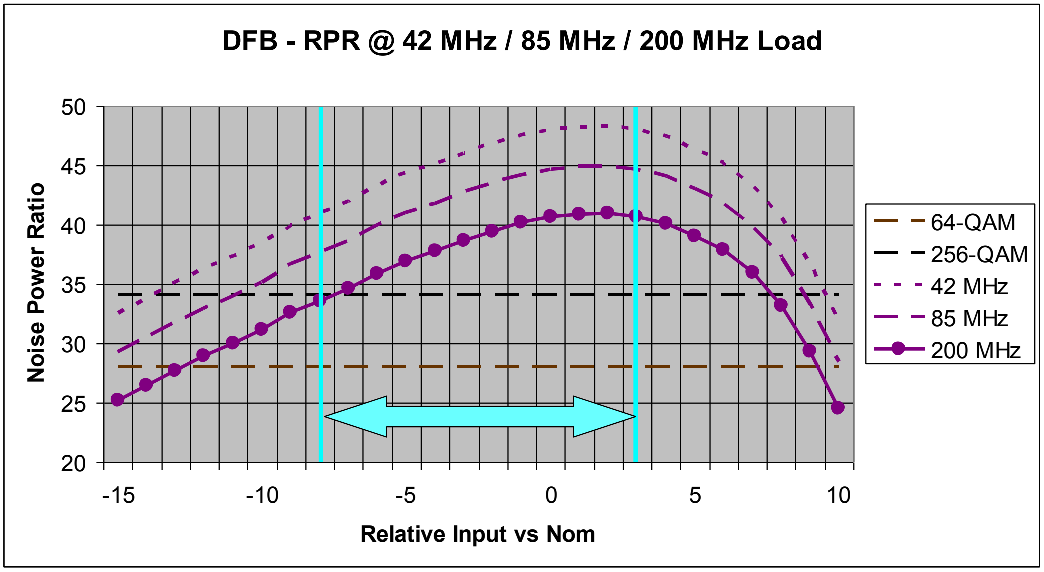


Figure 67 – Bandwidth Loading Effect, 42/85/200 MHz

possible in principle, there is likely to be legacy constraints to having the entire band available for a new, capacity-capable PHY to reach 1 Gbps.

However, this fact does point out that we are entering a new realm of possibilities on the return. Now, with de-combined Headends, 85 MHz of spectrum, modern HFC optics, and new CMTS receivers, and eventually new FEC, many new dB are becoming available toward theoretical capacity and lifespan.

As Table 38 points out, 1 Gbps requires that split to move up to about the 200 MHz range under DOCSIS upstream SNR constraints. 200 MHz is in fact well over 1 Gbps of theoretical capacity, but we assume DOCSIS remains in use for 5-85 MHz, and that the 85-200 MHz region is exploited more aggressively. With new modulation profiles enabled by new FEC, less than 200 MHz will be required, as has been previously discussed.

DOCSIS' maximum profile today (64-QAM@6.4 MHz) itself filling the band out to 200 MHz falls short of 1 Gbps. With 256-QAM, this would no longer be the case. In the case of using split technologies (5-85 MHz of DOCSIS and 85-200 MHz of something else), a shortcoming that could come into play is the inability of that architecture, or at least the added complexity, of supporting 1 Gbps of peak service rate across potentially different systems.

10.1.3.2 Supported by HFC Optics

An attractive advantage of a diplex-based return of 200 MHz or higher is the ability to use analog return optics. However, the additional bandwidth comes with a power loading SNR loss associated with driving a fixed total power into the laser over a wider bandwidth.

Figure 67 compares 200 MHz optical link performance, fully loaded, to 85 MHz and 42 MHz cases. As previously, the lines representing 64-QAM and 256-QAM are SNRs representing theoretical BER without the use of error correction. The power loading loss is easily predictable, as simply the dB relationship among total bandwidths. For the optical link at least, using typical

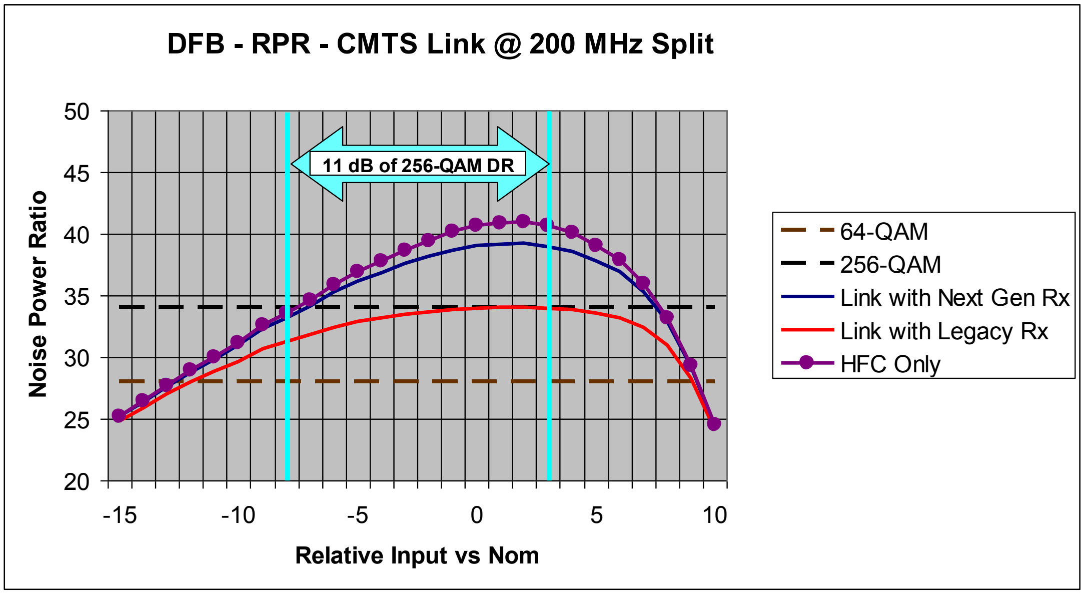


Figure 68 – Projected 256-QAM Dynamic Range Over 200 MHz Split

performance delivered by an analog DFB link, 10-11 dB of dynamic range exists across the HFC optics – a reasonable margin to accommodate alignment, drift, and plant behaviors, but borderline itself for robust, wide-scale roll-out, particularly given degradations that the link will inherent from the rest of the plant.

A comparison of the link using equivalent legacy CMTS receiver performance and modern, lower-noise receivers, is shown in Figure 68. Figure 68 helps to make the point noted in the beginning of this section. The minimum SNR limit assumed for DOCSIS is itself a very dated, and unfortunately conservative and constraining with respect to available capacity.

We now can observe in Figure 68 how the combined effect of the evolution of cost effective, high quality return optics coupled with low noise DOCSIS receivers is opening up new possibilities for extracting capacity from more capable upstream spectrum over wider band.

Based on Figure 68, the full low diplex migration approach has the flexibility of being supported over currently available linear optics. Note once again that we also observed DWDM lasers operating in Figure 20 over high split with NPR performance slightly better than the 1310 nm projection showed here under different link assumptions. This once again shows that today's HFC linear optics is at, or on the verge of, compliant performance for bandwidth efficient profiles over high-split, even without considering new FEC.

Furthermore, High-Splits that exceed current return path optical bandwidth, such as 300-400 MHz, could, in principle, be delivered over linear optics as well. The optics used would simply instead be forward path lasers, which would obviously be high performance.

The preferred, long-term, architectural direction for the long term is a solution based on digital transport over fiber to the node, such as Ethernet or EPON protocol based, to the node, and RF transport over coax. However, an approach based on a low diplex expansion does not require this architecture to operate, offering flexibility to the operator during the difficult transition phase of the network.

When such an architecture is available, the benefits of removing linear optical noise and distortion from the access link budget have very powerful capacity benefits to a low diplex, whose SNR performance is typically set by the optics.

10.1.3.3 Spectrum Evolution

If 85 MHz Mid-Split is a “natural” extension of the Sub-Split (42 MHz) for long-term growth, then a “natural” extension of Mid-Split for long-term peak rate support and FTTH competiveness is the 200-300 MHz High-Split. This concept is diagrammed in Figure 69, along with a summary of the pros and cons.

Unlike Mid-Split, a high split can achieve the 1 Gbps rate foreseen as possibly the next threshold in the upstream after 100 Mbps. And, in doing so, it does not suffer the very high RF attenuations that the alternatives that rely on frequencies above the forward band do. The exact upper band edge is a function of modulation profile, which again is tied to architecture and FEC.

This translates into more cost-effective CPE. As we have seen, implementation of today’s HFC optics is possible, as modern HFC optics is based on 5-200 MHz and 5-300 MHz RF hybrids. And, to reiterate, this architecture, too, would benefit from any migration in the plant that relies on digital fiber delivery and RF carried only in native form on the coaxial leg of the plant.

By maintaining fundamentally a diplex architecture, there is still but one guard band in the architecture, preserving use efficiency. Lastly, at the low end of the HFC spectrum, there would not necessarily be a compelling reason to require an OFDM system, unlike other portions of the band.

The channel quality would not necessarily demand a multi-carrier waveform, and it would have modest advantages at best in a clean channel environment anticipated. Extensions that further empower DOCSIS become more reasonable to consider without a fundamental change in the waveform used, silicon architecture, specification, or new technology learning curves.

At the same time, because the linear optical return architecture anticipates a broadband, noise-like signal, the addition of OFDM channels, even wideband, can be carried within the linear optical architecture as well if the high split band evolves to

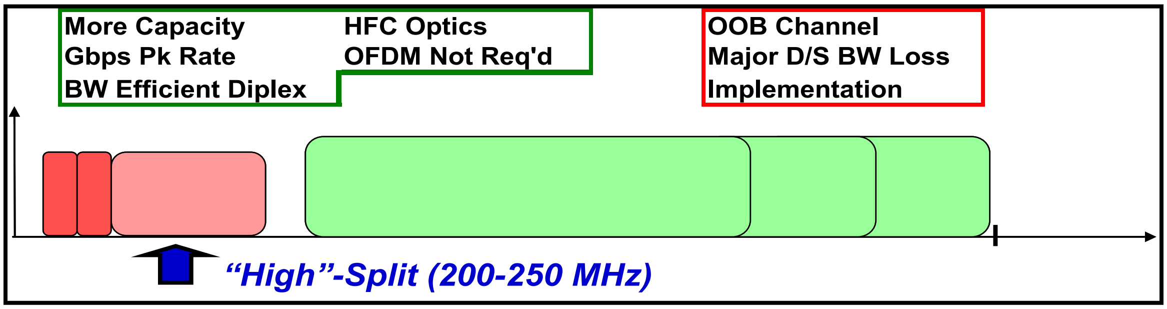


Figure 69 – High-Split Concept, Pros and Cons

include multi-carrier formats. Again, in comparison to other alternatives, this is an added degree of implementation flexibility.

The loss of the OOB downstream channel is an important consideration. However, the logic of this approach is that by the time it becomes necessary – again, likely at least 10 years down the road – the MSO has had ample opportunity to retire through natural attrition or actively manage down legacy STB relying on this OOB channel.

Again, knowing what steps are in place and coming over time, decisions can be made about handling legacy STB either through DSG or Home Gateways associated with an IPV transition.

10.1.3.4 Notable Obstacles

Unlike Mid-Split, High-Split is now a major imposition on downstream spectrum. However, it is expected that downstream spectrum will also undergo expansion over time as traffic in both directions continues to grow. There is already potential spectrum to be mined above the top end of the forward path in many cases, and it is anticipated that if the upstream is to continue to move "up" with high-split, there may be a need also to offset the loss of downstream spectrum by extending downstream as well beyond its current limitations.

By appending new spectrum to the end of the current downstream, this approach to exploiting new coaxial bandwidth is able to maintain a single diplex architecture. This concept is shown in Figure 70.

While this presents a potential solution from a capacity perspective, from a CPE perspective there are important limitations associated with legacy equipment. As the "Simulcast Bubble" winds down at the back end of this decade, models suggest that those savings will be able to compensate for the expansion of upstream into a high-split architecture.

However, under an assumption of persistent CAGR and a continued evolution of HD into even higher resolution formats, such savings will over time once again give way to spectrum management of a new phase of services growth. The window of savings, however, is an important component of a transition that includes the possibility of extending the forward spectrum. We will elaborate on the forward aspects in subsequent section.

10.1.3.5 High-Split Extension – Timing and Implications

The time frames required for a high-split migration are a key element of the strategy because of the intrusive nature of this magnitude of change, and the idea that we may wish to include as part of a transition plan the creation of new forward bandwidth. We touched on the expected timing of 1 Gbps solution in Section 2.6.

Even should the access network be evolved to enable a high-split in the 200-300 MHz band on-demand, such as putting the

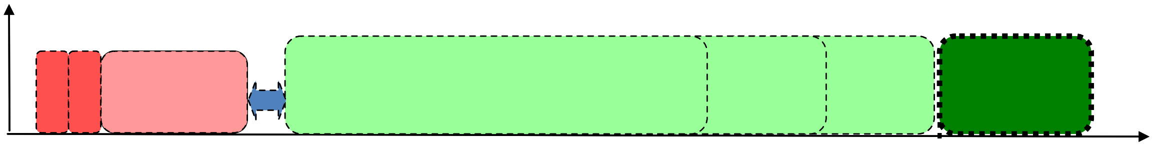


Figure 70 – Possible "Offset" Band Compensating for High Split

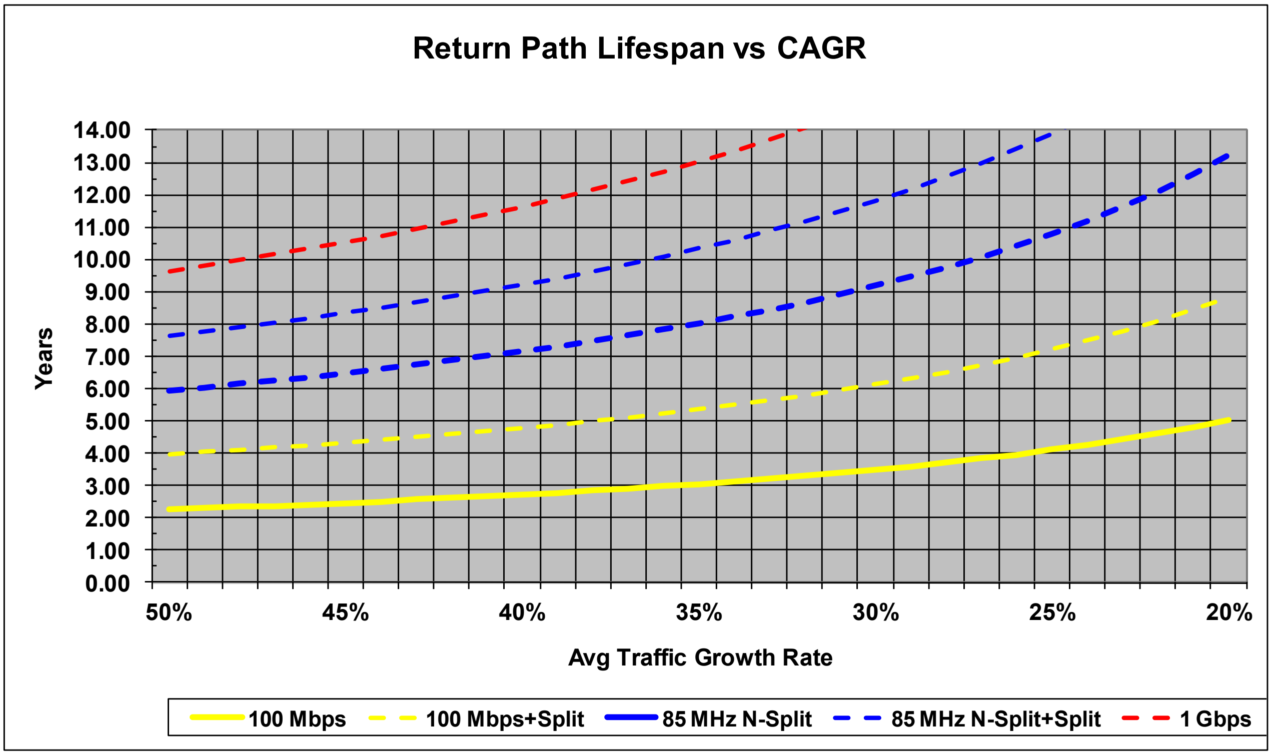


Figure 71 – Relative Lifespan and the Benefits of 1 Gbps

capability in when 85 MHz is deployed, the move to a high split has large impacts on the forward spectrum and return path transport that must be planned.

It is therefore important to get an idea of when we might need it. There are consumption and market pressure components of that, but let's view it in an apples-to-apples way with the prior analysis of the 85 MHz capability for extending return path lifespan. What does a Gbps of capacity imply for long-term traffic growth?

The answer to this question can be examined in Figure 71. It is an excellent illustration of how compounding works and the need to consider what it means if played out over the long term. It shows three threshold cases – 100 Mbps (A-TDMA only), 85 MHz Mid-Split and 1 Gbps (also with a split included).

Zeroing in on the gap between 85 MHz Mid-Split and 1 Gbps at 35% CAGR, we see that there exists about 2.5 years of additional growth after about 10.5 years of lifespan. When we think of "1 Gbps," this intuitively seems odd. Again, this is simply how compounding works. If we base analysis and decisions on the continuance of a compounding behavior paradigm, then the mathematical basis is quite straightforward.

With CAGR behavior, it takes many YOY (year-over-year) periods to grow from. For example, the 40 Mbps of upstream used by a service group today service today to the 440 Mbps that can be delivered by Mid-Split. That number, as Figure 68 shows, is 10.3 years of compounding at 35%. However, once there, the subsequent annual steps sizes are now quite large. That is the nature of compounding, resulting in what seem like small extra lifespan.

10.1.4 Summary

The spectrum migration shown and described above is repeated in Figure 72 and Figure 73. The role of the upstream migration phases in the larger picture of HFC spectrum evolution and the transition to an All-IP end-to-end system is shown in Figure 74 and Figure 75.

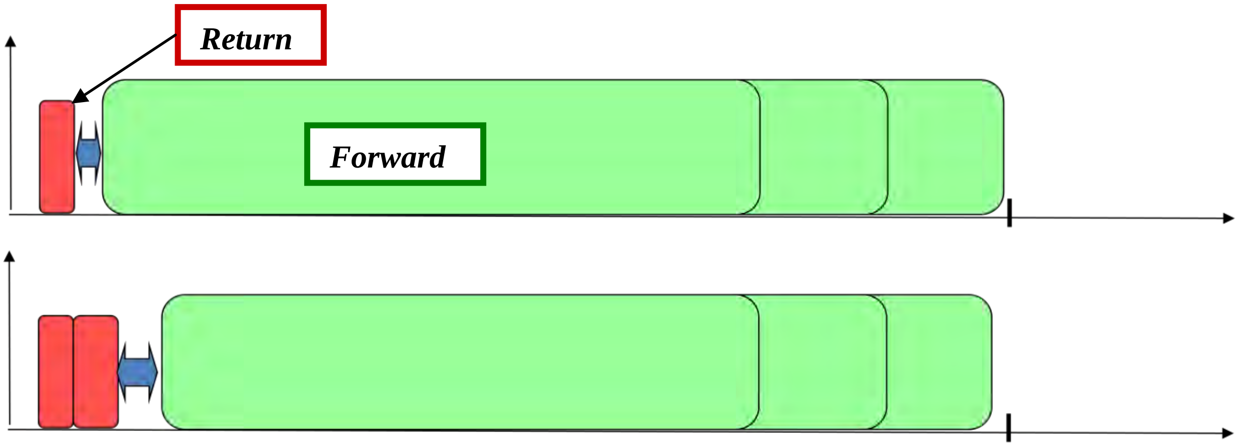


Figure 72 – Phase 1: 85 MHz Mid-Split

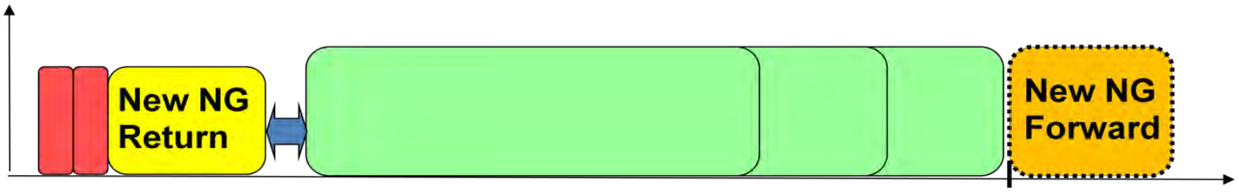


Figure 73 – Phase 2: 200+ MHz High-Split and Possible Relief Band Forward

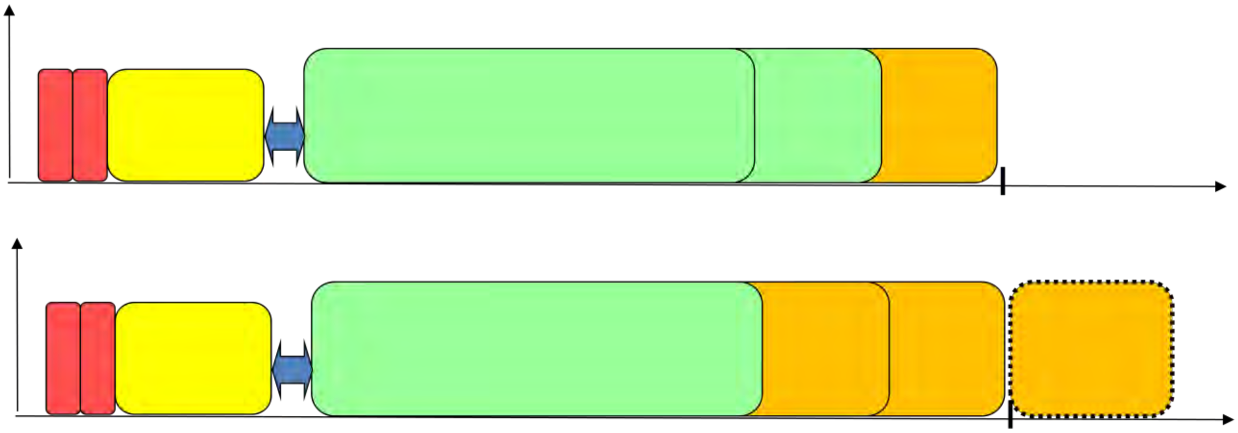


Figure 74 – IP Transition in Progress – Legacy Roll-Back

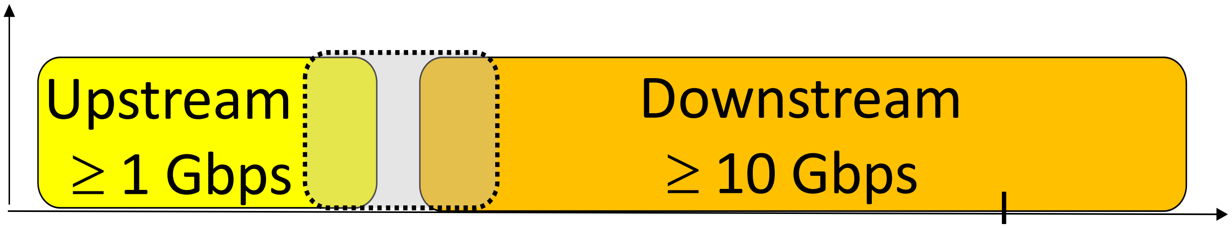


Figure 75 – Final State of All-IP Transition

Flexible/Selectable Diplex, Advanced PHY, Digital Transport-Based HFC Architecture, N+Small/N+0

10.2 Downstream Migration Strategy

10.2.1 Capacity and Lifespan Implications of IP Growth

Every individual HFC plant has evolved on an as-needed basis, and of course under CAPEX budget constraints that inherently come with a network of fixed assets expected to last a long time. As a result, HFC networks in North America have a range of top-end forward path bandwidths.

Typically, however, plant bandwidth is 750 MHz, 870 MHz, or 1 GHz – more so that former two. Absolute bandwidth is obviously important, but fortunately multiple additional tools are available to help manage downstream service growth, such as digital television (DTV), increasingly efficient DTV compression, more bandwidth efficient modulation formats, and switched digital video platforms (SDV). These are all complementary and are in addition to common network segmentation.

As cable advanced video services and data services have grown, however, it has become clear that powerful new dynamics are working against cable operators, and towards a capacity bottleneck in the downstream. The result has been a renewed interest in finding new spectrum, which to a first order directly translates to increased network capacity. Being aware that coaxial cable is not limited to any of the forward band limitations mentioned above, operators are exploring how to access what today is unexploited spectrum above these defined forward bands. There are no technology obstacles to its use, but significant legacy service, network, and equipment implications.

We have discussed in detail the capacity available in DOCSIS and DOCSIS NG as evolution phases take place. However, we have not discussed them in the context of the *available* HFC spectrum. While new DOCSIS capacity is powerful and important, most of the downstream spectrum today is locked down for video services. Finding new DOCSIS spectrum is a major challenge in the normal HFC band, and it is years away before we can exploit the extended bands. We can illustrate quite easily why finding new HFC capacity has become so important and difficult. Consider Figure 76.

Figure 76 projects two cases of IP traffic growth, modeled after the well-travelled Nielsen's Law approach to user bandwidth trends. In this case, it is taken in the aggregate, representing, for example, one service group or perhaps one node.

It assumes that eight DOCSIS downstream service this population today. This is represented on the y-axis, shown on a logarithmic scale because that is the nature of compounding growth. The axis is quite simple to translate in dB – 100 Mbps is 20 dB, 1 Gbps is 30 dB, and 10 Gbps is 40 dB. For eight DOCSIS channels (always using the transport rate in this example, since we are not quantifying service tiers), this works out to 25 dB as a starting point.

The trajectories proceed at 50% Compound Annual Growth Rate (CAGR), interrupted by service group segmentations (such as node splits). In this example, a simple, perfect split (in half) is performed mid-decade. A second, perhaps final, segmentation is done at the end of the decade that resembles an N+0 from a service group size perspective (40 hhp), although it is immaterial to the analysis whether there would physically need to be an amplifier in some particular plant geographies. We use N+0, as we subsequently discuss the implication this has for spectrum planning and capacity exploitation.

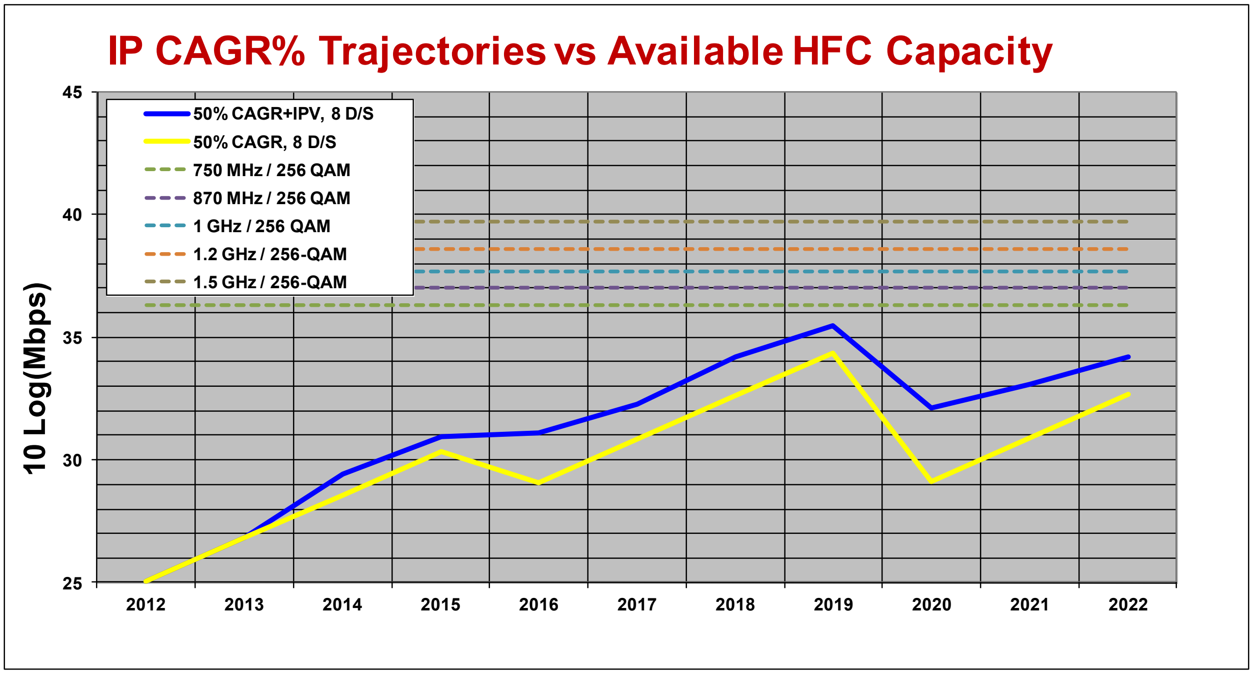


Figure 76 – HFC Downstream Capacity, IP Traffic Growth, and Segmentation

Finally, there are two trajectories because in one case we add dedicated IP Video channels to to IP traffic growth, in addition to the 50% CAGR itself. There is somewhat a philosophical discussion to be had about whether managed IP Video is the new engine of 50% growth (like OTT has been for years), or if CAGR plows ahead in addition to shifting the current video service onto the DOCSIS platform.

Here, the assumption is that blocks of DOCSIS carriers are added every other year beginning in 2014 – first four channels, then 8 channels, then 8 channels for a total of 20. It is a separate analysis how 20 DOCSIS slots represents an assumed video line-up that we will not go into here, but this has been analyzed and written about in many industry papers over the past 4 years.

Five thresholds are shown, consistent with five different assumptions of network bandwidth. In every case, it is assumed that the return bandwidth has been extended to 85 MHz, and the first forward channel is therefore in 109 MHz. It is also assumed, in the extended bandwidth cases of 1.2 GHz and 1.5 GHz, that 256-QAM can be supported.

This is a reasonable assumption – in fact minimally necessary to make turning that band on worth the effort – but obviously unproven at this point. Lastly, each of these thresholds can be incremented by about 1 dB (more) by making the assumption that 1024-QAM replaces 256-QAM (10 Log (10/8)). It was decided not to clutter this figure with those minor increments. But, as discussed, for DOCSIS NG, 1024-QAM downstream and up to 4096-QAM downstream are anticipated modulation profiles, with an objective for total downstream bandwidth of 10 Gbps (which is simply 40 dB in Figure 76 and Figure 77, however it is accomplished).

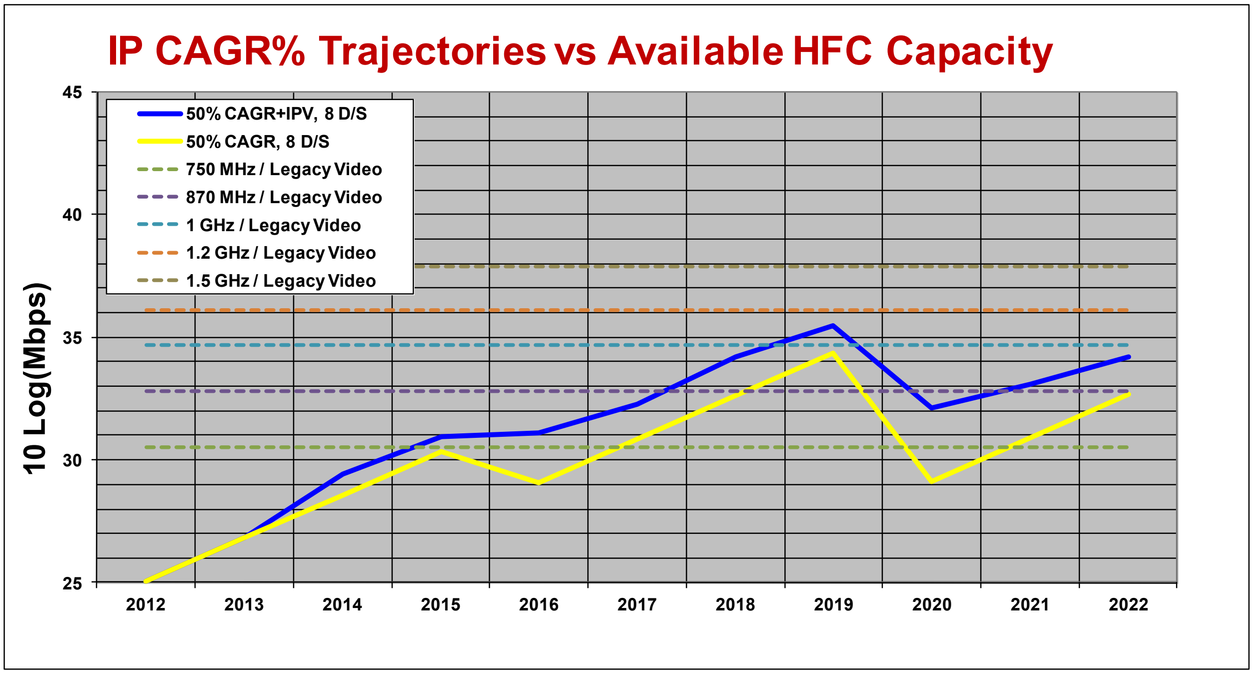


Figure 77 – Capacity, Traffic Growth, and Segmentation – Video Services Added

The thresholds are still based on the assumption of 6 MHz slots of 256-QAM, so represent "current" spectral usage efficiencies, and as such are conservative in that sense. The thresholds, thus, represent the integer number of 256-QAM slots, aggregated to a total based on 40 Mbps/per slot.

An obvious conclusion from Figure 76 would be that the HFC network is in fine shape to take on an extended period of aggressive growth. The network appears not threatened until (projecting to the right) the 2023-2024 time frame, worst case. Of course, there is something seriously missing from this analysis – current services.

Now consider Figure 77.

Figure 77 takes into account that most of the HFC spectrum is not available for new IP growth today. In fact, for most operators, have very little or no "free" spectrum to put new DOCSIS carriers in. When they need new ones, they shuffle other things around and use the tolls above to make it happen. This is much easier said than done as more spectrum, not less, is being consumed with the increasingly competitive environment around HD programming.

The programming line-up above assumes the following:

- Broadcast SD: 100 programs (10 slots)
- Broadcast HD: 40 programs (10 slots)
- SDV 24 slots: This increases the total programming to SD~300 and HD~150
- VOD 4 slots
- No Analog

Clearly, this is not particularly aggressive. First, it is assumed that there are no analog carriers – everyone's long term goal, but executed on by only a few. Also, not all operators are using SDV to this degree, the VOD count is modest, and objectives for HD are for 200-300 programs (not to be confused with "titles"). Finally, there is a real possibility that upstream congestion will require that this band be extended beyond 85 MHz, up to the 200 MHz range or beyond. This would significantly impose on available capacity.

And the result? A 750 MHz is in immediate danger without a service group split, and an 870 MHz network is not far behind. In all cases that do not go above 1 GHz, the "N+0" phase is required before the end of the decade to manage the growth.

The extra runway offered above 1 GHz is apparent – relatively modest for an extra 200 MHz (but this would offset a 200 MHz return at least), and substantial for a 1.5 GHz extension. In the context of the evolution of video services, then Figure 76 can be viewed as the capacities available when the full IP Video transition is complete, and no legacy analog or MPEG-2 TS based video services exist.

As such, they are not "phony" capacities – they merely represent the available capacity, under today's limitations of technology, at the point in time when the legacy service set is fully retired. In this sense, then, they are very valuable thresholds for guiding plant migration and bandwidth management.

A final note on the Figure 76 thresholds is to note that 1 GHz of ideal 1024-QAM bandwidth, at 10 bits/s/Hz efficiency, adds up mathematically to 10 Gbps. We almost achieved this only considering 256-QAM @ 1.5 GHz, and clearly would have done so under a 1024-QAM assumption (one more dB on added to this threshold).

This order of magnitude is important relative to competitive PON deployments. With respect to subscribers served, the PON port is shared by 32 or 64 subscribers. With cable, the access leg is shared by one node port as a minimum, or more generally one complete node. Today, a typical single node average is about 500 homes passed, and this is headed downward. At N+0, it will reside likely in the 20-50 HHP range. For cable then, the subscriber base sharing a 10 Gbps-capable node will be similar to 10 Gbps PON networks in the downstream.

10.2.2 Making Room for Gbps Upstream with New Downstream

Moving to the 85 MHz Mid-Split adds 43 MHz of return bandwidth, doing so at the expense of modest imposition on forward bandwidth. When factoring in the new guard band, possibly nine or ten forward path slots in the traditional analog band are eliminated. Mathematically, converting these channels to digital allows them to all fit into one slot.

As such, as analog reclamation continues, this forward loss does not represent a major capacity concern. The primary operational concern is that the nature of the channels in this region. They are often a basic service tier, and therefore cannot simply be transitioned into the digital tier and off of the analog tier, practically or contractually in some cases, as perhaps some of the longer tail of the analog service could.

Instead, some channel re-mapping and/or more aggressive deployment of digital adaptors would be required. In any case, given the powerful set of tools available to provide downstream capacity, 85 MHz does not present significant imposition on the forward bandwidth in terms of capacity loss.

In the case of a 200 MHz extension, however, this is no longer the case. Cable operators generally use all of their spectrum, and a changed such as high-split, even if it phased in, will call for some significant impacts to the downstream services line-up.

The issue is magnified further when considering that while we are looking to extract downstream capacity and give it to the upstream, the downstream itself continues to see rapid CAGR – more rapid and consistent that the upstream. This amount of lost downstream capacity will have to be replaced, and, in fact, capacity above today's available forward capacity will have to grow over time. 1 GHz worth of 256-QAM slots today adds up to about 6.3 Gbps of total transport capacity, and 7.9 Gbps by enabling 1024-QAM. A 300 MHz starting frequency for the downstream removes about 1.6 Gbps – too big to ignore. That means we must find new downstream bandwidth. In Section 4.5 to 4.7, we identified performance of spectrum above 1 GHz for upstream use, and argued that the obstacles to effectively using the band for upstream make it much more suitable for extending the downstream. Here, we elaborate on this possibility and the potential new data capacity available.

So, where would new bandwidth come from above today's forward band? Virtually any new (actually new, not reseller) plant equipment purchased today will be of the 1 GHz variety. This is clearly at odds with trying to use bandwidth above 1 GHz. Industry discussion around enabling new bandwidth is along three fronts:

(1) What bandwidth do 1 GHz devices actually have? We observed "1 GHz" Taps for out-of-band performance in Section 4.5. Because there is always design margin, is there " free," but unguaranteed, spectrum to exploit? Some operators already place channels above the "official" downstream bandwidth, perhaps at a lower modulation order for robustness, which indicates that there is obviously exploitable capacity in some cases.

It can be shown that some of the friendliest taps in the field have about 20% of imperfect excess bandwidth to mine before difficult to manage roll-off kicks in. Field testing of this grade of tap has been extensively performed. In live plant conditions, a typical tap cascade of nominal coaxial spacing showed useable bandwidth to 1160 MHz with high efficiency for wideband (50 MHz) single carrier QAM [1]. Not all deployed taps will have this amount of useful bandwidth. Of course, the best way to mine bandwidth in such difficult conditions would entail a different modulation approach, and this is particularly the case where discussion of multi-carrier modulation (OFDM) is often introduced for cable networks. Aside from the flexible use of spectrum it allows in periods of transition, and through its use of narrow QAM subcarriers, OFDM would more effectively extract bandwidth, and make more bandwidth able to be exploited.

(2) Some suppliers have developed a 1.5 GHz tap product line. However, there is not very much new build activity, so the market for such products has not grown. Extended bandwidth is also available for some taps already in the field by "simply" swapping out faceplates. This is very intrusive and time-consuming, but of course it is also much *less* intrusive and much *less* time consuming than a full tap swap-out.

Some suppliers have developed this technology specifically for existing plant (versus new build which could, in principle, purchase 1.5 GHz taps). The "swap out" approach yields taps with a specified bandwidth to 1.7 GHz. There is more bandwidth than the 1.5 GHz taps, but it comes at the expense of minor degradation in other specifications. However, field testing has been encouraging that these taps extend bandwidth to at least 1.6 GHz [1].

(3) Full tap swap outs for models that increase bandwidth to up to 3 GHz (or use in new builds). This, of course, is a very intrusive plant modification.

It is important to note that suppliers have not yet developed node or amplifier platforms, at least not in volume scale, that extend beyond 1 GHz. There are no technology reasons this could not be done, although there are likely major redesigns involved in most cases right down to the housing, circuit boards, and connectors.

This is viewed as unlikely to take place for RF amplifier platforms, but perhaps not so for nodes. As N+0 is potentially a logical "end state" for an HFC architecture, the ROI picture is somewhat clearer to make for equipment manufacturers. In addition, nodes have undergone generally more R&D investment than RF platforms have, as they have kept up with the optical technology evolution.

Many fielded RF platforms have not changed very much since they were originally designed, and have been had their bandwidth limits continuously pushed. It is unclear how many new MHz are easily available, and the range of RF platforms is much larger.

This limitation on the bandwidth of the RF amplifier is important in the context of accessing new bandwidth and understanding the enabling architectures to do so. We will elaborate and quantify aspects of this in subsequent sections.

10.2.3 Excess Bandwidth Calculations on the Passive Plant

The first place to look for more downstream spectrum is simply in the band that continues directly above today's forward path band edge. While this was shown to be a difficult band for an upstream service to efficiently and cost effectively support, it is much easier to consider as much for the downstream.

The downstream channel is already very linear, has a very high SNR, and these features of the access equipment are shared by the homes passed common to a piece of equipment in the plant. And, fortuitously, in many 1 GHz tap models there is that significant "free" bandwidth available.

Figure 78 shows the frequency response on the "through" port of the particular 1 GHz tap described in the field trials above that yielded an 1160 MHz net useful band edge. This port would be in series with other taps on the way to a connected home. The response on the tapped port also has essentially parasitic, low-loss properties over the first 200 MHz above 1 GHz.

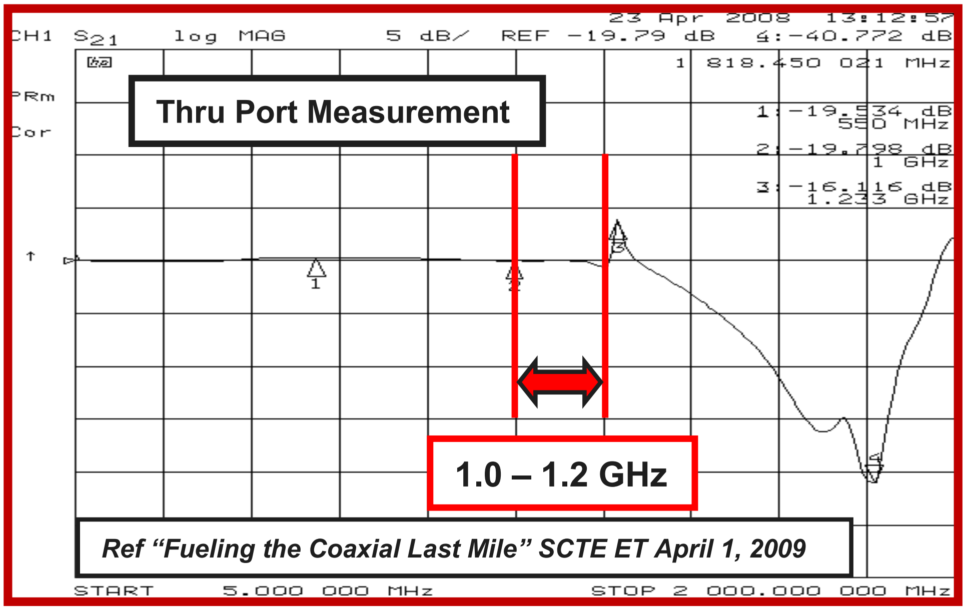


Figure 78 – 1 GHz Tap Frequency Response, "Thru"

Though not as perfectly flat, it creates no significant distortion burden to RF signals in the band, and in particular when considering that a new generation of OFDM technology will almost certainly be created to operate in that regions, and if so will run an adaptive bit loading algorithm.

The same is the case for some families of 750 MHz taps (available bandwidth exists above 750 MHz) and 870 MHz taps (available bandwidth exists above 870 MHz).

The amount of useful bandwidth and loss properties are vendor dependent, but cable operators already often use slots above these limits. Conveniently, as Figure 78 shows, the amount of available new bandwidth simply trickling over the top of the band is virtually the same the amount of bandwidth that would be removed from the forward by a 200 MHz high-split architecture.

With the support of the supplier community, CableLabs has undertaken an investigation to statistically quantify this excess bandwidth across Tap models and manufacturers so that operators can better understand in their specific plants what useful bandwidth is available, and how that changes with time with shorter cascades.

An important item to re-emphasize is that there is no guard band involved when this spectrum is operated as only a downstream extension, as there would necessarily be if upstream were to be deployed in this band. This "replacement" bandwidth amount provides adequate spectrum to facilitate new downstream capacity.

The ability to fully exploit this bandwidth in the passive plant obviously depends heavily on the band coverage of the actives themselves and the depth of the cascade. Clearly, this is where shortening cascades and "N+small" continue to payoff for HFC evolution.

The tapped port, of course, also contributes to the frequency response, and a sample of this port on the same 1 GHz tap model (2-port, 20 dB) is shown in Figure 79. The response on the tapped port also has essentially parasitic, low-loss properties over the first 200 MHz above 1 GHz.

Though not perfectly flat, it creates no significant burden to RF signals in the band, and in particular when considering a new generation of modem technology, such as multi-carrier. The same is the case for some families of 750 MHz taps (available bandwidth exists above 750 MHz) and 870 MHz taps (available bandwidth exists above 870 MHz).

It is clearly evident that the band between 1.0 GHz and 1.2 GHz is not flat, having about 2 dB of what can best be described as a broadband ripple in the response.

10.2.3.1 Excess Bandwidth SNR Model

In order to calculate the capacity associated with this "extra" bandwidth, we must numerically model this frequency response. This is easily accomplished for parasitic-type roll-offs, more so even that with classic RF filter responses such as diplexers.

We can, in fact, fit the attenuation response to some fundamental filter shapes and use those to calculate attenuation. And, by proxy, SNR for a fixed transmit power. In this case, the roll-off response can be fairly well represented by scaled versions of a 5th order Butterworth response, as shown in Figure 80.

Here, the thru attenuation (blue) of approximately 10 dB across the 1-2 GHz band, as well as the roughly 20 dB of attenuation over 600 MHz represented by the port (red), is represented. Note that increasing stop-band attenuation typically means correspondingly poor *return loss*, which is an RF reflection mechanism – a mechanism already part of DOCSIS, and that has become very sophisticated with DOCSIS 3.0. Of course, if a multi-carrier PHY is adopted in this band, it too is robust to this distortion, but through different means, such as use of a cyclic prefix.

Filter roll-off regions also typically correspond with regions of high group delay variation – another challenge taken on by the 24-Tap equalizer. For A-TDMA, however, there are limits to how successful the equalizer can be with combined micro-reflection, amplitude response, and group delay distortion.

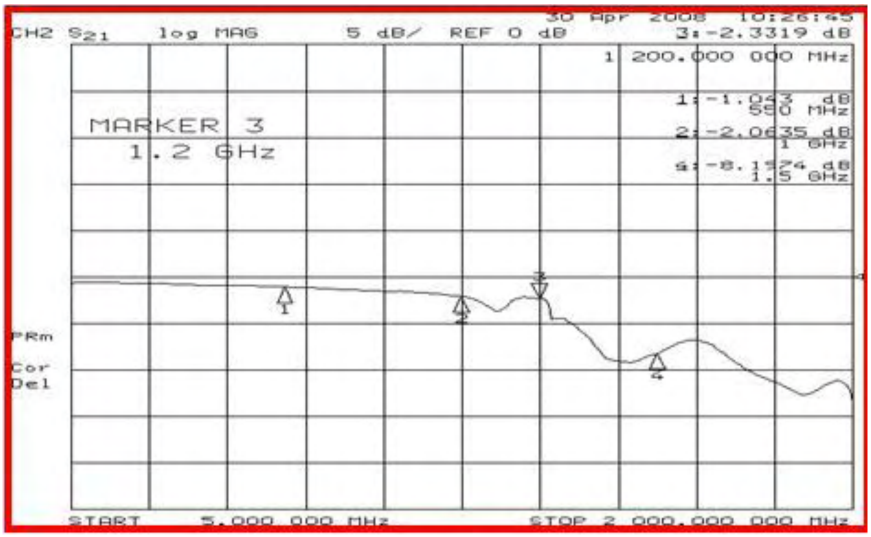


Figure 79 – 1 GHz Tap Frequency Response, Tapped Port to Home

Performance has been shown to be far, far beyond the conditions called out in DOCSIS specifications. Nonetheless, multi-carrier evolutions to the PHY minimize the potential concerns over operating in these regions as well. System parameters (subchannel widths, cyclic prefix guard times) can be used very effectively to overcome these obstacles where the channel performance degrades.

Consider the two narrowest bandwidth curves of Figure 80. These represent the composite frequency response of an N+0 cascade of five taps (N+5T, pink) or ten taps (N+10T, brown), and an accompanying length of coax governed by a typical attenuation model.

A subscriber at the end of a ten tap run will of course see nine thru responses and a tapped port (and quite possibly an active that would need to support this band or bypass it), and this response is represented by the brown curve. The pink curve represents a five tap scenario, which is a more typical run of taps between actives.

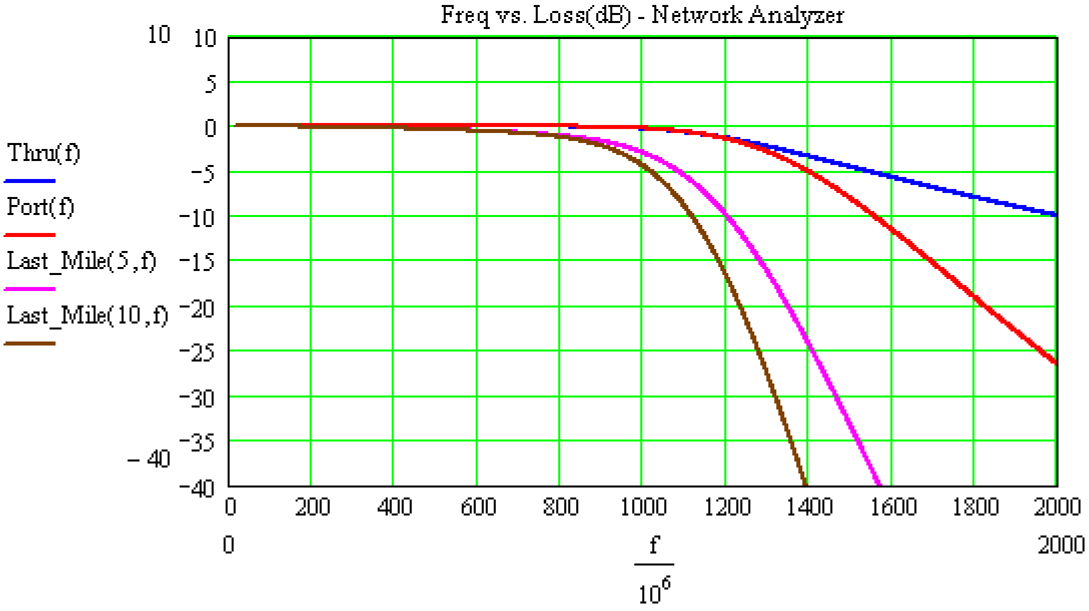


Figure 80 – Modeled Tap + Coax Performance

These attenuation curves for a cascade of taps, plus interconnecting coaxial runs, can be used to quantify the attenuation profile, and, given a transmit power profile (is it tilted or not), the SNR delivered from the network for a given power, and thus the capacity available as a function of new spectrum. We can thus see the efficiency with which this new part of the band delivers capacity.

10.2.3.2 Capacity Derived from Excess Spectrum

Figure 81 quantifies available capacity, assuming an HFC forward digital band starting SNR of 45 dB at 1 GHz in the HFC plant and using the frequency response of Figure 80. An HFC downstream link at the output of a node would be expected to deliver at least 51 dB of SNR as a common objective in the analog band, leaving the digital band 6 dB removed from that performance.

Thus, this represents an N+0 case ideally, but could also reasonably apply to a short cascade that includes RF amplifiers that pass this band with a flat response as long as there are not more than 5 taps in the series (the 10 Tap case is not shown in Figure 81). It also conservatively assumes a flat transmit response, and, while increasing in frequency, calculates the resulting capacity as this band edge moves to the right.

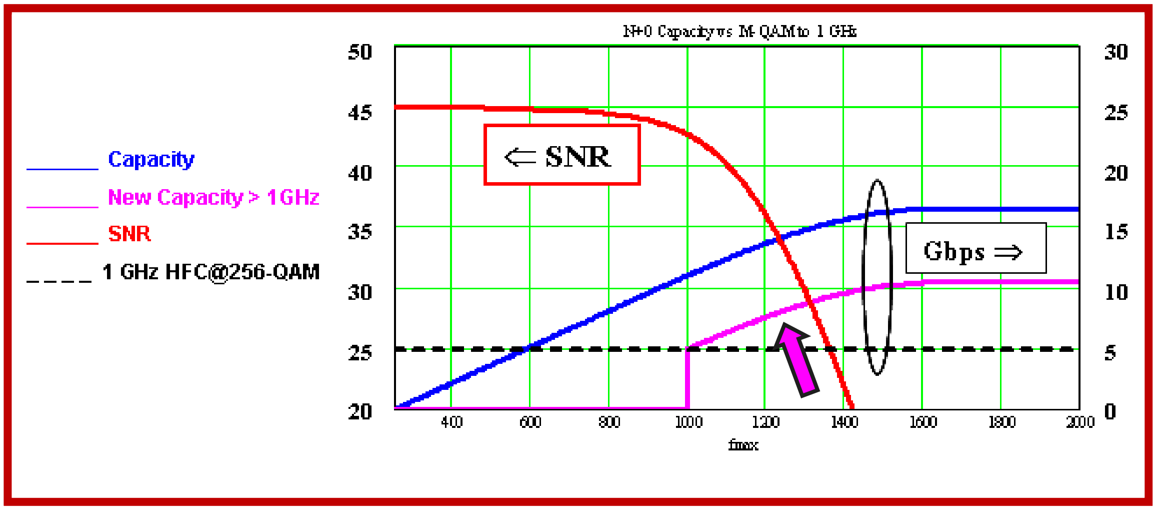


Figure 81 – N+0 Capacity vs. M-QAM to 1GHz

It is reasonable that an uptilt may be applied to compensate for the cable effect at least, but this would amount only to about 3 dB from one band edge to the other. Today's RF outputs are already tilted so as an extension of the payload this could be inherent.

The curves in Figure 81 show a full forward band throughput of 256-QAM , along with the theoretical capacity in Gbps (blue, right vertical axis), for a given maximum upper edge of the band shown on the x-axis. These capacities are shown along with the SNR vs. frequency delivered from a 5-tap cascade made up of taps such as that shown in Figure 78, and one coupled port from the same as shown in Figure 79.

The final trace (pink) recognizes the 256-QAM legacy spectrum as a given, already occupied bloc, and above that identifies new theoretical capacity potentially that can be exploited above 1 GHz in the passive segment as a function of the maximum upper frequency used.

Clearly, within the first 200 MHz above 1 GHz, more than a Gbps of capacity can be extracted. Also apparent is how much latent capacity still exists as the cascades shrink and open up new RF bandwidth potential, considering that 256-QAM is today's maximum modulation profile.

Of course, the expectation of 1024-QAM and perhaps even higher order modulations [1] are expected with the help of new FEC, allowing the "actual" to get closer to the capacity curve. Figure 81 also indicates that beyond 1.4 GHz there is diminishing return on new capacity as attenuation begins to take its toll on SNR.

For high SNR, such as those used in Figure 81, capacity is directly proportional to both bandwidth and SNR expressed in dB with very small error, a relationship observable in Figure 81.

10.2.3.3 Multicarrier Modulation Optimizes Channel Efficiency

Multicarrier techniques(OFDM)have made it possible to work through seriously impaired frequency response characteristics with high performance. As we observed in Section 7.3 “OFDMA, OFDM & LDPC”, the use of narrow subcarriers vastly simplifies the equalization function, and simultaneously provides the ability to consider each subcarrier independently in terms of the bandwidth efficiency of the modulation profile it can support on a dynamic basis.

Implementing multi-carrier technology for cable is a potentially attractive way to make use of the extended bandwidth of the coax, and because of this is a fundamental recommendation for the DOCSIS NG PHY. Much like xDSL before it, cable can leverage the powerful capabilities of OFDM techniques to most effectively use the current media, and this becomes more important as the use of the spectrum changes over time.

10.2.3.4 Excess Capacity Summary

In summary, here are plenty of available bits per second left to be exploited on the coax. It is expected that the DOCSIS NG PHY, using LDPC for most efficient use of SNR, and OFDM for most efficient use of unpredictable and changing bandwidth, will close the gap considerably on theoretical capacity over the HFC network. The most straightforward way to access this bandwidth is by continuing to migrate to fiber deeper, with a likely end state landing at an N+0 architecture of passive coax, and perhaps for practical purposes in some case N+1 or N+2.

Other useful elements of the migration include new RF technologies, such as GaN amplifiers that deliver more power at equivalent distortion performance can be used in multiple ways to enable this capacity to be accessed – allowing more economical deployment of N+0 long term (more hhp/node), using the additional RF drive capability to drive the new forward spectrum, or taking advantage of analog reclamation to deliver broadband performance based on QAM-only performance requirements.

Lastly, the same architectural option that delivers more capacity from the plant (N+0), bringing the last active and CPE closer together, works also from the receive end of the downstream link. Tied closely to optimal use of new spectrum is the ability to implement a point-of-entry (POE) home gateway architecture long-term.

This approach abstracts the HFC plant from inside the home, terminates downstream PHYs, delivers the bandwidth within the home on an IP network, and rids the access plant of having to overcome uncontrollable in-home losses and architectures.

10.2.4 Architectures for More Excess Bandwidth in The Passive Plant

As comforting as it might be that some plant segments already have some useable bandwidth above the specified top end of the equipment – used in some cases already for legacy extension – Figure 81 obviously behaves asymptotically because of the limitations of existing equipment. In the case evaluated above, it is due to the ultimate limitations of the 1 GHz Taps used in the analysis.

If this limitation could be addressed, then the blue and pink curves shown in Figure 81 would continue to climb, providing access to more capacity, and with only the inherent coaxial attenuation contribution to shaping of the frequency response.

While there is little appetite for the intrusive nature and cost of exchanging all Taps in the plant, an elegant solution to freeing up more very useful spectrum is one that allows more spectrum without a wholesale cut-out of the existing Taps.

Tap models, such as those developed by Javelin, Inc., that allow for only a faceplate change of the existing Tap housing have been on the market to support this concept for some models of Taps in the field.

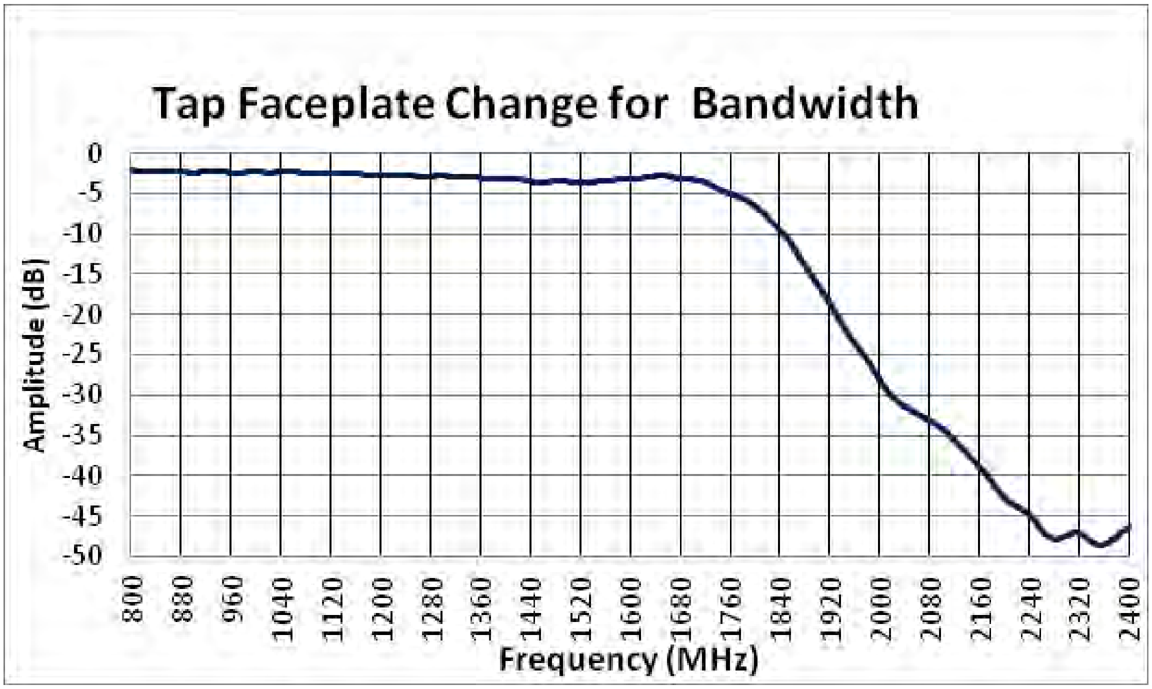


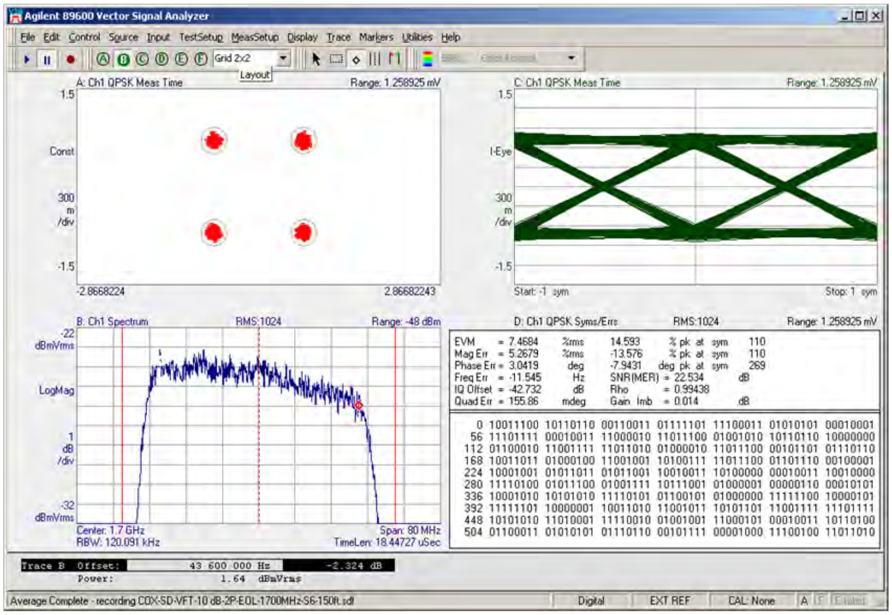
Figure 82 – Modifying Taps to Increase Bandwidth on the Passive Plant

This is a much more simplified and time-efficient process for a field technician, and thus potentially a manageable option to operators looking for the sweet spot of “quick fix” versus bandwidth extraction. Wholesale change-outs can extend the Tap bandwidth to almost 3 GHz.

Figure 82 shows a frequency response of a sample Tap that has had its faceplate removed for the purpose of having the bandwidth extended.

Figure 82 shows a well-behaved passive response to 1.7 GHz. It is straightforward to estimate the additional capacity this provides using Figure 81. The first 200 MHz of spectrum added slightly less than 3 GHz of new capacity to the forward path. The additional 500 MHz shown in Figure 82 under the same assumption increases the total new capacity available to a little more than 10 Gbps theoretically.

This is a compelling number, as it immediately brings to mind the ability of the properly architected and engineered HFC plant to deliver GEPON-like speeds to its subscribers, without the need to build fiber-to-the-home. Indeed, as pointed out in [1], exploiting all of the available coaxial plant instead of just the legacy spectrum allows HFC to be directly competitive with FTTH rates and services.



```
EVM        = 7.4684    %rms     14.593     % pk  at  sym    110
Mag Err    = 5.2679    %rms    -13.576     % pk  at  sym    110
Phase Err  = 3.0419     deg     -7.9431  deg pk  at  sym    269
Freq Err   = -11.545     Hz     SNR(MER) = 22.534          dB
IQ Offset  = -42.732     dB     Rho      = 0.99438
Quad Err   = 155.86    mdeg     Gain Imb = 0.014           dB

  0 10011100 10110110 00110011 01111101 11100011 01010101 00010001
 56 11101111 00010011 11000010 11011100 01001010 10110110 10000000
112 01100010 11001111 11011010 01000010 11011100 00101101 01110110
168 10011011 01000100 11001001 10100111 11011100 01101110 00100001
224 10001001 01011011 01011001 10010011 10100000 00010011 10010000
280 11110100 01011100 01001111 10111001 01000001 00000110 00010101
336 10001010 10101010 11110101 01100101 01000000 11111100 10000101
392 11111101 10000001 10011010 11001011 10101101 11001111 11101111
448 10101010 11010001 11110010 01001001 11000101 00010011 10110100
504 01100011 01010101 01110110 00101111 00001000 11100100 11011010

Trace B  Offset:          43 600 000 Hz          -2.324 dB
         Power:                 1.64  dBmVrms

Average Complete - recording COX-SD-VFT-10 dB-2P-EOL-1700MHz-S6-150ft.sdf    Digital    EXT REF    CAL: None    A
```

Figure 83 – Wideband (50 Msps) Characterization on Extended Tap BW

Even more simply, using just 1024-QAM, or one order of full modulation profile increase above 256-QAM (not full capacity), we need about 1.2 GHz of spectrum to aggregate to 10 Gbps of transport. Cable is not far from having the tools in place to achieve this already, and new LDPC FEC will make this actually quite simple to achieve.

Figure 83 shows a snapshot of the signal quality measured through an RF leg in the field made up of Taps of the type shown in Figure 82, transmitted *from* the end of a typical 150 ft drop cable (i.e. though passive, a measurement in the upstream direction).

There is some obvious droop at the band edge of this unequalized signal, with the drop cable contribution a primary culprit, but it is nonetheless easily corrected. The most important characteristic of Figure 83 has nothing to do with frequency response, but instead with the measured link loss from the end of the drop to the measurement station, sitting at the point where it would represent the first active in an N+0.

This is where “top split” architectures struggle to effective for return path applications. They must overcome in the 60 dB range – potentially worse when considering in-home variations – all tied simply to the relative attenuation characteristics of the low diplex band versus above 1 GHz.

The extended bandwidth taps relieve some of this through loss, but the impact on new CPE is significant in terms of generating broadband, linear, high RF outputs to overcome the loss and enable bandwidth efficient link budgets.

10.2.5 Summary

Many "1 GHz" Taps have significant, useable excess bandwidth above 1 GHz, although this is not guaranteed by specification. A practical cutoff point for family of Taps with the behavior shown in Figure 78 and Figure 79 for a 5-TAP cascade is between 1.16 GHz and 1.22 GHz.

It is expected that the same can be said above 750 MHz for "750 MHz" Taps and above 870 MHz for "870 MHz" Taps. However, because performance above 1 GHz is unspecified, different TAP models from different vendors are likely to vary in performance.

Faceplate replacement Taps represent a less-intrusive bandwidth extension option for the passive plant than 100% Tap replacement, and yield significant excess capacity.

The primary system issue is simply the RF loss entailed at these frequencies, and for this reason this capacity is most easily accessed for downstream use. The downstream channel already operates to 1 GHz, is highly linear across multiple octaves, delivers very high SNR for QAM, and is designed for broadband high power cost effectively to many users.

Each level of investment in bandwidth corresponds, as expected, to increased intrusiveness and operational expense. For some Tap models, there is virtually free bandwidth on the passive plant to at least 160 MHz above 1 GHz.

With the intrusiveness of a tap faceplate change, there is at least 700 MHz of new bandwidth made available. Finally, if all TAPs are completely replaced, bandwidth out to 2.75 GHz is freed up.

In all cases, standard 1 GHz HFC actives do *not* support the extended bands. And, in all cases, the rules governing RF loss versus frequency across the coaxial cable still exist and become the primary link budget obstacles to high order QAM transmission.

10.3 System Implications of HFC Evolution and Extended Bandwidth

There is already some flexibility in existing outdoor plant platforms. Modern nodes are very modular in nature and offer the flexibility to segment by port. Figure 84 shows the type of modularity most modern HFC nodes have today.

While amplifier platforms have seen less evolution than nodes in the past decade, there has been substantial investment in one area – fielded amplifiers today that can become nodes tomorrow through the swapping of internal plug-ins.

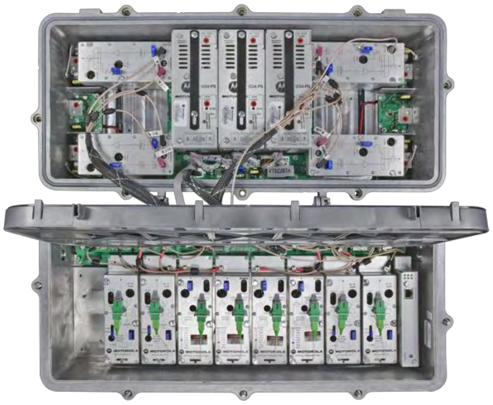


Figure 84 – Modern Node Platforms are Inherently Modular and Increasingly Flexible

This allows incremental bandwidth improvements as required within the context of the well-understood HFC infrastructure. Some suppliers have developed this capability for their entire RF amplifier portfolio, and it then becomes quite straightforward to envision at least a lower touch evolution to an N+0 deployment built around an existing plant.

Taking the idea of node splitting to it logical conclusion, it ultimately leads to a natural N+0 end-state architecture. It is the final incarnation at which the coaxial cable last mile medium remains, leaving this passive part of the network and infrastructure investment in place.

Now, since these deeper nodes will correspond with adding bandwidth and average bandwidth is about serving group size, practical geography (subscribers don't always tend towards a uniform physical density) may dictate that an active element is still required. And, getting to an N+0 by successively splitting nodes repeatedly until there is nowhere else to go is probably not the most effective way to accomplishing the objective.

Plant geography and diminishing returns on average bandwidth per SG due to imbalance are likely to make this approach and less effective than a managed transition plan, and likely more costly as well.

Note that the march of nodes deeper into the network to N+0 leaves high similarity at the block diagram level to FTTC architectures used in the telco domain. Of course, there are significant differences in signal types on the fiber (at least for now), what is inside the node, and in the electrical medium – copper pair or coaxial. At some point, and possibly within the window of this fiber-deep evolution, the fiber delivery may become more common, leveraging 10 GbE or EPON technologies in both cases.

Table 39 – Total QAM Power with *All* Analog Removed

	Additional QAM Level Available			
	870 MHz		1000 MHz	
Analog-QAM	870 MHz Uptilt		1000 MHz Uptilt	
Back-off	12 dB	14 dB	14 dB	16 dB
-6	2.8 dB	2.5 dB	1.9 dB	1.6 dB
-8	4.2 dB	3.8 dB	2.9 dB	2.5 dB
-10	5.7 dB	5.3 dB	4.2 dB	3.7 dB

10.3.1 Bandwidth and Power Loading

The highest order deployed QAM modulation today is 256-QAM, which delivers a 1e 8 BER at a 34 dB SNR, ignoring coding gain improvements for simplicity. Meanwhile, a modest analog channel requirement is on the order of 45 dB – or 11 dB different.

Some of that large margin is eaten up in the relative signal level back-off, used on the QAM load. Use of 64-QAM levels 10 dB below analog and 256-QAM levels 6 dB below analog are common – and yet still leave significant SNR margin (7 dB and 5 dB in the examples given). These digital offsets can be used as tools in the RF power loading plan, to a degree.

Because of the relationship between analog and digital power and their contribution to the total, when considering analog reclamation, additional power potentially becomes available for QAM signals. This added level means that they could absorb more attenuation from an SNR perspective.

Table 39 shows an example of the theoretically available increase in digital power on the multiplex, given that a fixed total RF output power is required for the mixed multiplex or for an all-digital load.

While this analysis is done for a full digital load, the analysis is easily adaptable to any number of analog carriers. For a small analog carrier count, the difference with "all-QAM" is relatively minor, because the limited set (such as 30) of analog channels are carried at the low end of the band, where their individual powers are smallest under commonly applied RF tilt. An example of stages of analog reclamation is shown in Table 40 for 870 MHz for comparison.

The case of "flat" would represent the change in the forward path multiplex sent across the optical link, while the uptilted cases represent the case out of the node or of

Table 40 – Power Loading Effects of Analog Reclamation - 870 MHz

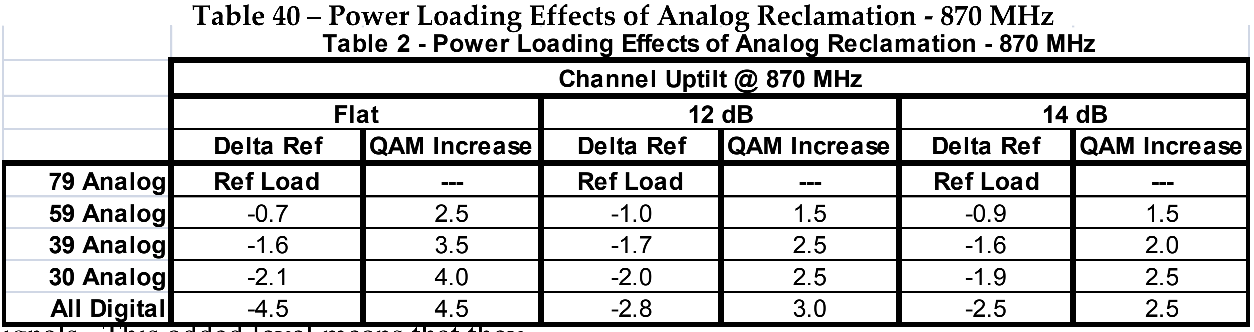


Table 2 - Power Loading Effects of Analog Reclamation - 870 MHz

	Channel Uptilt @ 870 MHz					
	Flat		12 dB		14 dB	
	Delta Ref	QAM Increase	Delta Ref	QAM Increase	Delta Ref	QAM Increase
79 Analog	Ref Load	---	Ref Load	---	Ref Load	---
59 Analog	-0.7	2.5	-1.0	1.5	-0.9	1.5
39 Analog	-1.6	3.5	-1.7	2.5	-1.6	2.0
30 Analog	-2.1	4.0	-2.0	2.5	-1.9	2.5
All Digital	-4.5	4.5	-2.8	3.0	-2.5	2.5

an amplifier where the RF level is tilted to compensate for cable attenuation versus frequency. Typically, it is the optical link which sets HFC SNR, and the RF amplifier cascade that is the dominant contributor to distortions.

What is clear from Table 39 and Table 40 are that the process of analog reclamation offers the potential; for SNR recovery. In the case of beginning with 79 analog slots and migrating to an all digital line-up, there is 4.5 dB of increased digital level available per carrier into the optical transmitter in theory, which can be converted to a better digital SNR.

10.3.2 Extended Bandwidth Loading

If the use of coax is to be extended to frequencies above 1 GHz, power loading will be affected accordingly for non-RF overlay approaches. For the sake of simplicity, we consider two cases:

1) Assume that the applied tilt will be required to extend this band according to the coaxial relationship previously discussed

2) Consider a flat signal band is delivered in the 1-1.5 GHz range, and new technology is burdened with overcoming the limitations of higher attenuation

Table 41 – Power Loading of Extended Bandwidth

	Analog BW	Digital BW	Digital Derate Relative to Analog (dB)		Digital BW Tilt (dB)		Relative Pwr
	MHz	MHz	550 MHz-1GHz	1-1.5 GHz	550 MHz-1GHz	1-1.5 GHz	dB
Reference	**550**	**450**	**-6**	**Unused**	**14**	**Unused**	**0.0**
Case 1	550	450	-6	-6	14	14	7.4
Case 2	550	450	-10	-10	14	14	3.9
Case 3	550	450	-6	-6	14	0	3.9
Case 4	550	450	-10	-10	14	0	0.9
Case 5	550	450	-6	-15	14	14	2.0
Case 6	265	735	-6	-6	14	14	7.2
Case 7	265	735	-10	-10	14	14	3.3
Case 8	265	735	-6	-6	14	0	1.2
Case 9	265	735	-10	-10	14	0	0.4
Case 10	265	735	-6	-13	14	14	2.2

We will use 1.5 GHz to be consistent with the above discussion on capacity and tap bandwidths. Example cases under these assumptions are shown in Table 41, which illustrates some key points. The starting point is the 1 GHz reference load of sufficient level and performance.

From a power loading standpoint, continuing the tilted response to 1.5 GHz adds a significant power load. However, variations to the tilt approach create a seemingly manageable situation (small dB's) from a power handling standpoint. Hybrids today are typically designed, through their external circuit implementations, to purposely roll-off.

Several 1-1.5 GHz RF loading implementations in Table 41 are relatively non-stressful. If the 1 1.5 GHz band is flat, the additional power load is between 0.4 dB to 3.9 dB. In the situation where the band is extended to 1.5 GHz in conjunction with analog reclamation leaving 30 channels in analog, the increase in total load is limited to 1.2 dB.

In order to maintain a tilted output to 1.5 GHz, an overall digital band de-rate of -10 dB instead of 6 dB keeps the power load hit to less than 4 dB. Given that this may be accompanied by perhaps anN+0 architecture, the 4 dB of power may be available while maintaining sufficient performance because no noise and distortion margin needs to be left for an amplifier cascade. This approach may be more costly in terms of added power, but it is more straightforward to implement a uniform frequency response in a single circuit, than one that tilts part of the band but not another.

A final set of cases that show reasonable loading increase are the 79 channel and 30 channel cases with the tilt maintained, but new derate applied in the 1-1.5 GHz band. To maintain a load increase of <2 dB, an additional 9 dB and 7 dB derate should be applied for 79 and 30 channels, respectively. However, considering the link budgets associated with HFC networks today, dropping the levels this low likely creates a challenge to most efficiently using this band, as this would is then lost SNR and lower capacity.

Summarizing, it appears that various implementation scenarios are eligible for maintaining a reasonable power loading situation while extending the band of the output to 1.5 GHz. This does not account for possible changes in hybrid capability for an extended band. The hybrids themselves have bandwidth up to 1.5 GHz, but the circuits they are designed into are purposefully limiting and optimized for today noise and distortion requirements over legacy bandwidths.

Table 42 – Performance Effects of N+6 to N+0 Conversion

Performance of 1 GHz Multiplex with 79 Analog			
Parameter	N+6	N+0 (nom)	N+0 (high)
CCN	48	51	51
CSO	56	64	62
CTB	58	70	67

10.3.3 Reduced Cascade Benefits

It is well-understood cable math how shorter cascades result in higher SNR and lower distortion, as the link degradation of adding a relatively short length of fiber is a favorable trade-off with a run of active and passive coaxial plant.

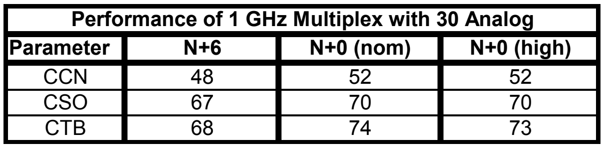
Let's look at a typical example and evaluate this cascade shortening impact. In this case, the link is a 1310 nm link in an N+6 configuration in its original state, and the noise and distortion performance calculated for a 1 GHz multiplex of 79 analog channels.

The link is then modified to an N+0, and the analysis re-run at the same nominal output levels. It was also run for a 4 dB increased output level mode, as the extension to N+0 architectures today may entail a higher output requirement to accommodate the likelihood that the plant geography is not well suited to 100% N+0, and recognizing that the removal of the RF cascade gives distortion margin back that may allow higher output levels. The results of this analysis are shown in Table 42.

Note the emergence of 3-4 dB of additional SNR (CCN or Composite Carrier to Noise). This is independent of any SNR gain due to increasing digital levels that may be possible with analog reclamation per Table 39.

Increasing QAM levels while adding QAM in place of analog is not a fixed dB-per-dB SNR gain, as adding digital channels adds contributors to CCN (composite carrier-to-noise). However, this conversion to CCN also creates a significant drop in CSO and CTB distortions, which are significant impairments for higher order QAM

Table 43 – Performance Effects of N+6 to N+0 Conversion



Performance of 1 GHz Multiplex with 30 Analog			
Parameter	N+6	N+0 (nom)	N+0 (high)
CCN	48	52	52
CSO	67	70	70
CTB	68	74	73

performance [1].

Table 43 shows the same parameter set and HFC architecture as used in Table 42, but with an analog channel count of 30. Note the significant improvements in analog beat distortions, as well as the SNR (CCN) behavior. Clearly, the added digital distortion that contributes to CCN is mitigated by the improvements obtained by eliminating the cascade effects.

10.4 Importance of the CPE in the DOCSIS NG Migration Plan

We are proposing that DOCSIS NG have a minimum of two (2) PHYs and a common MAC across these independent PHYs. These PHYs will be at least one of the existing DOCSIS 3.0 upstream PHYs and the downstream PHY. In addition there will be a modern PHY. The placement of DOCSIS NG CPEs in the homes that have both DOCSIS 3.0 and DOCSIS NG PHY provides an evolutionary migration strategy.

This will allow the MSO to use the legacy DOCSIS 3.0 PHYs while the cable operator grows the installed base of DOCSIS NG CPEs in their subscriber homes. At such time there are sufficient numbers of DOCSIS NG CPE deployed, the MSO may allocate a few channels to the new DOCSIS NG PHY.

By supporting legacy and modern PHYs within the same CM, the MSOs can smoothly transition to the modern PHY as the legacy CPEs decrease in numbers.

11 RECOMMENDATIONS

This section summarizes the recommendation of the authors. A more extensive explanation of each decision can be found the in the rest of this white paper.

11.1 Areas of Consensus

Compatibility

The recommendation is to define a backwards compatibility goal that would allow the same spectrum to be used for current DOCSIS CMs and new DOCSIS NG CMs.

In this context, co-existence refers to the concept that DOCSIS NG would use separate spectrum but coexist on the same HFC plant. Backwards compatibility would refer to the sharing of spectrum between current DOCSIS and DOCSIS NG.

One example of this strategy would require a 5 to 42 MHz spectrum to be used for four carriers (or more) of DOCSIS 3.0. At the same time, a DOCSIS NG CM would be able to use the same four channels (or more) plus any additional bandwidth that a new PHY might be able to take advantage of.

Upstream Spectrum

The immediate goal with DOCSIS NG is to get as much throughput as possible in the existing upstream 5 to 42 MHz (5 to 65 MHz) spectrum.

This goal recognizes that it will take time, money, and effort to upgrade the HFC plant. The initial goal will to see how more advanced CMTS and CM technology can extend the life of the current HFC plant.

The short-term recommendation for upstream spectrum is mid-split.

Mid-split can be achieved with today's DOCSIS 3.0 technology. If an HFC plant upgrade strategy could be defined that would allow a cost effective two-stage upgrade, first to mid-split, and then later to high-split, then the advantage of higher data rates can be seen sooner.

Conversely, if downstream spectrum is available, an HFC plant could be upgraded to high-split sooner, but would start by deploying mid-split DOCSIS 3.0 equipment.

The long-term recommendation for upstream spectrum is high-split.

High-split offers the best technical solution that should lead to the highest performance product at the best price. The logistical challenges that high-split encounters are not to be underestimated but they are both solvable and manageable, and significantly less imposing than a “top-split” approach.

Downstream Spectrum

The short term goal is to make use of any and all available tools to manage downstream spectrum congestion, such as analog reclamation, SDV, H.264 and deploy 1 GHz plant equipment whenever possible.

This goal includes an expanded upstream spectrum within the current operating spectrum of the HFC plant.

The long-term goal is to utilize spectrum above 1 GHz, and push towards 1.7 GHz.

Field measurements have shown that the spectrum up to 1.2 GHz is available in the passive RF link. Measurements also show that up to 1.7 GHz is available with modest plant intrusiveness. Spectrum above 1 GHz is unspecified, and inherently more challenging than the standard HFC band and thus should take advantage of advanced modulation techniques such as OFDM.

New US PHY Layer

The recommendation for DOCSIS NG upstream is to add OFDMA with an LDPC FEC.

There is considerable new spectrum with DOCSIS NG that only requires a single modulation. Although ATDMA and SCDMA could be extended, now is a unique time to upgrade the DOCSIS PHY to include the best technology available, which the team feels is OFDMA and LDPC FEC.

New DS PHY Layer

The recommendation for DOCSIS NG downstream is to add OFDM with LDPC FEC.

Using the spectrum above 1 GHz will require an advanced PHY such as OFDM. To minimize the cost impact on CMs, a cap could be placed on the number of QAM channels required. OFDM will also be used below 1 GHz, and likely supplant legacy QAM bandwidth over time.

PAPR

We do not anticipate PAPR issues with multicarrier modulation for the upstream or the downstream when compared with single carrier channel bonded DOCSIS.

It is recognized that PAPR for multi-carrier technologies such as OFDM is worse than a single isolated QAM carrier. However, as the number of SC-QAMs in a given spectrum are increased, multiple SC-QAM and OFDM exhibit similar Gaussian characteristics.

Higher Orders of Modulation

The recommendation is to study the option to define up to 4K QAM for OFDM in both the upstream and downstream.

These new modulations may not be usable today. However, as fiber goes deeper coax runs become shorter, and other possible architectural changes are considered (POE home gateway, digital optics with remote PHY), there may be opportunities to use higher orders of modulation. The DOCSIS NG PHY will define these options.

SCDMA Support in a DOCSIS NG CM

The recommendation is to not require SCDMA in a DOCSIS NG CM that employs OFDMA

It is generally agreed that OFDMA with LDPC will be able to replace the role that SCDMA and ATDMA perform today. Thus, in a DOCSIS NG CM, SCDMA would be redundant.

US MAC Layer Baseline

The recommendation is to use the SCDMA MAC functionality as a basis for designing the OFDMA MAC layer.

The SCDMA MAC layer is very similar to the ATDMA MAC layer that has allowed upstream scheduling and QOS services to be near seamless between the two current modulations. This structure is to be extended over OFDM so that the new PHY has a less impact on the rest of the DOCSIS system.

11.2 Areas of Further Study

Some of these decisions require additional information. Some of these decisions have most of the required information and just lack consensus.

High-Split Cross-Over Frequencies

Further study is required to determine the upper frequency of the high-split upstream spectrum and the lower frequency of the downstream spectrum.

At this time, we are not sure the right choice of upstream band edge to achieve 1 Gbps throughput with satisfactory coverage and robustness. This will depend upon the base modulation chosen, FEC overhead, and if there are any areas of spectrum that cannot be used. There will likely be a reference configuration that will pass 1 Gbps and other configurations that will run slower or faster.

There may even be a set of frequencies that matches a 1.0 GHz HFC plant, and a different set of frequencies that matches up to a 1.7 GHz HFC Plant.

There may also be the ability to configure the cross-over frequency in the HFC plant so that it can be changed over time with shifts in traffic patterns. Similar flexibility in the CM could also be considered.

ATDMA in the Upstream

Further study is required to determine how may ATDMA channels a CM and a CMTS should support in the upstream.

Many cable operators are already deploying three full-width carriers or four carriers of mixed widths between 20 MHz and 42 MHz. In order to fully utilize a 5 to 42 MHz spectrum, a DOCSIS NG CM would need to support these channels, so four is the minimum. Newer DOCSIS 3.0 CMs promise 8 upstream channels. It depends upon the market penetration of these CMs as to the impact on backwards compatibility.

Some networks may have migrated to an 85 MHz mid-split before any DOCSIS NG CMs are available, and these would then be A-TDMA channels. Timing of such activity might define minimum channel requirements for the NG CM.

The CMTS may need more QAM channels than the CM. The CMTS needs to have a spare ATDMA channel to support DSG. It also needs to have an ATMDA channel running at a lower rate to support DOCSIS 1.1 CMs. These may be in addition to the 3-4 channels for DOCSIS 3.0.

SCDMA in the CMTS

Further study is required to determine if SCDMA should be retained.

It is generally agreed that SCDMA does offer better performance below 20 MHz (in North America, higher in other countries with worse plant) than ATDMA. For DOCSIS 3.0, SCDMA may be required to get that extra fourth full-size carrier, and is an important component for maximizing the throughput available in 5-42 MHz band.

Retaining SCDMA in addition to ATDMA and OFDMA potentially adds product cost, development cost, and testing cost. This has to be weighed against any significant market penetration of SCDMA prior to DOCSIS NG being available.

One possible approach is to specify a small number of channels of SCDMA as mandatory and more channels optional. However, an overall objective is to try and get to only one or two PHY technologies in the CMTS silicon that would imply the elimination of SCDMA.

Early deployment of mid-split would also help negate the need for SCDMA, as that would provide the extra spectrum to relieve the congestion in 5-42 MHz

Advanced FEC for Single Carrier Systems

Further study is required to determine if LDPC FEC functionality should be added to enhance the existing upstream and downstream PHY.

The argument for doing this is that the bulk of new capacity comes from advanced FEC, and existing SC QAM that co-exists on the silicon should benefit from this investment to optimize efficiency in systems that will be operating single carrier mode for many more years. The argument for not doing this is to cap the legacy design and only expand capability with OFDM.

Expansion of Upstream ATDMA Capabilities

Further study is required to determine if ATDMA functionality should be extended with wider channels, more channels, higher order modulation formats, and improved alpha.

The argument for doing this work is that they represent simple extensions of DOCSIS 3.0, and field experience and RF characterization of A-TDMA tools suggests a high probability of success. The argument for not doing this is to cap the legacy design and only expand capability with OFDM, and that an OFDM implementation would be less complex.

Expansion of Downstream QAM Capabilities

Further study is required to determine if downstream QAM functionality, currently defined by ITU-T J.83, should be extended with wider channels and higher order modulation formats.

The argument for doing this work is that they represent simple extensions of DOCSIS 3.0 and field experience and characterization of A-TDMA SC tools suggests a high probability of success. The argument for not doing this is to cap the legacy design and focus on expanding capability only with OFDM, and that an OFDM implementation would be less complex.

US MAC Improvements

Further study is required to determine if any changes not directly related to OFDM are worth pursuing.

Current suggestions include changing the request mechanism from request-based to queue-based, elimination of 16-bit minislots, and not including request slots on each upstream carrier.

Modifications need to be weighed against increases in performance, decrease in cost, and the need for backwards compatibility.

12 ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of others.

From Cisco, the authors would like to thanks Ron Hranac, Fred Hirt, John Ritchie, and Bill Wall for worrying us about the aeronautical band and GPS frequencies.

From Arris, the authors would like to thank Scott Shupe, Dr. Derald O. Cummings, Dr. Tom Cloonan, Dr. Frank O'Keeffe, Dr. Zoran Maricevic, Dr. Marcel Schemmann, Brent Arnold, Bill Dawson, Venk Mutalik, Jeff Howe, and Dr. Ayham Al-Banna for their contributions.

From Motorola, the authors would like to thank Jack Moran, Rob Thompson, John Ulm, Gerry White, Dr. Amarildo Vieira, Mark Schmidt, Mike Aviles, Steve Frederick, Phil Miguelez, Dean Stoneback, Vipul Rathod, Chuck Moore, Mike Cooper, John Holobinko, and Dr. Sebnem Ozer.

From Intel, the authors would like to thank Dr. Bernard Arambepola who provided the OFDM simulation results.

13 REFERENCES

[DOCSIS 2.0]
CM-SP-RFIv2.0-C01-081104, DOCSIS Radio Frequency Interface Specification, CableLabs, August 11, 2004

[DOCSIS DRFI]
CM-SP-DRFI-I11-110210, DOCSIS Downstream RF Interface Specification, CableLabs, issued Feb 10, 2011

[DOCSIS MACUP]
CM-SP-MULPIv3.0-I15-110210, DOCSIS 3.0 MAC and Upper Layer Protocols Interface Specification, CableLabs, Issued Feb 2, 2011

[1] Dr. Robert L., Michael Aviles, and Amarildo Vieira, "New Megabits, Same Megahertz: Plant Evolution Dividends," 2009 Cable Show, Washington, DC, March 30-April 1

[2] White, Gerry and Mark Schmidt, 64-, 256-, and 1024-QAM with LDPC Coding for Downstream, DOCSIS 3.0 Proposal, Motorola Broadband Communications Sector, Dec 10, 2004.

[3] Howald, R., Stoneback, D., Brophy, T. and Sniezko, O., Distortion Beat Characterization and the Impact on QAM BER Performance, NCTA Show, Chicago, Illinois, June 13-16, 1999.

[4] Prodan, Dr. Richard, "High Return Field Test: Broadcom Frequency Response Model vs. Motorola Data," Report to CableLabs AMP Committee, April 11, 2011.

[5] Moran, Jack, "CableLabs High Frequency Return Path Passive Coaxial Cable Characterization Results," AMP Report, Nov 18, 2011

[6] Howald, Dr. Robert L, Fueling the Coaxial Last Mile, SCTE Conference on Emerging Technologies, Washington DC, April 2, 2009.

[7] Finkelstein, Jeff, "Upstream Bandwidth Futures," 2010 SCTE Cable-Tec Expo, New Orleans, LA, Oct. 20-22.

[8] Howald, Dr. Robert L., Phillip Chang, Robert Thompson, Charles Moore, Dean Stoneback, and Vipul Rathod, "Characterizing and Aligning the HFC Return Path for Successful DOCSIS 3.0 Rollouts," 2009 SCTE Cable-Tec Expo, Denver, CO, Oct 28-30.

[9] Howald, Dr. Robert L., "Maximizing the Upstream: The Power of S-CDMA" Communication Technology Webcast, Sept. 9, 2009.

[10] Howald, Dr. Robert L. and Michael Aviles, "Noise Power Ratio the Analytical Way," 2000 NCTA Show, New Orleans, LA.

[11] Miguelez, Phil, and Dr. Robert Howald, "Digital Due Diligence for the Upstream Toolbox," 2011 Cable Show, Chicago, IL, June 14-16.

[12] Dr. Robert Howald and Phil Miguelez, "Upstream 3.0: Cable's Response to Web 2.0," The Cable Show Spring Technical Forum, June 14-16, 2011, Chicago, Il..

[13] Stoneback, Dean, Dr. Robert L. Howald, Joseph Glaab, Matt Waight, "Cable Modems in the Home Environment," 1998 NCTA Show, Atlanta, Ga.

[14] Ulm, John, Jack Moran, Daniel Howard, "Leveraging S-CDMA for Cost Efficient Upstream Capacity," SCTE Conference on Emerging Technologies, Washington DC, April 2, 2009.

[15] Thompson, Robert, Jack Moran, Marc Morrissette, Charles Moore, Mike Cooper, Dr. Robert L. Howald, and "64-QAM, 6.4MHz Upstream Deployment Challenges," 2011 SCTE Canadian Summit, Toronto, ON, Mar 8-9.

[16] Thompson, Rob, "256-QAM for Upstream HFC", NCTA Cable Show, Los Angeles, CA, May 2010

[17] Thompson, Rob, Jack Moran, Marc Morrissette, Chuck Moore, and Dr. Robert Howald, "256-QAM for Upstream HFC Part II", The Cable Show, Atlanta, Ga, NOV 2012

[18] Woundy, Richard, Yiu Lee, Anthony Veiga, Carl Williams, "Congestion Sensitivity of Real-Time Residential Internet Applications", 2010 SCTE Cable-Tec Expo, New Orleans, LA, Oct. 20-22.

[19] North American Cable Television Frequencies, Wikipedia, http://en.wikipedia.org/wiki/North_American_cable_television_frequencies

[20] CEA-542-C, Cable Television Channel Identification Plan, Consumer Electronics Association, Feb 2009

[21] Chapman, John T, "Taking the DOCSIS Upstream to a Gigabit per Second", NCTA Spring Technical Seminar, Los Angeles, May, 2010.

[22] Chapman, John T., "What the Future Holds for DOCSIS", Keynote speech, Light Reading Conference, Denver, May 18, 2012

[23] Department of Commerce, "United States Radio Spectrum Frequency Allocations Chart", 2003, United States Department of Commerce, http://www.ntia.doc.gov/osmhome/allochrt.pdf

[24] "Capture Effect", Wikipedia, http://en.wikipedia.org/wiki/Capture_effect

[25] SCTE 40 2011, "Digital Cable Network Interface Standard", Society of Cable Telecommunications Engineers, Inc, 2011

[26] "Instrument Landing System", Wikipedia, http://en.wikipedia.org/wiki/Instrument_landing_system

[27] David J.C. MacKay and Edward A. Ratzer, "Gallager Codes for High Rate Applications" published January 7, 2003,

[28] ETSI EN 302 769: "Digital Video Broadcasting; Frame Structure, Channel Coding and Modulation for a Second Generation Digital Transmission System for Cable Systems"

[29] United States Nuclear Detonation Detection System (USNDS), http://www.fas.org/spp/military/program/nssrm/initiatives/usnds.htm

[30] Committee on Radio Astronomy Frequencies, http://www.craf.eu/gps.htm

[31] Grace Xingxin Gao, "Modernization Milestone: Observing the First GPS Satellite with an L5 Payload", Inside GNSS Magazine, May/June 2009, www.insidegnss.com/auto/mayjune09-gao.pdf

[32] Dennis M. Akos, Alexandru Ene, Jonas Thor, "A Prototyping Platform for Multi-Frequency GNSS Receivers", Standford University

[33] Emmendorfer, Michael J, Shupe, Scott, Maricevic, Zoran, and Cloonan, Tom, "Next Generation – Gigabit Coaxial Access Network" 2010 NCTA Cable Show, Los Angeles, CA, May 2010

[34] Emmendorfer, Michael J, Shupe, Scott, Cummings Derald, and Cloonan, Tom, "Next Generation - Cable Access Network, An Examination of the Drivers, Network Options, and Migration Strategies for the All-IP Next Generation – Cable Access Network", 2011 Spring Technical Forum, Chicago, IL June 14-16

[35] Emmendorfer, Michael J, Shupe, Scott, Cummings Derald, Cloonan, Tom, and O'Keeffe, Frank "Next Generation - Cable Access Network (NG-CAN), Examination of the Business Drivers and Network Approaches to Enable a Multi-Gigabit Downstream and Gigabit Upstream DOCSIS Service over Coaxial Networks", 2012 SCTE Canadian Summit March 27-28, Toronto, Canada.

[36] Emmendorfer, Michael J, Shupe, Scott, Maricevic, Zoran, and Cloonan, Tom, "Examining HFC and DFC (Digital Fiber Coax) Access Architectures, An examination of the All-IP Next Generation Cable Access Network," 2011 SCTE Cable-Tec Expo, New Atlanta, GA, Nov. 14-17.

THE GROWN-UP POTENTIAL OF A TEENAGE PHY

Dr. Robert Howald, Robert Thompson, Dr. Amarildo Vieira
Motorola Mobility

Abstract

Cable operators continue to see persistent annual increases in downstream traffic, most recently driven by aggressive growth in over-the-top video. Well-understood tools exist to manage the growth, including switched digital video (SDV), analog reclamation, service group splitting, improved encoding and transport efficiency, and RF bandwidth expansion. Beneath it all, however, the underlying RF transport approach has remained unchanged, relying on 1990's era ITU J.83 technology for PHY layer and FEC technology. Meanwhile, 15+ years of advancements in communications technology and processing power have since taken place. Many of these advances, which close the gap between Shannon theory and real-world implementation, are already being tapped in other industries. Some are now poised to enable cable to support a new generation of Gbps-class services and to mine completely the capacity of the coaxial last mile – a key element to guaranteeing an enduring HFC lifespan.

In this paper, we will present a comprehensive link analysis addressing the deployment possibilities of these communications technology advances over the HFC channel. We will focus in particular on the ability to support higher order QAM, such as 1024-QAM through 4096-QAM. We will discuss the role multi-carrier techniques (OFDM) could play and why. We will specify SNR implications to HFC, including considerations for fiber deep migration. We will describe the SNR repercussions of advanced QAM to CPE noise figure (NF), which is critical to understand as wideband, digitizing front ends replace analog STB tuners. In addition, we will dive deeper into the subtle link impairments that become potentially limiting factors as we push the boundaries of PHY technology on the cable plant. Previously less significant issues such as timing jitter and phase noise are magnified as constellations become increasingly dense. These items ultimately effect equipment requirements.

In summary, we will articulate and quantify the ability of the HFC network to support ever-increasing orders of bandwidth efficient modulation, and the impact these modern communications formats have on equipment and requirements.

INTRODUCTION

The industry is deeply engaged in long-term network planning, in recognition of the continuing growth of IP traffic and concern for the network's ability to support it. There are two key components to the problem. The first is simply determining if the infrastructure in place is physically capable of delivering on the growth, and, if so, for how long. There tends to be a consensus that the HFC architecture is capable enough, but that it has not been optimized as of today to ensure it [8, 11]. This brings us to the second part. If the answer to the first is yes, then how do initiate a transition plan from today's infrastructure to the architecture that does optimize what can be extracted from the network?

There are many spectrum and capacity management tools in the downstream. However, the operator has much less control over upstream congestion. Common to both downstream and upstream is the reliance on what is now aging, 1990's era, physical layer

(PHY) tools. In use on the downstream is the ITU J.83B Physical layer (PHY) and forward error correction (FEC) technology. The DOCSIS upstream is also QAM with a Reed-Solomon based FEC – powerful at the time but more powerful PHY techniques are available today. The result is less efficient use of cable spectrum that could be achieved with modern PHY tools.

Relying on 1990's era technology puts cable at a disadvantage. Perhaps the single most important long-term objective of the architecture transition is, in the end, to have created maximum bandwidth efficiency (and therefore maximum lifespan) cost-effectively. We will focus our attention on this one component of capacity management – more efficient use of spectrum – more bits-per-second-per-Hz (bps/Hz).

Capacity Levers

Note that theoretical capacity is a based on two variables – bandwidth (spectrum allocated in our case), and SNR. For high enough SNR, finding spectrum dominates the equation. The capacity equation can be simplified, and in so doing, it can be shown that capacity is essentially directly proportional to bandwidth, B and SNR expressed in decibels (dB):

$$C \approx [B]\,[SNR\ (dB)]\ /\ 3 \qquad (1)$$

This approximation is accurate asymptotically within 0.34% with increasing SNR. Because of the inescapable relationship of capacity to bandwidth, as cable looks to increase capacity, new spectrum is being sought after. Figure 1 is an example of a likely spectrum evolution [2], resulting in a final state of bandwidth allocations.

In the downstream, we are extending an excellent channel into an area where it will suffer more attenuation as a minimum. It will also likely have to deal with frequency response issues.

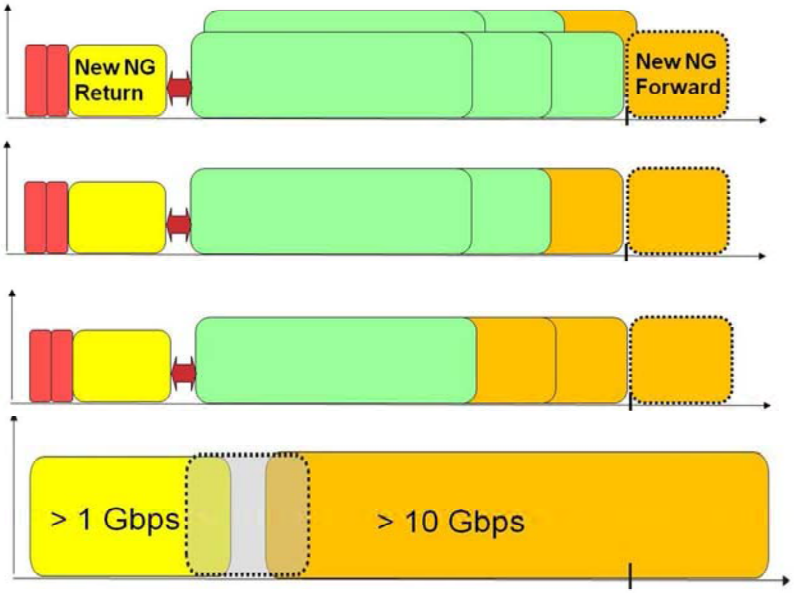


Figure 1 – Likely Cable Spectrum Evolution

In the upstream, we will be in some ways doing the opposite – extending a partially troubled channel into an area where we expect a much better behaved environment from which to extract new capacity.

We will be taking advantage of significant technology advances for enhancing the PHY. Much of what is being taken advantage of is continued advances in the real-time computing power of FPGAs and ASICs. The theoretical basis for modern PHY tools in some cases is, in fact, very old.

For example, Reed-Solomon coding itself was born in the 1959. Low Density Parity Check Codes (LDPC), the basis of today's most advanced forward error correction (FEC), is also quite old, first introduced in 1960. Information Theory is a linear algebraic discipline. However, the computing power to perform the algebraic operations required for efficient decoding of very large matrices, and using non-binary arithmetic (Reed-Solomon) came along much later.

Multi-carrier modulation (MCM, the generic name for OFDM and its variants – we will use both throughout) has a parallel history to advanced FEC in this sense. It was very difficult to implement, until the FFT version of the DFT came along, followed by computing power to calculate larger FFTs faster. IFFT/FFT algorithms form the core of the OFDM transmit and receive function. So, while we are talking about "new" technology, it is important to understand that these technologies are already very well grounded theoretically.

Of course, the single most important attribute of these advances is that they close the gap between theoretical Shannon capacity and real-world implementation – something the world has been trying to do since 1948. A simple crunching of today's HFC performance and throughput illustrates how far from this ideal we are today. Table 1 compares downstream and upstream as we use them today against the theoretical capabilities of the channel. We have accounted for code rate efficiency losses, but not framing, preamble, or other overhead unrelated to pure PHY channel transmission capacity.

Table 1 – Downstream and Upstream vs. Theory

	D/S	U/S
BW (MHz)	6	6.4
SNR (dB)	35	25
Capacity from (1)	70.00	53.33
Legacy QAM (no framing)		(t=10)
256-QAM	38.83	37.75
64-QAM	26.99	28.31
Delta Capacity		
256-QAM	**55%**	**71%**
64-QAM	**39%**	**53%**

As we can see in Table 1, today's commonly used modes – 256-QAM downstream and 64-QAM upstream – operate at 50-60% of capacity, and therefore are leaving a lot of bits on the table. With the help of new tools and supporting architecture evolution, some of the current limitations can be alleviated, putting cable in a position to deliver a new class of services and maximize the lifespan of its core architecture.

QAM LINK BUDGETS

Capacity Enhancements

Let's begin with the "SNR" part of (1). There are two elements of the SNR component of capacity. First, clearly, more capacity is available if higher SNR is available. Since it is related to SNR in dB, however, it is a compressing function, and its affect on capacity less effective in increasing C compared to spectrum. In practice 50% more spectrum yields 50% more SNR. This is also true for 50% more SNR in dB. However, in practice, 50% new SNR in dB, such as converting a 35 dB SNR into a 52.5 dB SNR, is not reasonable in most cases. Nonetheless, it is certainly the case that more SNR translates to more capacity, and architectures that create higher SNR are architectures that open up more potential capacity. This is why, for example, when fiber deep topologies are discussed, both average bandwidth per home (because of fewer homes per node) as well as a more robust, higher SNR channel for use are both important results. We will quantify architecture effects later in this section.

Part 2 of the SNR component of capacity is using the available SNR most efficiently. This is specifically where MCM and FEC advances come into play. A good way to understand the former is to use the “long” (but not longest!) form of (1).

$$C \approx (1/3)\sum_{\Delta f} [\Delta f]\ [P(\Delta f)\ H(\Delta f)\ /\ N(\Delta f)]_{dB} \quad (2)$$

This is the same information expressed in (1), just differently. Instead of bandwidth, we have used a summation over a set of small frequency increments, Δf. The sum of all Δf increments is the bandwidth available, B. Instead of SNR, we have identified the components of SNR – signal power (P), noise (N), and channel response (H) – each also over small frequency increments.

The capacity, then, is a summation of the individual capacities of chunks of spectrum. The purpose of (2) is to recognize that channels with changing SNR – such as any “new” bands to be exploited outside the normal cable bands – that may not have a flat response. In particular, above today’s forward band there will be roll-off with frequency. The capacity of this region can be calculated by looking at it in small chunks that approximate flat channels. More importantly, however, a technology that can actually *implement* small channel chunks can optimize each of those frequency increments to get the most capacity from them. This is the key advantage of MCM – very narrow channels, each of which can be loaded with the most bits possible. With a single, wide, transmission, it is difficult to achieve the same effect without very complex, and sometimes impractical equalization techniques and interference mitigation mechanisms. Thus, (2) effectively expresses why MCM is often better suited in channels with poor frequency response. MCM also accomplished this while overlapping these narrow channels. They are kept the independent through the orthogonality of frequency spacing – separated by the symbol rate of the sub-channels.

Figure 2 shows a capacity calculation of the forward band extension case described above. It plots the capacity including bandwidth above a 1 GHz network when the network is modeled as a lowpass roll-off, governed by the frequency response characteristics of 1 GHz taps [7]. The red curve shows the aggregate roll-off of five taps and a single coupled port, as well as interconnecting coaxial cable. An assumed 45 dB digital SNR at 1 GHz is used to calculate the capacity as signal power is attenuated above 1 GHz on a flat transmission profile.

The total capacity calculation if the entire forward band is taken into account (blue) is also shown, as well as the capacity over and above what is currently available in a 1 GHz network that is fully loaded with 256-QAM signals (pink). In both cases, the diminishing returns associated with the attenuation of current HFC passives – inherent implementation limitations, not barriers of technology – are obvious as the forward band goes above about 1.2 GHz. Analyses like Figure 2 point out why cable is bullish on the ability of the HFC network to support 10 Gbps data rates.

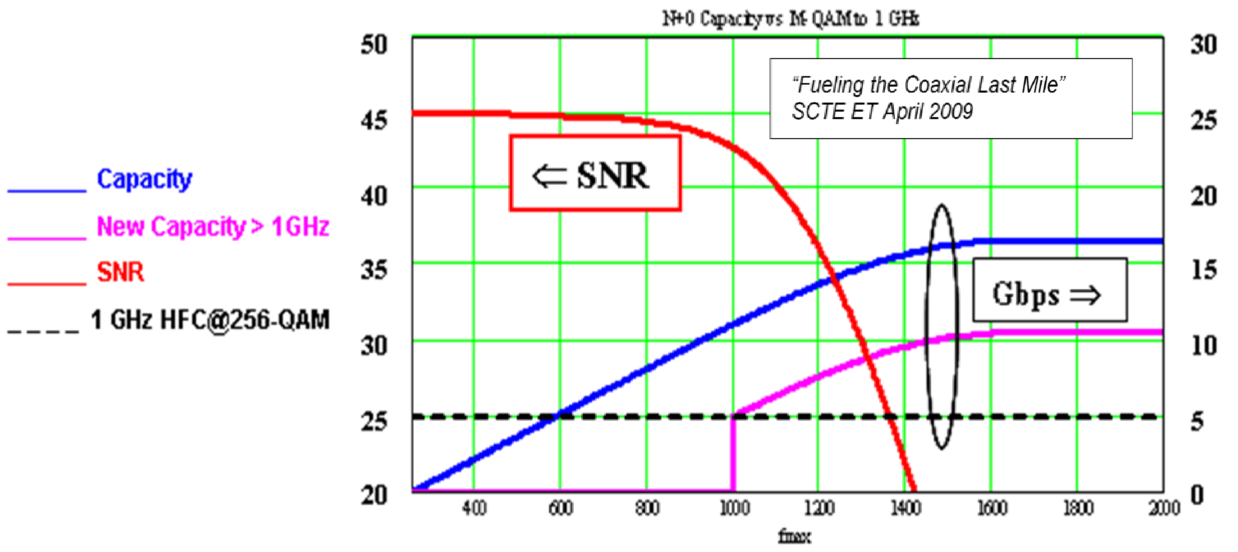


Figure 2 – Capacity Above 1 GHz on a Passive Coaxial Segment

Now let's consider the role of FEC. Using SNR most efficiently, from a coding perspective, is about finding the right codeword design. Major leaps in this capability have occurred in the past 20 years – Reed-Solomon (RS), Trellis Coded Modulation (TCM), Turbo Codes, and now LDPC. Again, the advances have mostly to do with the ability to process the complex decoding algorithms and the tools needed to design them specific to the application.

Coding theory has always been about trying to close the gap between the theory derived by Shannon and the reality on the wire or in the air. Quantifiably, this means simply getting more bps/Hz out of the same or lower SNR.

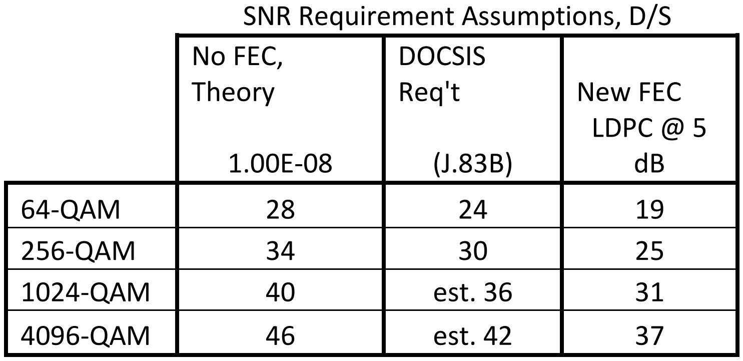
Defining SNR Thresholds

The impact on of advances in FEC is quite simple – it reduces the SNR required to achieve a particular modulation profile, increasing throughput. The SNR requirements for each QAM modulation profile without coding are theoretically well-founded. We will consider advanced FEC as we compare results of link analysis to determine what can be supported by a particular PHY and HFC architecture SNR, and to compare that to today's architecture and requirements. The thresholds that will govern the comparisons are shown in Table 2.

The three M-QAM BER columns are as follows:

1) 1e-8, No FEC
2) DOCSIS Specification (and extended estimates where QAM profile does not exist)
3) New LDPC-based FEC; assumption of 5 dB more gain

Table 2 - Downstream SNR Assumptions for M-QAM Profiles



SNR Requirement Assumptions, D/S

	No FEC, Theory 1.00E-08	DOCSIS Req't (J.83B)	New FEC LDPC @ 5 dB
64-QAM	28	24	19
256-QAM	34	30	25
1024-QAM	40	est. 36	31
4096-QAM	46	est. 42	37

Several important items must be noted with respect to Table 2.

- DOCSIS includes an allocation for implementation margin on top of an assumed coding gain impact. We are inherently carrying those implementation margins forward by using an LDPC gain factor and not an LDPC SNR versus QAM simulation.

- Coding gain may increase as M-QAM orders increase, but it is conversely more difficult to maintain a constant implementation loss across higher profiles. By using 5 dB, we are essentially calling these a wash. No effort has gone into infrastructure requirements to support, for example 4096-QAM, and hardware limitation can become exaggerated for these cases.

- There are no code designs selected using LDPC for North American cable. In Europe, DVB-C2, for example, has defined a range of code rates, and these are SNR reference points reported of the OFDM PHY + LDPC:

  256-QAM: 22-24 dB
  1024-QAM: 27-29 dB
  4096-QAM: 32-35 dB

  These numbers similarly require margin be applied in practice, but are nonetheless useful in understanding how efficient a network can be as other non-idealities of implementation are reduced.

- These are all AWGN-only SNR values, which is the fundamental construct of channel capacity.

- These are SNR thresholds for the downstream only. The first column, of course, is independent of downstream or upstream.

So, while the dB to use for a given profile can be debated, Table 2 gives us a ballpark starting point.

We similarly set thresholds to use for the upstream, which also includes margin for operations. Because of the variety of unknowns, the margin allotment is higher. Note that the upstream is not ITU J.83B, although it is Reed-Solomon-based with configurable error correction parameter. Thus, the RS code can be stronger or weaker, depending on configuration, although it is usually set at lower code rates (stronger).

As reference guidelines for upstream SNR, we choose what one particular operator uses as classification for a good upstream in terms of observable metrics. These metrics are based in part on upstream SNR (actually MER) reported. A good score for the upstream includes a consistent 30 dB reported SNR. We will assume this would represent a link capable of the highest order modulation profile at all times, which is 64-QAM today. This is what is reflected in Table 3.

Note that this threshold is actually above the no-FEC threshold for 64-QAM. This is simply the nature of the margin allotted to upstream as operated today. While the RS code offers a theoretical gain similar to the RS downstream, it is configurable from none up to strong. Also, the upstream channel has, in general, been difficult to fully exploit to date because of the range of impairments and field implementations. It is certainly reasonable to expect that, as service group sizes shrink, alignment practices improve as bonding becomes prominent, and other architectural changes take place (like a point-of-entry home gateway architecture to be discussed), the quality of the upstream will improve and the margin required to support a particular modulation profile reduced. We do not make any of assumptions about any new dB associated with those "what ifs" here.

Upstream traffic typically comes in small chunks, and therefore can only be supported by smaller block-sized LDPC codes. This leads to less coding gain for a given code rate compared to downstream. We round this difference up to an even 1 dB offset compared to downstream, or 4 dB of new upstream gain from LDPC. For upstream, then we use the assumptions shown in Table 3.

We will go forth with these SNR values as we investigate the implications to key components of the HFC architecture. Note that whether we are discussing legacy single carrier QAM or MCM systems, the SNR thresholds established in these tables are the same. These are based on AWGN performance. Thus, for both legacy QAM style and MCM, the link budget analysis below is applicable. However, it can be argued, in particular for the upstream, that use of MCM offers the opportunity to eliminate some of the margin currently allotted in Table 3, since this is based on experience with today's single carrier upstream channels. The reasoning here is that multi-carrier could be more resilient to some of the things that go into setting the upstream margin.

Table 3 - Upstream SNR Assumptions for M-QAM Profiles

SNR Requirement Assumptions, U/S

	No FEC, Theory 1.00E-08	DOCSIS w Upstream Margin Reed-Solomon	New FEC LDPC @ 4 dB
64-QAM	28	30	26
256-QAM	34	36	32
1024-QAM	40	est. 42	38
4096-QAM	46		

DOWNSTREAM HFC MIGRATION

Table 4 quantifies the delivered performance at the end of line for an HFC network based on a classic 1310 nm linear optical link as a function a modern RF cascade, such as GaAs-based RF. It is a typical mix of bridger (multi-port) amplifiers and line extenders where a cascade ensues. Table 4 includes an assumption of partial analog reclamation – a total of 30 analog carriers remain. Most MSOs have an analog reclamation plan, though many anticipate that they will leave a basic tier in place for a long time. We will analyze a full analog reclamation case as well.

Table 4 - Downstream Performance vs Cascade

	1 GHz, 30 Analog Carriers				
	CNR	CSO	CTB	CCN	QAM CCN
N+6	51	60	65	49	43
N+3	55	62	67	52	46
N+0	57	65	69	55	49

These numbers all relate to analog levels, of course, so in reference to digital they must be lowered by the amount of the digital de-rating. Mathematically, it is straightforward to show [7] that removal of analog frees up some RF power from the total load that could be re-allocated to digital loading. This varies from approximately 1-3 dB, depending on how many analog carriers remain (zero or 30) and the tilt used from RF amplifiers on the coaxial leg. We will not account for new RF power at this stage, but come back to this point as we discuss link budget closure. QAM SNR levels – 6 dB lower than the yellow column – are listed on the far right of Table 4 using a 6 dB de-rate. An important and expected result from Table 4 is the improvement in the CNR and CCN as the cascade becomes shorter.

Note that CCN stands for Composite Carrier-to-Noise, and captures all noise floor components – AWGN and digital distortions. It is the “true” SNR, although that is an imperfect label technically because of the contributions of distortion. However, the digital distortion contributors are many and largely independent, so a Gaussian assumption is reasonable. On a 6 MHz channel, a white, Gaussian assumption is also reasonable. For wideband channels, the “white” component may deviate, but this is exactly where MCM plays a role. By its nature, it again will make the channel noise, including CCN, “look” white (flat) in a sub-channel.

Using Table 2 requirements and Table 4 performance, and assuming the lower limit of input power is assumed delivered for a QAM channel at -6 dBmV, we can derive what noise performance is needed from the CPE to meet each threshold. This is shown in Figure 3. We have extended the CCN range downward on Figure 3 compared to Table 4 to represent perhaps deeper cascaded architectures than N+6, or systems running somewhat stretched or simply below the performance of Table 4 for a variety of other design reasons.

Also shown in Figure 3 are the above calculated CCN values of 43 dB (N+6), 46 dB (N+3) and 49 dB (N+0), identified and labeled using red vertical lines. Along with the maximum noise figures plotted in Figure 3 are noise figure values (black dashed lines) representative of common CPE platforms in the field, and of today’s vintage, which are lower noise designs. In the case of the “maximum noise figure” curves (color), QAM profiles are supported when the colored line identifying a modulation profile is *above* an example CPE NF threshold in black.

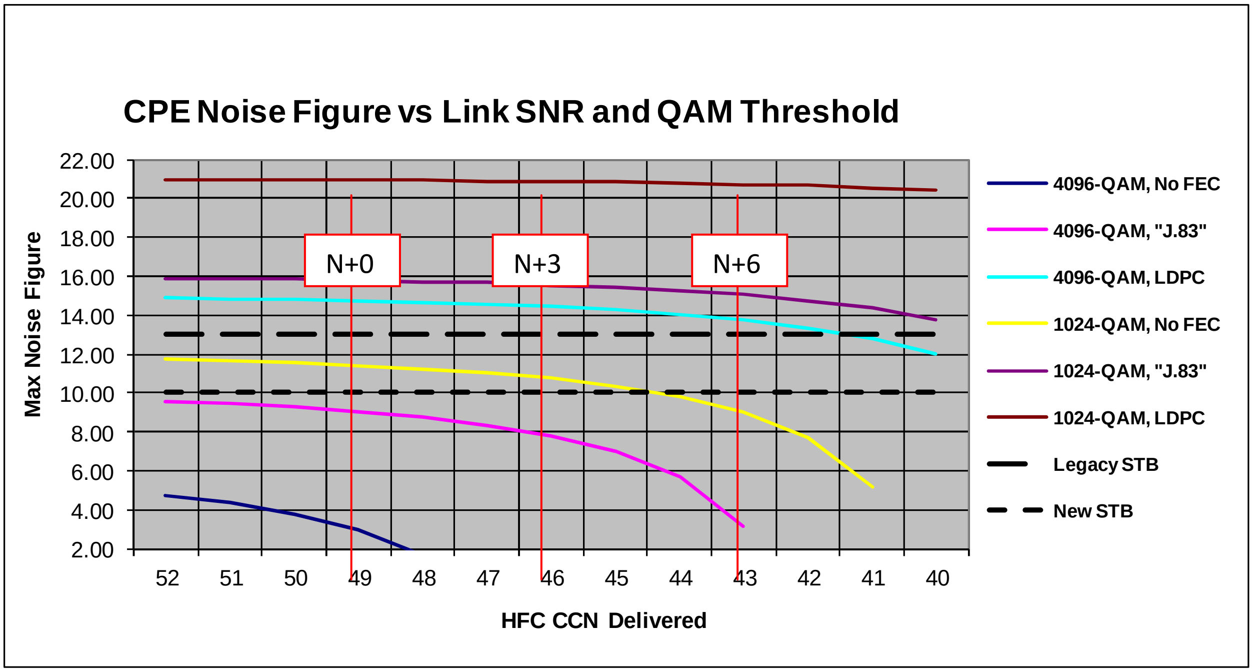


Figure 3 – STB Noise Figure Limit vs. Modulation Efficiency

Apparent from Figure 3 are three things with respect to the access network and home environment:

1) 4096-QAM is not achievable without introducing new FEC, based on today's linear optics and CPE performance. Even with new LDPC FEC, however, it is too marginal to be practical without sensitivity improvements of modern STBs. And, even with those improvements, there is just a little link budget to spare, and only if there is a fiber deep migration. Stretched architectures and high in-home losses could struggle – a QAM input below -10 dBmV to the STB instead of -6 dBmV would be insufficient for N+6, for example. Remember those possible dBs of power allocation gains of analog reclamation we identified earlier? It is cases like this where it becomes obvious that every dB of a link budget becomes critical in some cases for practical margins to be realized.

2) 1024-QAM with a J.83 flavor of FEC is achievable today with legacy STB performance, albeit there is also not much margin. For example, while 256-QAM performance requirements exist down to a -15 dBmV input in DOCSIS, this additional loss would not be able to be absorbed in the 1024-QAM link budget per Figure 3. On the other hand, 1024-QAM with LDPC is the one curve that is clearly and robustly supported – to levels as low as -13 dBmV for even existing STBs and below -15 dBmV for newer class boxes.

3) For HFC migrated to fiber deep architectures such as N+0 or N+(small) – left hand side of Figure 3 – there is little sensitivity of the NF curve to SNR variations for all cases except for a threshold using 4096-QAM without FEC, which is a non-starter.

The fact that 1024-QAM using J.83 is possible is consistent with the conclusions drawn in [9]. That analysis also pointed at the CPE noise as a potential link limiter to 1024-QAM today.

All in all from an SNR standpoint, though, updating the FEC will be instrumental to delivering higher modulation efficiency on today's quality of HFC architectures, and newer CPE will buy important link headroom to enable robustness.

Figure 3 captures access network and home. A missing component of this analysis is that Figure 3 inherently assumes a perfect transmit fidelity. In fact, of course, the DRFI specification governs transmit fidelity today. If we assume that plant linear distortions are properly handled in the receive equalizer (a good assumption), then we can consider the DRFI Equalized MER contribution of 43 dB. There is an implicit assumption that the MER is not dominated by a few discrete spurious components when we are using an SNR analysis. We will consider discrete interference in a separate section. Figure 4 shows the result with the DRFI requirement included.

While there are major differences in Figure 3 and Figure 4, the conclusions previously drawn do not vary very significantly. 4096-QAM is just more impractical than before, and the 1024-QAM "J.83" case has lost some of its margin, but remains on the bubble of workability as long as the HFC link is very good, such as the N+0 case.

Lastly, we point out that while we have captured the DRFI MER requirement in Figure 4, that requirement is obviously a minimum. Although broadband fidelity requirements are among the most difficult to meet, suppliers compete on key parameters and therefore product performance may be better in practice.

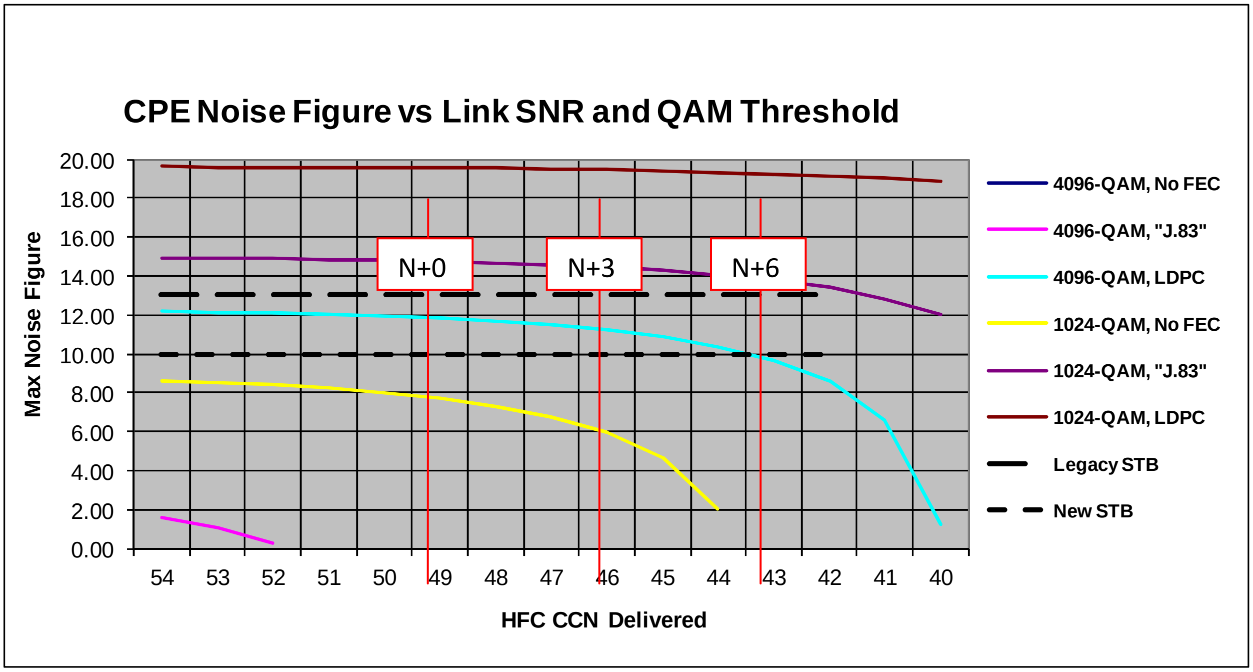


Figure 4 – STB Noise Figure Limit vs. Modulation Efficiency, DRFI MER

Multi-Wavelength Optics

Let's update the architecture now to include full analog reclamation and wavelength division multiplexed (WDM)-based linear optics commonly implemented today in multi-service fiber distribution architectures. WDM tools are becoming very valuable as operators take on Ethernet and EPON-based business services, avoid pulling new fiber where possible, and consider more consolidation of hub locations. With downstream loads moving to more QAM carriage and away from analog, the optical nonlinearities that make heavy analog loads difficult to manage over WDM become less imposing, and reaches can be extended.

Table 5(a-c) show three cases: 750 MHz, 870 MHz, and 1 GHz. In each case, a 1550 nm, ITU-grid-based, Analog/QAM transmitter is shown under nominal link conditions. Performance is also calculated with an 85 MHz upstream mid-split (slightly reduced forward load). This is a likely upstream evolution path for operators looking to exploit the full capabilities of DOCSIS 3.0 while minimizing the imposition on the downstream spectrum. The range of CCN's in Table 5 is within the ranges shown in Figure 3 and Figure 4, so these results are entirely applicable to the cases shown originally using Table 4.

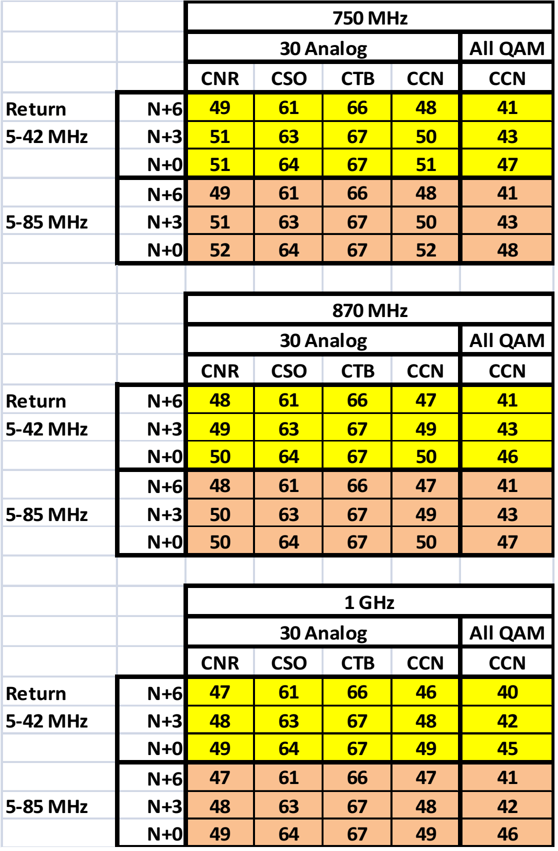
Compared to classic single wavelength 1310 nm delivery, advanced optics such as these are being implemented more often, and across a variety of service scenarios, so extensive effort has gone into characterizing them across load and band variations. The extended calculations of Table 5 offer a few interesting conclusions with respect to variations in performance.

1) As the architecture shortens to N+0, the CCN going from 30 analog to zero analog improves. Accounting for the QAM-relative 6 dB de-rate with analog, for example, would leave 45 dB of QAM CCN for 750 MHz and N+0 (5-42 MHz system), instead is 47 dB. This is indicative of the effect of the analog carriers, which are higher levels than the QAM, to have a major impact on the distortion mix because of the difference in channel power. As a result, without taking advantage of any load, a couple extra dB are available. This does not included any additional dB that may be available from allocating more per-QAM power as the analog load is removed. As previously discussed, there is potential for another 2-2.5 dB on the optical link (flat loading) available that would keep the total power load the same. The tilted RF link would see less benefit.

2) The 85 MHz architecture has minimal impact on QAM CCN (0-1 dB). It removes a small chunk of forward bandwidth, but not enough to have a measurable impact. This may change at 200 MHz or more of upstream bandwidth. That case is not shown here, however, as in that case we would also expect an extended forward band. The net result should benefit CCN, since sliding the entire band results in fewer octaves of coverage, and the number octaves is important for broadband RF distortion characteristics.

3) 750 MHz systems have a 1-2 dB of SNR compared to 1 GHz systems. This is a worthwhile amount of dB gained, but perhaps not a good tradeoff relative to the capacity lost for not having the spectrum available.

Table 5 – Performance vs. Architecture a) 750 MHz, b) 870 MHz, c) 1 GHz



		750 MHz				
		30 Analog				All QAM
		CNR	CSO	CTB	CCN	CCN
Return	N+6	49	61	66	48	41
5-42 MHz	N+3	51	63	67	50	43
	N+0	51	64	67	51	47
	N+6	49	61	66	48	41
5-85 MHz	N+3	51	63	67	50	43
	N+0	52	64	67	52	48

		870 MHz				
		30 Analog				All QAM
		CNR	CSO	CTB	CCN	CCN
Return	N+6	48	61	66	47	41
5-42 MHz	N+3	49	63	67	49	43
	N+0	50	64	67	50	46
	N+6	48	61	66	47	41
5-85 MHz	N+3	50	63	67	49	43
	N+0	50	64	67	50	47

		1 GHz				
		30 Analog				All QAM
		CNR	CSO	CTB	CCN	CCN
Return	N+6	47	61	66	46	40
5-42 MHz	N+3	48	63	67	48	42
	N+0	49	64	67	49	45
	N+6	47	61	66	47	41
5-85 MHz	N+3	48	63	67	48	42
	N+0	49	64	67	49	46

Home Architecture Evolution

New CPE are taking advantage of full band capture (FBC) A/D architectures, which avoid pre-digitizing tuners that can contribute to RF degradation and simplify CPE designs. It will also make them more flexible to evolve moving forward. A low noise amplifier (LNA) precedes the A/D conversion to achieve the necessary levels to efficiently operate an A/D. This architecture is shown in Figure 5.

Analog-to-Digital converters themselves are inherently high noise figure components for typical high-speed bit resolutions because of unavoidable quantization noise. Nonetheless, this architecture does offer added flexibility for front-end sensitivity, and systems are easily optimized by choosing an external LNA, with the effects straightforward to calculate and not frequency dependent.

A quick, nominal, example using Figure 5 illustrates the simplicity:

- Assume a 20 dB gain LNA with a 5 dB NF
- Automatic Gain Control (AGC) amplification to drive A/D converter
- 12-bit A/D (at least 11 effective bits)

A well-design input cascade (LNA + AGC + A/D), can maintain a net NF of 6-7 dB for low input signal levels (where the NF comes into play the most). Thus, with input losses such as diplexers, and design focus on achieving higher modulation efficiency, NF of 8-9 dB could be the next level of sensitivity in new CPE, slightly better than the 10 dB range available today.

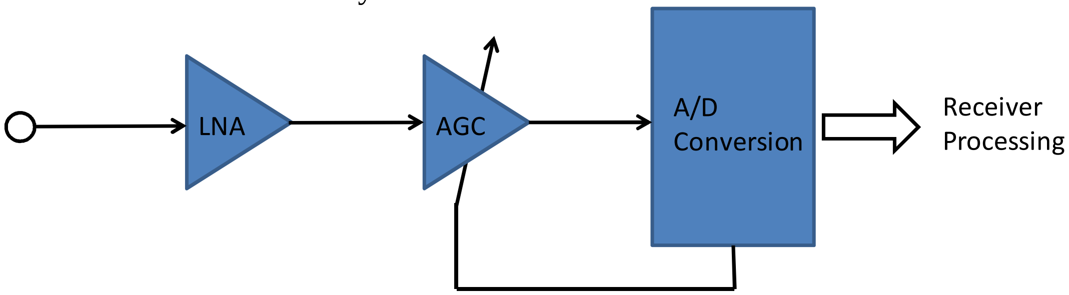


Figure 5 – Full-Band Capture Receiver Architecture

More important than the details of the receiver architecture, however, is that as part of the IP transition, operators are seriously considering investing in the next generation of home gateways based on a point-of-entry (POE) concept. Today, every subscriber is inherently *part* of the HFC access network, which makes for unpredictable results and ultimately money spent on truck rolls. The POE concept would have the cable drop go directly to an IP gateway (legacy support capabilities TBD). This IP gateway would completely abstract the inside of the home from the access network, and use only Home LAN interfaces to deliver content around the home – MoCA™, WiFi, and/or Ethernet interfaces delivering the bits over the last 100 feet. This has valuable benefits to RF losses in and out of the home for QAM receivers and upstream transmitters.

For the downstream, it amounts to benefits to the receiver SNR contribution, which can be substantial and meaningful at low input levels when advanced modulations are considered, as we have seen in Figure 3 and 4. Consider, for example, 20 dBmV tap port levels, 100 feet of drop (RG-6), 4-way splitting in the home, and 50 ft coaxial runs in the home (RG-59). At 1 GHz, we are losing RF power quickly:

STB RFin = 20 – 7 – 7 – 4 = 2 dBmV (virtual) or -4 dBmV.

If a few things break differently, the RF input will drop and the sensitivity of the receiver tested. A secondary splitter (-4 dB), extra drop length (-3.5 dB), and 15 dBmV design levels could challenge this link budget entirely, and house amplifiers may be called into play.

Now let's consider that all of the in-home loss is eliminated except for one splitter (an assumption for legacy considerations), as a POE architecture is apt to look. The loss is now the drop and one splitter, or 11 dB, a 7 dB savings. We can expect receiver NF degradation as the AGC design dials in more attenuation, but this is small relative to the improvement in signal power, so a higher SNR is obtained from the CPE.

Next, we consider the combination of the FBC architecture leading to lower noise CPE, and the POE concept, together as a "next generation" home architecture opportunity. By recalculating Figure 3 with 7 dB more of QAM power, and add a line representing the potential decrease in NF, we can recalculate NF margin to the various modulation profiles. This is shown in Figure 6.

We can quantify an example of possible NF degradation based on the Figure 5 architecture for this increased input level. Using a very simple front end cascade design, the degradation in NF calculated is less than 2 dB. It is probably much less than this, but this offers a boundary for particularly simple Figure 5 architecture. This still means at least a net SNR gain of 5 dB for the CPE contribution to the total SNR. A "degraded NF" case based on the above calculation would be identical to the 10 dB representing today's performance in Figure 6.

On Figure 6, the cascade depth marker lines used are the HFC CCN values taken from the fully loaded 1 GHz case, complete analog reclamation, and an 85 MHz Mid-Split upstream – i.e. Table 5c, orange shaded values.

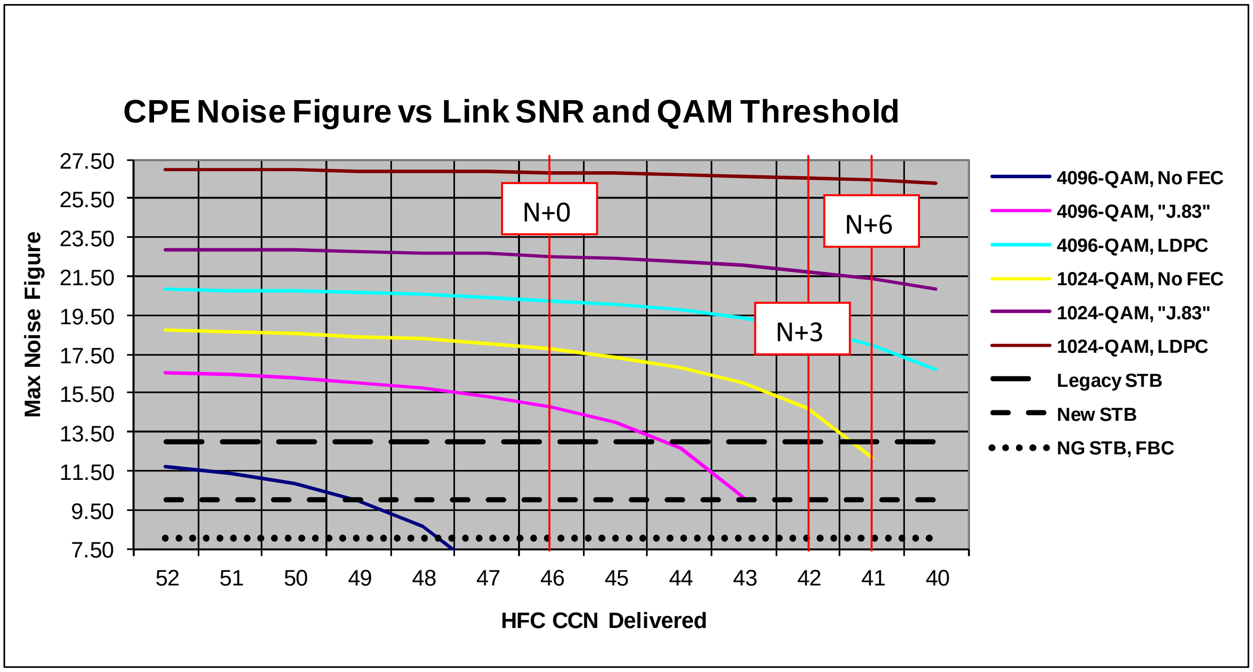


Figure 6 – STB NF Limit vs. Modulation Efficiency, POE Gateway

It would appear in Figure 6 that a tremendous amount of new margin has been created, and this certainly is the case with respect to the access network and home environment's ability to deliver the highest modulation profiles. The results suggest, for example, that "J.83" style 4096-QAM (pink) can be supported, at least on N+0 or equivalently high performing HFC links. Of course, now we are likely to see the biggest impact to the contribution of today's DRFI MER of 43 dB. The impact of this reality is shown in Figure 7.

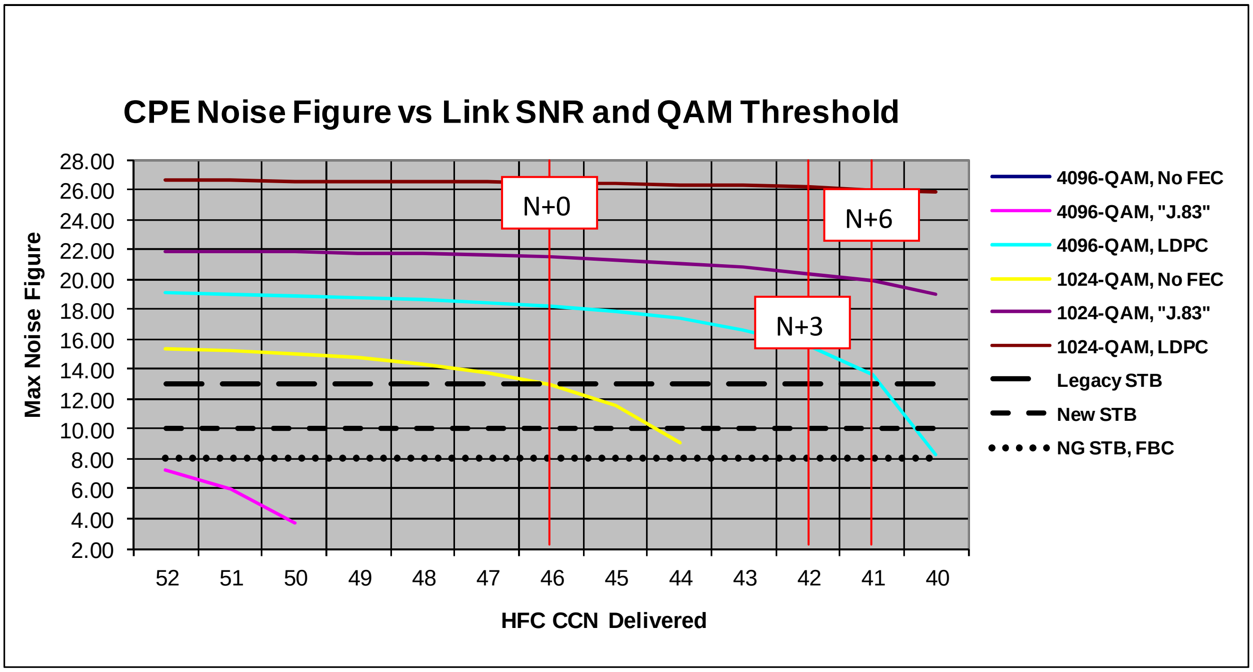


Figure 7 – STB NF Limit vs. Modulation Efficiency, POE Gateway, DRFI MER

We can draw the following conclusions when observing the full picture in Figure 7:

1) 1024-QAM is very comfortably supported from an SNR perspective with old or new FEC, cascade depth or forward band plan
2) Robust 4096-QAM would *require* advanced FEC, and a short HFC cascade – less than N+3 preferably as the curve of support is rapidly becoming sensitive to performance variation and running out of margin as the cascade lengthens.

It seems intuitive that 4096-QAM would require an updated FEC to be operational. Indeed, had it not been so, system architects would have previously considered increasing modulation profiles, as this would have indicated that HFC performance was sufficient. We can come full circle by considering the following:

1) HFC links were designed originally for analog video
2) Analog video, CNR delivered: ~45 dB
3) Digital CNR for a 45 dB analog: 39 dB
4) 4096-QAM with LDPC threshold (Table 2): 37 dB

This illustrates why we are now speaking of 4096-QAM as within range for HFC delivery, and in particular highlights the value of the advanced FEC in making this so. It also illustrates why 1024-QAM in "J.83" style is already on the verge (36 dB) [9].

Figure 8 plots one more example network architecture evolution – in this case removing the linear optical component and distributing the RF generation to the HFC node via digital optics. The assumption in Figure 8 is that this could be accomplished while maintaining DRFI-compliance (43 dB MER). Without the linear optics variation, of course, the "curves" are only a function of the RF cascade depth.

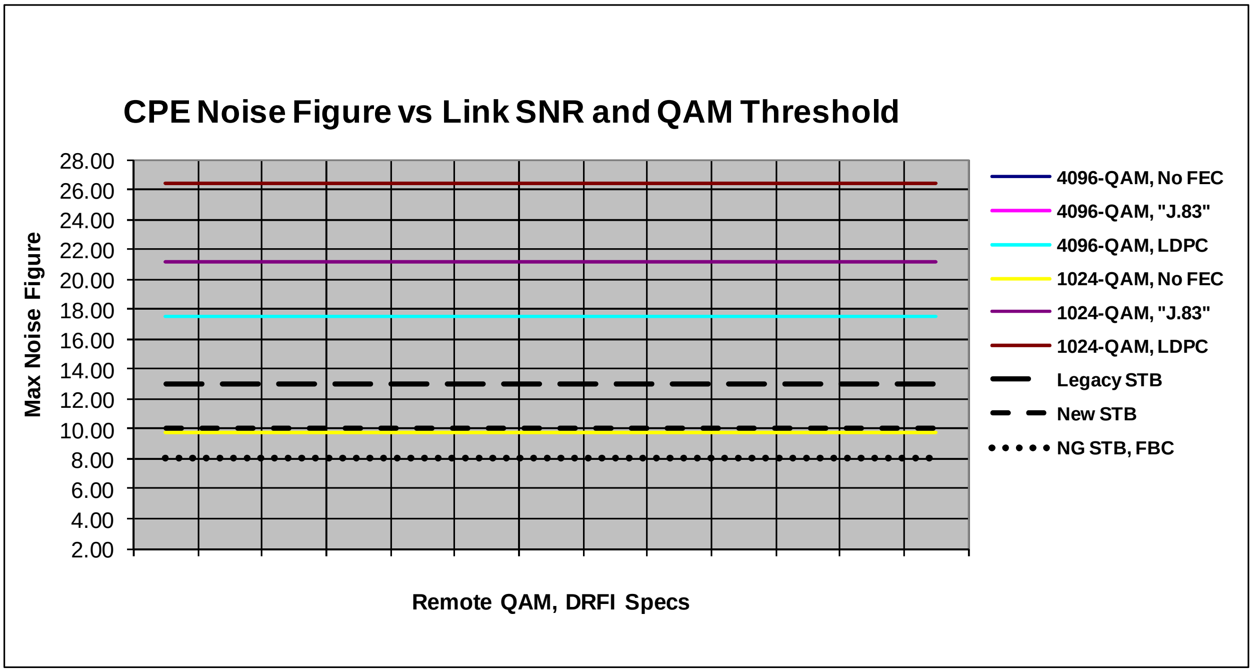


Figure 8 – STB NF Limit vs. Modulation Efficiency, POE Gateway, Remote DRFI, N+3

In Figure 8, we have assumed an N+3 cascade, using the contribution as calculated by comparing the N+0 and N+3 cases in Table 5c, and subtracting the difference. This calculation has some favorable uncertainty in it, because an "N+0" in fact includes the RF chain of the node itself, and the difference between N+0 and N+3 does not capture the effect of these gain blocks. However, Figure 8 gives a ballpark estimate of the performance of a remote QAM with DRFI performance driving an HFC cascade, and the ability this architecture has to support advanced modulation profiles. It also uses the assumptions of Figures 6 and 7 with respect to receiver and home architecture.

HFC UPSTREAM

Turning our attention to the upstream, the SNR threshold assumptions that will be used to illustrate capabilities were shown in Table 3.

In Figure 9, we show today's state of the art for a linear optical upstream. The ability to support 256-QAM upstream over 85 MHz mid-split architectures has been proven in the field, where throughputs of 400 Mbps were obtained [12, 19]. It was shown that DFB optics, coupled with higher sensitivity, higher fidelity DOCSIS 3.0 receivers, could robustly support a fully loaded 85 MHz upstream. A 12 dB dynamic range of sufficient NPR is shown. Dynamic range (DR) in the upstream is much more important than in the downstream. Network design is optimized and aligned precisely in the downstream, while upstream design and environmental variations, alignment techniques, as well as unpredictable RF channel conditions, require an SNR to be met over a range of input levels. Historically, DR on the order of 10 dB has been sought.

Measured packer error rates (PER) were taken at various input levels on the N+3 cascade used in Figure 9. These points are shown in Figure 9 where the yellow marker dots are along the blue noise power ratio (NPR) curve. These are where low PER was observed. The yellow dots on yellow trace are representative of DFB performance of transmitters such as many in the field today. The noise power ratio analysis of Figure 9 makes plain why robust performance was observed. The measured points clearly fall within an area of high NPR and SNR, where good performance would be expected, and with solid 6-7 dB of peak margin extending beyond the 256-QAM threshold identified.

Note that in all figures below, thresholds identified and not labeled as "LDPC" are the assumed "DOCSIS" thresholds of Table 3 (i.e. not the "No FEC" thresholds).

In Figure 10, we introduce the new PHY performance thresholds with advanced FEC from Table 3. In addition, we have included the net performance of the RF portion of the HFC by introducing a deep, combined amplifier cascade (orange). Because the return amplifiers contribute high SNRs individually, the effect here is minor, but we will see how this contribution may also increase as optics improves. Of course, over time, the expectation is that the RF cascade will shorten as well. Or, if the segmentation is virtual (combining in the node is removed), fewer amplifiers will "funnel" upstream. The net effect is the same – fewer amplifiers contributing noise to the return path, reducing the input noise power at the receiver and thereby increasing SNR.

It is clear from Figure 10 that "DOCSIS" coding for 1024-QAM (dashed red) would not be able to be supported from an SNR perspective. This threshold is simply above even the peak NPR. For 1024-QAM enabled with LDPC FEC, however, we can see that the threshold of operation is now exceeded by the combined HFC+CMTS link.

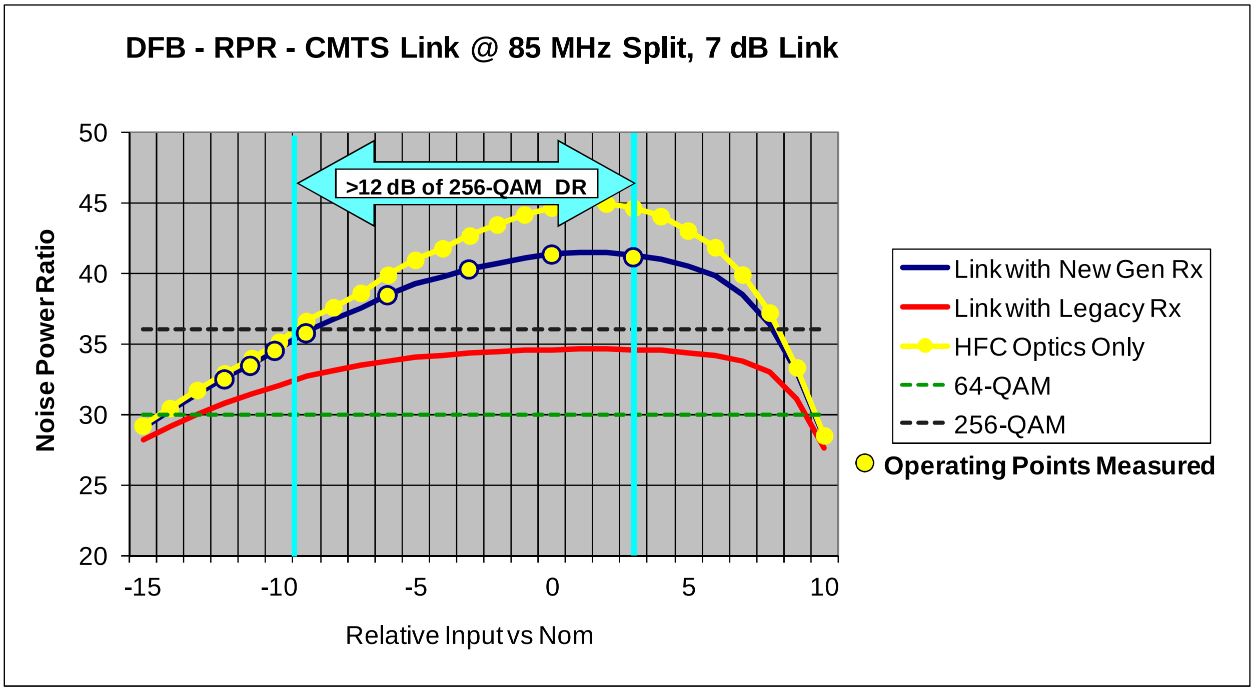


Figure 9 – 256-QAM "DOCSIS" over 85 MHz DFB Return Optics

Unfortunately, the dynamic range is reduced on the left hand side by 3 dB. It is also well understood and documented that, as QAM profiles become increasingly dense, the right hand side (soft distortion components, red arrow) also have a more significant impact, reducing dynamic range from the right. The above reduction from the right is an estimate based on prior work, which only considered up to 64-QAM [23].

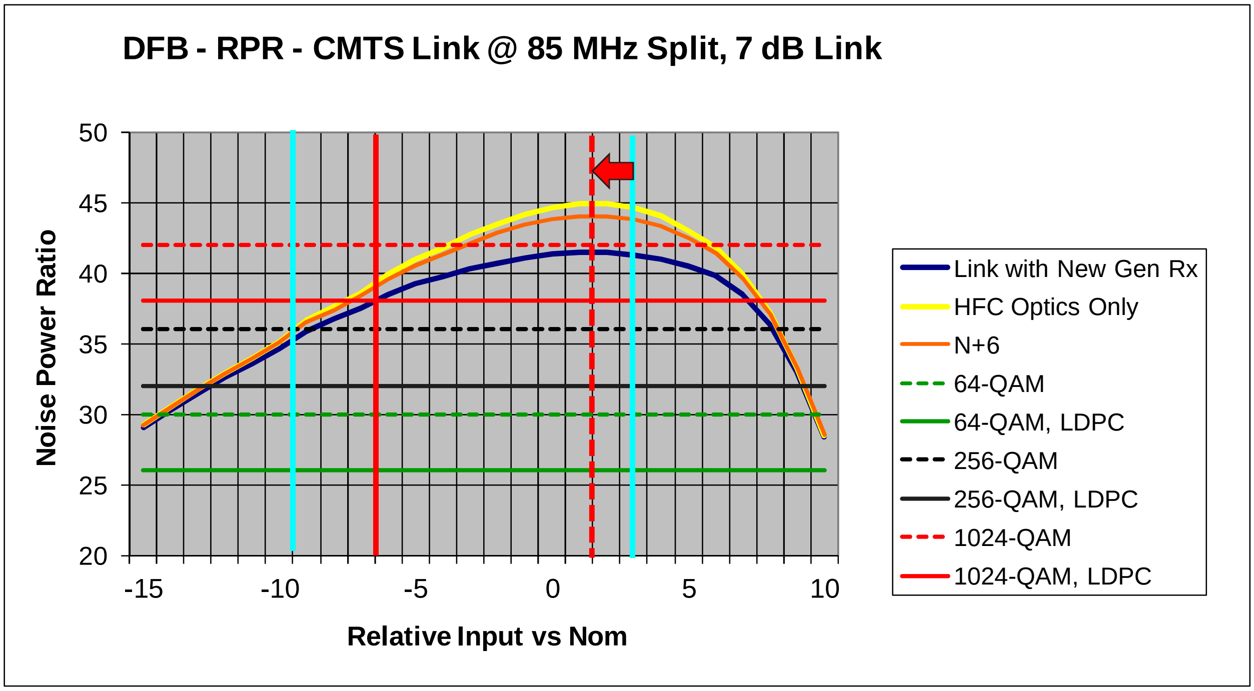


Figure 10 – 64/256/1024-QAM, "DOCSIS" and New FEC, 85 MHz DFB Optics

The net effect of the reduced range (8 dB DR) is that 1024-QAM with LDPC would be marginal in practice. It would likely work in some places, but inconsistently throughout a footprint without further network improvements. An analogous situation is the introduction of DOCSIS 3.0 64-QAM using FP upstream laser technology. FPs exhibit reduced performance compared to DFBs, and thus eat into the DR margin acceptable to run DOCSIS 3.0 64-QAM. Such is the case in Figure 10 for 1024-QAM with LDPC. The operating window is small, and this would likely be reflected in inconsistent performance.

In summary, with LDPC, it would be possible to get 1024-QAM working on well-behaved upstream channels that are aligned properly for laser input level, and only using new DOCSIS 3.0 upstream receivers with higher sensitivity. However, in practice, over a large range of plants, performance would be unreliable (and impossible at all if laser technology has not been updated and legacy receiver performance exists).

In Figure 11, we extend the Figure 10 case to a 200 MHz "high split" return – a long term approach under consideration by the industry to deliver more upstream capacity [2]. An implicit assumption is that the 200 MHz receiver could perform equivalently from a sensitivity standpoint as today's DOCSIS 3.0 receiver. Nonetheless, the pure power loading makes 1024-QAM completely impractical. The 256-QAM case without new FEC is now marginal, as the 1024-QAM case in Figure 10 was – in fact, it exhibits the same DR and signature relative to the performance threshold (highlighted in red). However, it is likely to be somewhat more robust that 1024-QAM simply because 256-QAM will be less sensitive to the types of things that the allocation of DR and margin is meant to protect against.

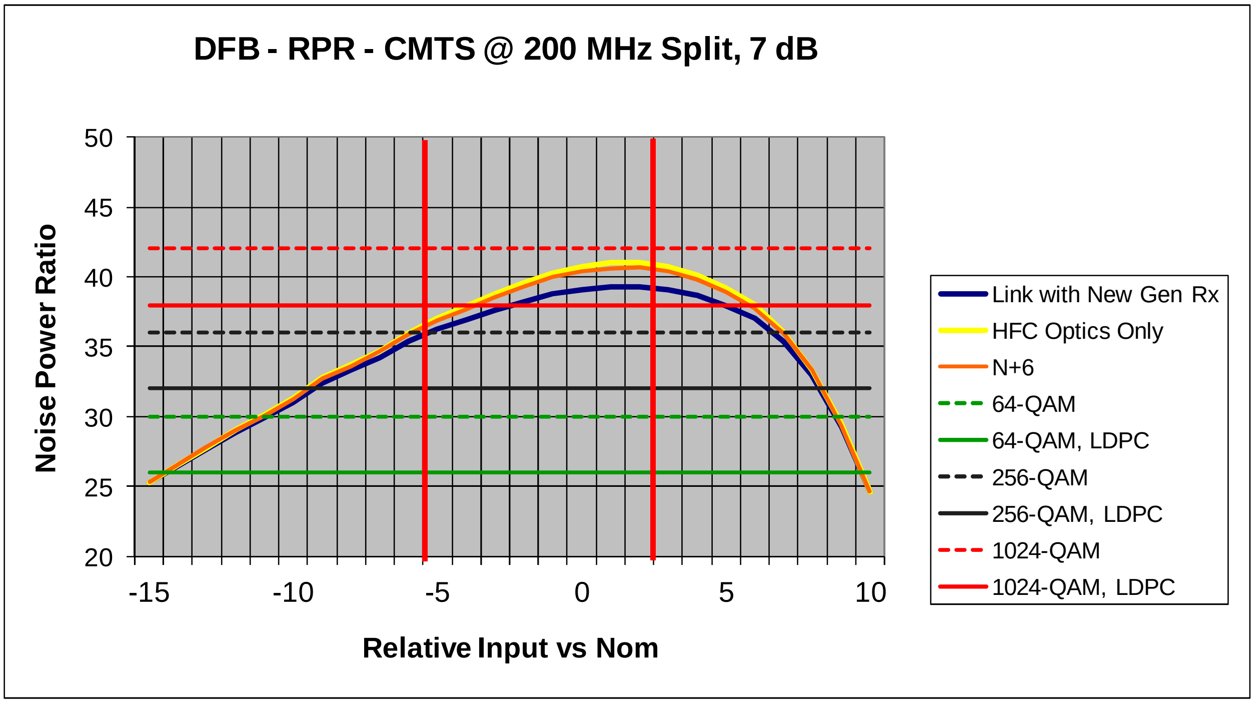


Figure 11 – 64/256/1024-QAM, "DOCSIS" and New FEC, over 200 MHz (Projected)

Enhanced Linear Optical Performance

The analysis in Figures 9-11 uses performance of DFBs like many that may be in the field, perhaps lower power (1 mw), where the upgrade from Fabry-Perot lasers (FPs) had already taken place. Today, because of the rising interest in mid-split architectures and more bandwidth efficient modulation, new development activity is focusing on analog and digital return solutions that optimize performance for extended bandwidths. Continued improvements in noise performance of both transmitters and receivers have also occurred over time. Recently measured performance of a mid-split bandwidth "DFBT3" (Motorola model, temperature compensated, 2 mw) to a return path receiver is shown in Figure 12.

In Figure 12, the performance curves from Figures 9 are also shown. A notable improvement over time in peak NPR is observable, and an associated dynamic range increase for a given threshold. Peak performance of 50-51 dB is observed – again making the point that a good DFB return path transmitter-receiver link looks very much like a 10-bit digital return link [16].

Also shown is this improved performance when combined with a DOCSIS 3.0 receiver (dashed blue). Here, it becomes clear that as the optics improves, the influence of receiver SNR contribution begins to have a larger effect. We will discuss this further later in this section.

In Figure 13, the DOCSIS and "next gen" thresholds using advanced FEC are evaluated against the improved mid-split performance identified and measured in Figure 12. Note that the performance shown is the linear optical return (yellow) only, along with the DOCSIS 3.0 receiver (blue). Thus, this is equivalent to an N+0 case, since there are no RF noise contributions from the plant in Figure 13.

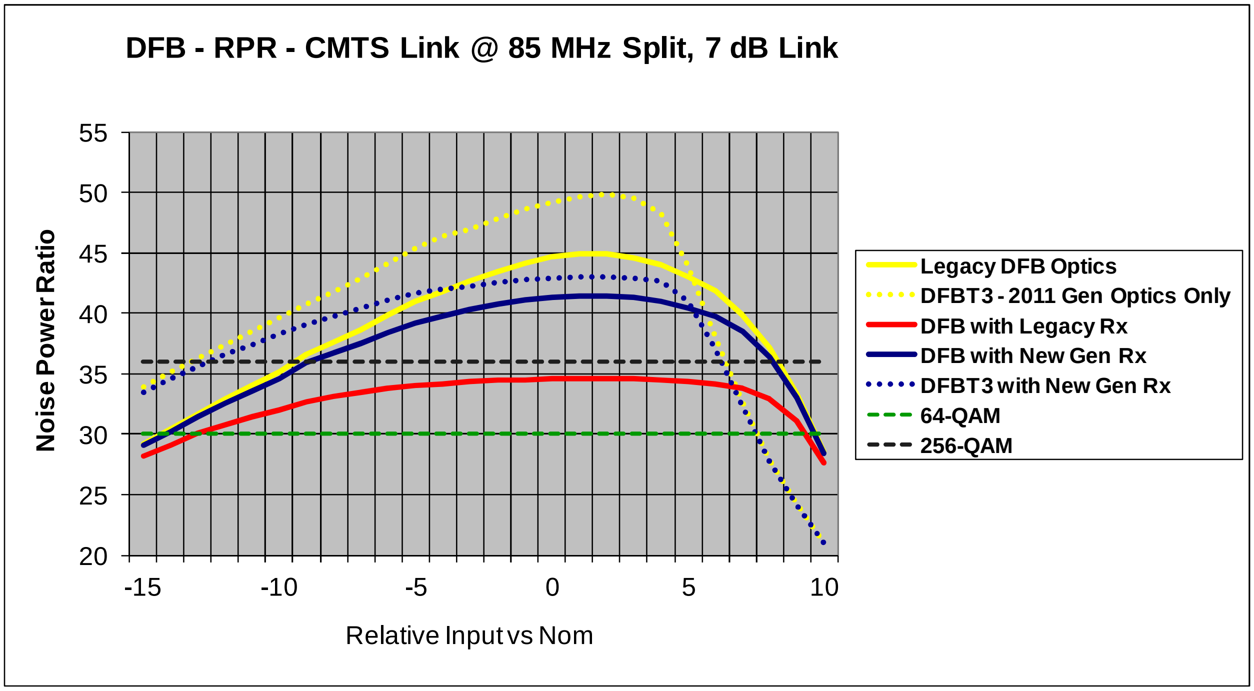


Figure 12 – Measured Mid-Split Performance, Modern DFBT3-RPR Link

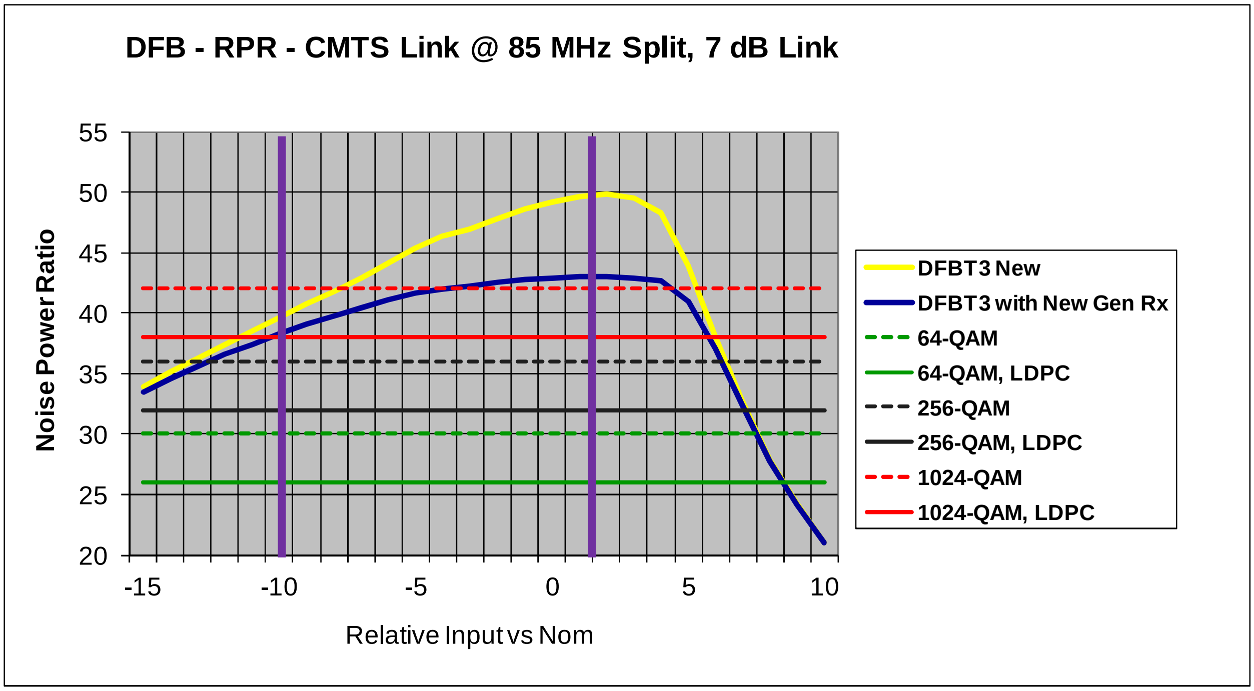


Figure 13 – High Performance DFBT3-RPR Mid-Split, All Thresholds

All of the thresholds are met in Figure 13, and the 64-QAM through 256-QAM cases comfortably supported. This is to be expected since this has been proven to be the case for 256-QAM today without any new FEC applied or improved HFC return performance.

For 1024-QAM, however, clearly "DOCSIS" PHY coding will not be sufficient using the Table 3 assumptions. The advanced FEC applied to 1024-QAM, however, does show promise. Indeed, it shows 11 dB of DR (purple markers) above the threshold identified for 1024-QAM. However, note that the margin above the threshold never exceeds 5 dB, and for half of the DR it is 4 dB or less. The receiver noise contribution, ably supporting 256-QAM (against a 64-QAM-maximum DOCSIS requirement, it should be added), comes into play here.

While the dynamic range appears robust, the low margin across the full DR range may make performance less robust than such a DR would otherwise indicate considering we are dealing with a more sensitive QAM profile. In other words, "peak" margin may need consideration as we move to more complex QAM profiles. Return path alignment is based on setting composite signal levels around a "sweet spot" close to where NPR is at its peak, but allowing for some margin of back-off to avoid the clipping and distortion region on the right hand side. At the very least, it would seem reasonable that the same peak margin available today for 256-QAM (6-7 dB, Figure 9) would be a good starting point objective for higher order scenarios like 1024-QAM.

Today's expectation of DR is built around volumes of 16-QAM and 64-QAM deployments, exclusively. Since the DR for 1024-QAM with LDPC in Figure 13 is about the same as we observed in Figure 9 for 256-QAM, we might expect 1024-QAM with LDPC to be operational under good upstream conditions. Similarly, since the peak margin available is reduced across its DR compared to Figure 9, and because other contributions not captured by NPR will effect 1024-QAM worse than 256-QAM, its performance is likely to be less rugged than the 256-QAM case in Figure 9.

Figure 14 adds an RF noise contribution assuming a deep cascade (N+6) and combined four ways. The degradation due to the quantity of RF amplifiers is seen (orange). However, it is clear in observing the blue curves – (optics + receiver) – with and without RF noise contributions from the cascade, that the limitation about 256-QAM is in the noise contribution of the upstream receiver, at least for nominal architectures and levels as they are implemented today. And, again, we are quantifying a receiver for 1024-QAM that had a requirement to meet 64-QAM performance objectives.

In summary, from an NPR/SNR perspective, with new LDPC-based FEC, 1024-QAM is possible on high performance upstream optical links. However, it may be operationally less robust when considered across a range of possible channel environments given the decreased peak margin and exaggerated sensitivity of 1024-QAM to other link impairments not captured by NPR analysis. It is hard to be certain, but we can get a window into the performance consistency through the ruggedness with which 256-QAM becomes implemented. 1024-QAM with LDPC should be "like" this but somewhat less robust. As might be expected, we will need to ask more of next generation receivers than sensitivity supporting 64/256-QAM if implemented over linear optical returns or today's vintage of digital returns.

In Figure 15, we apply the Figure 12-14 performance across a 200 MHz upstream. The assumptions are that identical optical noise performance can be achieved (shared over a wider bandwidth) and identical receiver sensitivity can also be achieved.

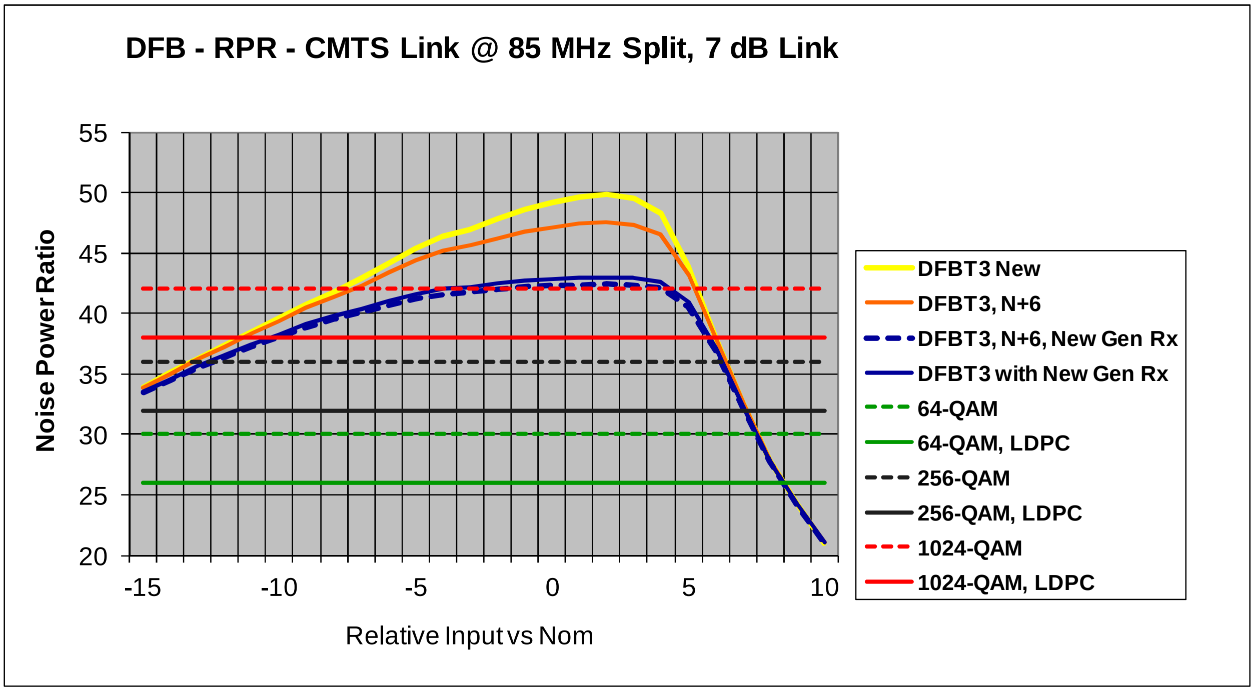


Figure 14 – High Performance DFBT3-RPR Mid-Split, N+6, All Thresholds

An encouraging conclusion is that, for this extended bandwidth case, 256-QAM of the "DOCSIS" variety appears to be operational at acceptable DR and a reasonable peak margin, at least under these assumptions of the wider band design. Since we do not have extensive field lessons for 256-QAM, the peak margin question raised for 1024-QAM could apply in this case as well. However, to be at least within reach of 256-QAM already as it exists today on an extended bandwidth return, based on today's linear optics, is an excellent side for the future of new broadband capacity for the upstream.

The 1024-QAM case now clearly has insufficient DR (purple) as well as small peak margin from the threshold – 4 dB maximum and less than that across the DR.

Figure 16 shows an enhanced return performance example, in this case based on digital return. Key advantages of digital return include ease of setup, and, from the perspective of this paper, NPR performance independent of link length. The return path performance of the digital return approach is almost entirely determined by the A/D converter resolution. Finding A/D converters that provide a high *effective* number of bits (ENOB) as bandwidths expand is the limiting performance component. The potential to require redesign for each bandwidth increasing increment is a disadvantage [16] of this approach.

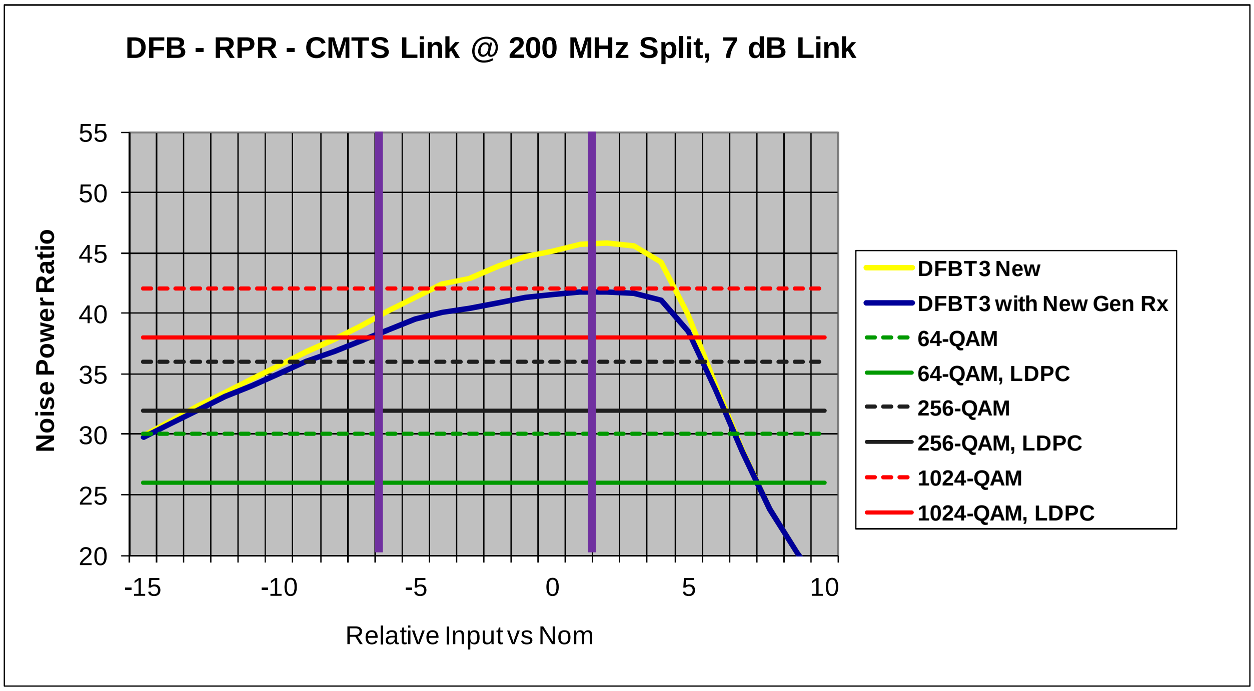


Figure 15 – High Performance DFBT3-RPR "High" Split, All Thresholds

Figure 16 uses measured performance of a solution that behaves like an ideal 10-bit A/D, or, equivalently, has an ENOB of 10 bits.

Three composite NPR curves that include RF and receiver contributions are shown with the digital return-only NPR performance:

1) Digital return, N+6 RF cascade, receiver (purple dash)
2) Low noise DFBT3-RPR link, N+6, receiver (yellow)
3) Original DFB-RPR (256-QAM analysis), N+6, receiver (blue)

The comparison indicates that the digital solution contributes a couple more dB of DR to the 1024-QAM with LDPC threshold on the left side of the NPR curve (the SNR side), which itself adds 4 dB of DR over the legacy solution. The extra couple of dB from the digital return could be the difference between robust or less robust. However, since the peak margin has not changed between the two (yellow and dashed purple), and is limited by receiver sensitivity, their performance will likely be similar – on the bubble of robust-enough margin.

A clear conclusion from these analyses is that support of 1024-QAM would require new FEC, and be aided by an improved receiver sensitivity. This is readily illustrated by observing the effect of a hypothetical receiver designed to support 1024-QAM (as opposed to DOCSIS 3.0, 64-QAM), with a receiver sensitivity improved 3 dB in so doing.

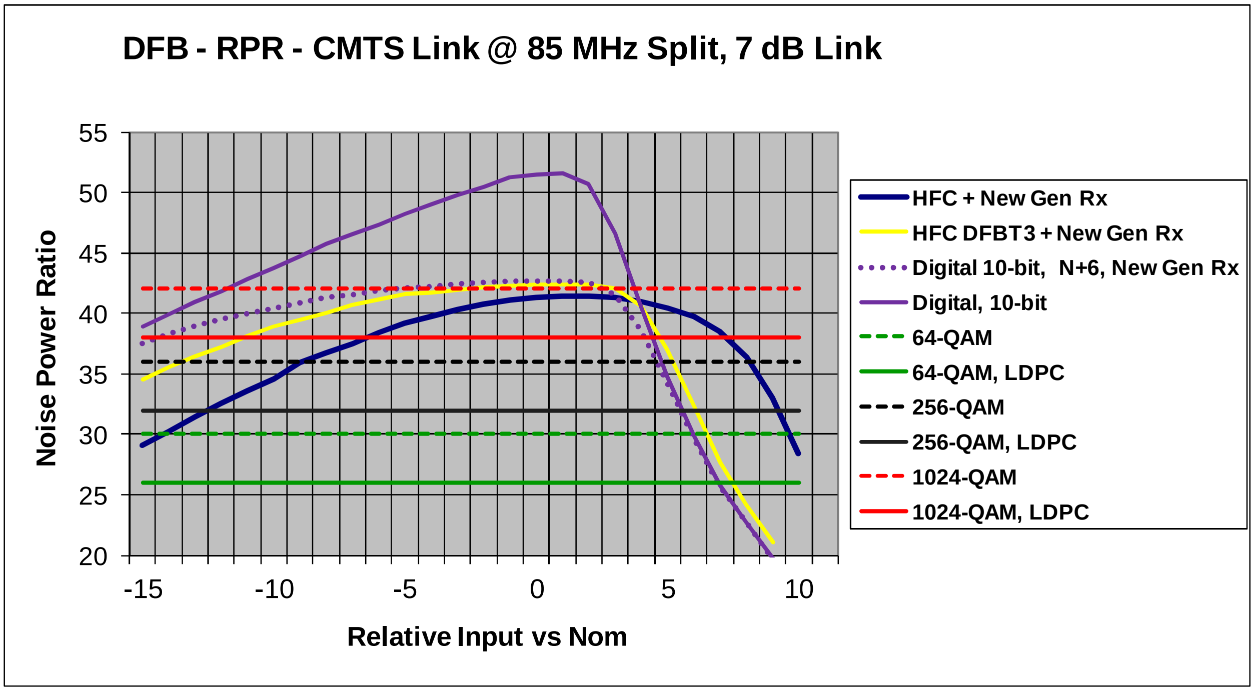


Figure 16 – 10-bit Digital Return & Linear Optics Cases, N+6, All Thresholds

Figure 17 shows the same composite NPR curves, minus the digital return-only one for this case of improved sensitivity. Figure 17 illustrates that for linear optical links or digital returns, it is clear that these would now have margin to support 1024-QAM with adequate DR. This could only be so if the 1024-QAM was accompanied by new FEC. And, there are still some open questions about whether peak margin standards should be considered, much of which may come as 256-QAM deploys under "DOCSIS" threshold conditions.

This leads to an overall conclusion that, for upstream, attention will need to be paid to upstream optical architectures, receiver performance, and possibly Headend architecture on the whole given that the optical receiver-CMTS connection today is a simple, almost forgotten pipe that can create low level inputs to CMTS ports and challenge their sensitivity. For all cases desiring to support 1024-QAM, LDPC-based FEC is a must-have, though itself not a sufficient condition.

<u>Remote Demodulation</u>

Future architectures may take advantage of distributed physical layers. The analogous concept in today's world is "CMTS in the Node." For a transition into new RF and IP technology or extension of DOCSIS, this may be easier to consider than it has historically been with DOCSIS. And, with Gigabit Ethernet and EPON optics available at low cost, it become very attractive to consider taking advantage of these standard interfaces, potentially eliminate linear optics from the plant, and improve performance all at the same time. Modular node platforms are now built to handle various plug-in optical and RF modules, so this approach is consistent with HFC node evolution.

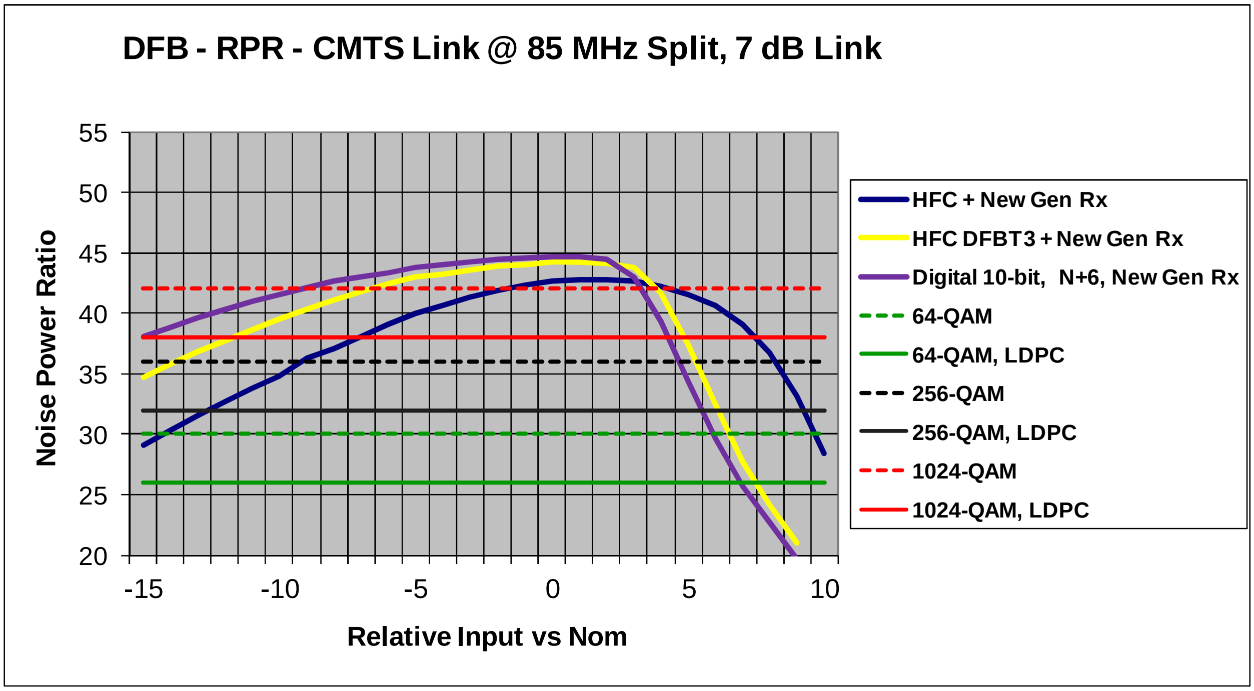


Figure 17 – All Cases of Return Optics, N+6, Improved Rx Sensitivity (3 dB)

If a remote receiver is placed in the plant, then we can simplify the analysis by removing the optical links from the equation. This can be shown using the same NPR curves as before, except now only the softly distorting amplifier limits the right hand side of the curves. While these can be characterized and specified, this is not usually done. It is nominally assumed that the upstream signal transmissions stay comfortably within the range of return amplifier linearity. The alignment of the total level of the load is critical at the upstream optical interface, but in the plant there is not the constraint of total power load to the extent that there is in the optics. Coupled with common noise figures of return amplifiers, the result is that very high SNRs are possible for a single amplifier relative to its noise contribution to the upstream cascade.

Instead of NPR, Table 6 shows a range of required Noise Figures of a module installed in node. We have made an assumption, based on our above discussion of peak margin above threshold for advanced QAM profiles, that a net SNR of 45 dB (7 dB of margin to 1024-QAM with LDPC) is the objective.

A range of port levels (low, mid, high) are shown, and the impact these levels would have on the remote receiver's NF requirement. The path loss between the coaxial input port and node module is accounted for, so this is the noise figure performance at the input to the receiver module. Provided low input levels do not reign, these are not particularly challenging. And, even at the lowest input level identified here, the NF is achievable with good design practices.

Table 6 – Calculating "Remote" Receiver Noise Requirements

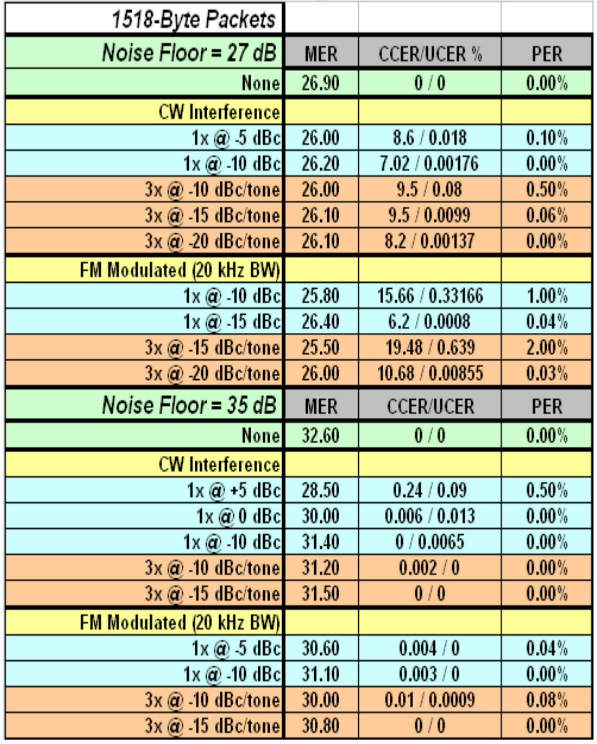
Signal Level		**Signal Level**		**Signal Level**	
Ports	10.0	Ports	15.0	Ports	20.0
Node Path Loss	5.0	Node Path Loss	5.0	Node Path Loss	5.0
Node Combine	6.0	Node Combine	6.0	Node Combine	6.0
Noise		**Noise**		**Noise**	
Amplifier (NF=8)	50.2	Amplifier (NF=8)	-50.2	Amplifier (NF=8)	-50.2
Cascade	6.0	Cascade	3.0	Cascade	3.0
Combine	4.0	Combine	1.0	Combine	1.0
RF Noise	36.4	RF Noise	-45.4	RF Noise	-45.4
SNR Req'd	**45.0**	**SNR Req'd**	**45.0**	**SNR Req'd**	**45.0**
Terminating NF	**6.3**	**Terminating NF**	**16.9**	**Terminating NF**	**22.0**

SIGNAL-TO-INTERFERENCE

Single carrier techniques to combat narrowband interference amount to attempting to notch out the offender's band through an adaptive filtering mechanism, and recover the modulated carrier around it as effectively as possible. Because removing the interference involves removing signal spectrum that subsequently must be equalized and detected, the effectiveness of the process is reduced as the interference becomes wider band, or, for a fixed signal-to-interference energy (S/I), if there are multiple interferers to handle.

Test results have been observed and reported with respect to the A-TDMA DOCSIS upstream in separate studies in recent years [13,22]. Table 7 shows thresholds of uncorrectable codeword errors observed for 64-QAM.

Table 7 – Ingress Thresholds for 64-QAM A-TDMA Upstream



1518-Byte Packets			
Noise Floor = 27 dB	MER	CCER/UCER %	PER
None	26.90	0 / 0	0.00%
CW Interference			
1x @ -5 dBc	26.00	8.6 / 0.018	0.10%
1x @ -10 dBc	26.20	7.02 / 0.00176	0.00%
3x @ -10 dBc/tone	26.00	9.5 / 0.08	0.50%
3x @ -15 dBc/tone	26.10	9.5 / 0.0099	0.06%
3x @ -20 dBc/tone	26.10	8.2 / 0.00137	0.00%
FM Modulated (20 kHz BW)			
1x @ -10 dBc	25.80	15.66 / 0.33166	1.00%
1x @ -15 dBc	26.40	6.2 / 0.0008	0.04%
3x @ -15 dBc/tone	25.50	19.48 / 0.639	2.00%
3x @ -20 dBc/tone	26.00	10.68 / 0.00855	0.03%
Noise Floor = 35 dB	MER	CCER/UCER	PER
None	32.60	0 / 0	0.00%
CW Interference			
1x @ +5 dBc	28.50	0.24 / 0.09	0.50%
1x @ 0 dBc	30.00	0.006 / 0.013	0.00%
1x @ -10 dBc	31.40	0 / 0.0065	0.00%
3x @ -10 dBc/tone	31.20	0.002 / 0	0.00%
3x @ -15 dBc/tone	31.50	0 / 0	0.00%
FM Modulated (20 kHz BW)			
1x @ -5 dBc	30.60	0.004 / 0	0.04%
1x @ -10 dBc	31.10	0.003 / 0	0.00%
3x @ -10 dBc/tone	30.00	0.01 / 0.0009	0.08%
3x @ -15 dBc/tone	30.80	0 / 0	0.00%

In the study summarized in Table 8, recent results for 256-QAM for a fixed PER objective of 0.5% and 1% for a high SNR condition are derived through testing. The SNR condition applied (SNR = 36 dB) is consistent with the Table 4 assumption on a robust SNR threshold to use for 256-QAM in the upstream, and the analysis in [22] further identifies this high SNR as one that increasingly enables ingress cancellation.

Table 8 – Ingress Thresholds for 256-QAM A-TDMA Upstream

		256-QAM			
	Level (dB, dBc)	UNCORR%	CORR%	PER%	MER (dB)
Baseline - AWGN	**36**	0.000%	0.000%	0.000%	37
Single Ingressor Case					
QPSK 12kHz 0.5%	3	0.254%	0.435%	1.060%	34
QPSK 12kHz 1.0%	1	0.447%	0.944%	2.300%	34
FSK 320ksym/s 0.5%	29	0.278%	0.032%	0.110%	35
FSK 320ksym/s 1.0%	27	0.633%	0.230%	0.810%	35
FM 20kHz 0.5%	2	0.128%	0.295%	0.750%	34
FM 20kHz 1.0%	1	0.187%	0.554%	1.260%	34
Three Ingressor Case					
CPD 0.5%	28	0.297%	0.041%	0.190%	34
CPD 1.0%	27	0.698%	0.144%	0.750%	33

Tables 7 and 8 are valuable indicators of the performance of ingress cancellation (IC). However, to understand how well it is working, it helps to know how robust the individual QAM profiles are to signal-to-interference ratio (S/I) to begin with. A system simulation was performed to evaluate this sensitivity. An example of 64-QAM with a single CW interferer is shown in Figure 18, while a 256-QAM example is shown with three non-coherent interferers in Figure 19. The familiar CW "donut" pattern is clear in Figure 18.

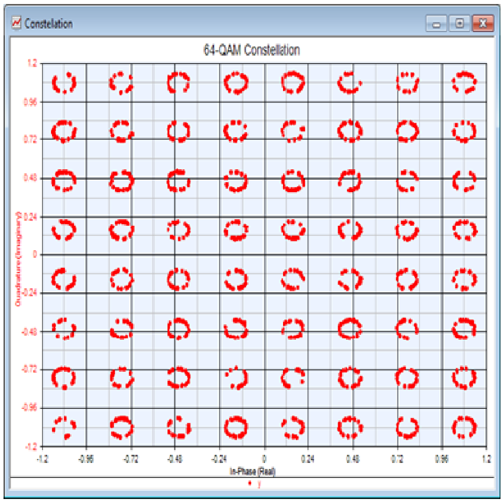


Figure 18 – 64-QAM with a Single CW Interferer

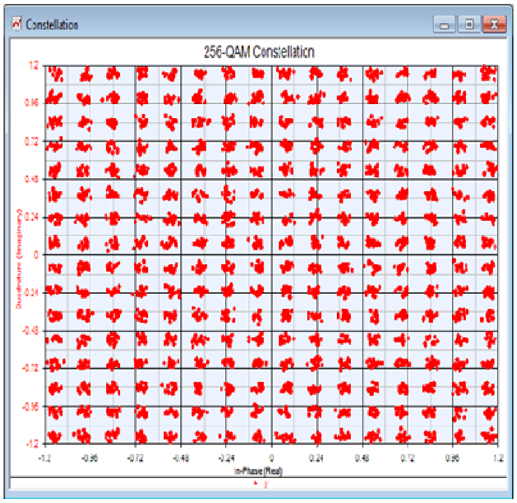


Figure 19 – 256-QAM with 3 CW Interferers

Both cases were evaluated to find thresholds of correctable low error rate. We use the more challenging three-interferer case as a reference. A subset of these simulation results are summarized in Tables 9a-d for 64/256/1024/4096-QAM. The modeling tool is described further in the Appendix.

Obviously, more dense profiles are more sensitive to S/I. The relationship of interest is to note that, for a given "DOCSIS" Table 3 SNR threshold, high enough to have robust performance and allow the IC to work, the relative S/I difference across profiles is also approximately 6 dB for when errors begin to be counted.

Now consider the IC performance based on Table 7 MER before and after IC. It can be calculated as providing roughly an effective 26-28 dB of cancellation for the case of multiple 20 kHz interferers. An estimate for the Table 8 case for 256-QAM and 0.5% PER using 256-QAM data at SNR = 36 dB, from the analysis table shown in Table 9b (est. 28 dB S/I), is that the IC is providing about 26 dB (S/I =2 to S/I = 28) of IC for a single 20 kHz interferer. Table 7 suggests that if the total power of the interference is the same, and it is narrowband, IC performance is close to the same for one interferer or three.

Table 9a-d – Ingress Thresholds for Uncoded QAM
a) 64-QAM b) 256-QAM c) 1024-QAM d) 4096-QAM

	64-QAM		S/I, N=3 Interferers									
		20	21	22	23	24	25	26	27	28	29	30
	25	1.80E-02	1.20E-02	9.10E-03	4.31E-03	3.45E-03	1.72E-03	9.35E-04	7.46E-04	5.01E-04	2.88E-04	2.09E-04
	26	1.90E-02	1.30E-02	7.08E-03	4.20E-03	1.63E-03	1.18E-03	5.63E-04	3.12E-04	2.07E-04	1.09E-04	7.90E-05
SNR	27	1.90E-02	1.00E-02	4.29E-03	2.19E-03	9.51E-04	3.26E-04	2.17E-04	1.22E-04	3.70E-05	2.00E-05	1.90E-05
	28	1.90E-02	9.70E-03	3.17E-03	1.03E-03	5.22E-04	1.36E-04	1.01E-04	2.20E-05	2.00E-05	1.10E-05	5.00E-06
	29	1.20E-02	4.47E-03	2.25E-03	1.31E-03	7.90E-05	1.30E-05	3.20E-05	1.00E-06	0	0	0
	30	6.32E-03	2.58E-03	1.17E-03	2.32E-04	1.78E-04	4.50E-05	7.00E-06	0	0	0	0
	31	1.40E-02	6.29E-03	2.10E-03	3.75E-04	7.00E-06	8.00E-06	0	0	0	0	0
	32	1.60E-02	1.13E-03	2.81E-04	3.90E-05	4.10E-05	2.00E-06	0	0	0	0	0
	33	2.69E-03	5.27E-04	1.05E-03	1.57E-04	2.00E-06	0	0	0	0	0	0
	34	2.36E-03	4.48E-03	3.10E-05	2.30E-05	1.00E-06	0	0	0	0	0	0
	35	9.62E-03	5.78E-04	6.60E-05	0	0	0	0	0	0	0	0

	256-QAM		S/I, N=3 Interferers									
		30	31	32	33	34	35	36	37	38	39	40
	30	4.02E-03	2.68E-03	1.90E-03	1.36E-03	8.94E-04	6.77E-04	5.22E-04	4.48E-04	3.53E-04	2.95E-04	2.86E-04
	31	2.86E-03	1.30E-03	9.74E-04	5.69E-04	4.25E-04	2.36E-04	2.20E-04	1.34E-04	1.11E-04	9.10E-05	8.60E-05
SNR	32	1.40E-03	9.41E-04	5.01E-04	2.03E-04	1.50E-04	7.90E-05	6.50E-05	3.50E-05	1.90E-05	1.50E-05	1.30E-05
	33	5.79E-04	3.93E-04	9.90E-05	1.18E-04	3.50E-05	2.00E-05	1.70E-05	1.20E-05	6.00E-06	4.00E-06	2.00E-06
	34	2.89E-04	2.76E-04	3.80E-05	2.00E-05	2.60E-05	2.00E-06	2.00E-06	0	0	0	0
	35	8.70E-05	1.09E-04	3.00E-05	1.10E-05	8.00E-06	0	1.00E-06	0	0	0	0
	36	3.80E-05	5.00E-06	7.00E-06	0	0	0	0	0	0	0	0
	37	3.70E-05	5.00E-06	1.00E-06	0	0	0	0	0	0	0	0
	38	1.50E-05	4.00E-06	0	0	0	0	0	0	0	0	0
	39	5.00E-06	0	0	0	0	0	0	0	0	0	0
	40	0	0	0	0	0	0	0	0	0	0	0

	1024-QAM		S/I, N=3 Interferers									
		40	41	42	43	44	45	46	47	48	49	50
	35	1.99E-03	1.60E-03	1.38E-03	1.10E-03	1.01E-03	8.36E-04	7.91E-04	8.35E-04	7.51E-04	6.92E-04	6.09E-04
	36	7.18E-04	6.75E-04	5.39E-04	4.05E-04	3.64E-04	3.26E-04	2.61E-04	2.13E-04	2.23E-04	2.04E-04	1.68E-04
SNR	37	4.34E-04	1.94E-04	2.18E-04	1.14E-04	8.40E-05	8.40E-05	4.60E-05	4.60E-05	6.20E-05	5.10E-05	5.40E-05
	38	1.33E-04	1.10E-04	5.00E-05	3.30E-05	2.20E-05	1.40E-05	1.40E-05	1.10E-05	7.00E-06	3.00E-06	7.00E-06
	39	2.60E-05	1.60E-05	1.50E-05	0	4.00E-06	0	0	4.00E-06	0	0	0
	40	5.00E-06	5.00E-06	2.00E-06	0	1.00E-06	0	0	0	0	0	0
	41	1.00E-06	0	0	0	0	0	0	0	0	0	0
	42	2.00E-06	0	0	0	0	0	0	0	0	0	0
	43	0	0	0	0	0	0	0	0	0	0	0
	44	0	0	0	0	0	0	0	0	0	0	0
	45	0	0	0	0	0	0	0	0	0	0	0

	4096-QAM		S/I, N=3 Interferers									
		40	41	42	43	44	45	46	47	48	49	50
	40	0.014	0.01	7.92E-03	6.40E-03	4.87E-03	4.24E-03	3.48E-03	2.83E-03	2.71E-03	2.28E-03	2.16E-03
	41	0.011	7.55E-03	5.34E-03	3.79E-03	2.63E-03	2.18E-03	1.56E-03	1.31E-03	1.11E-03	8.47E-04	7.26E-04
SNR	42	8.11E-03	4.64E-03	2.87E-03	1.98E-03	1.23E-03	9.98E-04	6.45E-04	5.35E-04	4.13E-04	3.17E-04	2.77E-04
	43	5.34E-03	2.49E-03	1.43E-03	1.26E-03	5.54E-04	4.43E-04	2.91E-04	1.32E-04	1.16E-04	6.60E-05	7.10E-05
	44	4.91E-03	2.18E-03	9.43E-04	3.59E-04	2.24E-04	1.02E-04	8.71E-05	3.60E-05	3.50E-05	2.50E-05	1.30E-05
	45	3.48E-03	1.53E-03	3.76E-04	2.03E-04	1.01E-04	2.50E-05	1.60E-05	8.01E-06	1.80E-05	5.00E-06	1.00E-06
	46	1.83E-03	1.04E-03	3.52E-04	1.55E-04	4.80E-05	1.20E-05	8.01E-06	1.20E-05	0	0	0
	47	2.04E-03	1.81E-04	1.68E-04	9.01E-06	1.10E-05	1.00E-06	0	1.00E-06	0	0	0
	48	9.81E-04	2.89E-04	2.70E-05	8.01E-06	1.00E-06	0	0	0	0	0	0
	49	7.19E-04	9.21E-05	1.00E-05	0	0	0	0	0	0	0	0
	50	4.51E-04	8.31E-05	6.00E-06	0	0	0	0	0	0	0	0

Based on the above, then, the IC is providing about the same amount of cancellation in these two cases tested, although there is much more SNR headroom in the 64-QAM case. However, the 64-QAM case also does not *need* additional suppression, so, while it may be available, no further IC adaption is required to deliver low error performance.

For the upstream, we are interested in extending what we have learned for 64/256-QAM to 1024-QAM as an advanced profile. Simulation results for the "DOCSIS" SNR threshold condition for 1024-QAM of Table 3 and a 27 dB S/I easily shows that uncoded error rates are horrendously high (0.1 to 0.01 range). They are so high as to possibly be unable to be corrected adequately if at all by FEC, or it would be not desired to rely so heavily on it. Better than 26 dB of IC would be required, and based on the above mentioned relationship, probably 6 dB better. It is unclear if today's IC can accomplish this for multiple interferers with bandwidth, but it likely will be necessary.

It is worthwhile to point out that for a single CW interferer, effective cancellation of about 35 dB was obtained based on Table 7. So, at least for this friendlier case, the IC function can be stronger.

Threshold Values with New FEC

We have already concluded in prior analysis that a 1024-QAM downstream will require LDPC, so let's consider how this plays out relative to S/I. Our revised threshold of 38 dB is for AWGN. For the upstream, where we allocate more margin, this is still a low error rate condition, as shown in Figure A-5 (appendix). A 40 dB SNR is the 1e-8 BER case for 1024-QAM.

The same modeling table above used for 256-QAM estimation indicates that for the 38 dB SNR case, an S/I of approximately 37 dB will leave a very correctable error rate in the 1e-3 to 1e-4 range. A 1e-5 threshold, based on Table 9c, suggests instead a 42 dB S/I. For the 37 dB objective, -2 dBc of interference successfully suppressed for 256-QAM in Table 9 then requires 35 dB of IC applied. This is *more* than 6 dB (by 3 dB) of additional IC given the 26 dB 256-QAM example. The same exercise for 4096-QAM, were it an upstream mode (using 44 dB SNR), would suggest a 42-45 (Table 9d) S/I, or 40-43 dBc of IC. This is 5-8 dB different than 35 dBc.

So, there is at this point only a range of additional IC expectations that seems to follow from these results. Intuitively, not being an AWGN impairment around which detection (and FEC) is optimized, we should expect that once the impairment is large enough to contribute to errors, the relationship would exceed 6 dB per modulation profile. Because of that and the sensitivity of FEC error rate curves to fractions of dB of SNR, it is probably a good starting point to consider a relationship such as 8 dB more IC capability until more granular modeling over a range of interference patterns and a larger sample size can be established.

Downstream Interference

In the downstream, a significant amount of study has already been performed to quantify the impact of CSO and CTB analog beat distortions on QAM. These distortions look like narrowband interferers, but the nature of their make-up (many independent distortion contributors falling close to one another) is that they have noise-like qualities, including an amplitude modulation component. This is meaningful for BER degradation. An example of a 256-QAM signal with analog distortion components from a 79-channel load is shown in Figure 20.

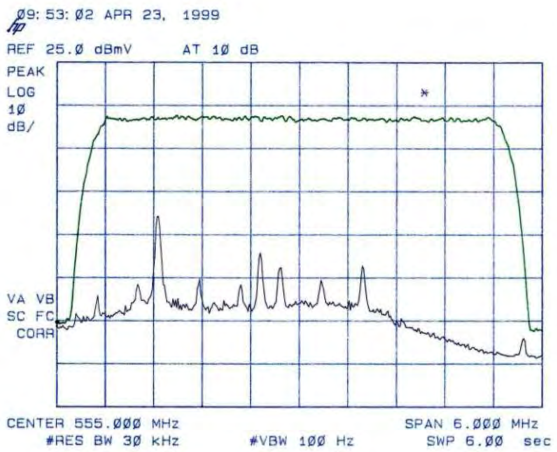


Figure 20 –256-QAM with CSO and CTB Distortions

The "donut" constellation we saw previously for 64-QAM in Figure 18 gives insight into why the beat distortion phenomenon when analog loads are prevalent matters from a QAM perspective. A CW interference example for 256-QAM at very high SNR is shown in Figure 21.

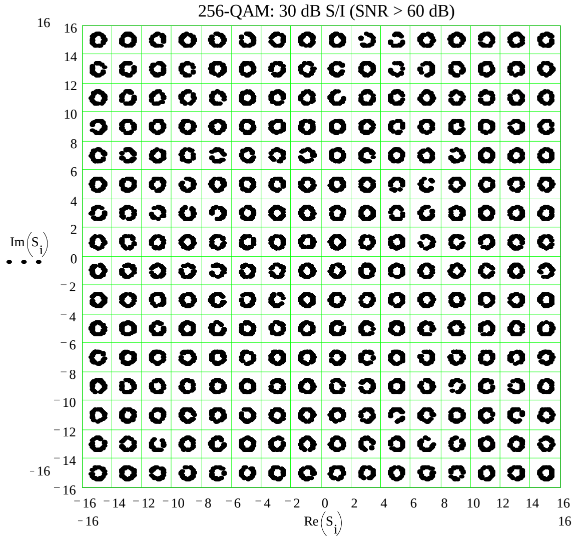


Figure 21 –256-QAM @ 30 dB S/I

In Figure 21, the SNR is 60 dB – an error free region in an AWGN-only environment, where 1e-8 occurs at SNR = 34 dB. Clearly, with only this CW carrier imposed on Figure 21, we still have an apparent error free environment. The peak-to-average ratio of a sinusoid is, of course, 3 dB – it has a constant envelope. If that envelope is not large enough to cause a decision error, then without noise there will be no decision errors even with the interferer.

The concerns with respect to CSO and CTB distortion beats are that, unlike CW or FM interference, they have a noise-like peak-to-average quality to them [20]. In fact, the envelope has a Rayleigh-like fit, which is representative of the detected envelope of a Gaussian process. This is shown in Figure 22 [20]. This amplitude modulating effect can be applied to the "donut" of Figure 21 – envision the circular symbol point breathing in and out. The nature of the distortion beat degradation is also that it is a narrowband process (10's of kHz) relative to the QAM bandwidth. Thus, a distortion beat sample will extend over many symbols in a row, and if it is high enough to induce decision errors, there is likely to be a burst of them.

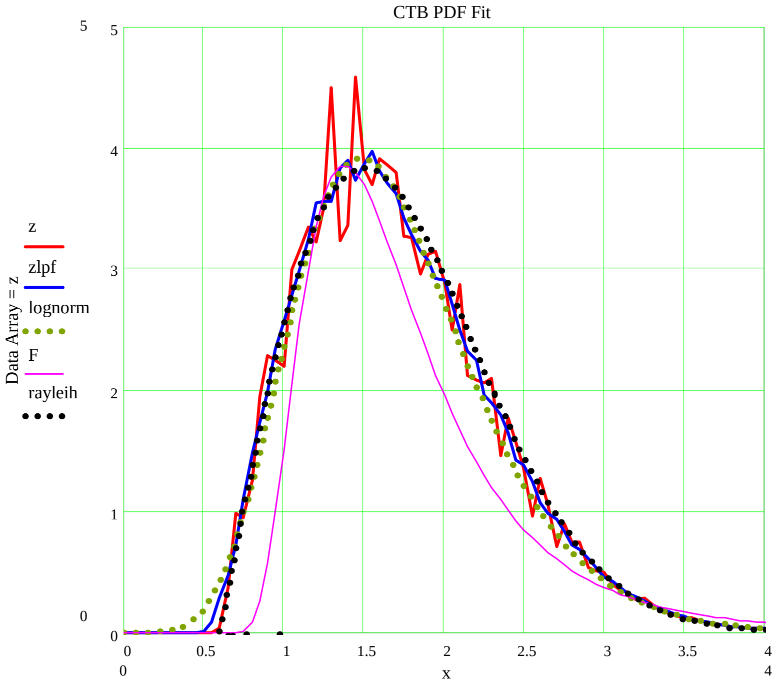


Figure 22 –Amplitude PDF of CTB

Fortunately, as we shall see and describe in more detail in the case of phase noise in the next section, in the downstream a powerful interleaver is available. It is capable of randomizing the errors, allowing the FEC to do its job better. This is quantified for phase noise, but the same dynamics apply in this case. A 20 kHz process has a 50 usec "time constant" of error generation, and a common (I=128, J=4) interleaver setting exceeds this by a factor of more than 5, easily distributing the errors into correctable codewords.

Analysis and test of the CSO/CTB impact for 1024-QAM has been performed [9]. A summary table of the results relative to these analog distortions is shown in Table 10. One of the key take-aways from that analysis – RF cascade depth as a function of analog carriers and amplifier performance – is shown in Figure 23.

For Figure 23, perhaps the single key result for purposes of this paper is that if analog video is reduced to 30 carriers, then the cascade depth that can be tolerated, under assumptions of 20 Log(N) degradation (which is overly pessimistic) is, for all practical purposes, unlimited for 1024-QAM using amplifier performance commensurate with today's plant equipment.

Table 10 – 1024-QAM Downstream Interference and Thresholds

	CW Interference		CTB Interference			
SNR	**Pre-FEC Error Threshold**	**Post-FEC Error Threshold**	**Pre-FEC Error Threshold**	**Post-FEC Error Threshold**	**Post FEC > 1E-6**	**Post FEC Broken**
50 dB	34 dB	33 dB	55 dB	55 dB	55 dB	45 dB
45 dB	35 dB	33 dB	55 dB	60 dB	55 dB	46 dB
40 dB	36 dB	34 dB	60 dB	60 dB	55 dB	49 dB
37 dB	38 dB	37 dB	60 dB	60 dB	55 dB	50 dB

Based on the CSO and CTB values we observed in Table 5, this is not surprising. The historical minimum acceptable value of 53 dBc works out to 47 dBc for 256-QAM, and this proved to be manageable with 256-QAM. With the 30-analog values of CSO and CTB ranging from 61-67 dB, 8-14 dB of better performance is occurring while a modulation profile increase of 6 dB is taken on for 1024-QAM. Also, Table 5 represents the performance at the worst case frequency. For CSO, these components tend to pile up at the low end of the band (analog). CTB tends to pile up in the middle, where QAM spectrum will be allocated – thus the focus on CTB below.

If we consider that 1024-QAM under an LDPC FEC has a threshold SNR of 31 dB based on our Table 2 downstream assumptions, then 60-something dB of CTB distortion, even with a noise like amplitude variation, would have little impact on an architecture delivering this SNR, or an SNR higher but not high enough to advance the modulation profile to 2k/4k-QAM. Basically, if 1024-QAM is workable with today's PHY, then better FEC can only make it better – it just may not achieve all of the new FEC gain of an AWGN-only noise. However, the interleaver, the relationship of the distortion levels to the CCN values in Table 5, the low pre-FEC thresholds observed in [9], suggest that it will achieve full benefits of the FEC.

We had already concluded in [9] that 1024-QAM was possible, and much more so at 30 analog carriers. What can we say about 4096-QAM?

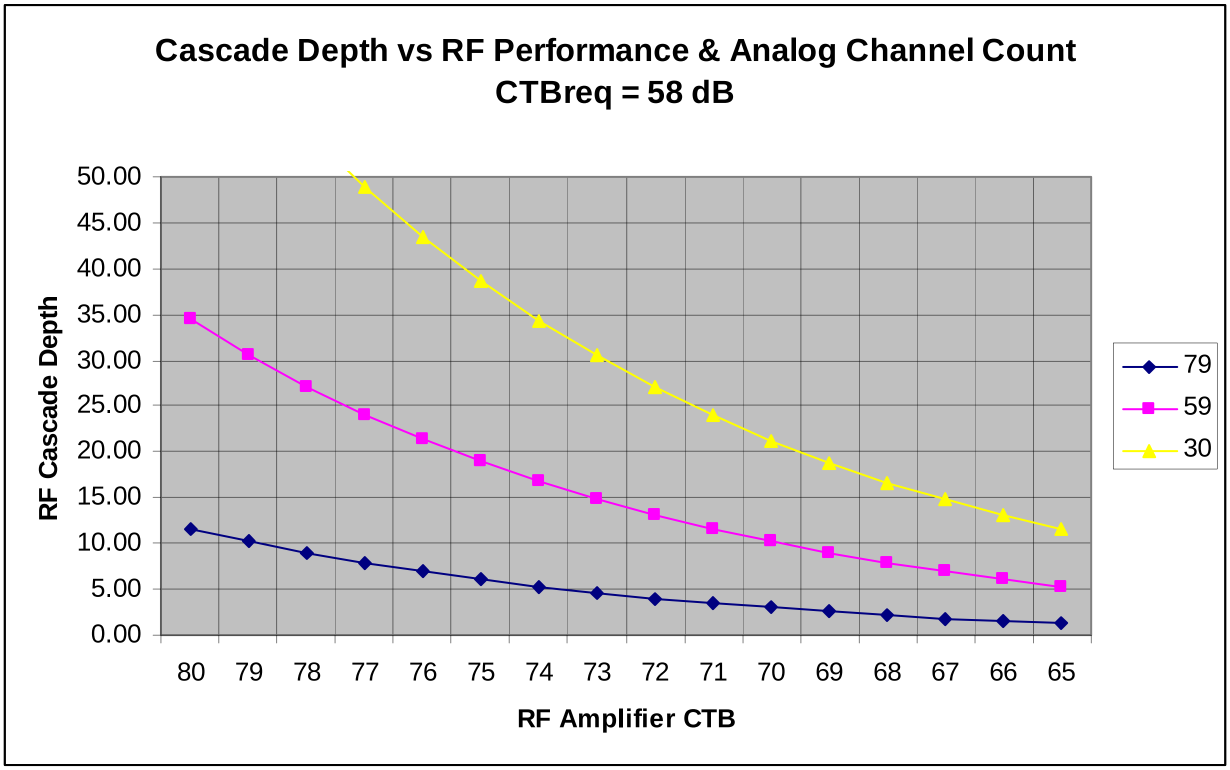


Figure 23 –RF Cascade Depth Limitations vs. Amplifier CTB for Analog Reclamations

First, a few items that may make us not care very much about that situation:

- When analog is fully removed, there are no longer any beat distortion components to worry about as interferers. We can then simply follow the CCN and the SNR analysis previously discussed.
- In the case of narrowband interference in the downstream, we could call on the upstream interference analysis above. It can only be conservative since the upstream must handle burst reception and adapt accordingly on a burst by burst basis. In the downstream, the receiver has the luxury of a constant input signal, a much simpler problem.
- Lastly, by the time we are deploying 4096-QAM, it is very likely that it is part of a multi-carrier downstream, and so narrowband interference analysis applies completely differently. We will discuss that later in the paper.

Figures 3-8 pointed out why the 4096-QAM case will require LDPC to be robust. It will provide the margin necessary to maintain performance against the 37 dB threshold established in Table 2. A high pre-FEC error count would ensue for 79-channel analog system at the 53 dBc minimum: The resulting narrowband interference would be 47 dBc on average, and peak to 33-35 dB, causing pre-FEC error rates worse than .01 (Figure A-5). They might be fixable, but this would not the ideal way to consume the FEC budget.

Therefore, 4096-QAM and full analog loading, we would suggest, is not a good combinations. 4096-QAM ought to be reserved for reduced analog loading or no analog loading. A caveat is that, as part of a transition plan of spectrum, such as Figure 1 indicates, may include new PHY above today's forward band, and beat mapping analysis would likely treat this region favorably in dBc of distortion reduction.

For reduced analog loading, CTB values of 66 dB (60 dB to QAM), peaking amplitudes would instead be 46-48 dB. The 1e-8 value for 4096-QAM without coding is 46 dB. So, it is likely errors will be counted, and corrected. They will be bursty, but the interleaving will arrange them nicely for the FEC. The situation is analogous to 256-QAM measurements with distortion in 2002 [20]. We now have 12 dB more of modulation profile sensitivity, and 13-14 dB better distortion values due to analog reclamation assumptions. Because they are noise like in effect on the constellation (MER is clouded), the similar relative relationship should yield similar results. This is also consistent with the expectation that 1024-QAM will perform very well in a reduced analog system from a distortion perspective.

PHASE NOISE

When QAM signals are put into the RF domain, they inherently have a phase noise mask applied to them. Phase noise is a measure of the spectral purity of the carrier signal itself. It is commonly measured by turning off the modulation on a waveform, and observing the level of noise surrounding the carrier at very close offset frequencies. At its most simplest, a perfect CW tone would be a single line in the frequency domain. In practice it cannot be perfect, and the amount that it is not perfect is quantified by this random phase modulation imposed at frequency offsets from the carrier frequency itself. The shape of the noise around a carrier is well understood, following many classic behaviors of semiconductor-based oscillators and the circuits that perform frequency synthesis to put modulated signals somewhere in the RF band. While there are many variables, in general, the higher the frequency, the worse the phase noise will be, the broader the tuning range, the higher the phase noise will be, and the finer the increment of tuning, the higher the phase noise will be. The shape is also known as a phase noise “mask.”

An example phase noise mask is shown in Figure 24 [10]. It illustrates some common characteristics – a close to carrier flat region and small peaking, and a region of 20-30 dB/decade roll-off. The flat region is often much wider than the example shown here. The “0” of the x-axis represents the carrier itself, and the data points plotted indicate offsets from the carrier where noise density is measured. The values at offset frequencies are given in dBc/Hz, and the total noise in a bandwidth is then the area under the curve of the mask, usually recorded as a dBc value or degrees rms.

Signal-to-Phase Noise Relationships

In this section we introduce some simple-to-understand M-QAM-phase noise relationships based on the qualitative descriptions above, some nomenclature, and a deeper understanding of the processes involved in determining its effects.

For converting dBc values of phase noise, or signal-to-phase noise ratios, to degrees rms, we have:

deg rms of phase noise = (180/π) sqrt[10^(-dBc of phase noise)]

The use of degrees rms is often very illustrative when we think about QAM constellations, as we shall see. There is a simple rule of thumb that keeps us from having to rely on the above equation. It is based on the recognition that a 35 dB signal-to-phase noise ratio is the same as 1° rms. Also, rms is a linear quantity, so doubling it is 6 dB.

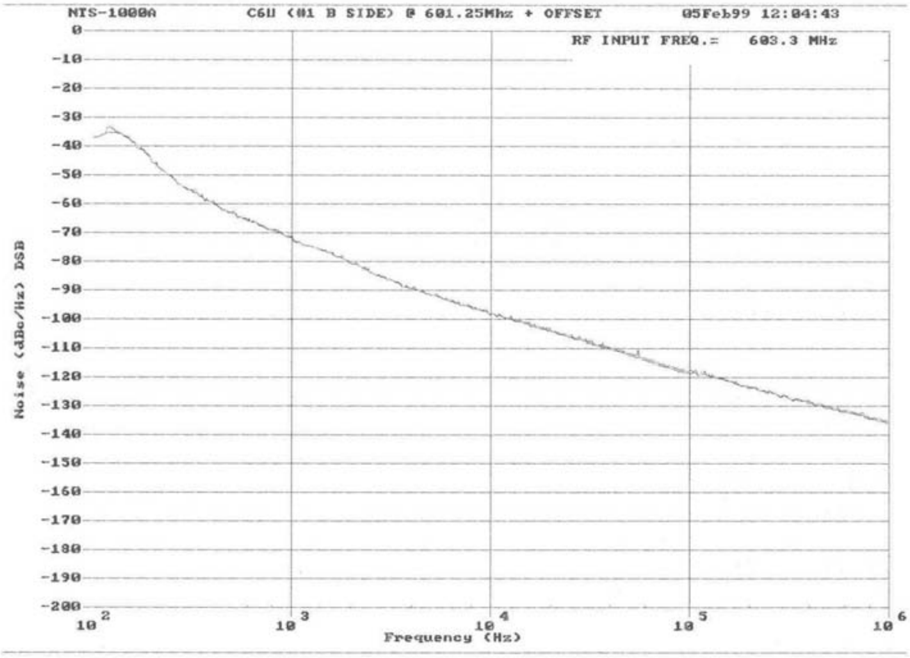


Figure 24 – Example Phase Noise Mask – RF Upconverter @ 601.25 MHz

For example, if -35 dBc = 1° rms, then

-23 dBc = 4° rms
-29 dBc = 2° rms
-35 dBc = 1° rms
-41 dBc = 0.5° rms
-47 dBc = 0.25° rms, etc.

Because of its rotational effect, phase noise affects QAM constellation points non-uniformly. Figure 25 [6] shows an example of essentially noise-free 64-QAM with 1° rms phase noise, using a mathematical tool on the left and a simulation environment (for error rate analysis) on the right. While the angle of rotation is the same for every symbol point, it is apparent and geometrically expected from polar coordinate mathematics how this impacts the outermost symbol points the most relative to breaching decision boundaries.

Compare Figure 25 with Figure 26, which shows the same 64-QAM symbol with only AWGN impairment, set at a 1e-8 BER. Note how the degradation due to additive noise is randomly distributed in I and Q dimensions, whereas the phase noise impact is exclusively angular.

Furthermore, the sensitivity of M-QAM gets worse with increasing M because of this non-uniform rotational effect. The shrinking of the distance to the decision boundaries for increasing M for a fixed average power puts makes the same amount of rotation more deleterious for higher M-QAM profiles. Figures 27 and 28 show a noise-free 256-QAM constellation with just 0.5° rms phase noise imposed, and a 1024-QAM constellation with a .25° rms phase noise imposed, respectively. The similarity in Figures 25, 27 and 28 of the relative rotation to decision boundaries, for the outer symbols in particular, is clear.

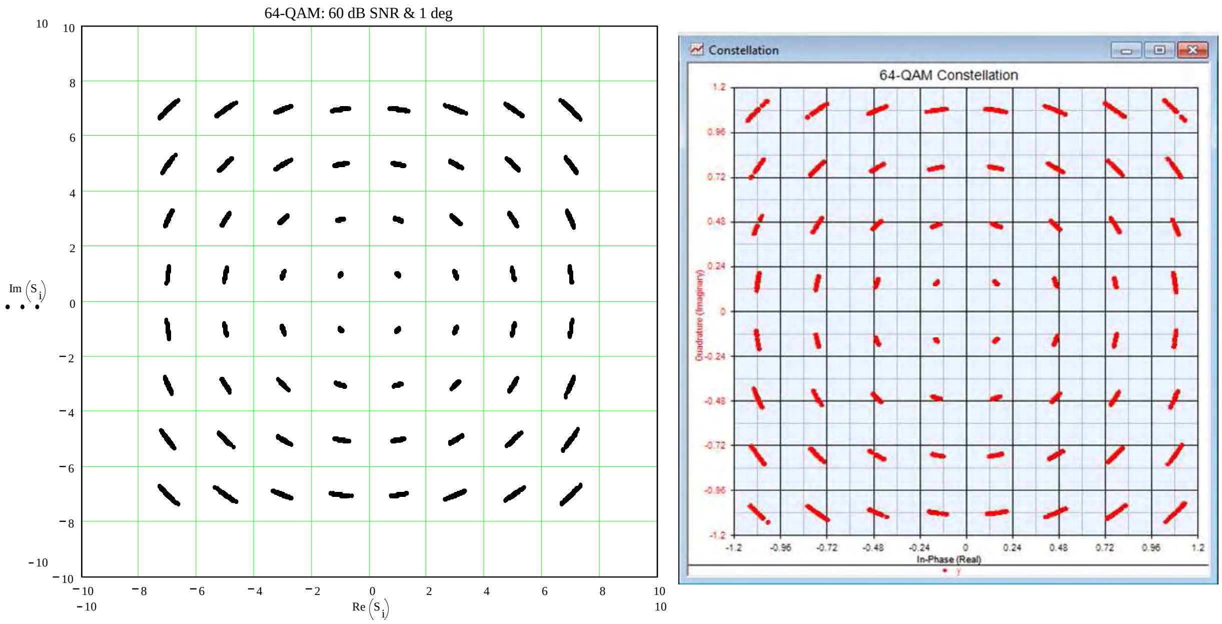


Figure 25 – 64-QAM, 1° rms Phase Noise (Analysis Tool, Simulation Tool)

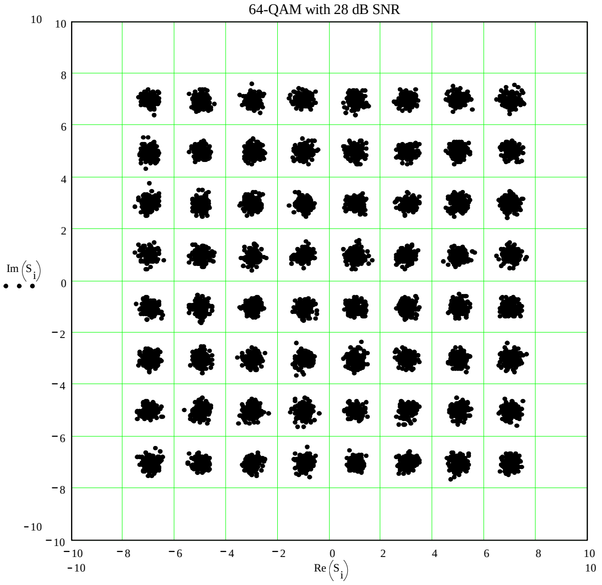


Figure 26 – 64-QAM @ 1e-8 Noise (AWGN) Level

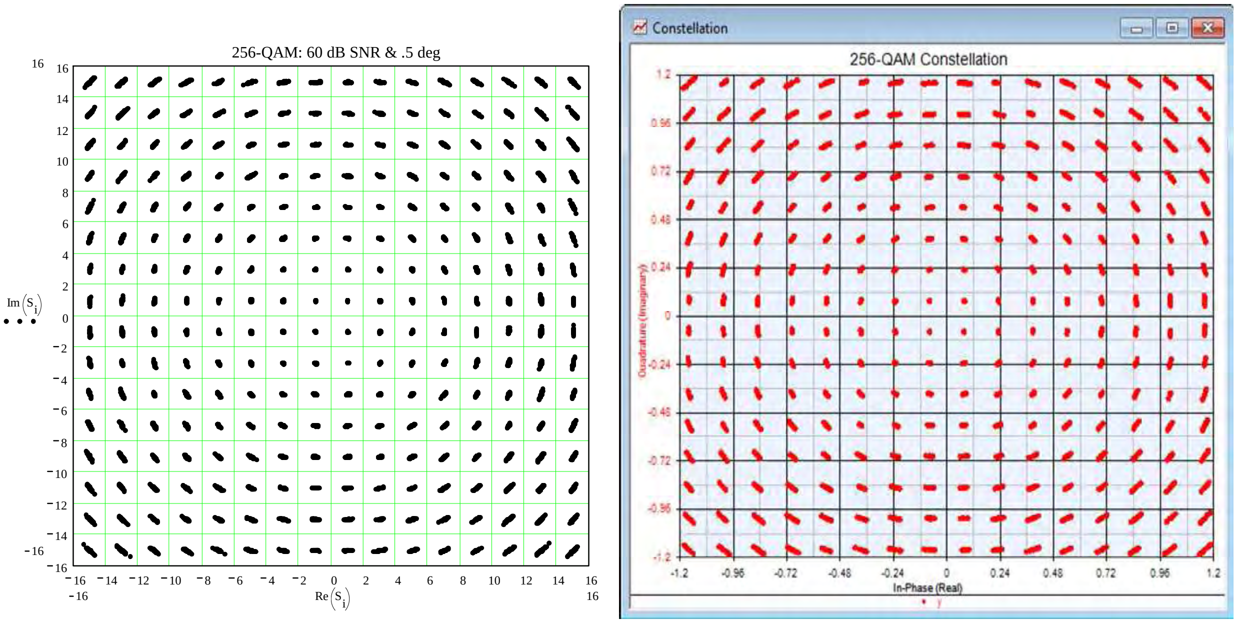


Figure 27 – 256-QAM, 0.5° rms Phase Noise (SNRφ = 41 dB), (Analysis, Simulation Tool)

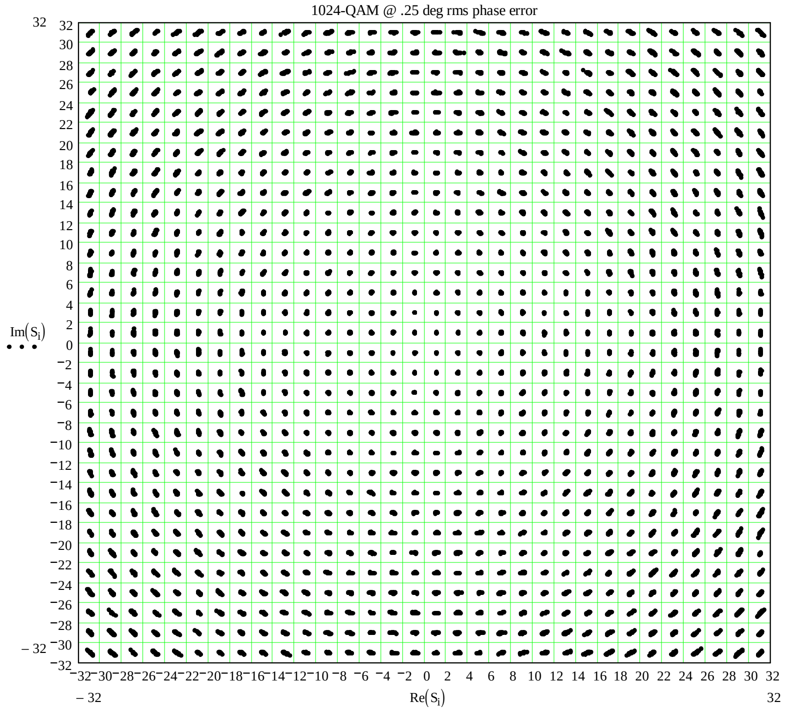


Figure 28 – 1024-QAM, 0.25° rms Phase Noise (SNRφ = 47 dB)

The nature of untracked phase noise is that it can lead to error rate floors at detection, because even without AWGN, if there is enough untracked phase noise after carrier recovery, it alone can cause symbols to cross boundaries. This is most likely for the outermost symbols, and these points can thus be used to determine limitations of phase noise necessary to eliminate flooring. More complex expressions are required to set thresholds associated with minimizing BER degradation [5, 6, 10].

It is a simple trigonometric matter to determine the rotational distance to a decision boundary as a function of M for M-QAM:

$$\varphi \text{ (decision boundary)} = \arcsin[(\sqrt{M} - 1) / M\sqrt{2}\,]$$

SNRφ Thresholds, M-QAM, and BER

Table 11 summarizes the phase error analysis across the modulation profiles of interest. It also identifies recommended levels of phase noise for minimal BER degradation, and levels beyond which the degradation curve shifts from a simple offset from theory to a more severe break from the normal steepness of descent of the BER waterfall curve.

Table 11 – Untracked Phase Noise Limits vs. M in M-QAM

	φ	dBc φ thresh	BER < 0.5 dB	SNR φ	BER on the Brink	SNR φ
16-QAM	16.8°	-10.7	1°	35.0	2°	29.0
64-QAM	7.7°	-17.4	.5°	41.0	1°	35.0
256-QAM	3.7°	-23.8	.25°	47.0	.5°	41.0
1024-QAM	1.8°	-30.1	.125°	53.0	.25°	47.0
4096-QAM	0.9°	-36.1	.0625°	59.0	.125°	53.0

The relationships shown can be deduced in part by recognizing that, since we are using a Gaussian statistical model for the jitter, the boundary merely represents a threshold on a normal curve that we can scale the rms (σ) to calculate its probability of threshold crossing. For example, a 16-QAM floor of about 1e-6 occurs for 4° rms, while for 64-QAM, a similar floor exists for 2° rms.

To exactly quantify allowable degradation with phase noise, AWGN and now phase noise can be combined together to create a composite "σ." However, they do not impact BER in a uniform fashion, as Figure 27 and 28 make apparent. Nonetheless, it is common approximation for lower order modulation formats to sum the AWGN noise and phase noise come up with a composite SNR. This simplification tends to understate the impact for high M, however.

Figure 29 shows a BER analysis that includes both phase noise (.25°, .35°, .5° rms) and AWGN contributions for 256-QAM [6]. The thresholds identified for 256-QAM in Table 11 are shown clearly on this chart by referencing the 1e-8 error rate threshold. The 0.25° rms value represents < 0.5 dB of degradation, while the 0.5° rms value has clearly is losing the characteristic waterfall shape, and on the verge of an error rate disaster.

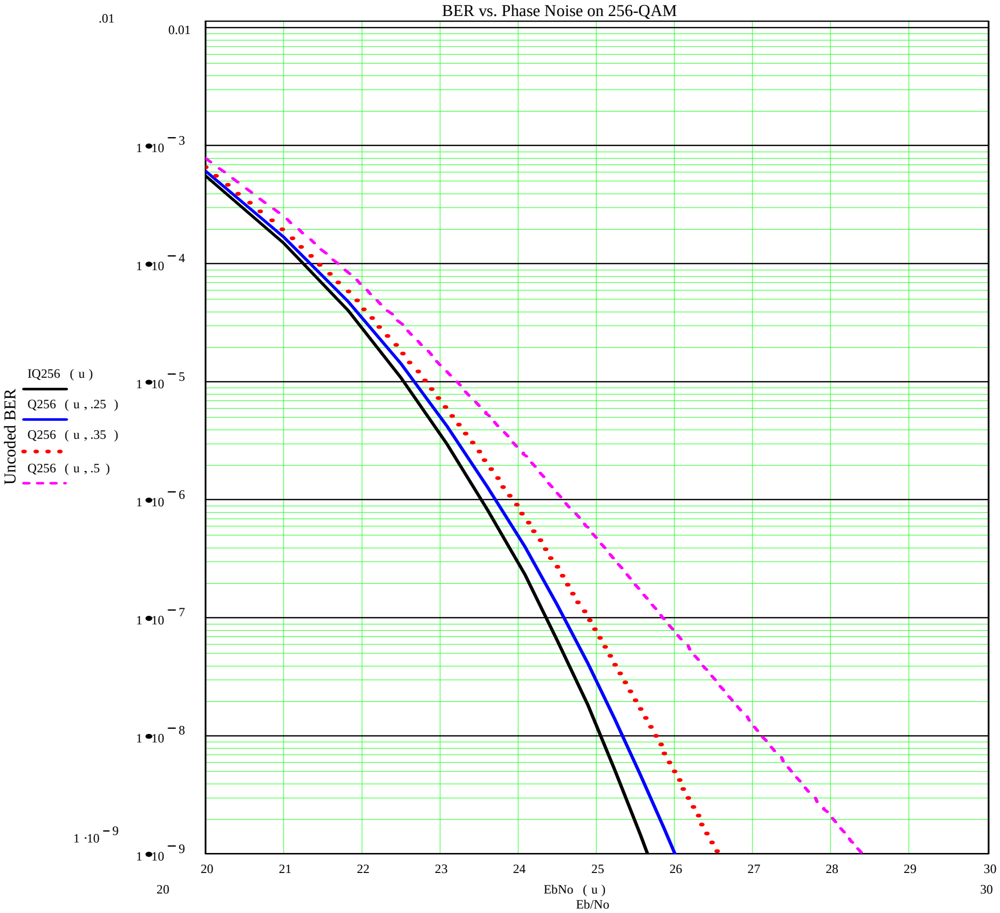


Figure 29 - 256-QAM BER with Phase Noise: .25°, .35°, and .5° rms

Note in Table 11 how the SNRφ required increases with increasing M. This relationship is similar to the relative relationship to AWGN. However, while a network architecture and new FEC may enable an AWGN performance improvement, the RF portion of the architecture that contributes to phase noise tends to remain in place and can be affected mostly in smaller ways by the tracking process. Redesign of RF equipment to achieve the same frequency agility objectives with improved phase noise is no minor proposition.

Offset M-QAM modulations and adjustments to decision boundaries in the face of a dominant phase noise impairment have been explored and this is addressed in [4] for the interested reader.

HFC Equipment Calculations

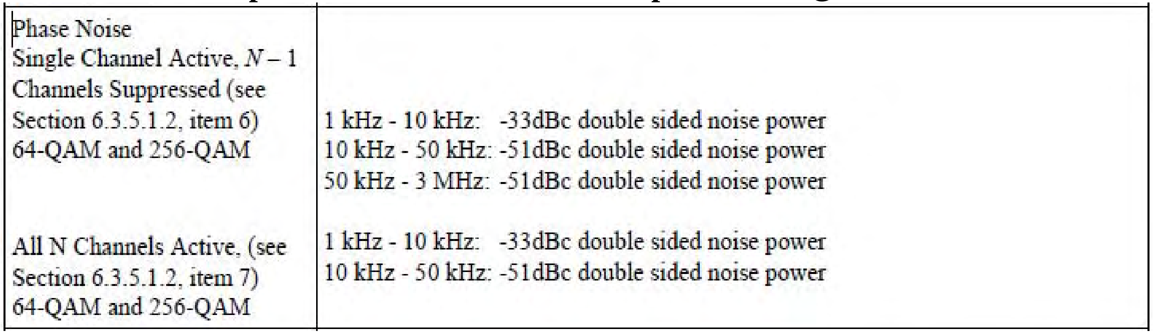
Phase noise is important because QAM, of course, encodes information in the phase of the symbol. A QAM signal contains "I" and "Q" orthogonal components, and the amplitude and phase applied to these identifies a point on a QAM constellation. This is why phase *noise* matters to M-QAM transport – noise in the phase domain translates to constellation position error or MER degradation in the signal space as we have seen in Figures 27 and 28.

The Downstream RF Interface Specification (DRFI), part of the DOCSIS portfolio of requirements, recognizes this and has a phase noise requirement, shown in Table 12.

All RF frequency synthesis or frequency conversion functions along the way contribute to the phase noise mask. The other typical major contributor in cable is the tuning function in the CPE. Though this function has been replaced in the RF circuitry sense by FBC technology discussed previously (wideband A/D conversion front ends), the clocking function of the A/D instead imparts the phase noise.

A modern wideband tuner built for digital cable and designed for compliance with ITU J.83A-C, is the Microtune MT2084. It has a specified phase noise requirement that serves as an excellent reference. These are shown in Figure 30.

Table 12 – Example Phase Noise Mask – RF Upconverter @ 601.25 MHz



Phase Noise Single Channel Active, *N* – 1 Channels Suppressed (see Section 6.3.5.1.2, item 6) 64-QAM and 256-QAM	1 kHz - 10 kHz: -33dBc double sided noise power 10 kHz - 50 kHz: -51dBc double sided noise power 50 kHz - 3 MHz: -51dBc double sided noise power
All N Channels Active, (see Section 6.3.5.1.2, item 7) 64-QAM and 256-QAM	1 kHz - 10 kHz: -33dBc double sided noise power 10 kHz - 50 kHz: -51dBc double sided noise power

Phase noise (SSB)
1 kHz offset -91 dB/Hz
10 kHz offset -92 dB/Hz
20 kHz offset -93 dB/Hz
100 kHz offset -105 dB/Hz
1 MHz offset -125 dB/Hz

Figure 30 – Sample Phase Noise Mask for RF Tuner in CPE

Since coherent QAM is used – meaning the carrier frequency and phase are recovered at the receiver in order to demodulate the signal and select which of the constellation points was transmitted, the final stage of "processing" of the phase noise mask occurs in the receiver. Carrier synchronization is performed by the carrier tracking subsystem. Modern designs use a decision-directed approach, which has been shown of the alternatives to have better noise performance, at least under low error rate conditions [18].

By tracking the carrier, a carrier recovery function is inherently also tracking the phase noise imposed on the carrier up to that point. However, it is a closed loop feedback system, and cannot track all of it without risk of other noise contributors, thermal and self-generated, from disturbing the stability of the recovery process. A feedback loop is in place which creates an error signal that is constantly adjusting the tuning oscillator to keep it aligned to the incoming signal. The feedback loop has a response time set by its loop bandwidth.

Without going into great detail about specific receiver architectures, the tracking occurs roughly up to the point of the loop bandwidth, and any RF-imposed phase noise beyond that is not tracked. There is an optimum bandwidth selection that considers these factors, input noise, and self-noise, among others. It is the total of untracked phase noise that contributes to MER degradation and possible symbol error. Note that DOCSIS specifications, in order to encourage innovation and competitive advantage among suppliers, allow flexibility on the receiver functions, where most of the complexity and sophisticated processing lie. As such, things such as loop tracking architectures or requirements are not defined, only end performance objectives under assumptions on channel conditions and other system assumption.

For the receiver designers, it is important to understand the associated transmit phase noise as defined in DRFI (CMTS or EQAM transmitter) and, for the upstream, in the DOCSIS PHY specification for cable modems:

-46 dBc, summed over the spectral regions spanning 200 Hz to 400 kHz
-44 dBc, summed over the spectral regions spanning 8 kHz to 3.2 MHz

Recall, the upstream has a range of symbol rates, beginning with 160 ksps (200 kHz) and increasing to 320 ksps (400 kHz) and so on in octaves up to 5.12 Msps (6.4 MHz). This range of symbol rates is reflected in the two requirements. We will assume wider symbol rates and thus the value of -44 dBc applies. Because of receiver design variations and the phase noise contributions from the receivers themselves, it is difficult to further quantify contributors to the phase noise process. We can say more will be added, some will likely be tracked out, and that the untracked jitter will have a lowpass structure to it.

We can at least, however, estimate what the specified requirements would mean to our M-QAM constellations, and estimate implications to receiver architectures using the requirements that are in place. Let's examine the effect of the combined DRFI specification and above tuner mask on the QAM profiles of interest. We have used the DRFI requirement in Table 12, and the tuner mask shown in Figure 30 to create a composite mask. This is shown in Table 13.

Table 13 – Composite Mask: DRFI + Tuner

Composite Mask (dBc/Hz)	
50 kHz	-90
100 kHz	-110
1 MHz	-130
3 MHz	-139
Total, dBc	-47

Assume that all of the mask beyond 50 kHz is untracked, and assume it is the dominant contributor to untracked phase noise after carrier recovery. It extends out to the symbol rate edge of 3 MHz (6 MHz double sided). This suggests a tracking bandwidth in the 50 kHz range, tied to other parameters [14, 15], and high SNR conditions in the carrier recovery architecture from self noise and input SNR.

As shown in Table 12 and 13, the composite mask is a 47 dB SNRφ, or .25° rms. The mask in Table 13 is shown on 64-QAM, 256-QAM, 1024-QAM, and 4096-QAM in Figure 31a-d.

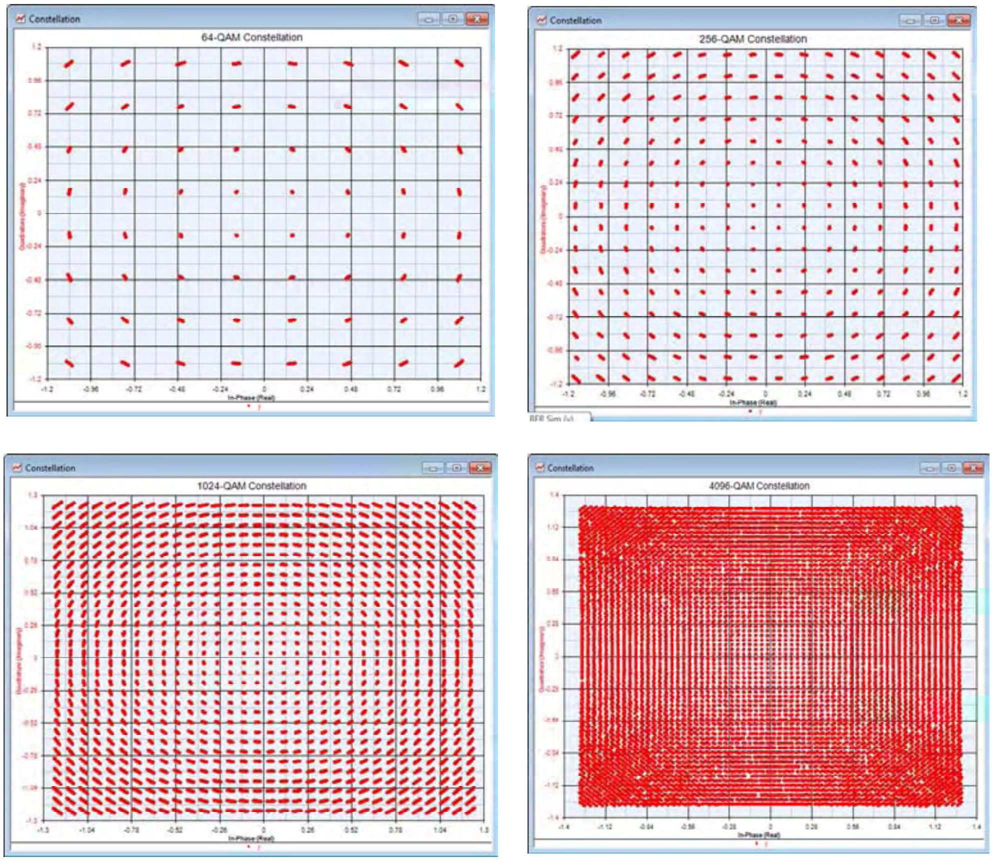


Figure 31(a-d) – Table 13 Mask Applied, Clockwise from Upper Left: a) 64-QAM b) 256-QAM c) 1024-QAM and d) 4096-QAM

In Figure 32, we have evaluated the uncoded BER for M-QAM profiles for M=64, 256, 1024 and 4096-QAM under the 47 dBc SNRφ conditions of Figure 31.

As Table 11 indicates, SNRφ = 47 dBc is the breakpoint between small degradation for 256-QAM, and the BER being on the brink of large degradation for 1024-QAM. The 4096-QAM case is untenable with this amount of untracked phase noise. This is all verified in Figure 32.

For 4096-QAM, which is clearly suffering and in practice would not be able to effectively hold the receiver locked for demodulation, it is difficult to see much in Figure 31 other than clouds of impossible-to-discriminate symbols. This case is shown again by itself in Figure 33, where you can begin to see some daylight between symbol points, mostly inner points. But, the symbol clouds at the edges are still massively intruding on each other's space to the point that they are becoming indistinguishable. This yields the very high symbol error rates, likely to overwhelm an error correction mechanism or carrier recovery subsystem.

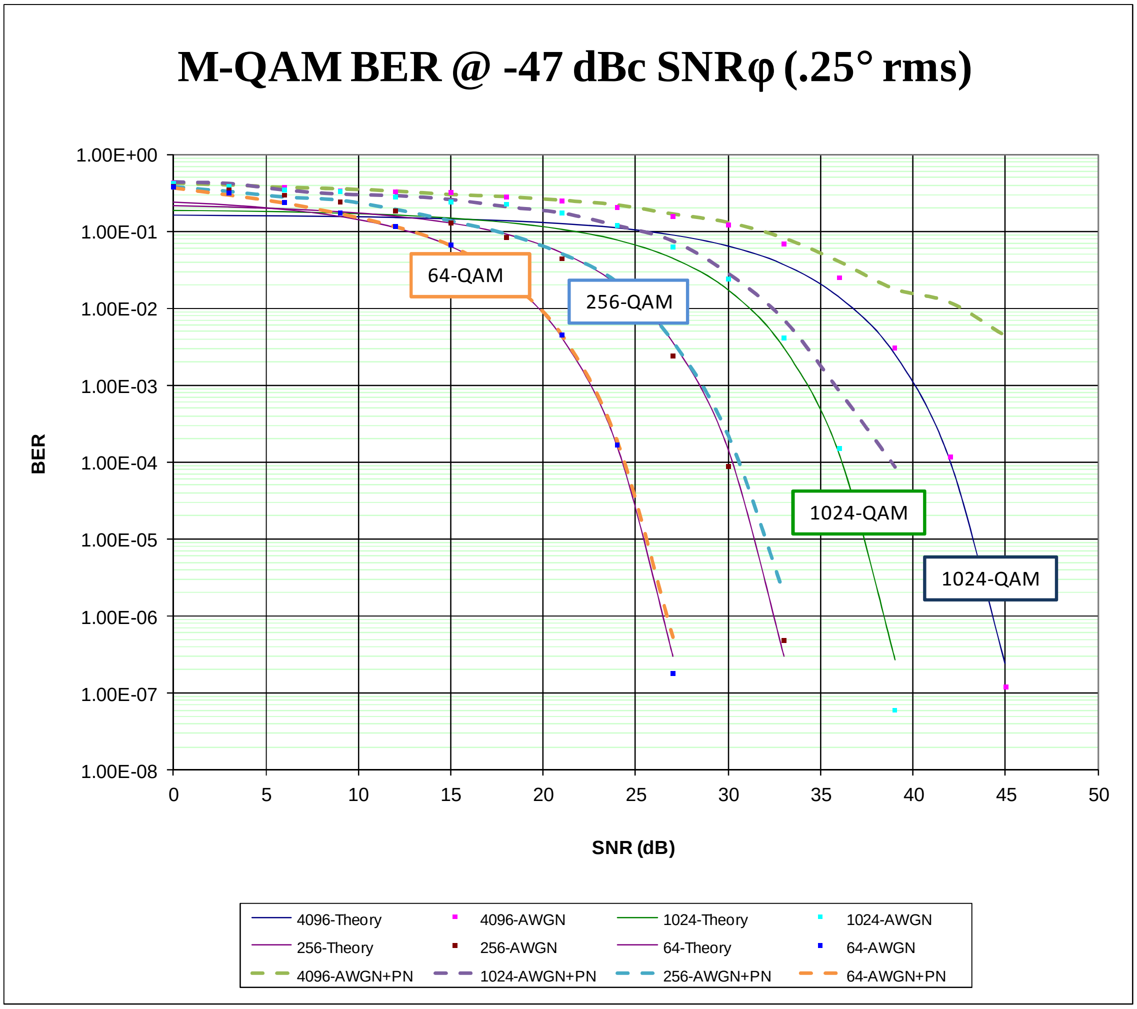


Figure 32 – M-QAM BER for SNRφ = 47 dBc (DRFI + Tuner)

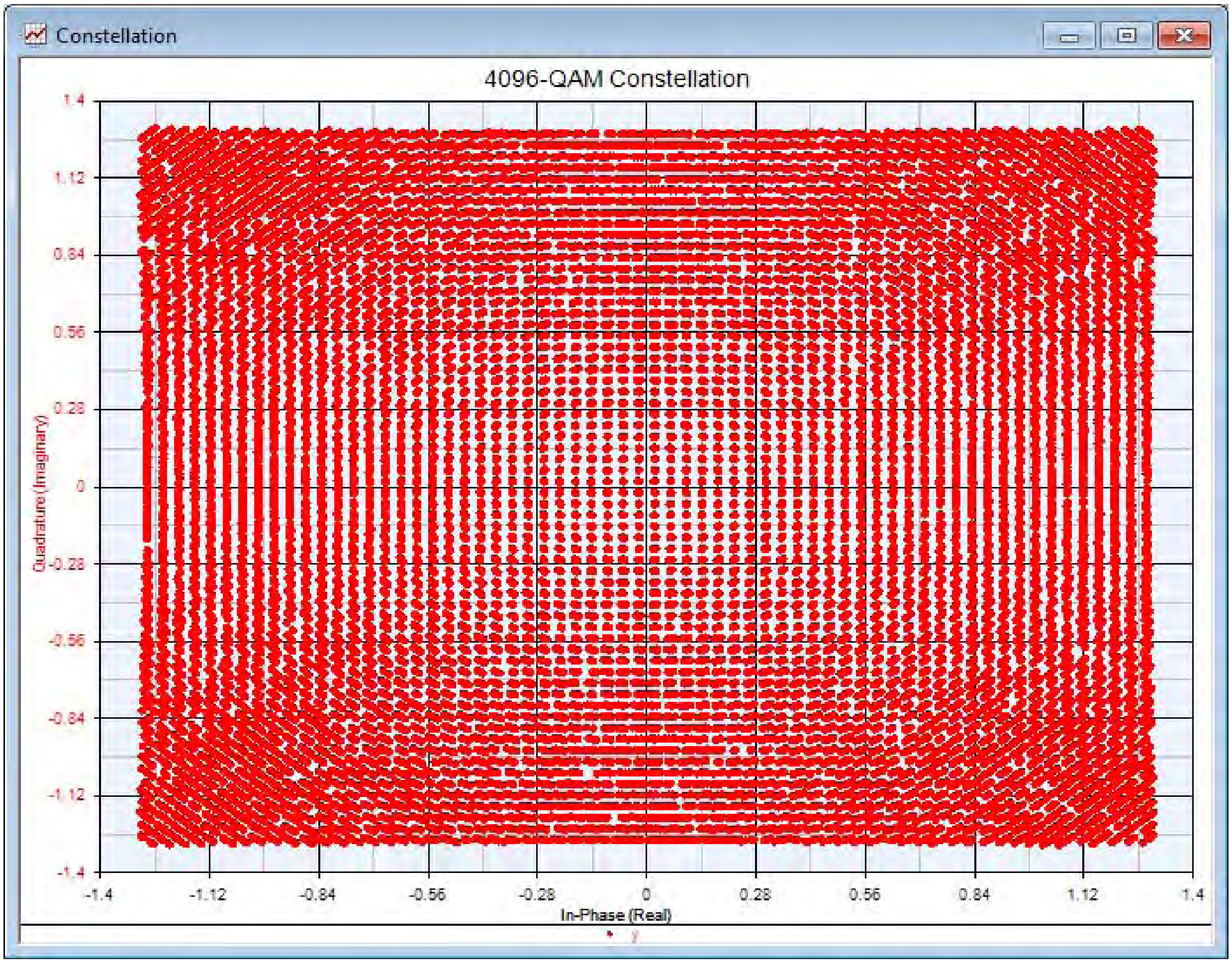


Figure 33 – 4096-QAM@ .25° rms (SNRφ = 47 dB)

According to the guidelines of Table 11, SNRφ = 53 dB is the brink of trouble for 4096-QAM, and a reasonable guideline for 1024-QAM that limits the degradation.

This 1024-QAM case, with SNRφ = 53 dB, is shown in Figure 34. The similar, relative MER characteristic compared to Figure 31b (256-QAM @ 47 dBc) is apparent. The 4096-QAM case for SNRφ = 53 dB is shown in Figure 35.

The BER evaluation for SNRφ = 53 dB is shown in Figure 36.

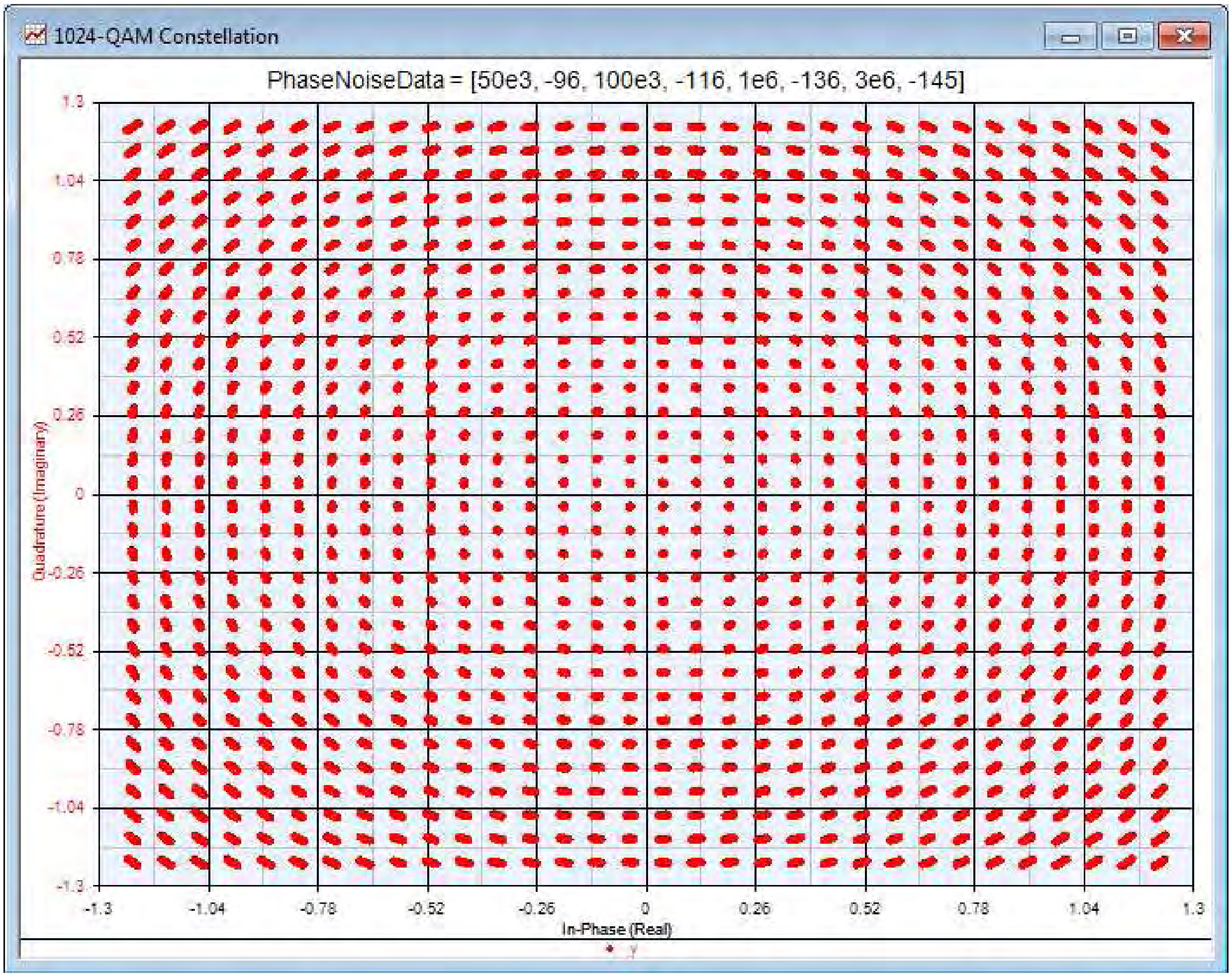


Figure 34 – 1024-QAM@ .125° rms (SNRφ = 53 dB) (Recommended)

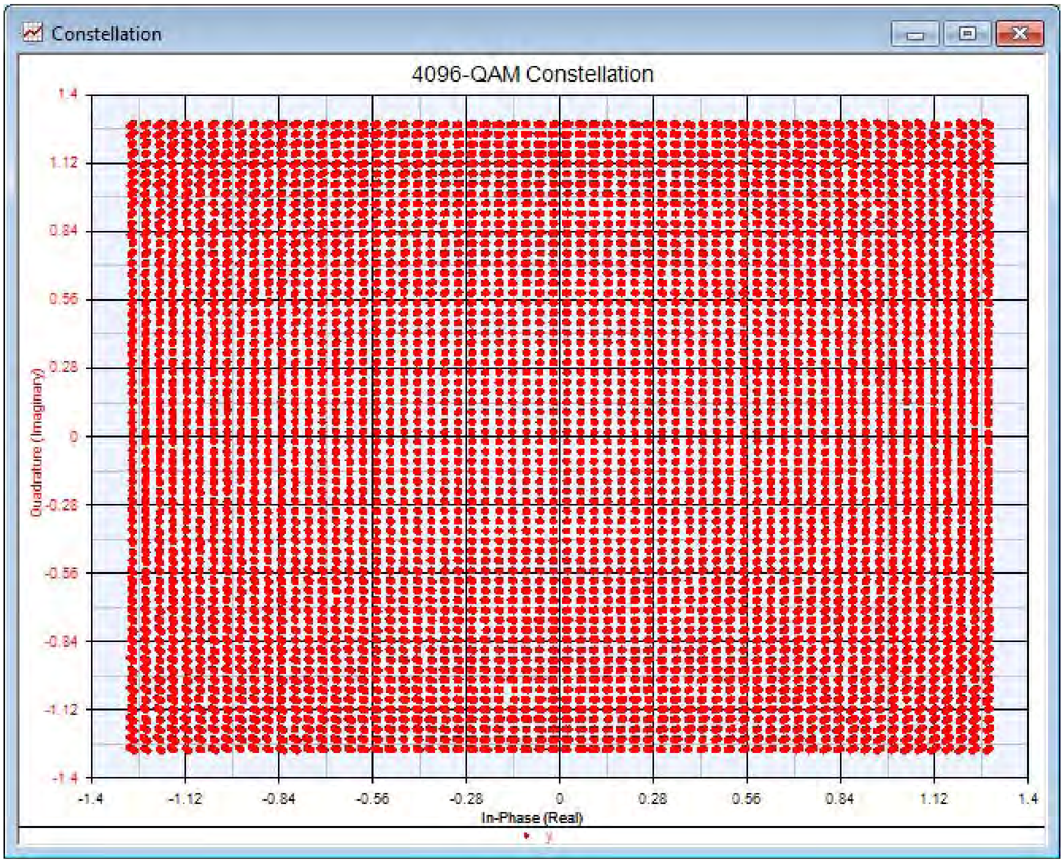


Figure 35 – 4096-QAM@ .125° rms (SNRφ = 53 dB)

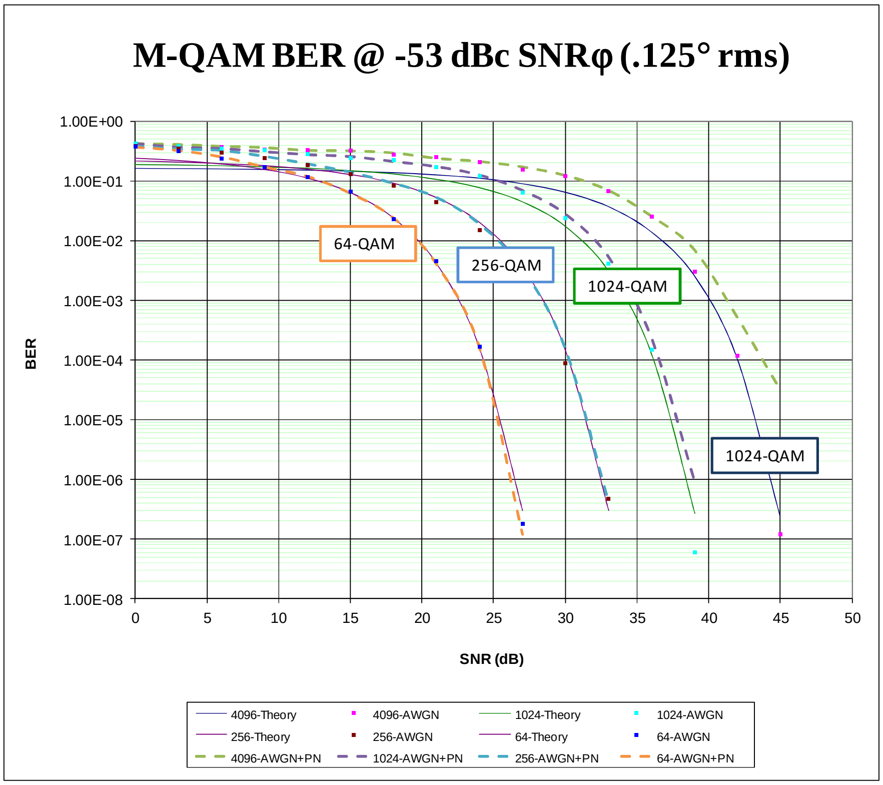


Figure 36 – M-QAM BER for SNRφ = 53 dBc

In Figure 36, we can see that 1024-QAM is now under control with modest degradation, and that 4096-QAM is on the edge of major BER performance degradation. This again is consistent with the recommendations in Table 11.

Finally, Figure 37 shows the constellation impact to 4096-QAM with the recommended maximum phase noise of SNRφ = 59 dBc, or .0625° rms. Of course, link phase noise is not going to adjust for the modulation profile, so the RF and tracking subsystem must be architected for the most sensitive modulation anticipated. The improved fidelity in Figure 37, in particular of the outer symbol points, illustrates why SNRφ = 59 dB is recommended for minimizing degradation against the theoretical performance curve.

Figure 38 shows the BER evaluation for the SNRφ = 59 dBc case, where it becomes clear that the 4096-QAM untracked rms phase noise recommendation of Table 11 is sufficient.

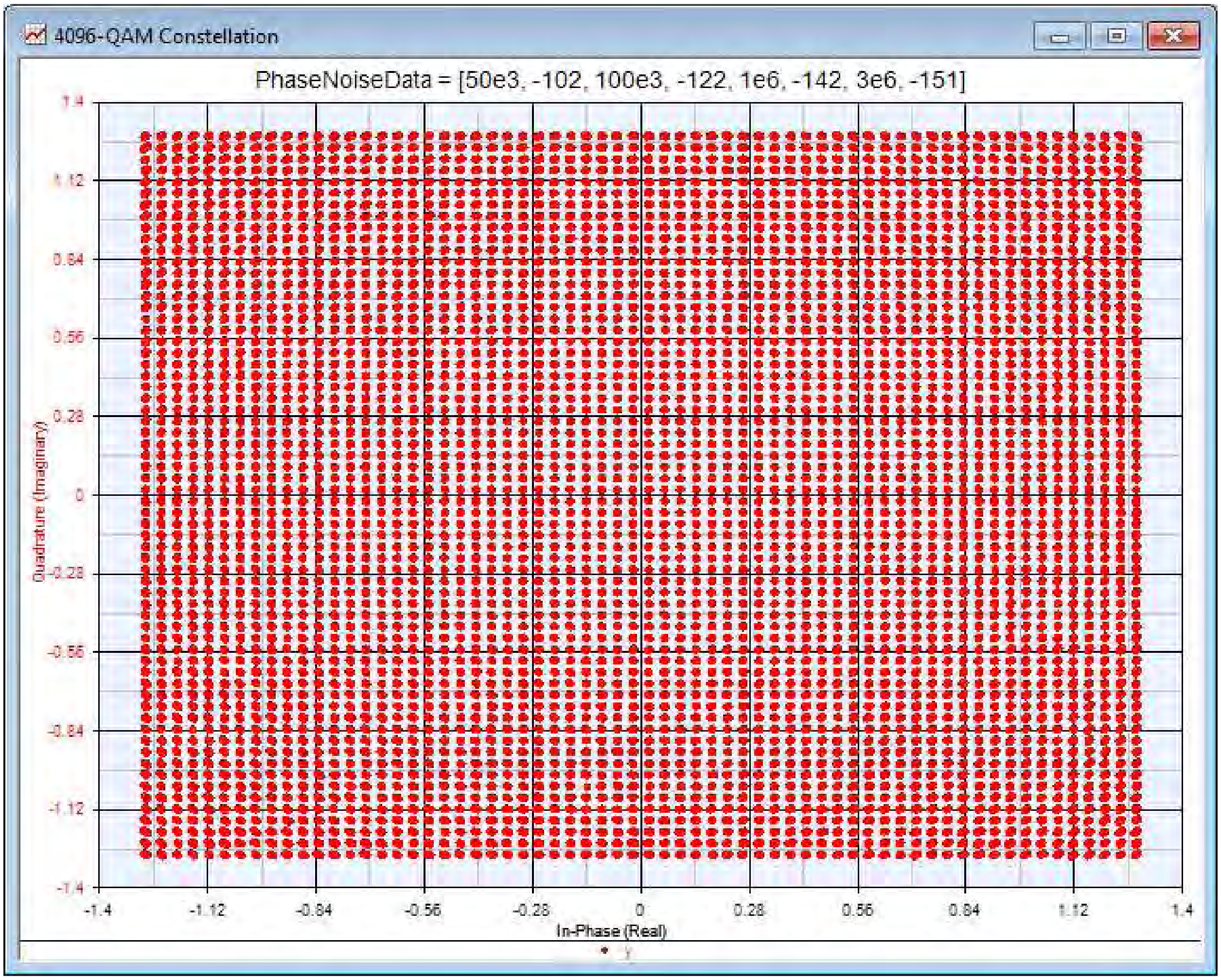


Figure 37 – 4096-QAM@ .0625° rms (SNRφ = 59 dB) (Recommended)

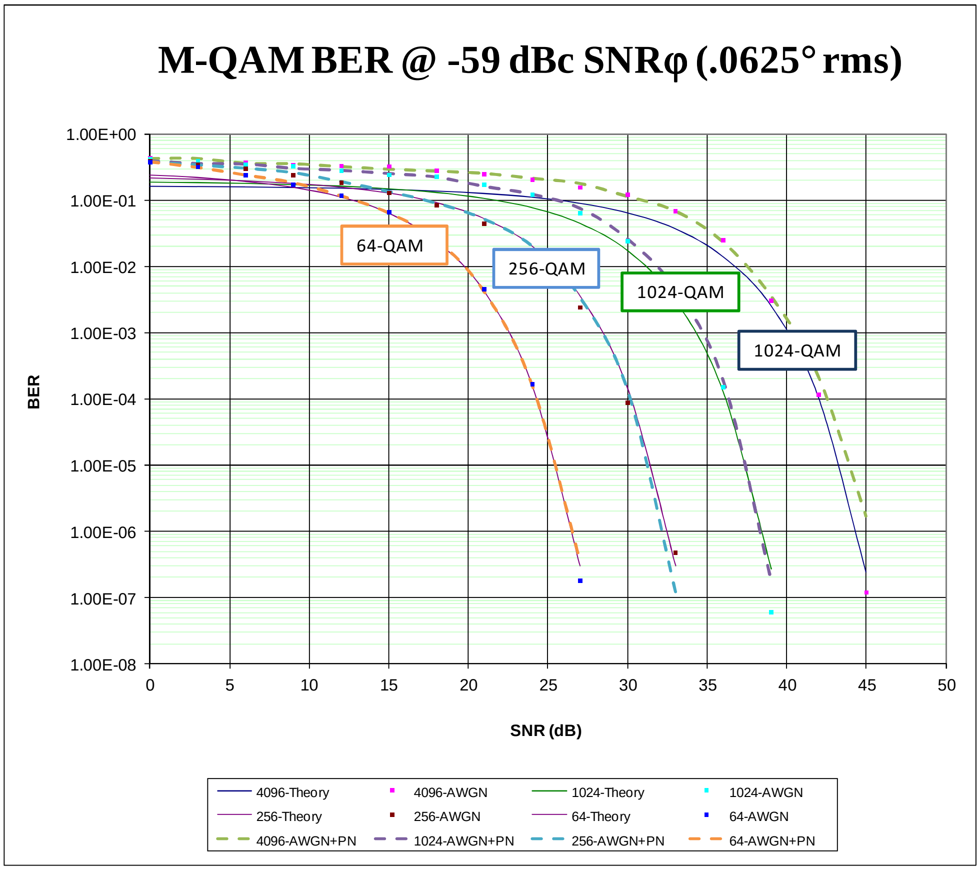


Figure 38 – M-QAM BER for SNRφ = 59 dBc

Post-Detection Processing

For high SNR systems, the loop bandwidth can be, relatively speaking, quite wide. However, it is nonetheless narrow compared to the symbol rates of single carrier QAM signals used in cable. This is important because it means that if there is enough phase noise to contribute to misplacing a symbol in the constellation, it will misplace potentially a large consecutive set of them for single-carrier systems. Because the loop bandwidth is much lower than the symbol rate, a sample of phase noise will be in about the same relative phase location for many symbols in a row – including when the sample is near a decision boundary or across one altogether. This is often referred to as the “slow” phase noise assumption, and is a common characteristic of single carrier QAM systems. The result is that phase noise, as an error mechanism itself, is bursty in nature. This puts pressure on the receiver to have burst correction either via FEC and/or interleaving. Reed-Solomon encoding is burst correcting, but the encoder in the J.83B downstream is only a t=3 symbol correcting design. It instead relies on the interleaver to provide a randomization of the symbol errors to make the RS decoding more effective, spreading out a burst of errors across codewords.

Fortunately, at least in the downstream, J.83B defines a very powerful, configurable, interleaver. It can configure burst protection from 66 usec to 528 usec (Level 2 mode with I = 128) at the expense of introducing latency. The lowest latency value is (I = 128, J = 1), where I and J describe the register structure used to feed Reed-Solomon codeword bits in and out. This setting provides 66 usec of burst protection at the cost of 2.8 msec of latency. Real time voice is the service that is typically most carefully watched for the latency budget, and 2.8 msec can be accommodated easily in a budget that targets around 50 msec typically one-way. I-128, J=4 is a recommended setting, contributing 11 msec of latency in exchange for 264 usec of burst protection.

The symbol rate of 5.36 Msps (256-QAM) works out to 187 nsec symbol periods. Using 50 kHz to represent the rate of the phase noise process, its “period” (it’s a noise process, so period is loosely used) is about 20 usec, or 107 QAM symbols for 256-QAM. The interleaver spreading exceeds 20 usec even for the lowest latency setting. Therefore, the interleaver is a very powerful helper against phase noise impairment - provided the native error rate is low to begin with.

The right-hand side column of Table 11 identifies rms phase noise thresholds that are at the edge of the native BER curve remaining stable. Because of the interleaving downstream, this column could be considered a target objective for the maximum allowable phase noise if it is within the budget of the FEC to support the error contributions from phase noise in addition to other channel impairments it may have been designed to protect against.

Measured performance is available for 1024-QAM in a pseudo “J.83” mode [9]. As shown in Table 14, pre-FEC errors are measured, and these are associated primarily with clipping and phase noise. In each case, however, post-FEC error rate is zero – meaning that the combination of the interleaver and RS FEC was able to completely eradicate any burst errors that may have been caused by the introduction of phase noise.

Unfortunately, in the upstream, we have potentially higher phase noise contributions specified, although it is specified over different ranges that may allow more of the transmit contribution to be tracked. This cable modem requirement was for SNRφ = 44 dBc. Upstream is likely to rely on lower orders of modulation, however, such as being limited to 1024-QAM. However, we have identified the 1024-QAM SNRφ threshold as 53 dBc, or 9 dB better than the cable modem requirement. This has important implications to the carrier recovery requirements in the burst receiver.

The upstream Reed-Solomon FEC is more powerful than the downstream, but still would not be capable of spanning a phase noise induced degradation of 50 kHz of noise bandwidth, much less as low as 8 kHz. This is more than five times the span, so represents about 5 times the number of symbols in error in a row at the highest upstream symbol rate – this likely outlasts the average burst size upstream entirely.

Table 14 – Pre-FEC 1024-QAM Error Rates with Zero Uncorrected Codewords

			1024-QAM Carrier Frequency		
			603 MHz	747 MHz	855 MHz
QAM @ -4 dB to Analog		MER	39.6	39.2	38.9
		BER	6.1E-08	1.12E-07	3.76E-07
QAM @ -6 dB to Analog		MER	39.0	38.9	38.6
		BER	1.5E-07	2.6E-07	2.5E-07
QAM @ -8 dB to Analog		MER	38.3	38.2	37.7
		BER	4.30E-07	2.02E-06	3.48E-06

As such, if impaired by phase noise, post-FEC results should register some low level of uncorrectable codewords. Since FEC is not a source of burst protection from a phase noise perspective, nothing is lost in moving from a RS-based FEC scheme to LDPC.

Note that SNRφ = 44 dBc is about .35° rms, which is plotted in Figure 29 for a 256-QAM – the state-of-the-art throughput available today [12]. In theory, this amount of untracked noise would lead to slightly less than 1 dB of degradation in an uncorrected BER curve. Based on the burst dynamics above, a post-FEC result would register some low level of uncorrectable codewords if there was a phase noise-induced BER contribution measureable. However, in [22], it is shown that there is error-free pre-FEC and post-FEC performance of 256-QAM upstream with SNR = 36 dB. However, this is consistent with the curve for .35° rms even if the SNRφ = 44 dBc is *all* untracked. Thus, it is not possible to learn whether or how much of an rms error reduction takes place in the carrier tracking process.

For purposes of upstream evolution, then, such as beyond 256-QAM, it is impossible to tell from 256-QAM performance, without additional measurements, whether there is adequate margin in the untracked rms phase noise to support 1024-QAM. However, without question, the current CM specification of -44 dBc over the specified bandwidth would be wholly inadequate without the ability to remove substantial induced phase noise in the carrier recovery process. This suggest these requirements may need to be updated to go beyond 256-QAM. The BER curve for 0.5° rms for 256-QAM in Figure 29 would be a reasonable approximation to the trajectory that the 1024-QAM BER would take for 0.25° rms, and this would of course get worse for 0.35°, meaning it would induce more than 2 dB of degradation at low error rates. Without interleaving, there would not be an opportunity to correct for this degradation in the upstream, so this bears consideration. Upstream phase noise for 1024-QAM may create the need for updated requirements.

In summary, to advance the modulation profiles, the phase noise requirements identified in DOCSIS and DRFI may need to be reconsidered to provide the spectral fidelity necessary to support very bandwidth efficient QAM transmissions such as 1024-QAM upstream and 4096-QAM downstream. In the upstream, 256-QAM has been shown to be supported today. It is inconclusive whether or not the phase noise margin contribution that is today adequate for 256-QAM is sufficient also for 1024-QAM. There is no mechanism in place to handle the burst noise environment phase noise-induced errors can create.

In the downstream, a measured post-interleaver, pre-FEC error floor suggests that there is some residual phase noise impact that is being handled well enough by these burst correcting mechanisms. Additionally, as shown in Figure 32, the error rate performance against the combined DRFI and tuner mask would be very poor for 4096-QAM – so badly so that the ability to manage an effective decision-directed tracking loop and successful decoding process is likely to be compromised. While the interleaver is very effective, there is an inherent assumption of low error rate to avoid overwhelming the interleaver-dispersed errors and the carrier recovery subsystem. Performance recommendations for total untracked rms phase noise for 1024-QAM and 4096-QAM are shown in Figure 34 and 37. Under these conditions, there would not be a heavy reliance on the interleaver, FEC budget, or concern about the sensitivity of decision-aided tracking robustness.

MULTI-CARRIER MODULATION

OFDM Applications to HFC

The industry is considering, as part of the IP transition, adopting a new RF waveform and fundamentally changing the access method away from a line-up of 6 MHz frequency domain multiplexed (FDM) slots. Wideband, scalable MCM or OFDM is being considered for the next generation of RF over HFC, for many of the reasons discussed previously about expanded bandwidths and RF channel uncertainties. There are many acronyms in use that describe an implementation of the same fundamental core concept: lots of narrowband carriers instead of one wideband carrier. Figure 39 illustrates the OFDM concept.

Historically, OFDM applications have been linked by a common thread – unknown or poor RF channels. Virtually all modern RF systems implement some form of MCM – 4G Wireless, MoCA, G.hn, HomePlug AV, 802.11n, and VDSL. The differences are based on the medium and channel conditions expected affecting the band of operation, subcarrier spacing, modulation & FEC profiles, bit loading dynamics, and whether the system is multiple access in the sub-channel domain (OFDMA). Table 15 lists some common Pros and Cons of OFDM.

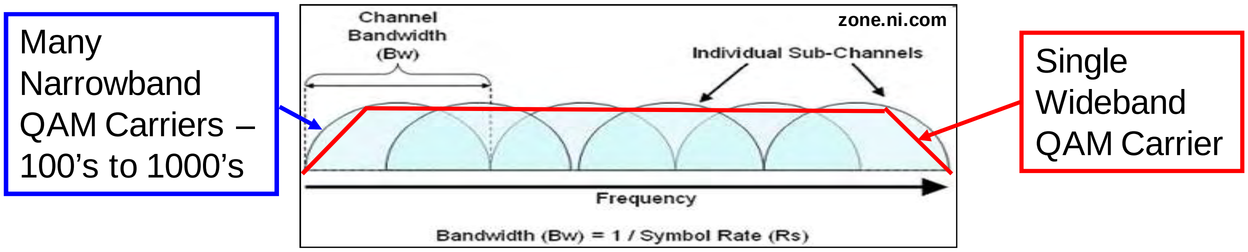

Many Narrowband QAM Carriers – 100's to 1000's
Channel Bandwidth (Bw)
zone.ni.com
Individual Sub-Channels
Frequency
Bandwidth (Bw) = 1 / Symbol Rate (Rs)
Single Wideband QAM Carrier

Figure 39 – Fundamental Concept of Multicarrier vs. Single Carrier

Table 15 – Pros and Cons of Orthogonal Frequency Division Multiplexing

Pro	Con (*or Comment*)
Optimizes Capacity of Difficult Channels	High Peak-to-Avg (CPE issue); (*PAR reduction schemes exist - adds OH*)
Simplified Equalization against Frequency Response or Multipath	(*Cyclic Prefix = Guard time OH*)
Robust to Narrowband Interference	Avoidance Approach – Throughput Penalty by Deletion or Mod Profile
Robust to Impulse Noise	*(Similar Principles as S-CDMA - Time Spreading and Parallel Transport)*
Modern Ease of Implementation – IFFT/FFT DSP functionality	Complexity Increase for Shaping and Wavelet schemes – trade-off C/I vs. ISI
Simple Co-Existence via Flexible Subcarrier Allocation (and Power)	Backward Compatibility with DOCSIS
More Spectrally Efficient Wideband Channel than FDM Can be Multiple Access (OFDMA)	Potentially More Sensitive to Synchronization Noise Such as Carrier Phase Jitter (loss of orthogonality)

The most powerful advantage of OFDM has been that it shines in difficult or unpredictable channel environments. With the increasing ability to do computationally complex operations in real time, OFDM implementation – once an obstacle – has become a strength through simple IFFT/FFT functionality that forms the core of the transmit and receive operations.

For HFC, of course, this primary advantage is worth a closer look. The HFC downstream is one of the highest quality digital RF channels available – it is very low noise, and very high linearity. Such channels benefit very little in performance from OFDM, and probably not enough to justify introducing a new waveform if modulation efficiency of today's forward band was the only thing at stake.

However, as discussed, operators are looking for places to exploit more spectrum, and the channel quality of extended coaxial spectrum will be less predictable. This makes OFDM well-suited to be introduced in this part of the downstream band above 1 GHz as shown in Figure 1. Then, as the IP transition moves ahead and legacy 6 MHz slots are eliminated, the spectral flexibility of OFDM through allocation of its subcarriers becomes an especially valuable transition tool.

The HFC upstream, of course, does have a troublesome part of the band at the low end of the spectrum to which OFDM is a good fit. Today, the solution available to exploit capacity here is S-CDMA, which is just now seeing growth in interest and field deployment as upstream spectrum become congested and there is nowhere else to go, but down (in frequency). Like S-CDMA, OFDM should be robust at the low end of the band if properly designed for the impulse and narrowband ingress environments in that region of the return.

Unlike the downstream, above the low end of the band, as the upstream is extended above 42 MHz, there is likely to be steadily *improving* channel conditions, at least up to the FM radio band of 88-108 MHz. As in the downstream, this part of the upstream band – the extended upstream – may have a less obvious need for the primary poor-channel performance value of OFDM. But, there are some unknowns, and the FM band looms. And, since DOCSIS carriers will exist for many, many years, the flexibility of spectrum allocation once again makes OFDM worthy of consideration in this changing environment.

A second well-earned "pro" for OFDM is the simplicity with which poor frequency response can be combated. We discussed this as part of the capacity discussion in the beginning of the paper. However, OFDM also makes difficult multi-path channels more manageable. The HFC network is prone to "multi-path" in the form of micro-reflections associated with impedance mismatches that occur naturally over time and unnaturally through the fact that, as discussed in the section on the POE home gateway, every home in the plant is also part of the access network. Unlike the mobile application, the "multi-path" is static or nearly so. Nonetheless, because the upstream is burst mode from a randomly located source, it has dynamic characteristics associated with the allocation of time slots to modems that are basically on a single frequency but have individually dependent, but unique channel characteristics. OFDM enables the simplification of the equalizer function in these cases.

Perhaps the most talked about disadvantage of OFDM is its inherently high peak-to-average-power ration (PAPR). An OFDM signal is a collection of independently modulated carriers, all sent at once. As such, the composite waveform has noise-like qualities. This is a potential RF concern, as it requires more linearity, or higher P1dB, in the transmit power amplifier stages compared to single carrier signals to ensure the waveform does not get clipped and distorted. PAPR is primarily an issue for CPE – more transmit power headroom translates to more hardware cost. The same is true in principle (the need for more headroom) for the OFDM receiver, but the receive side is rarely tested from a distortion standpoint, processing very low level signals such that the dB differences have much less impact to design and cost. Schemes that encode subcarriers in a way that reduce PAPR have been developed.

OSI Layer 1 Standard?

An emerging analogy for OFDM is to liken it for OSI Layer 1 what Ethernet and IP are for OSI Layers 2 and 3. A complete, modern Layer 1 PHY is emerging as Multi-Carrier QAM with LDPC-based Block Code. The combination of the two drives implementation very close to theoretical capacity, so there is little else to optimize. There is a natural convergence of solutions towards this combination to yield the highest throughput efficiencies for a given channel. System parameters around the OFDM implementation would vary by application as a function of channel characteristics, as do the block sizes used for the LDPC code. The large number of modern systems based on OFDM is another important factor driving towards a PHY layer "standard" approach.

HFC is no different in this regard – looking for optimal ways to extract capacity on channels that are ill defined or know to be potentially troublesome, while adapting around legacy signals. OFDM or MCM is an alternative suited to these objectives, and some variant is a likely final evolution phase of the coaxial last mile, as introduced in [7].

Channel Impairments and OFDM

As discussed, for AWGN channels, the results relative to SNR and architectures above applies directly. While the HFC channel is never "only" AWGN, this assumption applies well to the HFC spectrum that generally represents the "good" part of the spectrum. There are often modest linear distortions comfortably handled by straightforward time domain equalizer structures. OFDM may achieve high performance with less implementation complexity in these cases, but single and multi-carrier systems would otherwise perform very similarly. The same can be said for the upstream, although the adaptive equalizer complexity in the upstream is much greater, so the weight of a simplifying architecture may be of more value.

However, it is interesting to point out that the maximum attainable bit rate expression on a channel for a given SNR is approximately the same for multi-carrier and single carrier when equalized by a DFE – the approach used today for the DOCSIS upstream [1]. The key difference, again, is that as the channel gets more difficult in terms of frequency response, the theory holds up well, but scale of implementation favors multi-carrier, and more so the worse the channel conditions become. In both cases, feedback from receiver to transmitter fully optimized the bit rate attainable.

Let's take a look at some of the impairment scenarios we quantified for single carrier and discuss how they relate to multi-carrier.

Signal-to-Interference

Single carrier techniques combat narrowband interference by notching the band through adaptive filtering mechanism, as previously discussed. OFDM, on the other hand, deals with narrowband interference by avoidance. Subcarriers that are imposed upon by an interferer are notched out our dialed down to a more robust (less bandwidth efficient) modulation profile that can be supported, all as part of an adaptive bit loading algorithm. The effect is a capacity loss, but generally a modest one. Note that single carrier ingress cancellation requires overhead itself (lost capacity) in order to operate.

CW Interference

A simple example of the capacity loss for OFDM can be calculated by recognizing the sub-channel spectrum for OFDM when implemented in a pure FFT-based architecture, the simplicity of which being one of the reasons it has become so attractive. The spectrum of a single FFT-based sub-channel is shown in Figure 40. The roll-off of the Sin(x)/x response is slow – 6 dB per octave – so the "in-band" rejection can as a result be low when an interferer falls onto a sidelobe of a particular sub-channel. The first sidelobe has the commonly referenced 13 dB relationship to the main lobe response

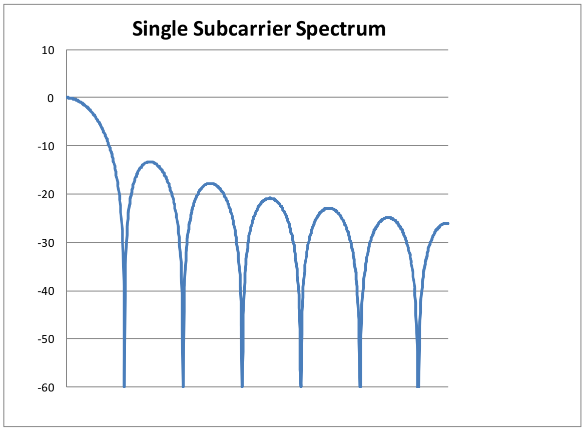


Figure 40 – FFT-Based OFDM Subchannel Spectrum

In Table 7, we noted that the A-TDMA ingress cancellation function had a limit of S/I = 10 dBc for zero corrected codeword errors for single CW ingress signal at an SNR = 35 dB. Three FM interferers could be as high as -15 dBc each (total S/I is still about 10 dBc). How would FM interference at -10 dBc affect the OFDM capacity? Figure 41 shows how several adjacent subcarriers appear in the midst of a CW interfering tone (not to scale of the numerical example).

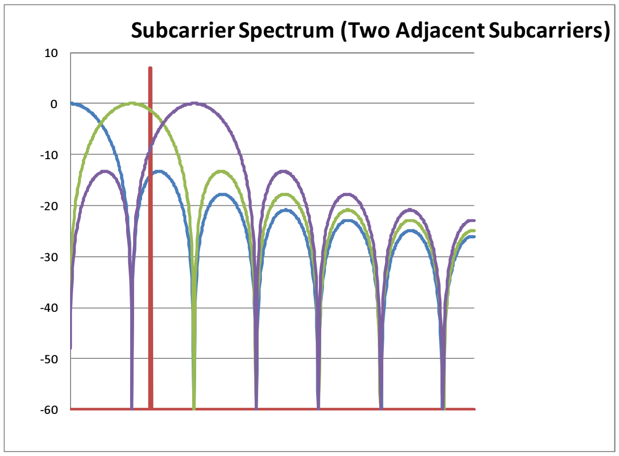


Figure 41 –CW Interference in an OFDM Channel

Let's assume channel SNR conditions are high, such that 64-QAM can be deployed. Going back to our 8k point FFT, each sub-channel is (1/8192) of the total, so the S/I on a per-sub-channel basis is (10-39) = -29 dBc on a subcarrier. Simulations performed such as were shown in prior results for 64-QAM indicate that a required 25 dB S/I for error free BER in this high SNR condition (35 dB). This would require 54 dB of rejection. If the interference coincides with a sub-channel frequency, then it does not interfere with adjacent sub-channels because of the same orthogonality properties that ensure that the sub-channels do not interfere with one another. However, this is unlikely. The worst case is it is just off center of a sub-channel so exposed to the envelope of Sin(x)/x roll-off of the spectrum in Figure 41. For rejection of 54 dB, this occurs at about 160 subcarrier indices away (320 total). If these sub-channels are all nulled, and all FFT sub-channels are used for payload, then the lost capacity is about 3.9%.

If instead of muting, for example, the adaptive bit loading tries to implement 16-QAM where possible, requiring only 20 dB S/I, then only about 160 sub-carriers *total* are lost, or 1.95% of capacity is lost to muting. There are then (320-160) 160 new subcarriers carrying 16-QAM, which works out to 0.65% of lost capacity, for a total of 2.6%. This is an improvement over muting all of them. Another subset could use 8-QAM, QPSK, etc. This is precisely how OFDM is handy for optimizing under varying channel conditions.

Alternatively, all of the above analysis was performed without considering new error correction. Our “new” SNR requirement is 26/32/38 dB for 64/256/1024-QAM, respectively. Simulations like those already discussed show that for these SNRs, we can arrive at S/I conditions that leave codeword error rates that are easily correctable (1e-4 or lower), for delivering low PERs. This would be a different set of dBc values to meet, but the same approach to the calculation of the carrier indices effected. This approach allots FEC “budget,” built around AWGN performance, to correcting for interference induced errors, which would come at the expense of SNR to some degree. Error rate curves are very, very steep compared to classic uncoded waterfall curves, so this type of analysis trade-off would require careful simulation and test.

Modulated Interference

Let’s assume the interference is FM modulated. Again referring to Table 7, zero correctable codewords required a 15 dB S/I for 64-QAM with high SNR. Now, however, at 20 kHz wide, its bandwidth is roughly that of an entire sub-channel, so looks like a noise floor increase.

Figure 42 shows the implications of an additive “narrowband” modulated interferer when applied to OFDM (not to relative dBc scale of S/I = 15 dB). On a per-sub-channel level, it represents -24 dBc. The main difference in the analysis approach is that we no longer need to refer to S/I behaviors to quantify how much rejection is needed.

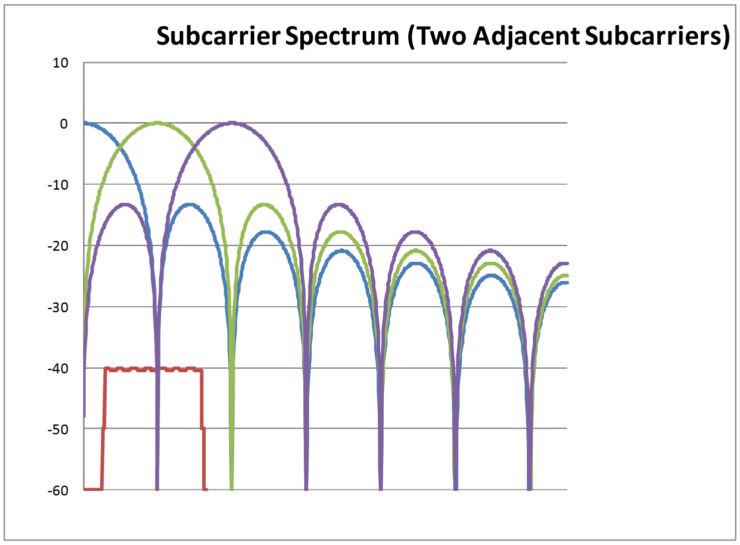


Figure 42 – Modulated “Narrowband” Interference for OFDM

We can treat it instead as a noise floor addition and refer to QAM profile performance against SNR. Modulation that creates a broad noise floor (relative to the sub-channel) and AWGN would not have the same precise effect, but that model is more relevant than a CW S/I model. Based on the thresholds used in Table 3 for upstream and 64-QAM (26 dB), the lost capacity would be close to the CW case. The differences would be in the S/I of a carrier index needing to reach 26 dB vs. 25 dB S/I, and the weighting that would apply for a broad spectrum applied versus a narrow carrier when passed through an FFT receiver.

As previously discussed, a reasonable argument can be made that the margin allotted to the QAM profiles in Tables 2 and 3, which are based on today's single carrier upstream channels, can be decreased *because* of a multi-carrier technique, as OFDM would be naturally more resilient to some of the items that contribute to the margin allotted.

Downstream Distortion Beats

We have noted that CSO/CTB interference has been a cause for concern for downstream QAM. We have also noted that, under the assumptions of analog reclamation, the CSO/CTB levels decrease dramatically. And it was noted that the bandwidth of these distortions is on the order of tens of kHz [20]. So, like the modulated interference previously described, this type of distortion is on the order of a sub-channel bandwidth for OFDM.

For full analog reclamation, the only number that matters becomes CCN, as all the distortions themselves become digital and spread across the spectrum in a noise-like fashion. Unlike the case of the FM interference above, beat distortion have an amplitude modulation component. Under the assumptions in Table 5, the worst case CTB indentified is 66 dBc. On a per-sub-channel basis, this becomes 27 dBc. Considering the noise-like peak-to-average in analysis makes sense for single carrier QAM because the distortion is "slow" in relation to the bandwidth, so peak samples exist for symbol after symbol. In the case of OFDM, the distortion bandwidth is on the same order of the sub-channel width (in this example), so noise averaging takes place just as symbol detection averaging does. As in the modulated interference case, we now can compare this 27 dBc to the modulated thresholds in noise environments to arrive at the impact to OFDM subcarriers.

Referring to Table 2, then, 256-QAM on a sub-channel interfered by this level of CTB would be supported in high SNR conditions (AWGN + CSO/CTB do not exceed the 25 dB shown). No capacity is lost in this case due to CTB for 30 analog carriers and 256-QAM. For 1024-QAM and 4096-QAM, however, threshold SNRs were identified as 31 dB and 37 dB. In both cases, only the sub-channel or two where the CTB falls will be impacted, since the spectral roll-off provides enough rejection (13 dB minimum) to meet these two SNR requirements.

This number of effected channels, too, can be calculated, as distortion noise "lumps" occur at periodic increments – two CSOs and two CTBs every 6 MHz (see Figure 20) – so there will be over 100 sub-channels imposed upon in the 192 MHz example discussed. This would be a maximum of 3.1% of the channels (128 = (192*4/6)). However, since we have shown that these channels could support 256-QAM, the capacity impact is only 0.62% for 1024-QAM and < 1.1% for 4096-QAM (less than 1.1% because some of the sub-channels could likely use 1024-QAM).

The same argument previously made about the FEC budget can be made here, which applies in particular for 4096-QAM. The SNR thresholds are based on AWGN, so it is the combination of AWGN and new noise contributors like CTB that should meet these thresholds. The "high SNR" assumption would then assume that this beat distortion is then the dominant effect in the sub-channels where it appears. When this is not the case, the offset for the addition of noise power must be made to guide the modulation profiles that can be supported.

For example, adjacent channel rejection of 13 dB from the 27 dBc example is 40 dBc. If our AWGN performance is 40 dB, supporting 4096-QAM with new FEC, then the two combine to 37 dBc, the 4096-QAM threshold identified in Table 2, and no capacity is lost in the adjacent channel. If instead we were already maximized at 4096-QAM with a 37 dB SNR, then the 40 dBc pushes the composite SNR closer to 35 dBc. In this case, a 2048-QAM profile may be required to have a robust channel performance.

Lastly, again, this combination of impairments, when coupled with sharp error rate functions that swing orders of magnitude on a dB of SNR difference, requires robust simulation to quantify precisely.

We nonetheless conclude that, in the case of a partial analog reclamation, OFDM should support the advanced modulation formats with little capacity degradation to do distortion interference.

Other Multi-carrier Approaches

Various shaping techniques and use of wavelets for orthogonality have been studied to reduce the effect of narrowband interference. However, these add complexity, and waveforms that provide narrower frequency response are inherently creating longer symbols in the time domain, so negatively affect performance on dispersive channels. By defining expected channel conditions, an optimum balance of time domain and frequency domain robustness can be implemented.

For an OFDM system for HFC, there is still some homework to be done on the system optimization side to determine the sub-channel spacing, shaping, and channel conditions anticipated and specified as new HFC bands become part of the RF channel definitions. We can count on improvements in SNR and downstream distortions, but updated frequency response and impairment models need a careful examination to ensure the HFC flavor of OFDM is optimized for its channel the way other OFDM-based systems in the wireless and wireline world have been optimized in their applications.

Phase Noise

OFDM creates an interesting scenario with respect to phase noise degradation. A core component of the analysis for the single carrier case discussed previously was built around recognizing that symbol rates for single carrier QAM generally exceed the frequency offsets where phase noise is prevalent. We used the example of 50 kHz of phase noise "bandwidth" to point out that the assumption for degradation is "slow" phase noise. A phase noise sample that rotates a constellation tends to be in the same place over many symbols in a row. If that phase error is close to a decision boundary or across it, there is likely to be a consecutive burst of errors.

Of course, with OFDM, we are using many, many narrow sub-carriers. For example let's use 192 MHz as a maximum OFDM bandwidth – consistent with coexisting with 6 MHz and 8 MHz forward path channel line-ups. An 8k FFT implementation for OFDM would mean that subcarriers are approximately 23.4 kHz apart. For a 16 k FFT, it would be 11.7 kHz.

Compare these to the 50 kHz phase noise "bandwidth" we were using earlier. This phase noise mask then extends *beyond* the sub-channel QAM symbol rate and beyond the main lobe of the OFDM spectrum implemented via FFT. Clearly, this is no longer a case of a phase noise process that is "slow" compared to the QAM bandwidth.

Let's look at an example phase noise spectrum after carrier recovery. We saw a sample spectrum in Figure 24 of phase noise that would be imposed on a QAM carrier by an RF frequency conversion. For slow phase noise, the exact shape is not as important – it is the total rms noise that matters, because the phase noise power is dominated by "slow" or low frequency energy. As such, we referenced "dbc" values of total phase noise based on requirements currently in place. Some of the may be tracked out at the receiver, but in all cases for single carrier it can be characterized as "slow" except perhaps for the lowest upstream symbol rates, which are rarely implemented today.

For OFDM, however, the spectral content of the phase noise mask matters. Consider an example post-carrier recovery mask shown in Figure 43. This is the characteristic lowpass shape of untracked phase noise. It has components associated with RF phase noise imposition, additive noise, and self-noise of the carrier recovery process itself.

Now let's take a look at how this type (two examples) of mask might look against an OFDM sub-channel spectrum. This is shown in Figure 44.

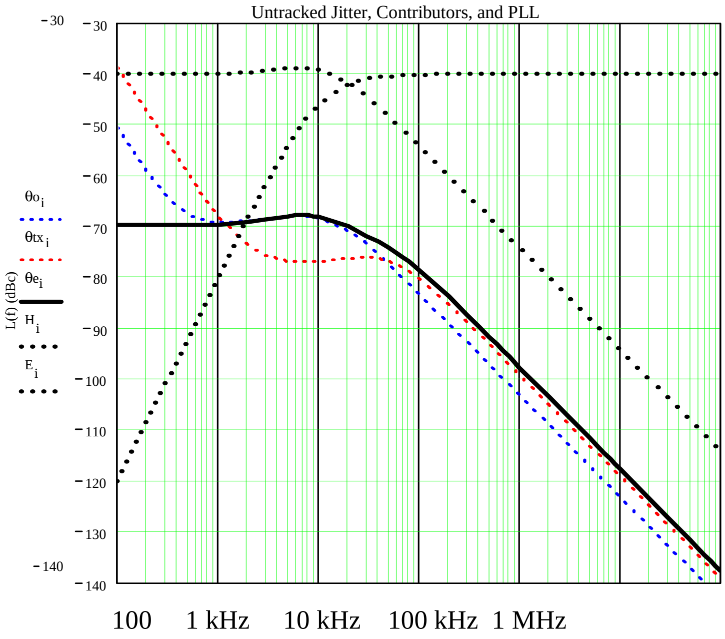


Figure 43 – Example Untracked Phase Noise Spectrum

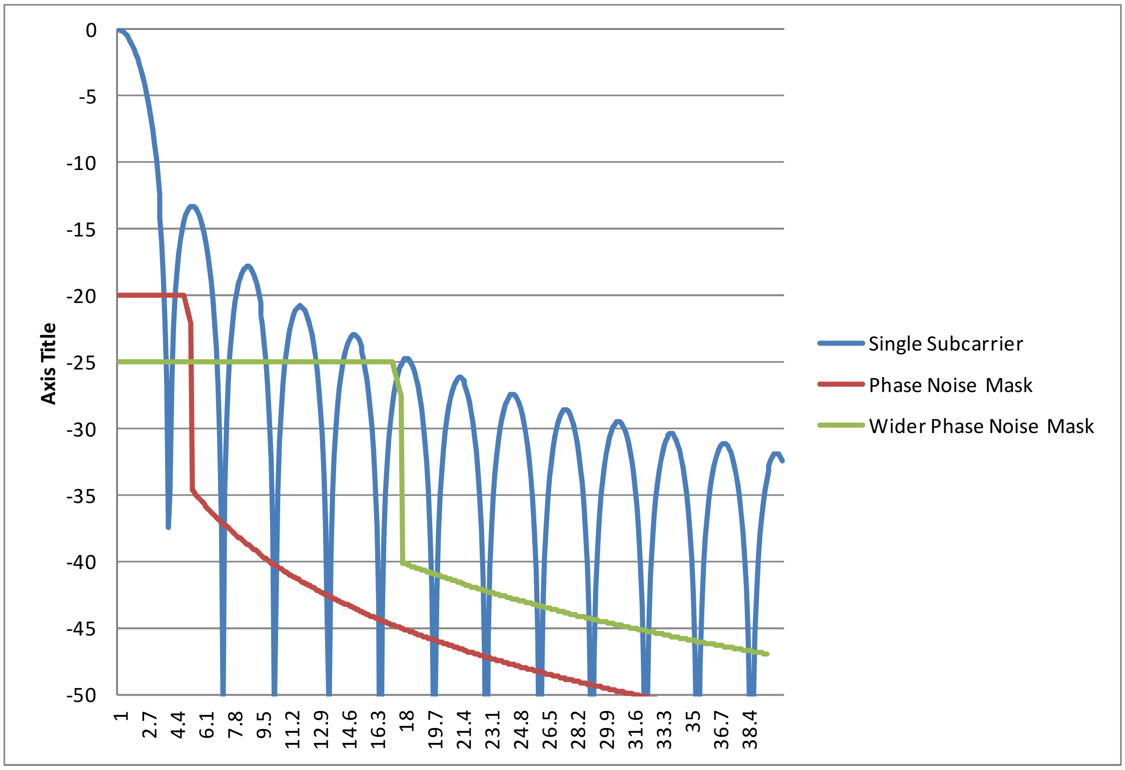


Figure 44 – Untracked Phase Noise vs. OFDM Sub-Channel

Two examples are shown. The red mask, for example, represents how the spectrum of Figure 43 relates to the 16 FFT discussed previously, with its roughly 12 kHz sub-channel spacing. Under this scenario, the 50 kHz of phase noise bandwidth we used earlier to discuss single carrier degradation is shown in green. Every OFDM sub-channel is effectively demodulated with the noise imposed by the phase noise mask.

However, as Figure 44 reveals, phase noise contributes two kinds of degradation to OFDM. There is an error common to all subcarriers related to the "in-band" effects – what for single carrier is the "slow" phase noise. Only, for OFDM, there is a good chance that the slow assumption is no longer valid, advantageously so in fact. It depends on the untracked mask – the shape or spectral occupancy of the phase noise is now important. Moderately varying or "rapid" phase noise allows some averaging over the bandwidth that the symbol is integrated over, and an average of a zero mean process is a better scenario than a single amplitude phase error sample.

However, there is also a component of phase noise that contributes to Interchannel Interference (ICI) as the masks cross into other sub-channel bands *because* of the relative relationship of phase noise mask to subcarrier spacing. This phase noise effect is additive looks nature, just as AWGN (mathematically easy to demonstrate sing the small angle assumption: exp(jφ) ≈ 1+jφ). In Figure 44 it is also clear that the noise in an adjacent band is a function of the phase noise spectrum itself (the shape) weighted by the Sin(x)/x response it leaks into. Not obvious from Figure 44 is that all of the subcarrier phase noise spectra combined create the full ICI effect. Because of this, and because the noise level is monotonically decreasing, the middle sub-carriers are the most effected by ICI due to phase noise.

The SNR degradation due to phase noise for OFDM has been calculated in many papers, and is simplified in [17] for a basic coherent receiver architecture as

$$\text{SNR(penalty, }\varphi) = 1 + \text{SNR} * \varphi_{\text{rms}}$$

For less than 0.5 dB of degradation, this simply reflects that the rms phase noise would be about 10 dB better than the SNR itself – a common relationship when analyzing additive impairments, again verifying the ICI component of phase noise degradation for OFDM.

Comparing this to the assumptions in Table 11, we see that this is about 3 dB better per modulation when compared to the single carrier, no error correction, slow phase noise case. As discussed, slow phase noise and its angular rotation effect is more painful than an averaging of that noise or an additive effect such as in OFDM. The common phase error on all channels is less degrading when some of the energy is outside of the symbol bandwidth, and the energy that contributes to angular rotation is now no longer slow by definition. This appears to overcome the effect of additive noise contributions that leaks across and create ICI.

The study of these two effects has led to substantial research on the sensitivity of OFDM and studies of the proper carrier recovery approach for OFDM frequency and phase synchronization and manipulation of phase noise processes by transmission and tracking systems. In fact, you can find literature that indicates OFDM is *less* sensitive to phase jitter, or *more* sensitive to phase jitter. And, in fact they can both be correct because the nature of the relationship of phase noise to the QAM carrier has changed, and new variables come into play. The assumptions about those relationships affects the results, and a comprehensive analysis for an HFC version against phase noise mask requirements such DRFI would be necessary for 1024-QAM upstream and 4096-QAM downstream to be effectively deployed.

AND THAT'S NOT ALL FOLKS

We have offered some guidance here on requirements and impairments for new modulation profiles and access techniques, but also recognize that more information is required to make solid requirements and recommendations in many cases. Even so, we have not covered all of the potential angles of the analysis. As more work goes into defining advanced PHY profiles for HFC, we will consider yet the next level of details. This includes items such as new isolation requirements for new service on old and old service on new. And, discussion of the equalizer complexity issues of single carrier, or the pro-con trade-offs of different multi-carrier approaches. We have discussed carrier synchronization in depth, but not timing synchronization. Symbol degradation is a quantifiable problem by quantifying timing jitter relationship relative to the eye diagram and pulse shaping used. We also have not discussed timing requirements that become complex in OFDMA. All of these are important topics for future discussion, along with new depth and insights on what we have discussed here as more variables become known and information complete.

SUMMARY

In this paper, we have discussed HFC architectures and key variables for downstream and upstream in order to allow an increase in spectral efficiency and maintain robust performance. We have provided guidelines for system parameters and discussed specifications of equipment today and the implications of the requirements to support for long term bandwidth efficiency objectives. We have investigated the component parts from optical links to RF links, to CPE, and into the home itself. We have explored options for network architectures that extend beyond today's bandwidth and carrier access methods, and quantified how such shifts in network design may affect these choice. We have analyzed how these modern multi-carrier methods may be affected by HFC conditions today and moving forward. We have broken them down into downstream scenarios and upstream

All in all, the outlook is hopeful for cable operators to be able to exploit modern tools and enable more spectral efficient use of the network, prolonging its already healthy lifespan. However, indications are that some important changes to business-as-usual may be in store to ensure the required robustness on the most advanced modulation profiles – the silicon itself of course, but also outdoor plant architectures, potential requirements changes to create the fidelity conditions that support the modulation efficiencies of interest, changes in service delivery at the interface to the home, and comprehensively defining RF channels that heretofore have not been used by cable, but would necessarily be so to deliver on the capacity needs of the future. Indeed, there is much to do, and based on lifespan projections of service mixes ahead, now is proper time to be game-planning the transition.

APPENDIX – SIMULATION TOOL

Modeling Environment

Agilent's SystemVue was used to conduct system-level analysis of M-QAM with combinations of AWGN, static narrowband interference, and phase noise. The SystemVue model was comprised of random data source, transmitter, channel, receiver, and a data sink. Data streams at the source and sink were compared bit-for-bit to approximate system impact in terms of Bit-Error-Ratio (BER). Parameter sweeps were conducted for the relative RF levels for both AWGN and interference contributions within the channel.

Modeling QAM

Construction of the baseband signal first required a random sequence generator, such as a PN15 data pattern, operating at the system bit rate. As an example, the system bit rate for 4096-QAM is 12 bits/symbol multiplied by the symbol rate, in this case 5.360537 symbols/second, resulting in a system sample rate of approximately 64.33 Mbps. The random bit sequence was then fed into a symbol mapper, whose states were organized such that errors associated with misinterpreting a received symbol as an adjacent symbol would result in only 1 bit error in the symbol. After mapping, both the In-Phase and Quadrature (I and Q) baseband signals were filtered using root-raised-cosine (RRC) filters. Below are the impulse and frequency response plots associated with the RRC filters with an alpha of 0.12 (downstream excess bandwidth).

Figures A-1 through A-4 show the time domain pulse, the baseband pulse spectrum, and the RF spectrum (SNR = 28 dB) and constellation (SNR = 28 dB), respectively.

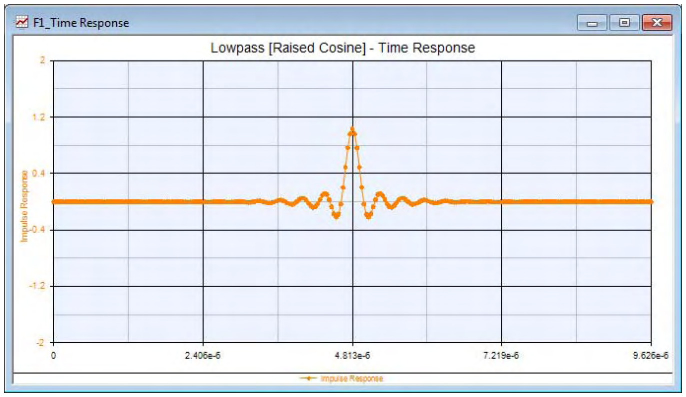


Figure A-1 – Time Domain Root Raised Cosine (RRC) Pulse Shape

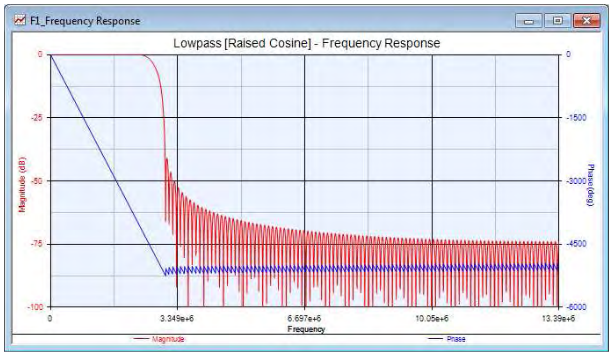


Figure A-2 – Root Raised Cosine (RRC) Spectrum

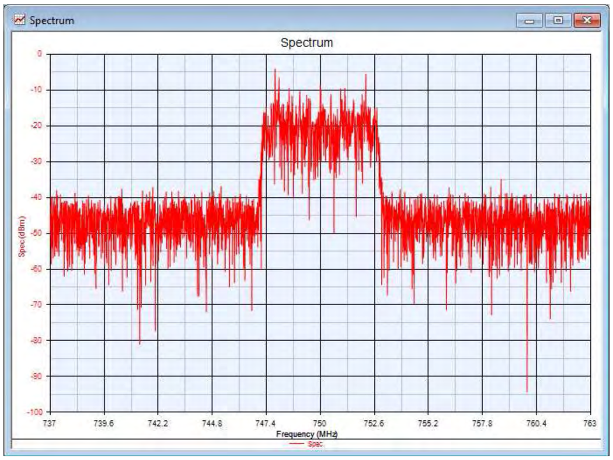


Figure A-3 – RF Spectrum with AWGN – SNR = 28 dB

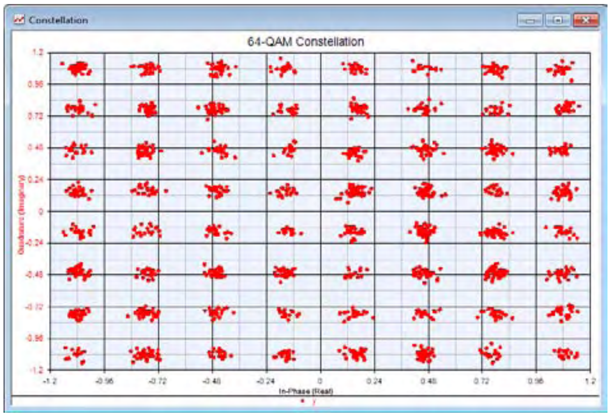


Figure A-4 – 64-QAM Constellation with AWGN – SNR = 28 dB

At the receiver, the signal is demodulated, RRC filtered and de-mapped using the same structures described above in reverse order. The response of 64, 256, 1024, and 4096-QAM to varying SNR measured against theory for uncoded transmissions is verified in Figure A-5. It can be seen that the simulated results track closely with theoretical expectations. This exercise provides the model basis for now extending channel impairments to items such as phase noise and narrowband interference.

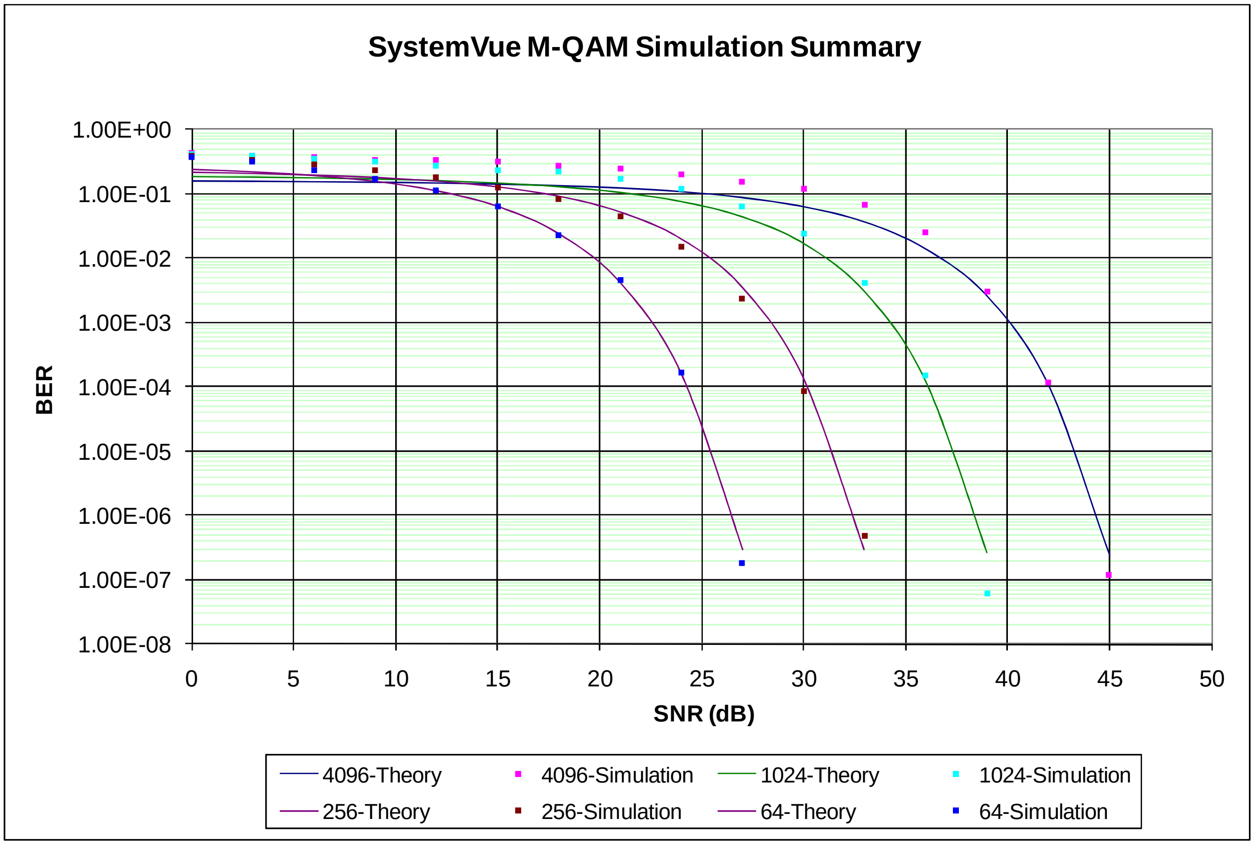


Figure A-5 – Simulated BER vs. Theoretical

REFERENCES

[1] Bingham, John C, *Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come,* IEEE Communications Magazine, May 1990.

[2] Chapman, John, Mike Emmendorfer, and Dr. Robert Howald, *Mission Is Possible: An Evolutionary Approach to Gigabit-Class DOCSIS*, 2012 Cable Show, Boston, MA, May 23-25.

[3] Howald, Dr. Robert, *Boundaries of Consumption for the Infinite Content World*, 2010 Cable-Tec Expo, sponsored by the Society for Cable Telecommunications Engineers (SCTE), New Orleans, LA, October 20-22, 2010.

[4] Howald, Dr. Robert, The Communications Performance of Single-Carrier and Multi-Carrier Quadrature Amplitude Modulation in RF Carrier Phase Noise, UMI Dissertation Services, 1998.

[5] Howald, Dr. Robert, *The Exact BER Performance of 256-QAM with RF Carrier Phase Noise*, 50th Annual NCTA Convention, Chicago, IL, June 10-13, 2001.

[6]Howald, Dr. Robert, *Fueling the Coaxial Last Mile*, 2009 Society for Cable Telecommunications Engineers (SCTE) Emerging Technologies Conference, Washington, DC, April 3, 2009.

[7] Howald, Dr. Robert, *Looking to the Future: Service Growth, HFC Capacity, and Network Migration*, 2011 Cable-Tec Expo Capacity Management Seminar, sponsored by the Society for Cable Telecommunications Engineers (SCTE), Atlanta, GA, November 14, 2011.

[8] Howald, Dr. Robert**,** Michael Aviles, and Amarildo Vieira, *New Megabits, Same Megahertz: Plant Evolution Dividends*, 2009 Cable Show, Washington, DC, March 30-April 1.

[9] Howald, Dr. Robert, *QAM Bulks Up Once Again: Modulation to the Power of Ten*, SCTE Cable-Tec Expo, June 5-7, 2002, San Antonio, TX.

[10] Howald, Dr. Robert L. and John Ulm, *Delivering Media Mania: HFC Evolution Planning*, 2012 SCTE Canadian Summitt, March 27-28, Toronto, ON, Canada.

[11] Howald, Dr. Robert and Phil Miguelez, *Upstream 3.0: Cable's Response to Web 2.0*, The Cable Show Spring Technical Forum, June 14-16, 2011, Chicago, IL.

[12] Howald, Dr. Robert L., Phillip Chang, Robert Thompson, Charles Moore, Dean Stoneback, and Vipul Rathod, *Characterizing and Aligning the HFC Return Path for Successful DOCSIS 3.0 Rollouts*, 2009 SCTE Cable-Tec Expo, Denver, CO, Oct 28-30.

[13] Lindsey, William C. and Marvin K Simon, Telecommunication System Engineering, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[14] Mengali, Umberto and Aldo N. D'Andres, Synchronization Techniques for Digital Receivers, Plenum Press, New York, 1997.

[15] Miguelez, Phil, and Dr. Robert Howald, *Digital Due Diligence for the Upstream Toolbox*, 2011 Cable Show, Chicago, IL, June 14-16.

[16] Piazzo, L and P. Mandarini, *Analysis of Phase Noise effects in OFDM Modems*, Technical Reprt No. 002-04-98, INFOCOM Dept. University of Rome "La Sapienza", May 1998.

[17] Proakis, Dr. John G, Digital Communications, McGraw-Hill, New York, 2001.

[18] Robuck, Mike, *Cox, Motorola lay claim to new return path speed record*, CedMagazine.com, March 01, 2011.

[19] Stoneback, Dean, Robert Howald, Tim Brophy, and Oleh Sniezko, *Distortion Beat Characterization and the Impact on QAM BER Performance*, 1999 NCTA Convention, Chicago, IL, June 13-16.

[20] Stott, J., *The Effects of Phase Noise on COFDM*, EBU Technical Review, Summer 1998.

[21] Thompson, Robert, *256-QAM for Upstream HFC Part Two*, 2011 SCTE Cable-Tec Expo Atlanta, GA, November 15-17, 2011.

[22] CableLabs, Inc., *Return Laser Characterization Techniques - Results and Recommendations*, Engineering Report, September 21, 1999.

[23] Data-Over-Cable Service Interface Specifications Physical Layer Specification (DOCSIS-PHY), CM-SP-PHYv3.0-I08-090121, January 21, 2009, Cable Television Laboratories, Inc.

[24] DOCSIS Downstream RF Interface Specification (DRFI), CM-SP-DRFI-I10-100611, June 11, 2010, Cable Television Laboratories, Inc.

ACKNOWLEDGEMENTS
The authors would like to thank Stuart Eastman for his invaluable contributions to this paper.

HFC NETWORK CAPACITY EXPANSION OPTIONS

Jorge D. Salinger
VP, Access Architecture
Comcast Cable

NOTE: The concepts and proposals presented in this paper are for discussion purposes only and do not reflect actual plans from Comcast. Similarly, all examples presented are only provided for illustrative purposes.

Abstract

MSOs are deploying more narrowcast capacity than ever before, and there is no evidence of a change in this trend.

- *DOCSIS® 3.0 is widely deployed, with 4 and 8 downstream channel bonding groups becoming the norm. A continual annual growth of 40-60%, observed industry-wide for over 10 years, would require many more channels over time*
- *8-channel service groups for video on-demand (VOD) are commonplace. Growth rate increasing due to both higher usage and higher bitrate (high definition)*
- *10, 20 or even more channels for switched digital video (SDV) are frequently used for longer tail content*
- *Growth in business service applications requires additional increased capacity*
- *The advent of IP video services and network-based digital video recorder (DVR), which are anticipated to be very popular amongst current and potential subscribers, will compound the need for additional narrowcast capacity.*

The effect of the above trends, combined with the need to simultaneously support a full set of legacy broadcast services, including digital, analog and/or both, would likely require additional hybrid fiber-coax (HFC) network capacity.

While it seems conceivable that a transition from legacy and broadcast services to an all-narrowcast/IP services infrastructure could be established, the industry as a whole is looking for options that would provide additional capacity to support simultaneous uses, and increased capacity beyond such transition. These options include:

A. *Traditional service group segmentation*
B. *Move quadrature amplitude modulation (QAM) generation downstream into the network*
C. *Implement higher modulation physical layer (PHY) and/or more efficient media access control (MAC) protocols*
D. *Increase HFC downstream capacity beyond currently deployed, and/or move split to higher spectrum for increased upstream capacity*
E. *Develop technology that would operate in unused portions of the spectrum, and even unleash spectrum above current top range (e.g., above 1 GHz)*

Each of the above options has benefits and drawbacks. Each approach offers different

network engineering and operational simplifications and complexities. And, the relative improvements in offered capacity versus cost and customer impact can be significantly different.

This paper will provide, from an operator's perspective:

1. *A technical overview of each of the options outlined above, describing how each would be deployed and evolved over time,*

2. *The key benefits/drawbacks for each of the options, including engineering and operational pros and cons for each option,*

3. *Possible implementation approaches for various applications, including residential and commercial services.*

TYPICAL HFC NETWORKS TODAY

Most MSO's hybrid fiber-coax (HFC) networks have been designed to either 750 or 860 MHz of spectrum capacity. If not fully utilized, it is expected that use of their capacity will be increased to the point of exhaustion as the use of DOCSIS® increases for the higher high-speed data (HSD) service tiers, additional high-definition (HD) programs for both broadcast (BC) and especially narrowcast (NC) services such as video on demand (VOD) and switched digital video (SDV) are deployed, or new services such as internet protocol (IP) video and network-based digital video recorder (n-DVR) are added.

Proportionally few HFC networks have been deployed to operate up to 1 GHz, although all equipment available today can support the use of spectrum up to 1 GHz and even 3 GHz for some components.

In recent years the growth in, and demand for, HD programming has resulted in the need for allocation of large numbers of EIA channels for HD services, both for BC and NC, which has filled every available portion of the spectrum. This is especially true for BC, where large numbers of programs are offered in HD format, while simultaneously the need for distributing the standard definition (SD) version has persisted. This has resulted in the need for use of 3x to 5x the number of EIA channels than previously required. For example, a typical digital multiplex including 10 to 15 programs would require an additional 3 to 5 EIA channels for the HD equivalent streams, even assuming the newer, more sophisticated multiplexing schemes available in the market. Of course not every program is available, or still sought by subscribers, in HD format. But very large numbers of them are, including 100 to 150 BC programs.

The above is also applicable to a great extent in systems utilizing SDV technology for distribution of its content. The difference is that the SD version of the program is not distributed unless a subscriber is requesting it, which reduces the marginal increase in capacity. Assuming that all programs are distributed in only one format, which is certainly a valid expectation for programs of low viewership, then the increase in capacity for a conversion from SD to HD would just be the increase in capacity required for the transmission of the HD program without requiring the simultaneous use of bandwidth for both formats.

Additionally, considerable spectrum is needed to deploy high-capacity narrowcast legacy video services, especially n-DVR, and a full-array of HD video-on-demand services. For the former, initial observations suggest that network requirements for n-DVR may be as high as 4x to 5x that of VOD, and that peak utilization overlaps, at least partially, with that of peak use for other narrowcast services.

Finally, the growth in HSD services shows no sign of letting up. Network operators have observed an increase use of HSD service capacity for well over a decade now, which amounts to a year-over-year compounded growth of 40% to 60%. The applications have changed throughout this time, but the demand has continued to increase at the same relentless rate.

In fact, such increase in demand for HSD capacity shows no evidence of decreasing. Should that trend continue, MSOs would be in a position to increase access network capacity through either one or more of the existing capacity tools and/or through one or more of the new capacity tools outlined in this paper.

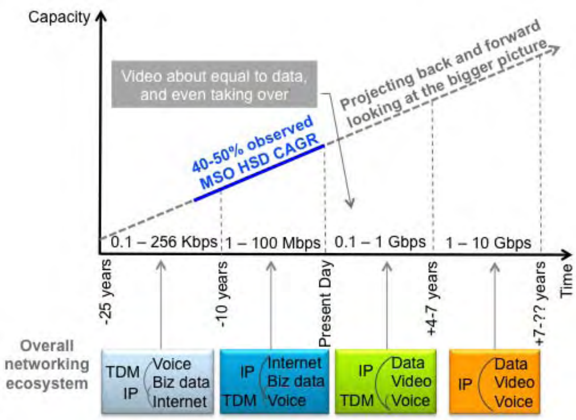


Figure 1: Example of HFC capacity utilization over time

How does this compare to other operator's data services and a longer period? As shown in Figure 1, projecting the MSO's HSD service growth back in time to when Internet services started as shown in the diagram, 25 years ago services should have been about 100 bps. This coincides with the history of telephone modems from 110 and 300 baud modems from the mid-80s, to 56 Kbps/V.42, into ISDN services.

This demonstrates that the growth seen in MSO's HSD services is typical over a much longer period of time, rather than an exception observed by MSOs in recent years.

GROWTH PROJECTIONS

From all of the above, it then follows that, should the usage growth pattern continue at the past experienced pace, networks will be required to provide HSD services in the range approximating 1 Gbps within the next few years. This growth, coupled with the surge in HD video formats, and more personalized narrowcast services, will result in a significant growth in NC capacity, as shown in Figure 2 below.

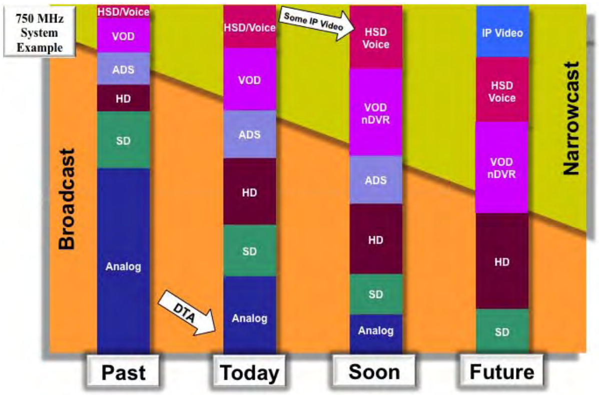


Figure 2: Example of narrowcast service growth over time

To support this growth, MSOs have deployed, or are considering deployment of, bandwidth reclamation tools such as SDV for digital broadcast, digital terminal adapters (DTAs) for analog services, or a combination of both. These tools have been extremely valuable to MSOs, which have seen their operational complexity and cost to be well justified.

In the case of SDV, early predictions several years back from industry analysts projected that the efficiency of SDV would reach 40% (e.g., programs requiring 10 EIA channels could be carried in 6). This has proven to be understated, since it was based on the use of SDV for reduction in bandwidth required for existing services. As SDV's role in the network grew, the efficiencies have been even greater, especially as SDV has been used to introduce niche services that have low viewership and would have otherwise been difficult to deploy.

The benefit of DTAs has been just as, or perhaps even more, striking. MSOs deploying DTA devices are able to eliminate the need to distribute the analog channels in the network. Even if DTAs are distributed to top analog tier customers, such as only to subscribers of the traditional expanded basic subscribers, such deployment would reduce a channel line up from perhaps 50 EIA channels dedicated to 50 analog programs to perhaps as little as 4 EIA channels dedicated to transport the 50 programs in their equivalent digital transport. Using the same comparison method as the above SDV case, this is a >90% efficiency. If extended to the entire analog tier the efficiency gains are very significant.

Despite the availability of these tools, they are not universally applicable. With respect to SDV, in general it is not likely that all broadcast programs will be switched since experience shows that many broadcast programs are constantly viewed by someone in the service group during peak hours, which will leave a large portion of the spectrum still used for broadcast. Similarly, not all analog channels can be removed in the short term due to operational and/or cost constraints.

Additionally, while many MSOs will use one or both tools, in general these tools won't be used by every MSO for all applications. Finally, there are also significant potential gains to be achieved from the use of advanced video CODECs (AVCs) and variable bit-rate (VBR). In the case of AVCs, coding efficiencies of approximately 50%, depending on implementation and content type, can be obtain with H.264[1] and/or MPEG-4 Part 10[2]. And the use of VBR promises to result in a capacity efficiency gain of as much as 70% versus CBR[3]. The combined gains from using both approaches could be very significant.

However, these are difficult tools to take advantage on the network since proportionally relatively few legacy set-tops still support AVCs and VBR, especially the latter. These tools will likely enjoy significant support in newer, IP-video based services equipment moving forward.

But, this approach will require additional capacity on the network. This is especially true when considering that the deployment of these advanced video services will result in an additional simulcast of video programs, at least initially, which is expected since its deployment will not at least initially replace the currently deployed services. Furthermore, ubiquitous support for such devices would require considerable spectrum if the legacy services are maintained for an extended period, as it is expected since legacy devices are and will continue to be deployed. Moreover, this increase in simultaneous use of advanced, IP video services while maintaining legacy services will be especially impacting over time as its penetration increases.

All of the above, coupled with the success experienced by MSOs in recent with business services, will likely require the deployment of IP capacity beyond what can be supported today, requiring the development of tools for increased efficiency in the use of spectrum and/or unlashing of additional spectrum in the HFC network. The following sections of this paper will enumerate ways in which this can be achieved.

OPTIONS BEING CONSIDERED

[1] ITU-T Recommendation H.264: 2005, Advanced Video Coding for generic audio-visual services

[2] ISO/IEC 14496-10: 2005, Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding

[3] Capacity, Admission Control, and Variability of VBR Flows, CableLabs Winter Conference, February, 2009

Let us review the categories of options being considered throughout the industry, and evaluate how each one fulfills the above desirable targets. In the process, let us review the key implementation aspects of each option, leaving for another opportunity the details of the options and on how these could be deployed.

The categories of options are:

1. Traditional service group segmentation

2. Move QAM generation downstream into the network

3. Implement PHY and MAC improvements

4. Increase HFC downstream capacity beyond currently deployed, and/or move split to higher spectrum for increased upstream capacity

5. Develop technology that would operate in unused portions of the spectrum, and even unleash spectrum above current top range (e.g., above 1 GHz)

1. Traditional service group segmentation

This option is readily available and has been in use for many years. It consists of decombining service groups (SGs) when possible, or dividing nodes into smaller groups when decombining SGs is no longer viable.

Traditionally SGs have consisted on a number of nodes combined together in the cable headend, and nodes include a number of homes passed and corresponding subscribers. Therefore, service group segmentation normally is achieved initially by separating nodes into smaller SGs, and when SGs consist of a single node these are segmented further by separating a number of the homes in a node into a new, separate node.

For example, assume that a SG consists of 2,000 homes passed (HHP), which results from combining 4 nodes, each with 500 homes passed. The SG decombining could be initially achieved by dividing the SG into 2 SGs, each consisting of 1,000 HHP. The segmentation could continue by separating each of the 4 nodes into a separate SG, consisting of 500 HHP/SG. Beyond that, SG segmentation would include "breaking up" each of the nodes into a smaller group by adding 1 additional node, creating nodes (and SGs) consisting of 250 HHP.

The following a key options for SG segmentation:

I. SG decombining is generally achieved by adding equipment in the cable headend. This re-uses the spectral HFC network capacity in smaller SGs.

II. Node segmentation requires the same additional equipment in the headend, but also requires that additional nodes, and/or fiber be installed in the plant.

And, the following are key factors to consider regarding SG segmentation:

A. SG segmentation usually involves the same decomposition in the upstream (US) and downstream (DS).

B. The relative cost of SG segmentation is higher for node segmentation than for SG decombining. This is because the work requires for the former requires the installation of additional nodes and/or fiber in the network (node splits), which in some cases is substantially more expensive. Conversely, in general SG decomposition is significantly less expensive than node segmentation.

C. However, when additional peak capacity is needed, such as in high-speed data (HSD) services, the SG segmentation is not a viable solution since it does not inherently add peak capacity.

2. Move QAM generation downstream into the network

This option would require including the PHY, part, or all, the MAC, or all of the CCAP functionality into a line-gear device which would be installed in the HFC network. Depending on the functionality being 'remoted' into the HFC network and the desired interoperability, this option would require the creation of specification. Connectivity back to the headend would be achieved via a baseband laser, such as point-to-point Ethernet, as opposed to an analog modulated laser as used now in HFC network.

The advantage of this approach is the migration to a baseband laser, and the operational simplifications that this entails. This approach would also result in additional capacity given the inherent segmentation that would be implemented. And, given the reduction in noise sources (e.g., removal of the analog laser, shortening of the links especially upstream, and reduction in the number of components), it should be possible to achieve higher order modulation rates than are possible to achieve with the PHY located in the headend.
From an operational perspective, however, the proliferation of intelligent devices that would need to be maintained, upgraded, and supported, might result in an increased complexity.

3. Implement PHY and MAC improvements

Clearly, cable systems today are capable of supporting higher order modulations, resulting in greater bit transmission capacity in the same spectrum. For example, it is considered possible to support 1,024 QAM downstream modulation in current cable systems. In fact, it should be possible to support even higher downstream modulations such as 2,048 and perhaps even 4,096 QAM. In addition, it should be possible to support 256 QAM in the upstream, and perhaps even higher order modulation rates. These improvements would come at a cost of higher signal-to-noise requirements, which are believed possible to achieve in today's cable systems.

Additionally, given advances in CPU performance in DOCSIS components, both in the CPE and the CMTS, it should be possible to replace the currently used Reed-Solomon forward error correction (FEC) for Low-density Parity Check (LDPC) FEC. This change is expected to provide an improvement in bitrate equivalent to 2 bits/Hz.

Additionally, it appears that it would be beneficial to migrate to multicarrier modulation techniques, such as orthogonal frequency division multiplexing (OFDM) for the downstream, and orthogonal frequency division multiple access (ODFMA) for the upstream, as opposed to the currently used single-carrier approach.

OFDM and OFDMA offer superior performance and benefits over the older, more traditional single-carrier QAM modulation methods because it is a better fit with today's high-speed data requirements. The use of OFDM, and OFDMA, has become widespread and their implementation well understood in recent years, which was not the case when DOCSIS was initially conceived 15 years ago OFDM when it was extremely difficult to implement with the electronic hardware of the time. These techniques remained a research curiosity until semiconductor and computer technology made it a practical method in recent years, and extensively used for cellular and Wi-Fi transmission. OFDM, and OFDMA, is perhaps the most spectrally efficient method discovered and implemented so far.

4. Increase HFC downstream capacity beyond currently deployed, and/or move split to higher spectrum for increased upstream capacity

From a headend equipment perspective, this option is generally readily available to MSOs. However, CPE equipment would have to be implemented to support the new enhanced upstream and/or downstream spectrum.

From a network perspective, this option involves the change of the diplexers throughout the network such that the frequency division crossover is moved from the 42-50 MHz up to a higher portion of the spectrum, plus the simultaneous expansion of the network capacity to 1 GHz via a retrofit of the active components with minimal changes to the plant spacing and passive components. However, from an operational perspective, this option requires perhaps the most operational change to existing services, such as the removal of analog channels in that portion of the spectrum. That may not be possible for many MSOs that are either required to maintain support for analog TVs directly (e.g., without DTAs), or are unable to remove the analog channel for contractual reasons, or some combination of the above two reasons.

Even if removing the analog channels is possible, this option seems to require the installation of CPE filters in most or perhaps all home CPE devices (e.g., TVs, VCRs, etc.) to both protect that portion of the spectrum from emissions from such home devices and to protect the devices themselves from the levels of transmission of the new CPE that would use that portion of the spectrum for transmission.

And, even if removing the analog channels and deploying the necessary filters were possible, this solution alone provides limited additional US capacity in the network, as follows:

- A move of the split to 65 MHz provides an additional capacity of just 15 MHz, which less than doubles the current capacity. By all accounts, this is a change not worth embarking on.

- A move of the split to 85 MHz almost triples the US capacity, and the simultaneous expansion of the DS network capacity to 1 GHz would add a net 15-30 new DS QAMs (this calculation considers the combined effect of expanding the capacity of the network to 1 GHz from 860 MHz or 750 MHz respectively, and the loss of DS spectrum with the move of the split into the current DS region).

- The shift of the split up to the 200 MHz is also being considered, but while this change would provide much more US capacity, it would reduce the next number of DS capacity significantly and would require the change of large numbers of non-DSG STBs (most of the STBs deployed to date) because the existing and extensively deployed OOB carriers would become inoperable since the region of the spectrum these utilize would be used for the US. Additionally, this change has other plant implications, such as the US equipment currently deployed would not support such extensive US, and thus a new HFC return strategy/equipment would be required.

5. Develop technology that would operate in unused portions of the spectrum, and even unleash spectrum above current top range (e.g., above 1 GHz)

Unlike option #4, this approach involves equipment not currently available. Instead, implementation of this option will require the development of network components and corresponding equipment that would make use of the existing forward spectrum but would use an unused portion of the spectrum, above 750, 860 MHz, or even 1 GHz. This new technique, which we will call High Spectrum Overlay, would require new equipment that could be built in the form of a new network gateway that could be installed in the headend, or in the vicinity of the node, or even deep within the HFC network. This new equipment would provide the 'translation' from the optical transmission generated at the headend into electrical signals, and RF transmission from the location of the converter to the coaxial portion of the HFC network.

This approach would increase US and DS capacity considerably, likely providing multiple Gbps of net additional US and DS capacity. In the process it leaves legacy services and existing CPE untouched.

However, this approach will require considerable equipment development before it would become available for deployment. Such equipment would use spectrum above that being used today for both additional US and DS capacity.

This option could be implemented in three fundamental ways: where the network gateway is located in the headend, or where the network gateway is deployed in the vicinity of the node, or where the network gateway is deployed throughout the HFC network.

In the first case, the RF signals would have to traverse the entire HFC network, including the forward and return analog modulated lasers and receivers, thereby being limited to the spectrum manageable by the analog modulated lasers and receivers.

In the second, the RF signals would traverse the various amplifiers within the coaxial part of the network, but would not require of transmission via the analog modulated lasers and receivers.

And, in the third, the network gateway would be installed in the vicinity of each active component where advanced services are to be provided. This option is known as a Passive High-Spectrum Overlay system. Therefore, this option would require the deployment of additional fiber beyond what's already installed in the network, namely between the existing node and each of the active components in the HFC network. In that way WDM would be used to carry baseband signals up to the node, from which traditional PON technology would be used to interconnect each of the new network gateways back to the HE.

Any modern HFC network should support a Passive High-Spectrum Overlay. Figure 3 depicts an initial deployment of Passive High-Spectrum gateways, for which EPON equipment is deployed in the headend, a separate optical wavelength is used in the trunk fiber to carry the EPON signals up to the node (shown in dashed blue lines), additional fiber is deployed in the distribution portion of the network (shown in solid blue lines), and new Network Gateways that provide optical-to-electrical signal conversion are installed to provide the overlay within an HFC segment between amplifiers.

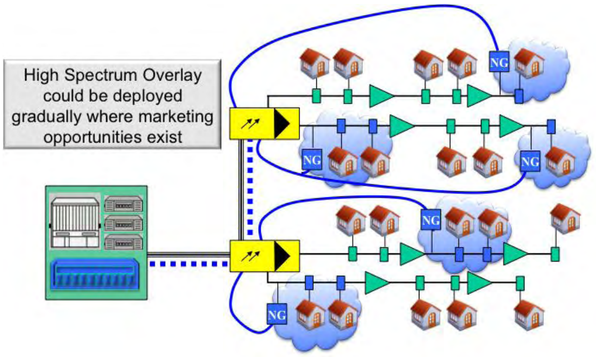

High Spectrum Overlay could be deployed gradually where marketing opportunities exist
NG
NG
NG
NG
NG

Figure 3: Initial High Spectrum Overlay

This approach should not be construed as resulting in a Node + 0 HFC cascade reduction. This is because the cascade of HFC actives is not modified. Instead the RF output of the gateways deployed in the HFC network and operating above 1 GHz are combined with the RF signals existing in the coaxial network which operate below 1 GHz, much in the same way as narrowcasting a set of signals on a per service group basis where the other signals are broadcasted to the set of service groups.

The following categories of work would need to be performed in the plant in order to achieve the above:

- WDM could be used from the headed to the location of the node to reuse the existing long-haul fiber.
- To provide the remaining optical link from the node to the location of each active, additional fiber would be over-lashed to the distribution coaxial hardline cable, which is generally a short to medium length span.
- Finally, in order to pass RF signals above 1 GHz on the distribution network, it is likely that a proportion of the tap faceplates would need to be replaced, although it is expected that the tap housing will likely support these new faceplates, and that only faceplates serving subscribers and upstream from it would need to be replaced.

Assuming a high-bandwidth optical network from the headend to the network gateway, such as 10 Gbps EPON, and a high-order modulation and encoding scheme, it is expected that a transmission achieving 8-10 b/Hz might be possible, therefore resulting in a combined US/DS payload transport capacity of approximately 3-5 Gbps.

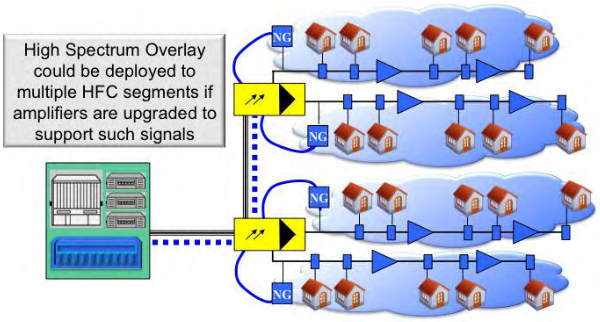

High Spectrum Overlay could be deployed to multiple HFC segments if amplifiers are upgraded to support such signals

Figure 4: Multi-segment High Spectrum Overlay

Figure 4 depicts the case of a deploying Network Gateways at node locations. This option would require less fiber, but would necessitate a rebuilt with amplifiers that would pass the new RF signals.

POSSIBLE IMPLEMENTATIONS

In evaluating the possible approaches outlined above, and taking into account the technologies available to date, it makes sense to consider the following implementations: enhancement to DOCSIS for residential applications, and development of a new transport alternative to EPON over HFC for commercial applications. Naturally, despite the primary target services, either of the two technologies could be used for either or both services.

DOCSIS Enhancements

Given the success and widespread use of DOCSIS-based services to date, and the advent of the technical advances outlined above, it seems plausible to consider the following enhancements to DOCSIS to enable additional capacity and a more efficient use of HFC spectrum:

- Use of higher-order (1/2/4K QAM) and modulation techniques (OFDM/OFDMA) to improve throughput and simultaneously reduce spectrum utilization by as much as 50%,

- Replace the current Reed-Solomon FEC technique with a more modern Low Density Parity Check (LDPC) FEC, which would improve overall efficiency by as much as 25%,

- Enable use of additional spectrum for the US, beyond the current 5-42 MHz, up to 100 MHz or even higher spectrum, to increase US transmissions by a 3x to 5x factor, and

- As capacity is enhanced, consider simplifications of the DOCSIS protocol that may reduce implementation complexity, accelerate the availability of newer implementations, and reduce costs.

Implementation of the above new functionality will have to be done taking into account backward and forward compatibility to maximize the benefit for current and new equipment.

New HFC transport for EPON

Similarly to the enhancements now available for DOCSIS, it seems possible to implement a new transport for EPON over coax. Envisioned in the past as a component of Comcast's Next Generation Access Architecture[4], a new transport for EPON Protocol over Coax (EPoC) is now under development at IEEE. This new transport will make it possible to provide EPON services to end-devices attached via cable operator's coax network rather than only via fiber cable.

The work currently underway, known as an IEEE 802.3 Study Group, is intended to demonstrate the feasibility of implementing a coax transport for EPON using technologies and approaches similar to those that would be applicable to DOCSIS. Once completed the work of the Study Group, a Task Force would be formed to define the new PHY for EPON over coax.

This work would lead to the availability of a coaxial-attached alterative to EPON, which would enable MSOs to deploy EPON services to customers already served by its HFC network. This should result in a more economical and operationally simpler way to provide Metro Ethernet (MetroE) services to business customers without having to deploy fiber to each potential customer premise.

OVERALL ACCESS ARCHITECTURE

The new edge platform devices currently under development by vendors, as specified by the CCAP architecture, will support either of the approaches described above. The CCAP architecture already supports the modularity necessary to upgrade line cards progressively as new technologies become available.

For the case of the enhancements to DOCSIS outlined above, it seems reasonable to expect that the current downstream line cards could be updated via field programmable gate array (FPGA) programming changes, such as Hardware Descriptor Language (HDL) or Register Transfer Level (RTL) programmable changes. In the case of the upstream, it is expected that new line cards could be developed that would take advantage of the new technologies.

Furthermore, the CCAP architecture provides support for EPON, such that even EPoC is supported in the overall access architecture.

SILICON DEVELOPMENT

One important consideration in evaluating the benefits of each approach is the need and availability of silicon components, or on the flip side the need for its development.

[4] What is CMAP? Jorge Salinger and John Leddy, CED Magazine, February 2010

This is critical for the following fundamental reasons:

a. When silicon exists the availability of the system solution is quicker, whereas when it needs to be developed the timeline is significantly longer, and

b. If silicon devices, or at least some of their components, are used for multiple purposes, especially for multiple industries, then their production increase rapidly and costs decrease considerably.

Some of the new technology enhancements will likely require silicon development, but others would not, for which technology design decisions would be important.

CONCLUSIONS

Additional HFC network capacity will be required for narrowcast services for both residential and commercial service applications in years to come. The expected growth appears to be quite large.

New technologies are now becoming available that would make it possible to achieve higher throughout and more efficient use of spectrum. This includes higher-order and more modern modulation techniques, more sophisticated forward error correction, and the use of more spectrum than currently utilized.

This paper presented an analysis of these technology options and their corresponding pros and cons, and outlined how these technologies could be used to enhance the current transport options available to MSOs, such as DOCSIS, and to create new infrastructure options, such as EPoC.

ACKNOWLEDGEMENTS

Thanks to my colleagues Sam Chernak, Doug Jones, Saif Rahman, John Bevilacqua, Chris Bastian, and Dave Urban from Comcast for their contributions to the discussions leading to the proposals and conclusions presented in this paper.

Also, many thanks to my colleagues from CableLabs®, the many vendors and advisors of Comcast, and especially fellow MSOs, for the many discussions leading to the materials presented.

OPTIMIZING FAIRNESS OF HTTP ADAPTIVE STREAMING IN CABLE NETWORKS

Michael Adams
Chris Phillips
Solution Area Media
Ericsson

Abstract

This paper describes a novel approach to traffic management for HTTP adaptive streaming that optimizes fairness across multiple clients and increases network throughput. Readers of the paper will gain an understanding of the network impacts of implementing HTTP adaptive streaming, and how network management techniques may be applied to optimize fair bandwidth allocation between competing streams.

Benefits for the network operator include:

- *enforcing fairness in the network (without resorting to techniques such as deep packet inspection),*
- *managing and ensuring consumers' overall quality of experience, and*
- *preventing network instability that can be caused by competing clients in a shared access network.*

Benefits for the consumer include:

- *a more consistent overall quality of viewing experience, and*
- *the ability to simultaneously use multiple devices within the home.*

The concepts described in this paper have been prototyped to show improvements in fairness and overall network throughput without placing special constraints on the client implementation (which is typically outside of the operator's control). The results are being published here for the first time.

BACKGROUND

There is a great deal of interest in HTTP adaptive streaming because it can greatly improve the user experience for video delivery over unmanaged networks. Adaptive streaming operates by dynamically adjusting the play-out rate to stay within the actual network throughput to a given endpoint, without the need for "rebuffering". So, if the network throughput suddenly drops, the picture may degrade but the end user still sees a picture.

Although adaptive streaming was originally developed for "over-the-top" video applications over unmanaged networks, it also brings significant advantages to managed networks applications. For example, during periods of network congestion, operators can set session management polices to permit a predefined level of network over-subscription rather than blocking all new sessions. This flexibility will become more and more important as subscribers start to demand higher quality feeds (1080p and 4K).

HTTP adaptive streaming is the generic term for various implementations:

- Apple HTTP Live Streaming (HLS) [1]
- Microsoft IIS Smooth Streaming [2]
- Adobe HTTP Dynamic Streaming (HDS) [3]

Although each of the above is different, they have a set of common properties (see Figure 1):

- Source content is transcoded in parallel at multiple bit rates (multi-rate transcoding). Each bit rate is called a profile or representation.
- Encoded content is divided into fixed duration segments (or chunks), which are typically between two and 10 seconds in duration. (Shorter segments reduce coding efficiency while larger segments impact speed to adapt to changes in network throughput).
- A manifest file is created, and updated as necessary, to describe the encoding rates and URL pointers to segments.
- The client uses HTTP to fetch segments from the network, buffer them and then decode and play them.
- The client algorithm is designed to select the optimum profile so as to maximize quality without risking buffer underflow and stalling (rebuffering) of the play-out. Each time the client fetches a segment, it chooses the profile based on the measured time to download the previous segment.

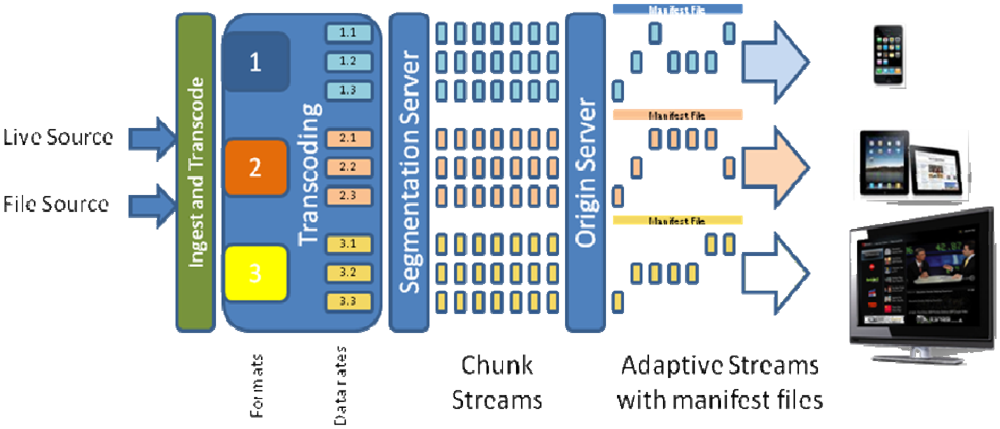


Figure 1: Ingest, transcoding, segmentation and adaptive streaming.

MPEG DASH

MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH) is certain to become a significant force in the marketplace [4]. While HLS uses the MPEG-2 transport stream format (which is widely deployed in most conventional digital TV services), Smooth Streaming and MPEG-DASH use an MPEG-4 Part 14 (ISO/IEC 14496-12) transport format known as fMP4 or ISO MP4FF.

Smooth Streaming and MPEG-DASH include full support for subtitling and trick modes, whereas HLS is limited in this regard. MPEG-DASH enables common encryption, which simplifies the secure delivery of content from multiple rights holders and to multiple devices.

Another key difference is the way in which MPEG-DASH and Smooth Streaming play-out is controlled when transmission path conditions change. HLS uses manifest files that are effectively a playlist identifying the different segments so that, for instance, when path impairment occurs, the selection of the URL from the manifest file adapts so that lower bit-rate segments are selected. In Smooth Streaming the client uses time stamps to identify the segments needed and thus certain efficiencies are gained. Both HLS and Smooth Streaming handle files in subtly different ways, each claiming some efficiency advantage over the other. Both use HTTP, which has the ability to traverse firewalls and network address translation, giving it a clear advantage over RTSP, RTMP and MMS.

Adaptive Streaming Standardization

There are a number of initiatives aimed at large parts of the overall solution for streaming video. A document called *MPEG Modern Transport over Networks* was approved at the 83rd MPEG meeting in January 2008, which proposed a client that was media aware with optimization between the transport and content layers to enable video to traverse networks in an adaptive manner. However, at that time, its focus was on the widespread adoption of a variant of AVC/MVC called SVC (Scalable Video Coding) that would allow the client to generate acceptable video from a subset of the total aggregated transport stream.

Subsequently, at the 93rd meeting, the focus was changed to HTTP streaming of MPEG Media called *Dynamic Adaptive Streaming over HTTP* (DASH) using 3GPP's Adaptation HTTP Streaming (AHS) as the starting point. MPEG has standardized a Manifest File (MF), a Delivery Format (DF), and means for easy conversion from/to existing File Formats (FF) and Transport Streams (TS) as illustrated in Figure 2.

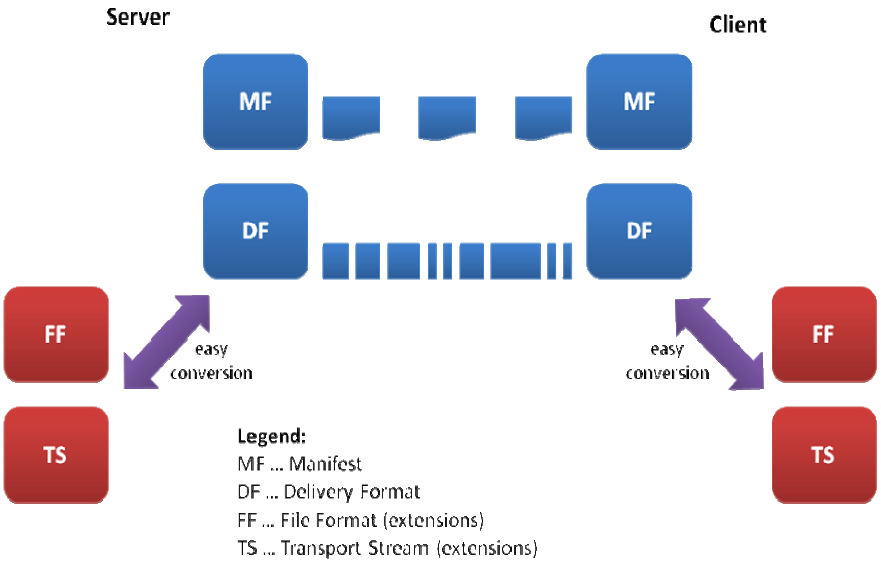


Figure 2: Dynamic Adaptive Streaming over HTTP (DASH).
Courtesy: Christian Timmerer, Assistant Professor at Klagenfurt University (UNIKLU)

The MPEG-DASH standard got Final Draft International Standard status in December 2011. MPEG-DASH has the potential to simplify and converge the delivery of IP video, provide a rich and enjoyable user experience, help drive down costs and ultimately enable a better content catalog to be offered to consumers, because more revenues can be re-invested in content, rather than paying for operating overheads. It will help streamline and simplify workflows and enable operators and content providers to build sustainable business models to continue to deliver the services that consumers demand.

FAIRNESS IN ADAPTIVE STREAMING

HTTP adaptive streaming clients implement a “greedy” algorithm, in which they will always seek to achieve the maximum bit rate available. This can lead to instability, oscillation and unfairness, where some clients will win and others will lose in times of network congestion [5], [6].

Reference Architecture

Figure 3 illustrates a typical arrangement where HTTP adaptive streaming is used to deliver video and audio programming to a device in a subscriber's home. Note that a CDN is typically used to replicate segments within the core network and this is assumed to have infinite bandwidth. At the edge of the network, the bandwidth is constrained by:

1. The downstream path over DOCSIS.
2. The wireless network path to a Wi-Fi connected device.

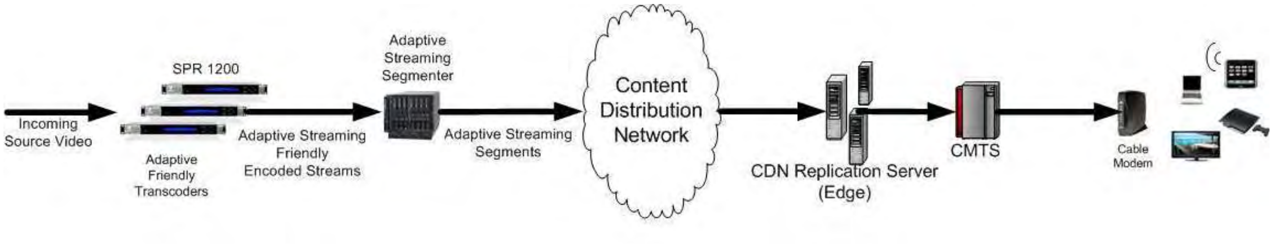


Figure 3: Reference architecture

Prototype System Description

Figure 4 shows the prototype system that was developed to analyze the behavior of standard HLS adaptive streaming clients on a shared access network.

Packet scheduling is determined by the bandwidth monitor/allocator. It also creates a virtual pipe and constrains all packet delivery within it. The virtual pipe can be dynamically resized while the system is running. Two scheduling algorithms can been implemented within the virtual pipe; best-effort and weighted fair queuing. Best-effort implements first-come, first-served packet delivery. The weighted fair queuing algorithm schedules transmission according to the virtual pipe size and compares the amount of data transmitted through various classes to ensure that each class is allocated its fair share.

The bandwidth monitor/allocator also logs bandwidth utilization data for use by the real-time statistics monitor and the graphing module. This data is used to visualize the behavior of the system and to understand the behavior of the adaptive streaming clients.

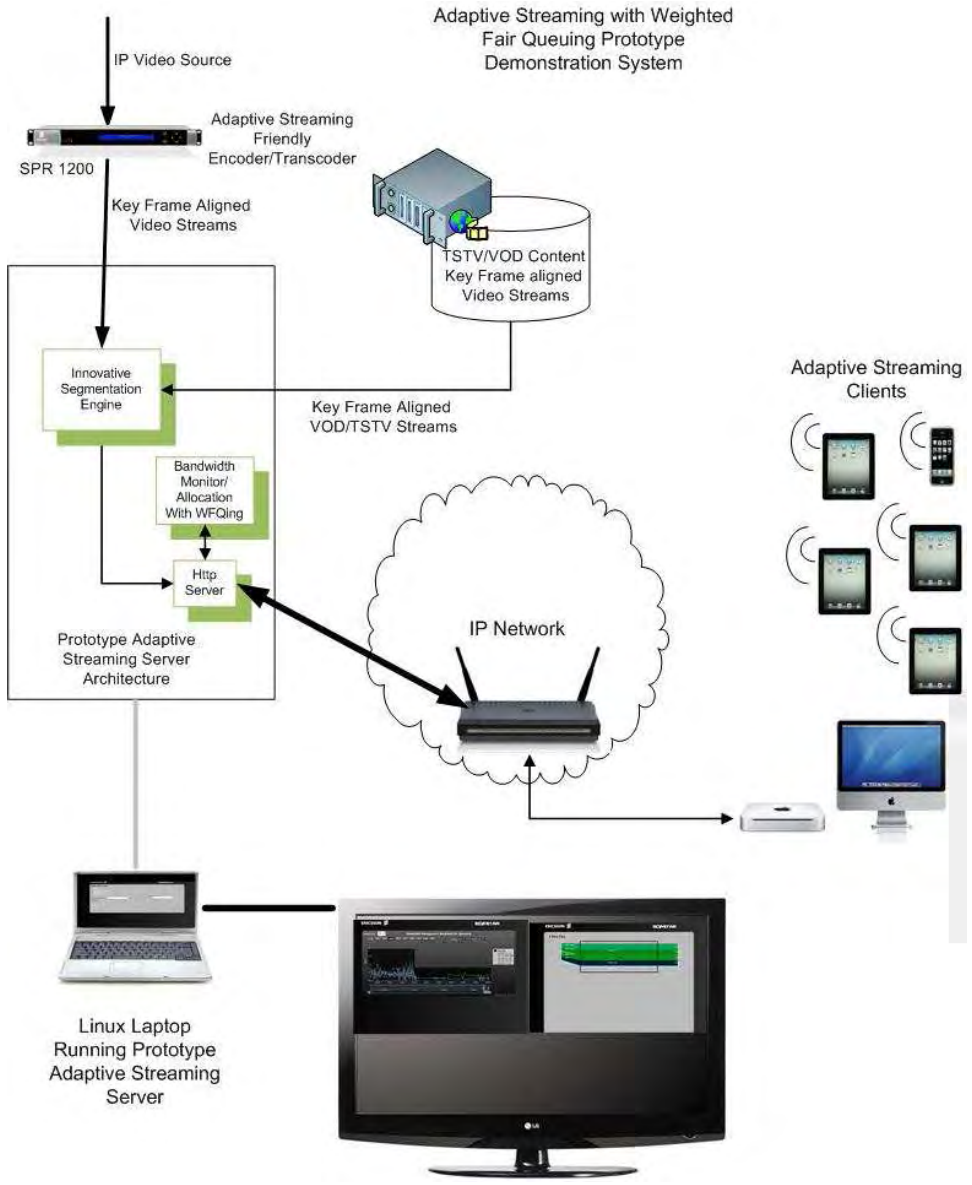

Adaptive Streaming with Weighted Fair Queuing Prototype Demonstration System
IP Video Source
SPR 1200
Adaptive Streaming Friendly Encoder/Transcoder
Key Frame Aligned Video Streams
TSTV/VOD Content Key Frame aligned Video Streams
Innovative Segmentation Engine
Key Frame Aligned VOD/TSTV Streams
Adaptive Streaming Clients
Bandwidth Monitor/ Allocation With WFQing
Http Server
IP Network
Prototype Adaptive Streaming Server Architecture
Linux Laptop Running Prototype Adaptive Streaming Server

Figure 4: Prototype system

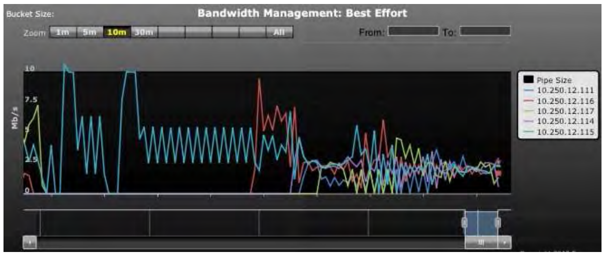
EXPERIMENTAL RESULTS

The experimental approach followed was to compare results of the best-effort (no traffic management) behavior with that of weighted fair queuing. The first best-effort run was repeated with identical parameters except for enabling the weighted fair queuing algorithm in the second and third runs. A fourth and fifth run were done with a different set of encoding profiles to investigate the effect that this would have on the behavior of the adaptive streaming clients.

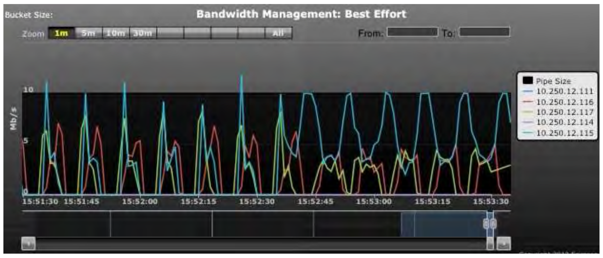
Run 1: Best Effort

Start Time	15:30:00 GMT
Pipe	10 Mbps
Content	How to Train your Dragon
Profiles	560, 660, 760, 860, 1000, 2000, 4000 Kbps
Clients	Mac Mini (115), iPad (112), iPad (114), iPad (111), iPhone (117)
Server	Best Effort

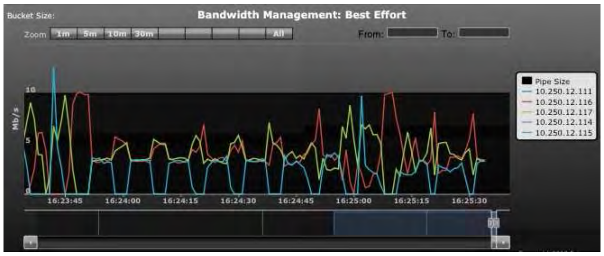
1. Each client started in sequence; all clients settled to 2 Mbps profile (Graph 1).
2. After 10 minutes iPad (114) goes to 1 Mbps profile; 9 Mbps pipe utilization.
3. Stopped iPad (114); 8 Mbps or 100% utilization (Graph 2).
4. Stopped iPad (111); 6 Mbps or 60% utilization (Graph 2).
5. After approximately 10 minutes, Mac Mini (115) jumped to 4 Mbps profile; still at only 80% pipe utilization.
6. Eventually, after approximately 30 minutes, system achieved 10 Mbps or 100% utilization (Graph 3).



Graph 1: Five clients started in sequence.



Graph 2: From four to three clients, 60% utilization.



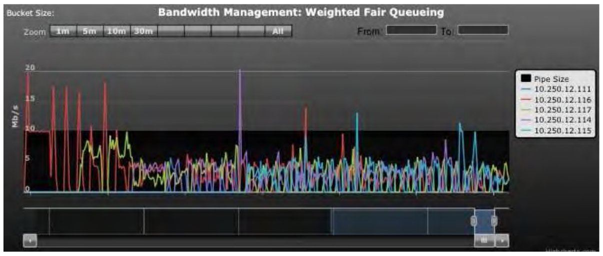
Graph 3: Final state achieved: 100% utilization.

This result was unexpected. It appears that the three clients (Graph 2) became locked in a synchronous pattern and as a result measured a lower segment-download rate than expected. As a result, none of the clients moved to a higher-rate profile, even though adequate bandwidth (4 Mbps) remained unused. Eventually (Graph 3) this pattern corrected itself, but not until approximately 30 minutes had elapsed.

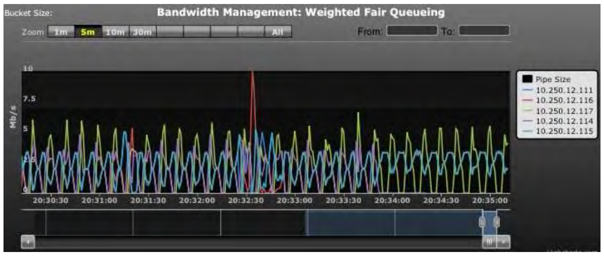
Run 2: Weighted Fair Queuing

Start Time	15:27:00 GMT
Pipe	10 Mbps
Content	How to Train your Dragon
Profiles	560, 660, 760, 860, 1000, 2000, 4000 Kbps
Clients	Mac Mini (115), iPad (112), iPad (114), iPad (111), iPhone (117)
Server	Weighted Fair Queuing
Weighting Factor	All clients set to 1

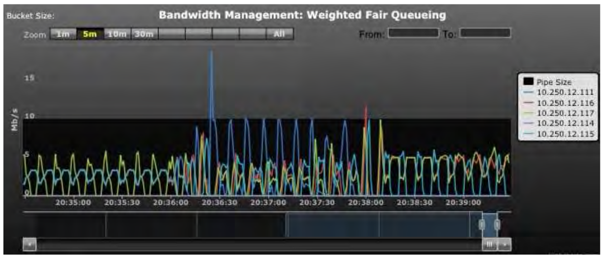
1. Each client was started in sequence (as before); same result as best-effort case (Graph 4).
2. iPad (111), iPad (116), and Mac Mini (115) all occasionally reached the 1 Mbps profile. Pipe stays at very close to 100% utilization.
3. Stopped iPad (114); immediately remaining clients achieved 100% pipe utilization. After 2 min iPad (111) reached the 4 Mbps profile (Graph 5).
4. Stopped iPad (111); pipe at 90% and then increased to 100% utilization (Graph 6).



Graph 4: Five clients, 100% utilization.



Graph 5: Four clients, 100% utilization.



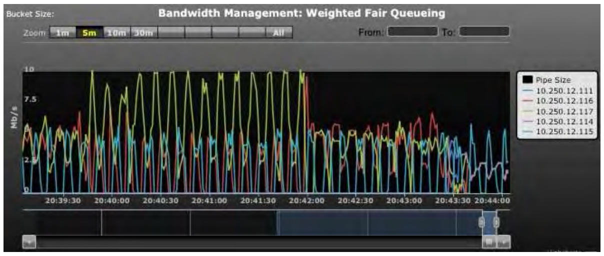
Graph 6: Three clients, 100% utilization.

It is apparent from Graphs 5 and 6 that the throughput of the system is much higher than in the best-effort case. In all cases, the pipe was utilized at, or close to, 100%. This may be understood by considering the scheduling algorithm at the HTTP server sequences through each partial segment download. Hence the clients take turns to achieve a more efficient segment download and therefore request a higher profile than in the best-effort case. The pipe throughput is maximized by the scheduling algorithm.

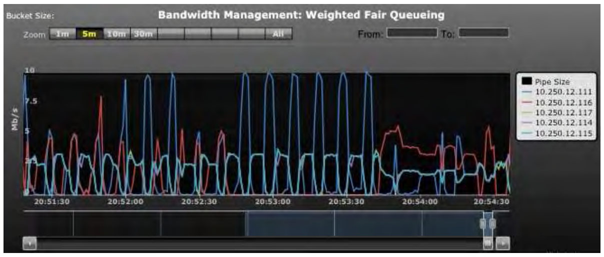
Run 3: Weighted Fair Queuing

Start Time	20:39:00 GMT
Pipe	10 Mbps
Content	How to Train your Dragon
Profiles	560, 660, 760, 860, 1000, 2000, 4000 Kbps
Clients	Mac Mini (115), iPad (112), iPad (114), iPad (111), iPhone (117)
Server	Weighted Fair Queuing
Weighting Factor	iPad(116) = 2, 3, 4

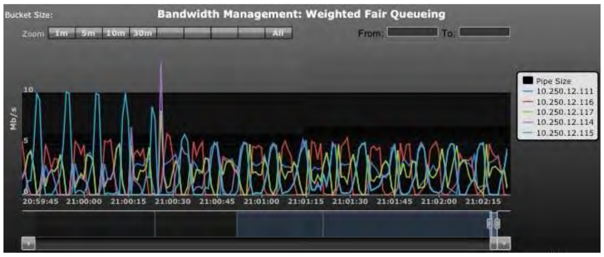
1. All clients started - no premium factor applied (Graph 7).
2. iPad (116) stopped.
3. Premium factor 2 applied to iPad (116) and started - quickly ramped to 2 Mbps (Graph 8).
4. Premium increased to 3. 115 went to 4 Mbps (momentarily).
5. Premium increased to 4. 114 went to 4 Mbps (Graph 9).



Graph 7: Fair network queuing.



Graph 8: Weighted fair queuing: iPad (116) weighting factor = 2.



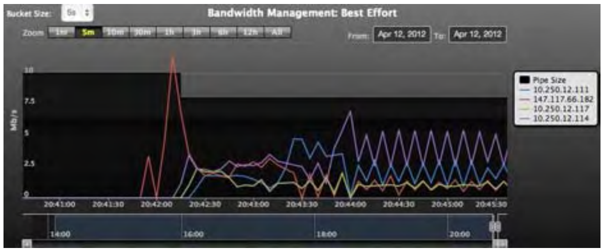
Graph 9: Weighted fair queuing: iPad (116) weighting factor = 4.

The weighting factor allows the premium client to achieve a higher profile than in Run 2, but it did not achieve the highest profile, probably because it was such a large jump in bit rate from 2 Mbps to 4 Mbps. Therefore, it was decided to re-run the test with a closer set of profile rates.

Run 4: Best Effort

Start Time	20:42:00 GMT
Pipe	8 Mbps
Content	Promo Reel
Profiles	400, 600, 910, 1200, 1600, 2000 Kbps
Clients	iPad (111), Mac Mini (182), iPhone (117), iPad (114), iPad (116)
Server	Best Effort

1. iPad (111), Mac Mini (182), iPhone (117), and iPad (114) started (Graph 10 and Figure 5).
2. 5th client iPad (116) started at 20:46:25 (Graph 11 and Figure 5).



Graph 10: Best effort, four clients.

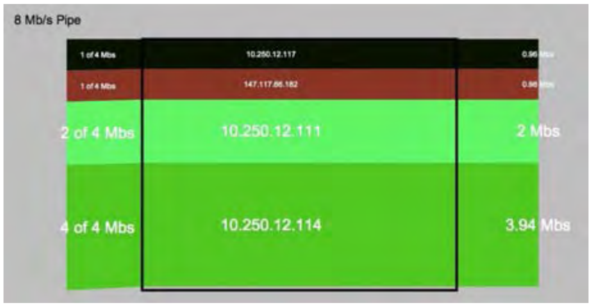
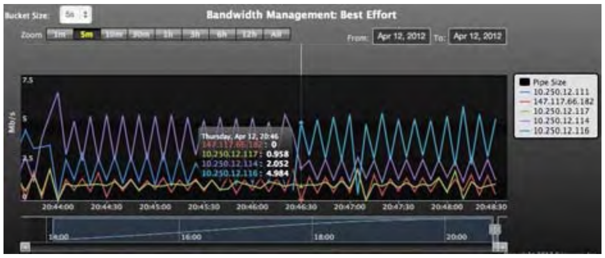


Figure 5: Best effort, four clients



Thursday, Apr 12, 20:46
147.117.66.182: 0
10.250.12.117: 0.958
10.250.12.114: 2.052
10.250.12.116: 4.984

Graph 11: Best effort, fifth client started at t = 20:46:25.

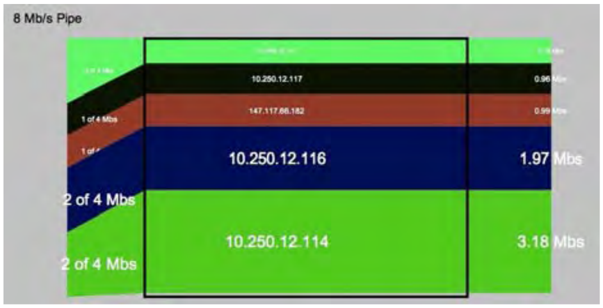


Figure 6: Best effort bit-rate allocation.

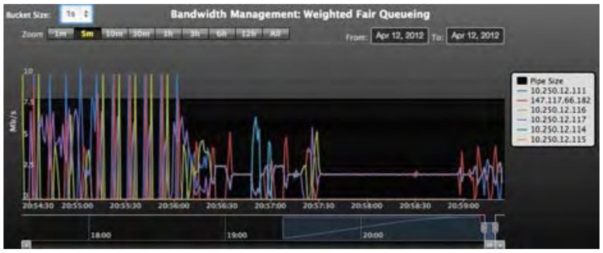
In this case, the bandwidth was shared unfairly. We hypothesize that the clients that first achieved the highest profile (2 Mbps) were able to maintain an unfair share of the pipe because of a positive feedback effect.

Run 5: Weighted Fair Queuing

Start Time	20:54:30 GMT
Pipe	8 Mbps
Content	Promo Reel
Profiles	400, 600, 910, 1200, 1600, 2000 Kbps
Clients	iPad (111), Mac Mini (182), iPhone (117), iPad (116), iPad (114)
Server	Weighted Fair Queuing
Weighting Factor	iPad(114) = 3

1 iPad (111), Mac Mini (182), iPhone (117), and iPad (114) started (Graph 12 and Figure 7).

2 iPad(114) with weighting factor of 3 started, and quickly ramped to 2 Mbps (Graph 13 and Figure 8)



Graph 12: Four clients, fair network queuing.

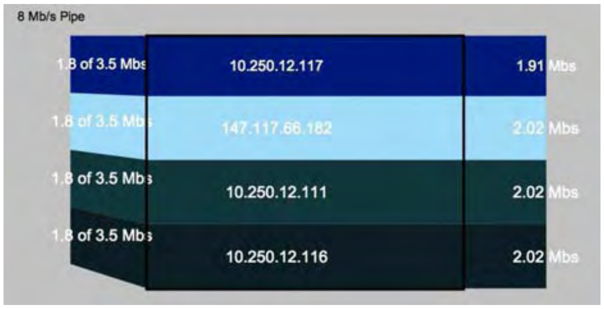
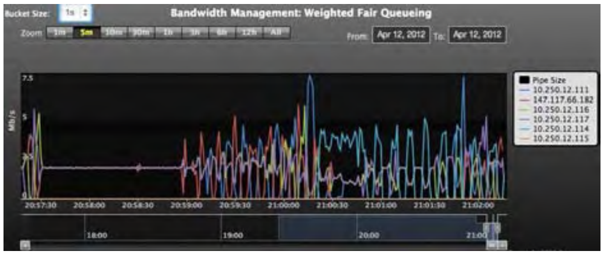


Figure 7: Four clients, fair network queuing.



Graph 13: Five clients, WFQ iPad (114) weighting factor = 3

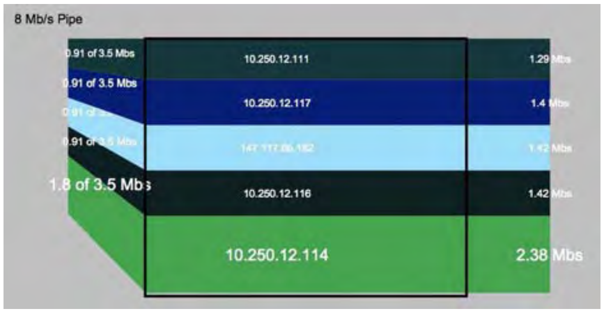


Figure 8: Five clients, WFQ iPad (114) weighting factor = 3

In this case, bandwidth was allocated equally (Figure 7) with four equally weighted clients. Subsequently, a greater share of bandwidth was allocated to a premium client (Figure 8) with a weighting factor of 3.

CONCLUSIONS

Implementation of a bit-wise, round-robin scheduler at the HTTP server can be used to effectively enforce fairness amongst HTTP adaptive streaming clients. In addition, a weighting factor may be established for a “premium” client to ensure that it experiences greater throughput during periods of access network congestion.

It appears that the weighted fair queuing mechanism [7] implemented in the prototype system is effective because it operates at the HTTP server layer, which is at a higher layer in the network stack than TCP congestion control alone [8], [9], and because it operates over a significantly longer time frame. If a client tries to “cheat” the system by requesting a higher profile than can be sustained by the network, it will only impact its own performance and not that of other clients.

IMPLEMENTATION OPTIONS

In order for the weighted fair queuing algorithm to be effective it must be implemented at the point in which traffic converges on a shared link in the access network. In the case of a DOCSIS access network, each downstream service group must be treated separately. The algorithm is implemented at the HTTP server (at the edge cache) and a virtual pipe must be created to each downstream service group (as illustrated in Figure 9).

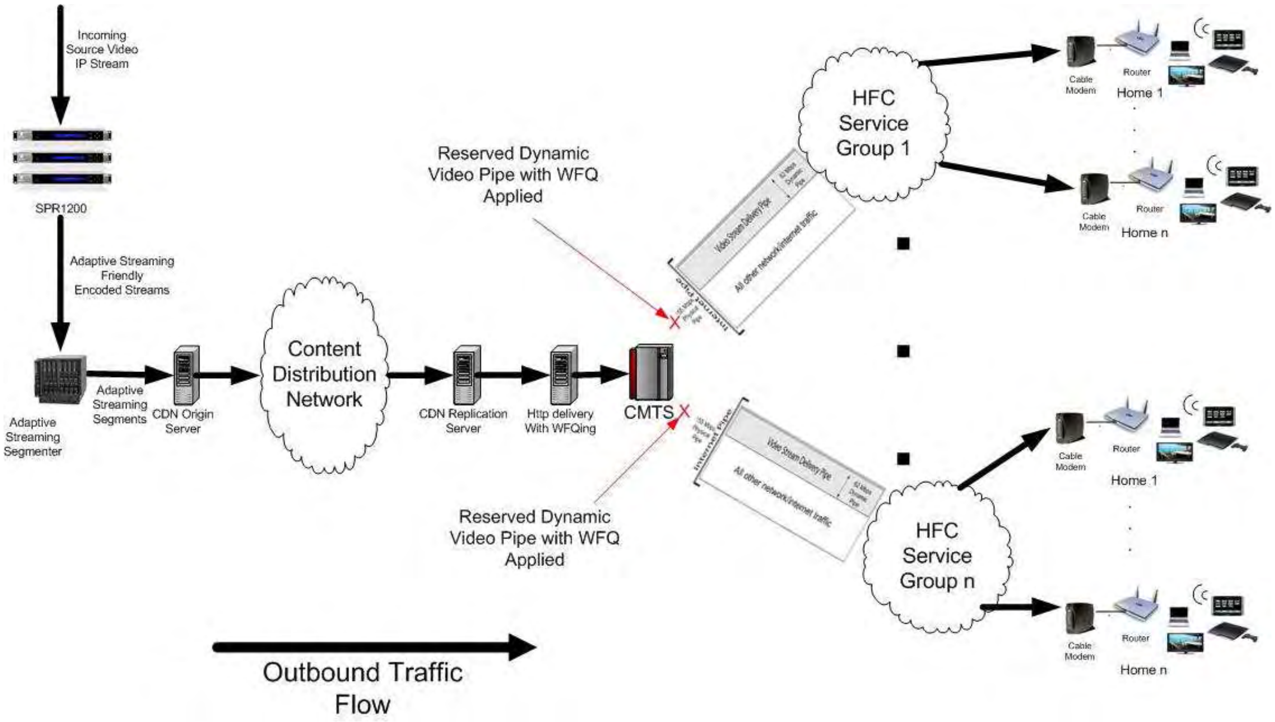

Incoming Source Video IP Stream
SPR1200
Adaptive Streaming Friendly Encoded Streams
Adaptive Streaming Segmenter
Adaptive Streaming Segments
CDN Origin Server
Content Distribution Network
CDN Replication Server
Http delivery With WFQing
CMTS
Reserved Dynamic Video Pipe with WFQ Applied
Internet Pipe
Video Stream Delivery Pipe
All other network/internet traffic
HFC Service Group 1
Cable Modem
Router
Home 1
Home n
Reserved Dynamic Video Pipe with WFQ Applied
Internet Pipe
Video Stream Delivery Pipe
All other network/internet traffic
HFC Service Group n
Cable Modem
Router
Home 1
Home n
Outbound Traffic Flow

Figure 9: Implementation in DOCSIS Network

References

1. R. Pantos and W. May. HTTP Live Streaming. IETF Draft, June 2010.
2. A. Zambelli. IIS smooth streaming technical overview. Microsoft Corporation, 2009.
3. Adobe HTTP Dynamic Streaming. http://www.adobe.com/products/hds-dynamic-streaming.html.
4. ISO/IEC 23009-1:2012 – Information Technology – Dynamic Adaptive Streaming over HTTP.
5. Saamer Akhshabi, Ali C. Begen, and Constantine Dovrolis. An Experimental Evaluation of Rate-Adaptation Algorithms in Adaptive Streaming over HTTP. Mac MiniSys'11, February 23–25, 2011, San Jose, California, USA.
6. Bing Wang, Jim Kurose, Prashant Shenoy, and Don Towsley. Multimedia streaming via TCP: An Analytic Performance Study. Department of Computer Science, University of Massachusetts.
7. Martin J. Fischer, Denise M. Bevilacqua Masi, and John F. Shortle. Simulating The Performance of a Class-based Weighted Fair Queuing System. Proceedings of the 2008 Winter Simulation Conference.
8. Robert Kuschnig, Ingo Kofler, and Hermann Hellwagner. An Evaluation of TCP-based Rate-Control Algorithms for adaptive Internet streaming of H.264/SVC. Institute of Information Technology (ITEC) Klagenfurt University, Austria.
9. Luca De Cicco and Saverio Mascolo. An Experimental Investigation of the Akamai Adaptive Video Streaming. Dipatimento di Elettrotecnica ed Elettronica, Politecnico di Bari.

A Comparison of Economic and Operational Tradeoffs for the Deployment of Broadcast, Multicast, and Unicast Infrastructures within an IP Video Environment

Carol Ansley, Jim Allen, Tom Cloonan
ARRIS

Abstract

As the Cable Industry evaluates the incorporation of IP Video as the next stage of video delivery, an important consideration is the need for analogues of the current Broadcast, Switched, and Unicast protocols within the IP Video deployments. Initially, it was assumed that the new IP Video world would look much like the current Legacy Video world, with its own architecture based on a triplet of protocols- one for Broadcasted (always-on Multicast) video, one for Switched (Multicast) Video, and one for Unicast Video. As the industry has continued to traverse the complex learning curve, this fundamental understanding has come into question.

Arguments have been made for the elimination of Broadcast, based on the idea that a Multicast deployment would provide increased network efficiencies. An opposing viewpoint is that a small Broadcast tier coupled together with a Unicast tier might provide greater network simplicity by eliminating the need for (and complexities of) a Multicast tier. This paper will use simulations based upon subscriber behavior to explore design approaches for several possible deployment scenarios. The analysis would consider network efficiency, possible economic factors, and possible feature interactions in an effort to help guide MSO decisions as they move forward towards future IP Video deployments.

ON VIDEO EVOLUTION

Legacy video services have long been the core of the basic cable service offering. In the distant past, these services were offered using NTSC-based analog programming, with one program per 6 MHz channel (in North America). In the past twenty years, this service has been augmented (and in some cases, replaced) with the arrival of MPEG-based digital video services transporting digital program streams over Quadrature-Amplitude Modulated (QAM) 6 MHz channels. This new service capitalized on advanced coding and compression techniques that permitted ten or more standard-definition program streams to be temporally multiplexed into a single 6 MHz channel (in North America).

In a legacy video environment, there are typically two distinct video service types offered to subscribers:

a) Linear video services
b) Video on Demand (VoD) services

Linear Video Services

Linear video services have been a part of the cable networks since their inception. Linear video services provide the “normally scheduled” program line-up to subscribers, with transitions between programs usually occurring at half-hour increments throughout the day. A program has a pre-assigned, scheduled time-slot when it is transmitted, and as many viewers as are interested can watch the broadcast program feed at the same time. Over the years a multicasting technology called Switched Digital Video (SDV) also evolved, reducing bandwidth demands by enabling program delivery only when a subset of one or more subscribers wanted to watch a program. Thus, we can define the two common methods used for the delivery of Linear video services:

a) Broadcast - each of the Linear program streams is transmitted from the head-end over the HFC plant to all of the subscribers. As a result, all programs consume bandwidth at all times, whether being viewed or not. However, Broadcast offers the benefit that only a single copy of the program needs to be transmitted into a particular Service Group when multiple users are viewing the program and a single headend signal can be split to accommodate any number of service groups within the same ad zone.
b) Switched Digital Video - only the Linear programs that are currently being viewed by one or more subscribers within a Service Group are transmitted from the head-end to that Service Group. Linear programs that are not being viewed are not transmitted, so bandwidth savings result relative to Broadcast techniques of transmission. These bandwidth savings do not come for free, because they do require a two-way protocol to exist between the client devices and a head-end management system. The actual magnitude of the bandwidth savings over straight broadcast depends on many factors, which are discussed later in this paper. Like Broadcast, SDV offers the benefit that only a single copy of the program needs to be transmitted into a particular Service Group regardless of the number of users viewing the program.

Video On Demand Services

VoD services permit offerings such as standard Video on Demand, Network Digital Video Recording (nDVR), and Start-Over. VoD services offer subscribers access to an extended library of stored video content. These “extra” programs are traditionally provided as a free or fee-based service. Viewers can select a program from the Video on Demand content library at their convenience. They can start and stop the program as they wish, and often trick modes such as fast-forward, rewind, and pause are available. Since it is unlikely that two subscribers will choose to watch the same VoD program at exactly the same time, no effort is made to broadcast or multicast VoD content. It is simply unicast to the single user who has requested the content. Unlike Broadcast and SDV feeds, each new VoD selection must be sent individually to each new viewer, so there is a one-for-one utilization of bandwidth for each new stream.

IP Video Services

Just as MPEG-based digital video services were used to augment and in some cases replace analog video services over the past twenty years, a new technology is now being viewed by the cable industry as a potential augmentation (or eventual replacement) for MPEG-based digital video services. This new entrant capitalizes on the recent advances in DOCSIS technology and advances in video delivery over IP.

IP Video delivers encoded and compressed video program content from origin servers to client devices by inserting the audio and video information into the payloads of Internet Protocol (IP) packets that are then passed over

IP networks. IP Video architectures have the potential to enable support of new end devices and new revenue opportunities based on personalized advertising or expanded video services.

As MSOs begin to architect new video delivery systems to take advantage of IP Video techniques, the video delivery models that have been successfully utilized in the legacy video delivery world come first to mind. As such, one would expect that MSOs might consider IP Video as a delivery system for all of the following service types:

1. IP Video VoD services
    a. Standard VoD
    b. nDVR
    c. Start-Over
2. IP Video Linear services
    a. IP Video Linear Always-On services (similar to legacy Linear Broadcast services)
    b. IP Video Linear Switched services (similar to legacy Linear SDV services)

It should be clear that the various IP Video VoD services will likely be delivered using point-to-point IP Video unicast delivery. These services will likely be based on the latest IP/TCP/HTTP transport technologies, the dominant protocol stack used for unicast IP Video Streaming. The increasing use of non-television devices to access this content also suggests that reuse of popular Internet technology would be advantageous, when it fits with the unique cable industry infrastructure.

What is less clear is the most efficient method or methods to implement the delivery of traditional Linear Video services over IP. There are many ways to emulate Linear Video delivery within an IP environment so that, in the end, the video is ultimately delivered via IP packets to client devices within the subscriber's home. Some of the possible techniques that are currently under consideration are enumerated here:

a) Point-to-point, unicast IP/TCP/HTTP packet streaming from head-end origin servers (or caching servers) over DOCSIS to each individual subscriber client device, requiring lots of point-to-point connections to be established for popular programs
b) Dynamic, point-to-multipoint, multicast IP/UDP packet delivery from head-end origin servers (or caching servers) over DOCSIS to any subscriber clients that join the multicast stream
c) Always-On, point-to-multipoint, multicast IP/UDP packet delivery from head-end origin servers (or caching servers) over DOCSIS to any subscriber clients that join the multicast stream
d) Legacy Linear MPEG-TS transmission on the HFC plant, with IP Encapsulation in a residential Media Gateway and unicast IP/TCP/HTTP Video delivery within the home network, re-uses the existing MPEG-based delivery infrastructure and reduces IP Video architectural complexity

It should be noted that of the various techniques listed above, the first three utilize IP delivery techniques to the home. The final technique utilizes a unique hybrid approach, where the content is sent via traditional MPEG-TS delivery over the HFC network, but then uses an IP unicast stream for final delivery over the home network. This last technique, while valid for consideration as a method for IP video delivery overall, will not be discussed further in this paper. We will concentrate on discussion of IP Video architectures that use IP in the transport arena.

Another layer of complexity that is not addressed in this paper is the Quality of Service (QoS) architecture for IP Video. Within DOCSIS, there is a substantial QoS infrastructure that can preserve or improve the performance of individual IP Video flows with respect to the overall volume of IP traffic. Subscribers today are accustomed to always-on TV service from their perspective, SDV is typically engineered to be indistinguishable from legacy broadcast delivery. An important part of the IP video architecture will involve the decision to preserve, or not, the current levels of video service reliability and the implementation of that decision. While it has some relevance to the protocol topics discussed in this paper, the QoS topic is complex enough to warrant another discussion focused on that topic.

With IP technologies, each technique for transport (unicast, multicast, broadcast) has its own set of advantages and disadvantages. For example, Unicast is relatively simple to deploy since it is based on variants of basic HTTP transactions. It currently being used by several MSOs to provide some IP-based subscribers with the a limited equivalent of Linear services, but a unicast approach can be wasteful of the limited HFC bandwidth if any unicast program is actually sent to more than one subscriber at the same time. Unicast delivery also suffers from a "simulcast effect" that may exist if early deployments of IP video begin while legacy video distribution to legacy STBs is also in place, as the same programs will need to be simulcast across both distribution systems.

Multicast, in dynamic or static varieties, can be more complex to deploy since it may require an additional headend server to support bandwidth management, similar to SDV. Depending upon the current configuration of a headend's routers, switches and CMTSs, they may also require upgrades to support multicast protocols. If these multicast or broadcast techniques are eventually utilized, they will yield bandwidth savings on the HFC plant due to the fact that multiple viewers of a single stream within a particular Service Group will not require extra replications. A dynamic multicast approach is more bandwidth efficient than a static Always-On approach. The multicast approaches may also suffer from the "simulcast tax" mentioned above, but the overall bandwidth cost of that simulcast may be reduced by the inherent advantages of dynamically switching a program in only when it is actually to be viewed.

The rest of this paper attempts to quantify and explore the many tradeoffs associated with these technologies.

UNICAST IP VIDEO DELIVERY

It is important that we clearly define the protocols that we have analyzed in the paper for delivery of unicast IP Video. If it is mapped into the layers of the Open System Interface (OSI) model, IP Video clearly uses IP as its Layer 3 (Network Layer) protocol. However, it can use any one of two different Layer 4 (Transport Layer) protocols: Transmission Control Protocol (TCP) or User Datagram Protocol (UDP). TCP is a connection-oriented protocol that provides guaranteed packet delivery, flow control, and congestion control for the data transport, whereas UDP is a simpler, connectionless protocol that provides none of the advanced services of TCP.

During the early days of IP Video (in the early-1990s), the content was initially delivered to the home using Ethernet/IP/UDP/RTP encapsulations. Custom players and custom servers were usually utilized. It worked fairly well, but it did run into issues with in-home NAT boxes and congested networks.

The original IP Video Download protocols were used for over a decade, and they are still used (to some extent) today. However, the latest improvements in IP Video delivery began to be utilized in the middle of the 2000 decade. This new approach is oftentimes called HTTP-based Adaptive Streaming.

HTTP-based Adaptive Streaming has come to be used quite extensively in most applications that require a unicast IP Video stream to be delivered from a single origin server to a single client device. The application program typically uses TCP transport services for downloading fragments of the video content file by invoking Hypertext Transfer Protocol (HTTP).

HTTP-based Adaptive Streaming replaced the single HTTP GET message of the first Downloading protocols with a series of repeated HTTP GET messages, with each HTTP GET message requesting a different, small chunk (or fragment) of the video content file. As a result, only the video content that is to be viewed is actually requested, so the problems associated with wasted bandwidth are minimized. In addition, since the video fragments tended to be fairly short in duration (2-10 seconds was typical), it was easy to efficiently support simple trick modes. The short-duration fragments also made it possible for the clients to rapidly identify network congestion and adjust their HTTP GET messages to request higher or lower resolution fragments that could be accommodated by the available network bandwidth at any instant in time. These rapid adjustments in the resolution (and bit-rate) of successively-requested video fragments came to be known generically as HTTP-based Adaptive Streaming.

Current Unicast IP Video is essentially a TCP-based, HTTP pull model, with unicast packets only being sent from the source to the destination whenever the destination requests the content (with HTTP GETs). Other than the routing tables that help to steer the packets, no other state information is required within the intermediate network elements to ensure correct transmission of the packets between the source and the destination. The typical control plane protocols and data plane exchanges for Unicast IP Video are illustrated in **Fig. 1** and **Fig. 2**.

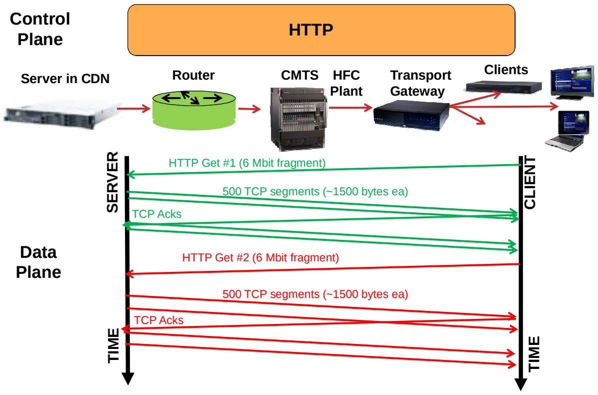

Control Plane
HTTP
Server in CDN
Router
CMTS
HFC Plant
Transport Gateway
Clients
SERVER
CLIENT
HTTP Get #1 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
Data Plane
HTTP Get #2 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
TIME
TIME

Figure 1 - Unicast over Transport Gateway

Fig. 1 illustrates an example unicast architecture where clients send HTTP GETs directly to the head-end video server, and a Transport Gateway merely passes the upstream and downstream packets between the HFC plant and the Home Network.

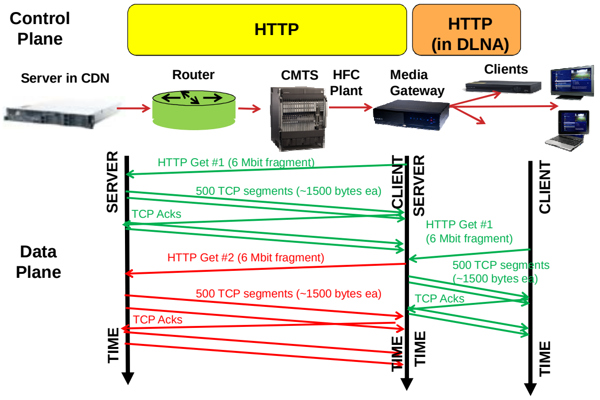

Control Plane
HTTP
HTTP (in DLNA)
Server in CDN
Router
CMTS
HFC Plant
Media Gateway
Clients
SERVER
CLIENT
SERVER
CLIENT
HTTP Get #1 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
HTTP Get #1 (6 Mbit fragment)
Data Plane
HTTP Get #2 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
500 TCP segments (~1500 bytes ea)
TCP Acks
TIME
TIME
TIME
TIME

Figure 2 - Unicast over Caching Media Gateway

Fig. 2 illustrates an example unicast architecture where clients send HTTP GETs through a DLNA network within the home to a server application running on the Media Gateway within the home, and the Media Gateway also has an HTTP client application running on it that would have (hopefully) previously used HTTP GETs to request and cache fragments from the head-end video server. If the content was not cached, then the Media Gateway would simply relay the HTTP GET upward towards the head-end video server.

Caching operations in the network may be added to reduce traffic on the back bone network, but do not appreciably add to overall architectural complexity.

MULTICAST IP VIDEO DELIVERY

The use of Multicast for IP Video delivery improves bandwidth efficiencies over Unicast video delivery on the HFC plant as well as on the MSO's back-office network. This fact is simply illustrated in **Fig. 3**, where we have assumed that two users (Client #1 and Client #2) are both accessing the same linear video content at the same time. For comparative purposes, we will assume that the bandwidth associated with this video content is 7 Mbps. If delivered using Unicast, then the resulting bandwidth consumed in both the HFC plant and the MSO back-office network is 14 Mbps, since two separate streams containing the video content must be propagated through the network. If delivered using Multicast, then the resulting bandwidth consumed in both the HFC plant and the MSO back-office network is only 7 Mbps, since only a single stream containing the video content is propagated through the network- both CM #1 and CM #2 receive and pass the stream on to their respective clients, resulting in the inherent "replication" of the stream near the stream destinations.

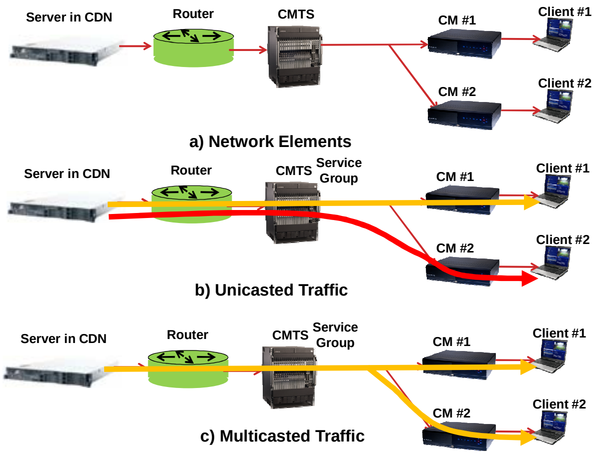


Figure 3 - Unicast vs. Multicast

While Multicast IP Video delivery is more bandwidth efficient for streams that are simultaneously viewed by more than one recipient, it is also more complex to manage than unicast IP Video delivery. This added complexity is primarily due to the fact that multicast IP Video requires additional protocol support in the intermediate network elements and the client devices.

Multicast IP Video is quite different from Unicast IP Video. Since there are multiple destinations receiving the Multicast IP Video feed, the TCP-based, HTTP pull model used in

Unicast IP Video cannot be utilized for Multicast IP Video.

As a result, a UDP-based push model is used for IP Multicast, with packets being sent from the source to the multiple destinations without HTTP GETs or TCP ACKs being required. While one could (in theory) send the IP Multicast to all possible destinations, that approach would be quite wasteful of both bandwidth within the network and processing power within all of the destinations. As a result, standard IP Multicast solutions limit the scope of destinations to which the multicast streams are sent. In particular, the multicast streams (which are identified by a particular Multicast Group IP Address as the Destination Address within the IP packet header) are only sent to destinations that have formally requested that the stream be transmitted to them. This formal request is typically made within a LAN using the Internet Group Multicast Protocol (IGMP) for IPv4 systems and using the Multicast Listener Discovery (MLD) protocol for IPv6 systems.

In both cases (IGMP and MLD), the destination desiring access to the content within a multicast stream would typically use the appropriate protocol to send a "Join Message" (a.k.a. a Membership Report or a Multicast Listener Report) that would be broadcast to the router(s) in its LAN. If/when the destination desires to no longer receive that particular multicast stream, it can optionally send a "Leave Message" (a.k.a. a Leave Group or a Multicast Listener Done). In order for routers to stimulate destinations to report that they are joined to a particular group, they would typically send a "Query Message" (a.k.a. a Membership Query or a Multicast Listener Query). Routers must maintain state information indicating which of their ports have listeners that have indicated a desire (via Join Messages) to receive each multicast stream. The routers then must forward packets associated with each particular multicast streams to the ports that have listeners associated with that multicast stream. Routers typically communicate their desire to receive a multicast stream from other routers using one of several possible multicast routing protocols, including PIM-SSM, PIM-SM, PIM-DM, DVMRP, MOSPF, MBGP, and CBT. The routers involved in multicast address exchanges must be capable of communicating using a common multicast routing protocol.

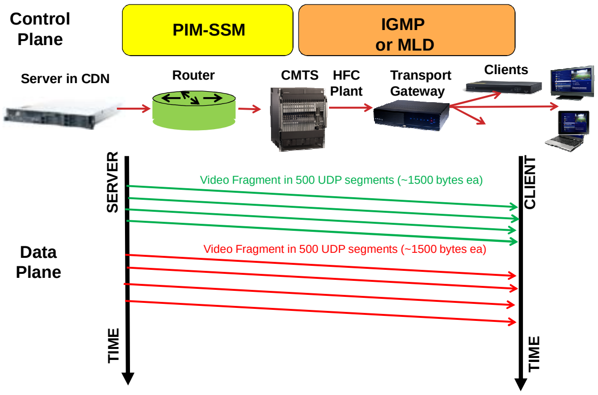


Figure 4 - Multicast with UDP

There are many different architectures that one can envision for the deployment of a Multicast IP Video delivery system- several of them are illustrated below. **Fig. 4** illustrates an example architecture with a data plane that uses UDP transport of multicast IP Video packets from the head-end multicast server to the clients in the home, with the packets passing through a Transport Gateway within the home. The control plane within **Fig. 4** uses IGMP or MLD to establish the multicast path between the clients and the CMTS, and it uses PIM-SSM to establish the multicast path between the CMTS and back-office routers and back-office multicast server.

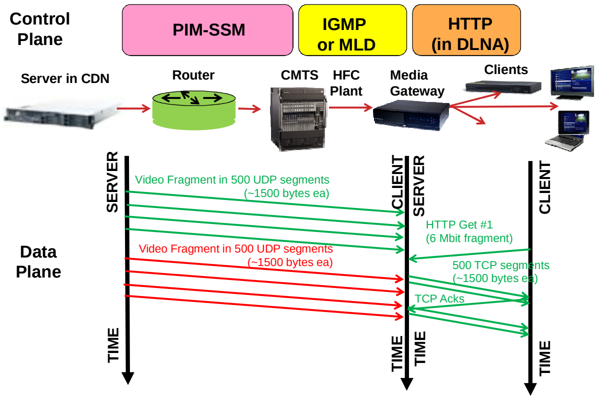

Control Plane
PIM-SSM
IGMP or MLD
HTTP (in DLNA)
Server in CDN
Router
CMTS
HFC Plant
Media Gateway
Clients
SERVER
CLIENT
SERVER
CLIENT
Data Plane
Video Fragment in 500 UDP segments (~1500 bytes ea)
Video Fragment in 500 UDP segments (~1500 bytes ea)
HTTP Get #1 (6 Mbit fragment)
500 TCP segments (~1500 bytes ea)
TCP Acks
TIME
TIME
TIME
TIME

Figure 5 - Multicast with UDP & Conversion to Unicast TCP

Fig. 5 illustrates an example architecture with a data plane that uses UDP transport of multicast IP Video packets from the head-end multicast server to the Media Gateways in the home, with the video content file being re-constituted by the Media Gateway. The Media Gateway then acts as an HTTP server to distribute the video content over a unicast HTTP/DLNA connection to the HTTP client within the Home Network. The control plane within Fig. 5 uses HTTP/DLNA within the Home Network, it uses IGMP or MLD to establish the multicast path between the Media Gateway and the CMTS, and it uses PIM-SSM to establish the multicast path between the CMTS and back-office routers and back-office multicast server.

The widespread deployment of SDV has also proven in many parts of the multicast technology that are directly applicable to multicast IP Video distribution. Many optimizations that were developed to make SDV robust and efficient can be extended to the IP Video world to enable IP Multicast to be successful. One example technique would be the automatic joining of all available SDV multicast streams by an Edge QAM even before any specific end device has chosen one of those programs. This pre-join speeds up the acquisition of a new program by a client, as the Edge QAM merely needs to be instructed by an SDV server which stream to activate on which QAM and PIDs. This feature is not a part of traditional multicast as practiced by IT professionals, but it an obvious improvement directly applicable to a CATV IP Video architecture. The analogous IP Video feature would instruct the CMTS to join all IP Video multicasts, which would let a device activate a new stream with just a transaction with its CMTS. Since the CMTS manages its own bandwidth constraints, the SDV Server's bandwidth allocation might be transferred entirely to the CMTS, resulting in no new network elements for multicast.

COMPARING UNICAST AND MULTICAST IP VIDEO

The primary benefit of Multicast IP Video delivery is its basic ability to reduce the bandwidth required to deliver video content to multiple destinations when two or more of those destinations are viewing the same content at the same time. Many MSOs already treat bandwidth on their HFC networks as a critical and precious resource as multiple services compete for that bandwidth and the situation can only become more contentious as HSD continues to increase its requirements and video any time anywhere continues as well. If these trends continue, then the primary benefit of Multicast IP Video may prove to be very important.

Unicast IP Video delivery also has a place, even with its bandwidth usage, since it can provide a simple deployment model for early stages of IP video deployments, when the concentration of IP video users in any one service group is low. Trends within the universe of any time any place video distribution may also tend to accelerate the usage of network DVR and other unicast services, which will increase the amount of natively-unicast traffic.

Depending upon the stage of the IP Video deployment and the deployment choices connected with other related areas, such as network DVR, the answer of what may be the most efficient may vary depending upon whether one considers network bandwidth, operational/deployment costs, and service flexibility. When considering a real world deployment, the answer may even be that a mix of technologies will be required to ensure that MSOs can obtain an optimal efficiency from their HFC plant for video delivery.

Some of the issues to be considered are listed below.

1. Common protocols for multicast IP Video and any optimizations over DOCSIS should be available in an open forum, similar to the TWC ISA or Comcast NGOD SDV specifications
2. Any optimizations for Unicast IP Video that allow robust performance for first screen viewing should be provided in an open forum for maximum benefit
3. Current encoding methods will require multiple choices for unicast and/or multicast stream delivery, new encoding choices, such as SVC, could improve multicast efficiency for multicast and stream management for unicast
4. Reliability concerns in the IP Video packet delivery
5. Distributed Denial Of Service attacks by hackers on head-end equipment
6. Multicast must be tied into a Connection Admission Control algorithm to identify overload conditions when a new Multicast stream cannot be set up

While the issues listed above should be carefully considered, it is important to note that many technical and architectural proposals have already been created to mitigate most of the issues. Granted, some of these proposals require more complexity to be added to the equipment, but they nevertheless provide solutions to the problems.

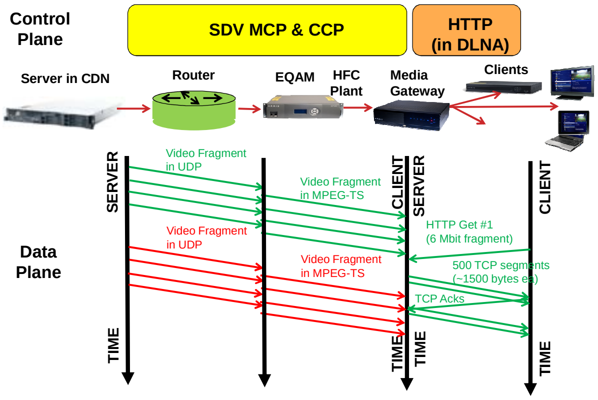


Figure 6 - Multicast with UDP & MPEG-TS & Conversion to Unicast TCP

Fig. 6 illustrates an example hybrid architecture with a data plane that uses UDP transport of multicast IP Video packets from the head-end multicast server to the head-end EdgeQAM, MPEG-TS transport from the head-end EdgeQAM to the Media Gateways in the home. The Media Gateways re-constitute the video content file. The Media Gateway then acts as an HTTP server to distribute the video content over a unicast HTTP/DLNA connection to the HTTP client within the Home Network. The control plane within **Fig. 6** uses HTTP/DLNA within the Home Network, and it uses SDV-oriented protocols like the Channel Change Protocol (CCP) and the Mini-Carousel Protocol (MCP) to establish a video stream flow between the back-office server and the Media Gateway.

SIMULATION RESULTS

The fact that Broadcast, Unicast and Multicast are closely related protocols allows us to simulate their respective behavior using a common simulation base – modeling both Broadcast and Unicast as special cases of the more general Multicast model. Of these three protocols, however, only Unicast natively permits trick modes such as pause and replay – a quality that, though it may be quite valuable to viewers, has no correspondence in the other two protocols. We have focused, therefore, on modeling only properties (listed below) that can be used to describe all three protocols.

Broadcast can support an arbitrarily large number of viewers (when the downstream program capacity is sufficient to carry every program in the lineup). Unicast, on the other hand, can support an arbitrarily large number of offered programs (when the downstream capacity is sufficient to dedicate a separate program channel for every viewer). Multicast, however, possesses both of these properties and is also able to provide bandwidth-efficient service even when either of the two above constraints on the downstream program capacity cannot be met.

Viewer Modeling Parameters

Two attributes of a video delivery network lie largely outside the control of the MSO. These properties can be measured but not controlled by the MSO. Numerical values for these parameters are best attained through careful analysis of actual viewer tuning behavior. These attributes are:

1. Acceptable Tuning Blockage Probability
2. Program Viewership Popularity

Customers will ultimately decide with their feet how often (relative to competing providers) they are willing to tolerate being denied a program that they have requested. Video service providers, however, are forced to make a reasonable guess at exactly what this limit of viewer tolerance might be, as we know of no applicable field study in this area. Throughout this paper we have assumed viewers will be satisfied if they are denied a program selection request no more than 0.1% of the time (or once per 1000 tuning requests).

It is also the customer population that determines the relative popularity of each of the programs offered in the lineup. Modeling this property of the viewer population can present a significant challenge since relative program popularity varies substantially with time-of-day, day-of-week and with the demographics of the neighborhood served by the service group.

While neither Broadcast nor Unicast services are sensitive to program popularity, the relative popularity of programs in the offered lineup plays a significant role in Multicast by determining how many programs can be expected to be multiplexed onto a limited amount of downstream bandwidth.

Fortunately the dynamic nature of a Multicast protocol causes it to automatically adapt to changes in relative program popularity (both temporally and also between service groups). This means that it is not so important to know exactly which programs are most popular – only that we know in a general sort of way.

A number of studies have suggested that if we first sort a program lineup by market share, from most to least popular, then the popularity or market share of a program (n) can be approximated using a Power Law Distribution, shown here:

$$P_n = n^{-\alpha} / \sum_{n=1}^{N} n^{-\alpha}$$

In this equation $\boldsymbol{P_n}$ represents the probability that a randomly chosen viewer is currently watching program n (from a lineup with a total of N possible choices). The parameter, alpha (α), can take a value only between 0 and 1 and should be chosen to provide the best fit to actual field data. Like any Probability Density Function (PDF) the total area under the curve must always be zero. Figure 7 shows the shape of typical Power Law Distributions for various population sizes. Although the Power Law is at best an approximation of an actual program lineup popularity, we have found that an alpha value around 0.8 provides a fairly reasonable first order approximation of many actual field measurements. Except where explicitly stated otherwise, we have used this value in this paper.

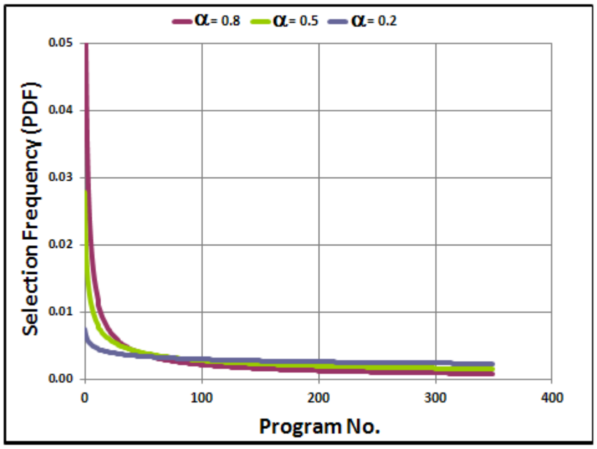


Figure 7 - Simulated Popularity Curves

The next figure illustrates normalized program popularity curves from 4 sample Service Groups. Each service group had about 400 programs available and had between 300 and 500 settop boxes. The curve was developed by accounting for all channel dwell times across 1 week.

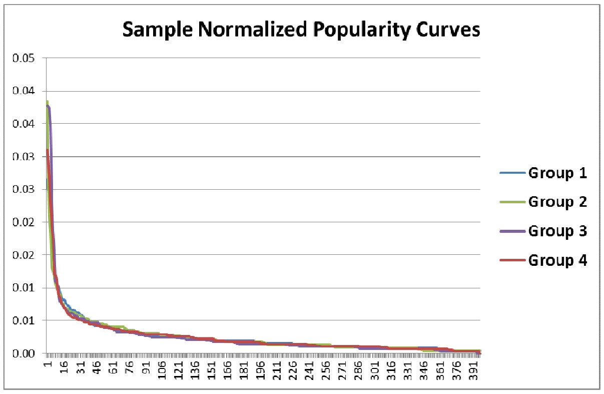


Figure 8 - Sample Popularity Curves

These curves illustrate that the Power law approximation holds up fairly well across a range of service group sizes.

<u>Network Modeling Parameters</u>

In addition to the viewer modeling parameters discussed above, three more attributes are required to model characteristics of the video network that are very much under the control of the MSO. These are:

1. Number of Offered Programs
2. Downstream Program Capacity
3. Number of Viewers in a Service Group

The challenge for network designers is to optimize these parameters to provide the maximum level of service to the viewers at an affordable equipment cost. These are not, however, three independent variables. Once any two of these three variables are chosen the value of the remaining parameter is dictated by the values chosen for the first two under the constraints imposed by the level of blocking deemed acceptable and the popularity profile of the offered program lineup.

Of these three attributes, the service group size will normally be the property that varies the most between network nodes and is least likely to be precisely determined at network design time.

This paper uses software simulations, employing Monte Carlo techniques, to model and chart the relationships among these attributes. Results of these simulations are shown in the following sections.

Downstream Program Capacity

Figure 9 shows simulation predictions for the downstream program capacity (as a function of the size of the service group) required to provide viewers with a lineup of 200 programs using each of the three video delivery protocols. For the purposes of this simulation, all programs are assumed to require the same amount of bandwidth. The chart contains Multicast curves consistent with a 0.1% blocking probability for three different values of alpha – showing the sensitivity of Multicast performance prediction over quite a wide range of values.

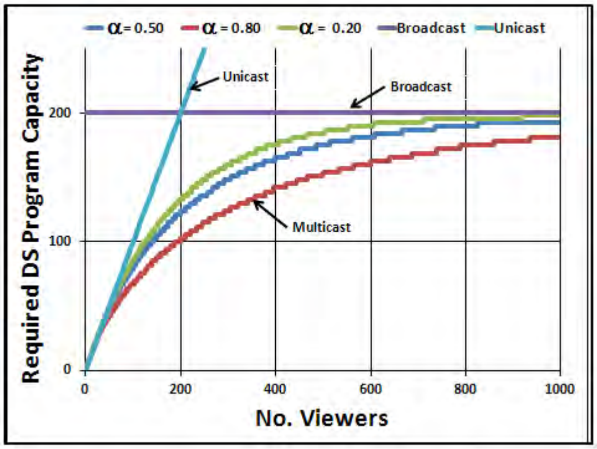


Figure 9 - Required Capacity vs. No. Viewers

Broadcast, of course, always requires a constant amount of downstream capacity (sufficient to carry a single copy of every offered program). Unicast, on the other hand, requires a separate downstream program channel for each individual viewer. The curve for the Multicast service is asymptotic to Unicast for very small service group sizes (very small numbers of viewers are likely to each select a different program). As the service group size gets very large the Multicast curve becomes asymptotic to the Broadcast service (since every program in the lineup will likely be selected by at least one of the very large number of viewers).

It is in the intermediate service group sizes that Multicast can be seen to require less downstream capacity than either of the other protocols. The vertical distance between the Multicast curve and either of the other protocols represents the downstream channel capacity that can be saved by using Multicast rather than the other protocol.

Program Lineup Size

A chart like the one in Figure 9 can tell us the relationship between downstream capacity and service group size for a known program lineup. Often, however, it may be that downstream program capacity is constrained a priori and we would like to know the relationship between the service group size and the number of programs that we could provide in the program lineup.

The next figure assumes that downstream capacity is available for only 100 simultaneous video programs with the resulting relationship between the service group size and the number of programs that could be offered.

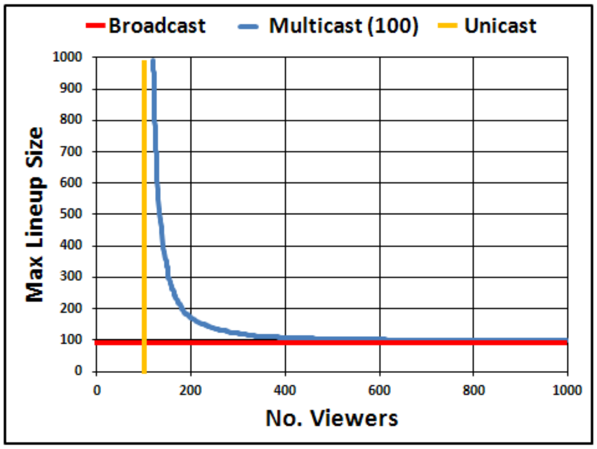


Figure 10 - Maximum Lineup Size

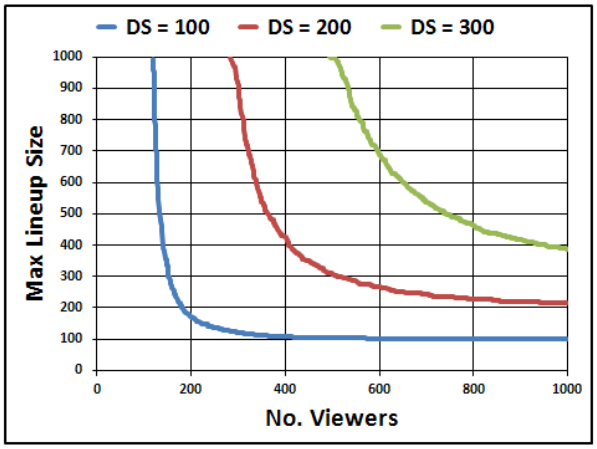


Figure 11 - Lineup Size vs. Downstream Program Capacity

Again we see that the Multicast curve is asymptotic to Broadcast service for very large service group sizes and to Unicast for small service group sizes, still assuming the same 0.1% blocking probability. The straight horizontal line corresponding to Broadcast service shows that Broadcasting always requires a separate program channel for each offered program, but can support an arbitrarily large service group. The straight vertical line corresponding to Unicast reveals that Unicast can support an infinite number of offered programs (when the downstream program capacity is greater than the number of viewers in the service group) but cannot support even a single viewer more without failing to meet the required blocking probability.

The vertical distance between the Multicast and Broadcast curves shows how many more programs Multicast could support in the program lineup (as a function of the service group size). The horizontal distance between the Multicast and Unicast curves, on the other hand, shows how many more viewers could be in a Multicast service group (as a function of the number of offered program choices).

Figure 11 shows the same curve (for 100 downstream Multicast program channels) but adds two more curves – for 200 and 300 Multicast downstream program channels. These curves seem to indicate that the power of Multicast service (i.e., the distance from the Multicast curve to either of the asymptotes) increases significantly for larger numbers of downstream program channels and for larger program lineup sizes. Curves for smaller numbers of downstream program channels (like Figure 9) closely hug both asymptotes with only a fairly narrow range of service group sizes in which Multicast shines relative to the other protocols.

This behavior suggests that a network evolution plan that begins by transferring a small number of Broadcast programs onto a small amount of downstream Multicast bandwidth may not immediately experience the full advantage that might come later when a larger program lineup is offered via Multicast. This finding also has significance for the importance of improvements in coding efficiency. As the number of programs that can be efficiently carried within a given network bandwidth increases, this analysis suggests that the increase in multicasting gain will be non-linear. For example, if the number of programs carried in a given bandwidth can be doubled, taking a service group from a ceiling of 100 streaming programs to 200 streaming programs, the actual offered lineup could increase from 150 linear programs to 900 linear programs for 300 viewers while still maintaining the same blocking ratio.

Extension with Actual Data

Because of the extensive deployment of SDV in some markets, there is a large body of data that can allow a comparison of simulated results with real-world behavior. The information in the section comes from SDV deployments in several different regions. Because of the variations due to local conditions, it is not always possible to find perfect matches to the simulations. The data in this paper was chosen to represent average conditions, and may represent data that was averaged over many service groups.

As was observed in the previous sections, simulations predict that the size of a service group and the number of offered programs can significantly influence the program popularity behavior which is directly related to the efficiency of various multicast/unicast/broadcast implementations. In actual deployments, there is a limited dispersion in the sizes of service groups. Service groups that are very large or very small are difficult to gather significant amounts of data on. The next figure, Figure 12, compares a small group of service groups and generally confirms the logical assumption that the size of a service group has an effect on the number of programs that it will consume in the aggregate. These groups do show, however, that the effect is not linear; there is not a 50% decrease in the number of active programs when the service group is 50% smaller. This finding agrees with the simulation shown in Figure 9.

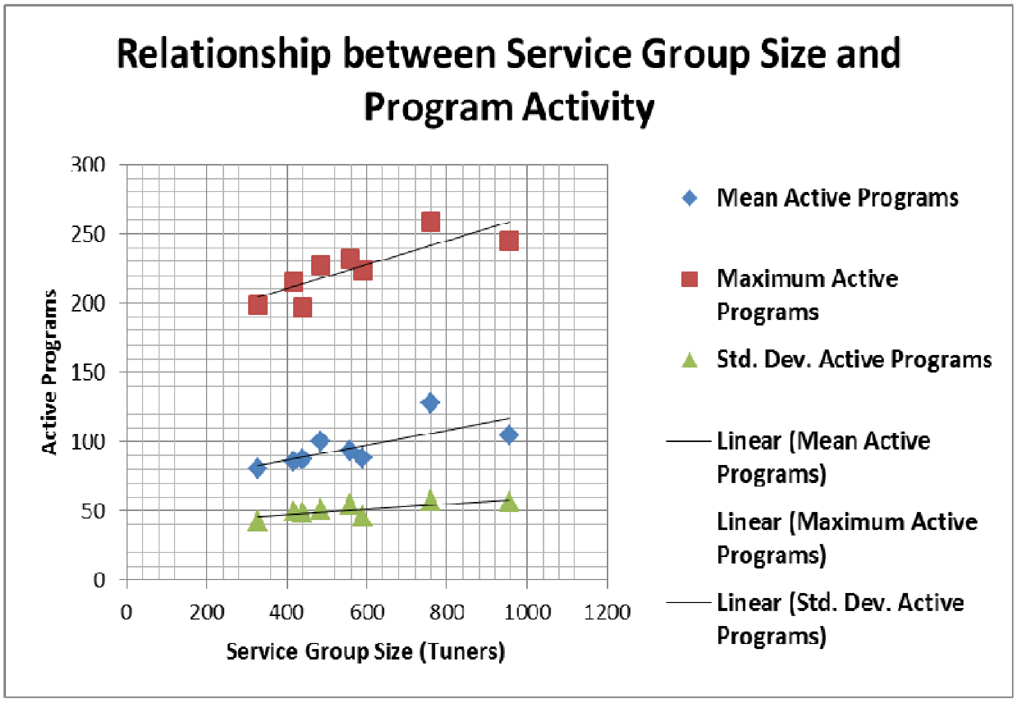


Figure 12 - Comparison of Service Group Size and Viewership

Based on some actual viewership data that included peak tuning activity across many hundred service groups, an interesting dichotomy was observed. The relationship between the size of the program lineup and the percentage of programs that had at most one viewer was strongly correlated, implying that for a given size of service group the addition of more programs to a channel lineup will tend to add mostly unicast instances. For the same study, larger program lineups had a weaker, though still positive, correlation with multicast viewing instances. In contrast,

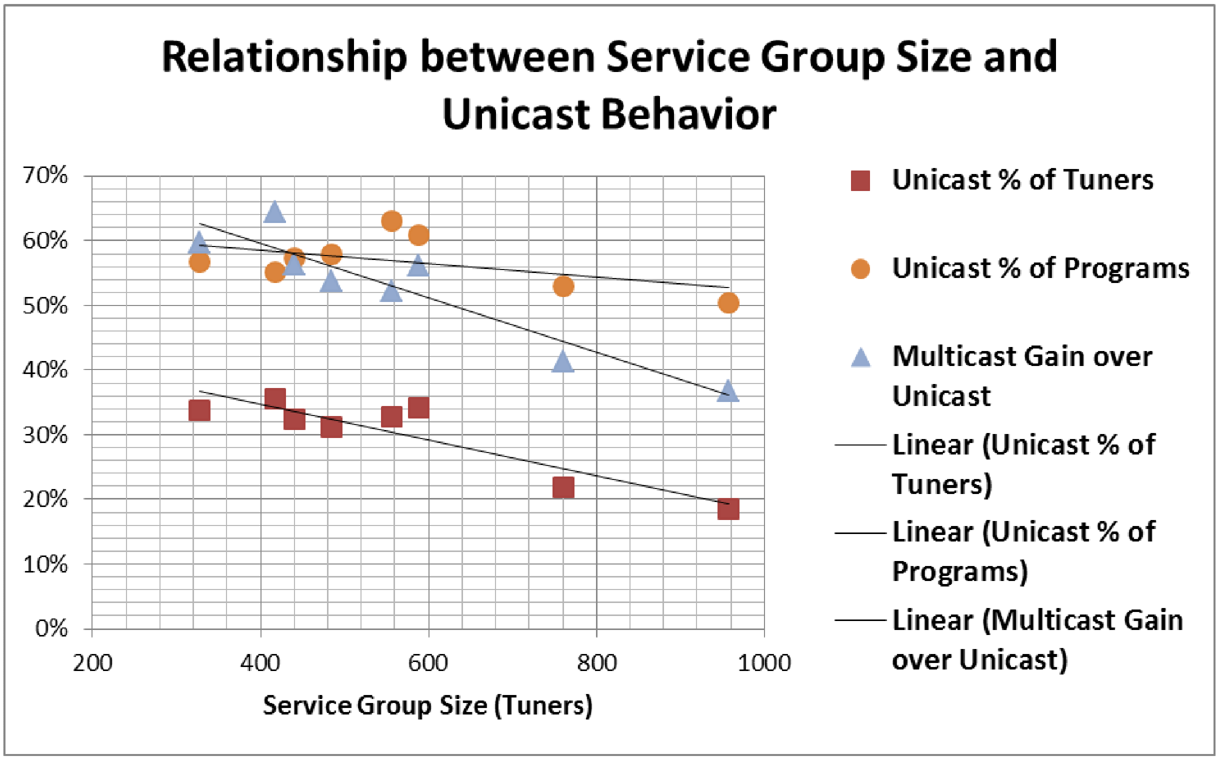


Figure 13 - Relationship of Service Group Size to Unicast Behavior

comparing increased numbers of viewers per service group for the same size of program lineup showed a strongly negative correlation with the number of unicast instances. In other words, as the number of viewers in a service group increased, they tended to watch similar programming to the other subscribers, which reduces the overall unicast percentage. This observation was compared against the test block of service groups and a similar pattern was seen in Figure 13. The service groups all had similar program lineups, and the percentage of unicast traffic declined as the number of subscribers grew in the service groups.

These observations, taken together, suggest that there is an optimum service group range that balances the number of viewers and the program content available to them.

Comparison of Deployment Scenarios

Another important area in which real world data can provide important information is the relative network impact in real time of implementations of the various protocols we have been discussing. The diagrams in this section were taken from a detailed analysis of the channel change logs of 8 service groups chosen at random.

A week’s worth of channel change logs from 8 different service groups were analyzed and used to drive various network simulations. The service groups were chosen to be roughly representative of common configurations. The wide variety of network and node configurations means that any extrapolations must be taken with a grain of salt, but they may still prove useful illustrations of the performance of different proposed systems.

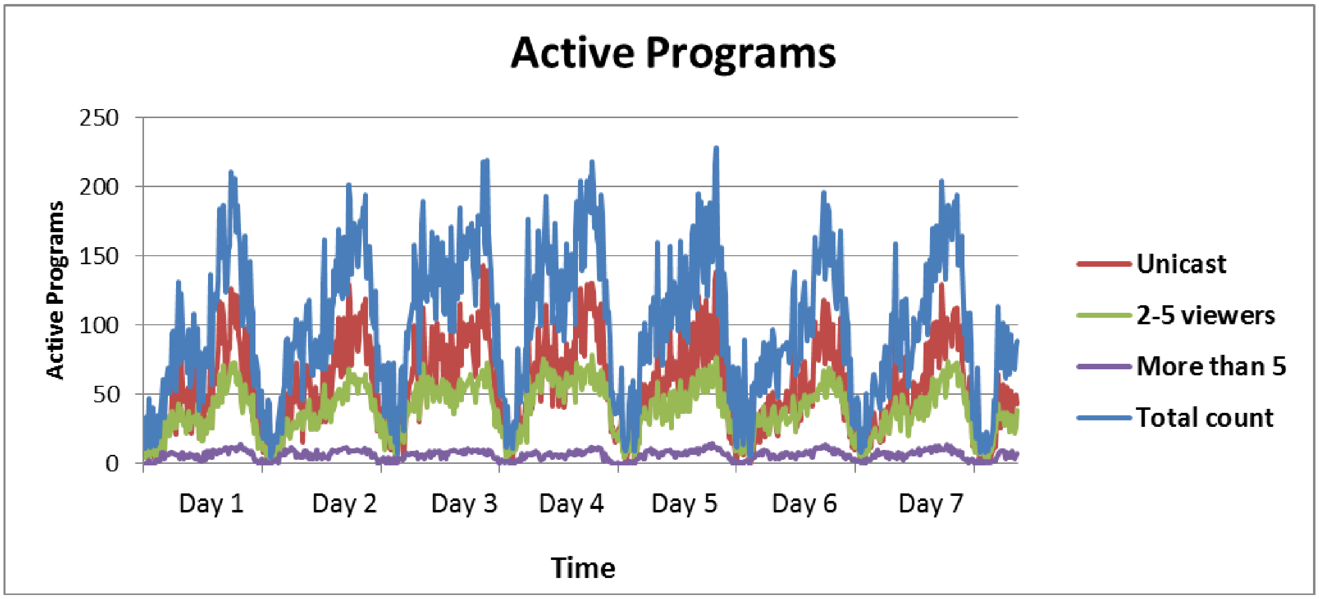


Figure 14 - Classifying Programs by Viewership

The channel change logs allowed the simulation to play out a week's worth of channel change events in various scenarios to see if the resulting network would be practical.

First to be considered is the question of the practicality of an all-unicast solution compared to an all-multicast solution. A reasonable way to study this problem is to study the distribution of viewers to programs. The 8 service groups referenced behaved similarly. One service group's results are used for illustration below, but the other service groups showed very similar results.

When one considers the distribution of viewers per program, the programs watched by only one viewer constitute the majority as shown in Figure 14. Across the SGs studied the percentage of programming viewed by only a single tuner peaked between 50% and 63% as shown in Figure 13, Unicast % of Programs.

But the dominance of Unicast in the program view is a bit misleading if one is considering an actual unicast deployment. If one considers the same period with the same

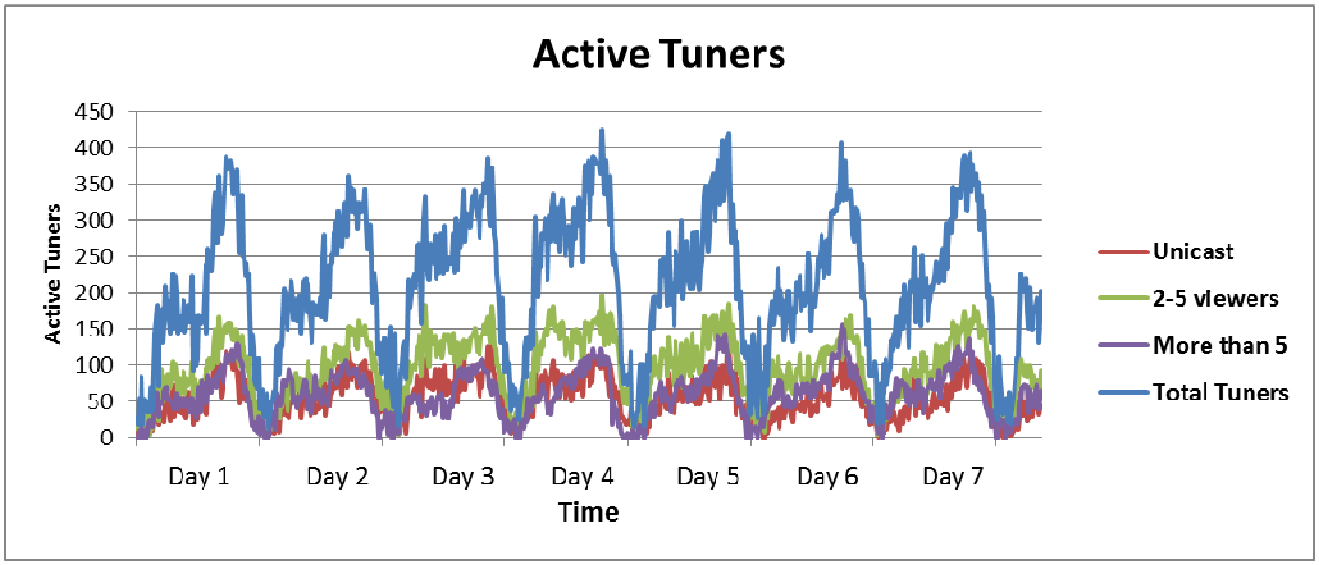


Figure 15 - Distribution of Tuners versus Other Tuners

service group, but instead studies the actual number of tuners attached to each program, a different picture emerges. The actual viewers, tuners really, are split fairly evenly across the different categories used in the graphs. From the larger group of SGs, the peak percentage of tuners that were alone in viewing a program ranged between 18% and 34%, as shown in Figure 13, Unicast % of Tuners.

Turning back to the sample service group, Figure 15 clearly shows that while the majority of streams, particularly during primetime, only have a single viewer, the majority of the viewers are actually on channels with more than one viewer.

This result implies that to move to an all unicast model for IP video requires substantially more bandwidth than a model using multicast. On average, for the service groups used as examples, an all unicast model would require 55% more bandwidth than an all multicast model. Using our example service group again, in Figure 16 the difference between an All Unicast and All Multicast model can be seen.

One other option that deserves consideration is a model that combines a static multicast tier, emulating broadcast, with a unicast tier. This combination could allow a reduction in the complexity of an IP Video deployment by simplifying the network engineering required since the static tier could be processed to improve its compression statistics, and possibly that scenario would require less protocol support that would be unique to CATV.

Using the sample service groups and the tiering shown before, the programs that had only been unicast were identified, and it was assumed that the rest of the program lineup was broadcast. That scenario was 27% more efficient, on average, than a full broadcast model. A full multicast model would have allowed 77% bandwidth reduction over broadcast. Another scenario was considered where any channels that had had at most 2 viewers were left as unicast, with the rest broadcast. This scenario offered a 50% bandwidth reduction over broadcast with performance close to that of multicast during primetime.

In Figure 17, several scenarios are compared

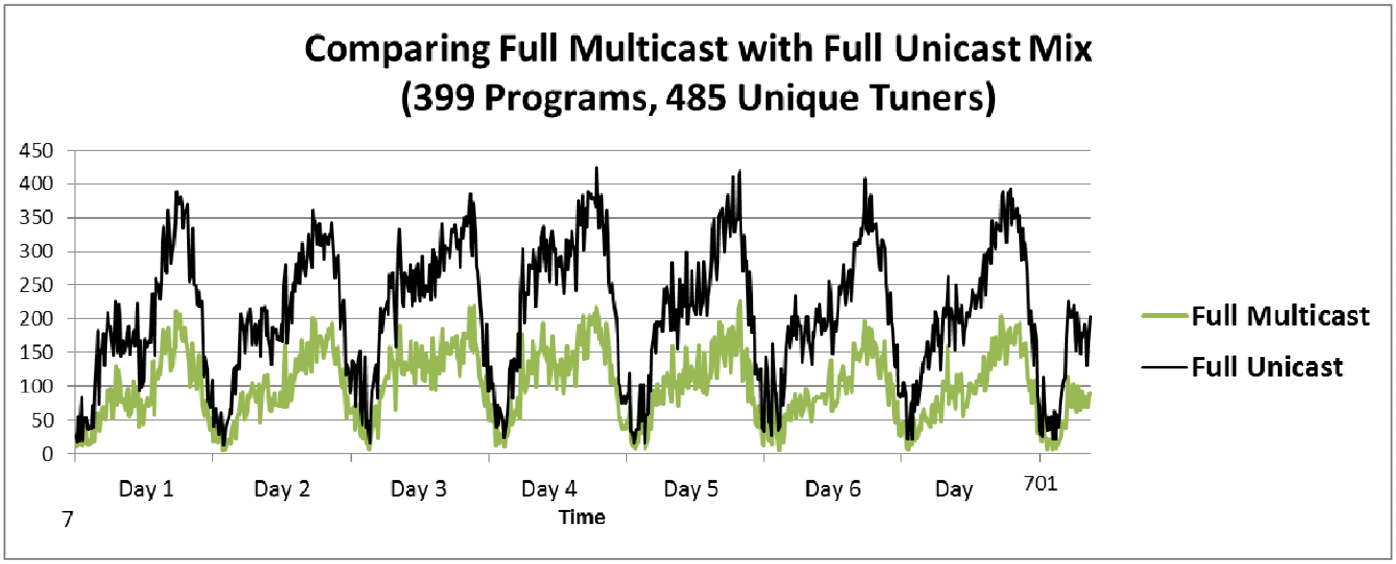


Figure 16 - Comparison of Multicast and Unicast

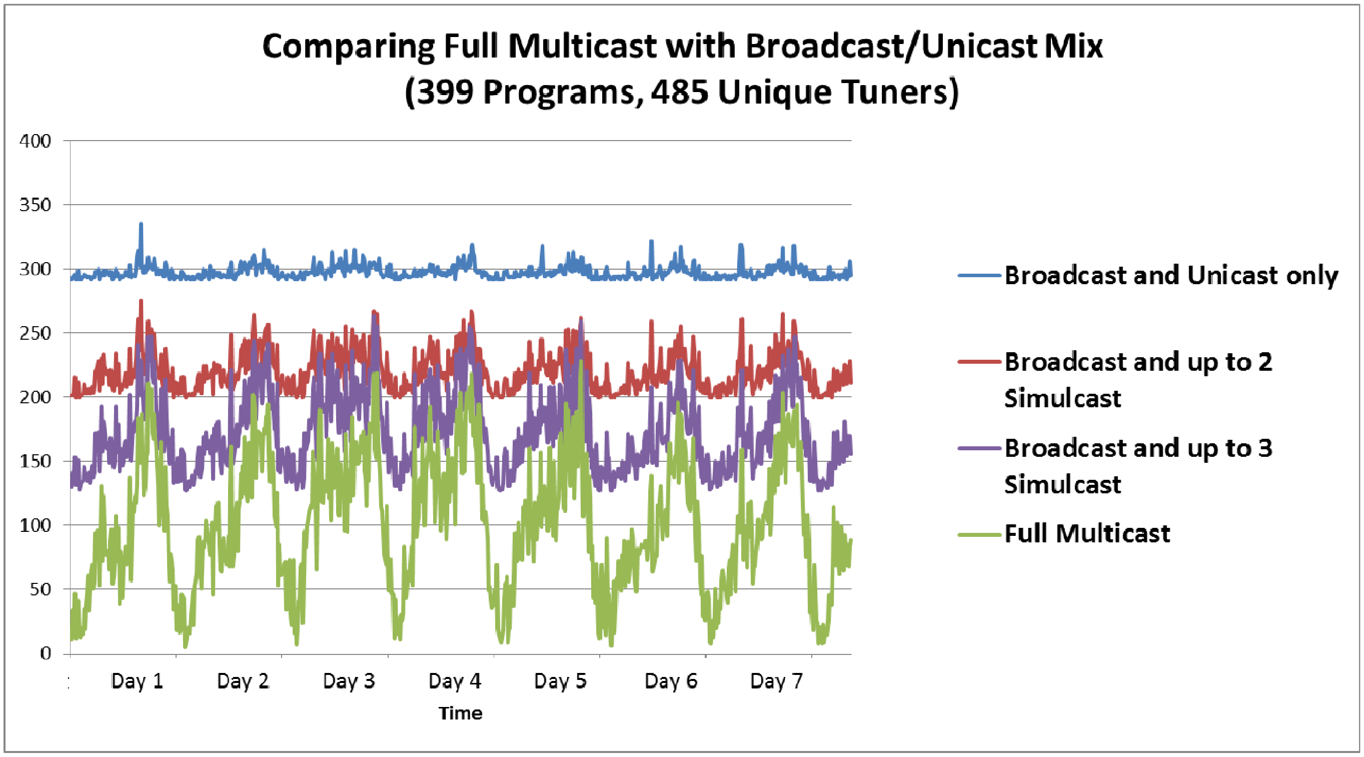


Figure 17 - Comparison of Multicast with Hybrid Broadcast/Unicast Model

using the example service group again. A more nuanced picture emerges from consideration of this figure. During primetime, the most popular channels are almost always playing, so utilizing broadcast or multicast has little effect during that time, so long as the most popular channels are correctly identified for broadcast. Allowing the least popular channels to be unicast does improve bandwidth utilization over broadcast, and as the channels committed to unicast increase the efficiency of this scheme approaches that of full multicast. Multicast offers a lower average bandwidth utilization, but its benefits are most apparent outside of primetime, traditionally the most congested time of the day in residential areas.

The value of the tradeoffs within the choice of IP Video distribution protocols is difficult to quantify. Pure broadcast's simplicity is counter-balanced by its total lack of network bandwidth efficiency. Pure unicast delivery is more complex than broadcast, but with only a small increase in DS bandwidth efficiency over broadcast. The two-way nature of the popular unicast video delivery protocols also uses more upstream bandwidth than either broadcast or multicast. Full multicast distribution offers the best bandwidth efficiency to reduce outside plant expenditures, but has not been extensively deployed past the headend and may pose unknown challenges.

Other Considerations

Some concerns have been raised about the practical limits of channel change times using multicast. DOCSIS3.0 multicast specifications involve fairly complex scenarios wherein a CM/STB must send a request to the CMTS to join a multicast group, and the CMTS must attempt to join the multicast group, then respond to the CM. An IP Video CM could conceivably have to change its DS bonding group and worst-case even reset to reach a new multicast stream.

While these scenarios are possible within the specification's limits, a sensible IP Video architecture can make many simplifications and improvements by observing the choices the successful SDV architecture has made to improve its performance. For instance, the time it takes to join a multicast group cold, so to speak, was recognized as a potential problem within SDV. The solution that was developed within the SDV architecture was to have the EQAM join as many multicasts as it would potentially source over its channels. The CMTS, occupying the same network position as the EQAM for IP Video, is equally capable of joining multiple multicast groups, thus eliminating potential router latency from the aggregate channel change time.

The analyses in the foregoing sections have assumed that subscribers will continue to behave mostly as they do today. A critical part of that assumption relates to the behavior of content providers and the regulatory landscape. If the content providers were to change from their current course and deemphasize linear programming and promote a more VOD-style consumption of their content, similar to that provided by most over-the-top providers today, then any assumptions made based on extrapolations from the behavior of today's subscribers would become moot. Most analysts have not predicted that sort of change any time soon due to primarily commercial factors, but a radical change is always a possibility driven by a new application or possible new regulations.

The Path Forward

As the operators move toward incorporating IP video into their day-to-day operations, the availability of both unicast and multicast protocols within the IPTV 'toolbox' may prove to be quite valuable.

For early low-volume deployments, unicast delivery offers a simple first step. It enables experimentation with alternative user interfaces, and hybrid STB/cloud architectures, without the complications of a volume deployment. For some networks with very small effective service groups this technology may continue to be cost-effective even as the network approaches saturation.

As IP Video deployment moves out of limited trials and into larger deployments in more traditional larger service groups, multicast can be employed to enable a cost-effective deployment of services that still fall into today's linear model. Depending upon the tradeoffs possible between network bandwidth, service group size and program popularity mix, as well as the popularity of new services such as network DVR, there is not a single answer as to the most efficient IP Video distribution model. An MSO that has moved to small service group sizes for other reasons may be able to utilize a mix of unicast and multicast with good results. An MSO that has not lowered the size of its average service groups may well decide to make more extensive use of multicast to get the most efficiency out of its network. An MSO with many commercial customers that could use the non-prime-time bandwidth freed up by multicast may also choose to implement a full multicast solution.

As the content distribution model evolves past the traditional linear program distribution, unicast may return to prominence if few users tend to watch the same thing at the same time. Some events, like sports or breaking news, may still attract enough viewers to leave multicast a place on the table even then.

In Summary

IP Video delivery over unicast protocols has flourished on the Web, but for the CATV application of bulk delivery of programming over a pipe with limited bandwidth, the unicast model tends to break down due to the sheer volume of users.

Broadcast has been great for CATV for many years, but as the number of programs has proliferated and the required variety of resolutions for those programs has grown as well, the sheer volume of programming selections has tended to exhaust the available bandwidth.

Multicast, perhaps in combination with unicast, may offer a robust solution, similar to the use of SDV in conjunction with VOD in current MPEG distribution network. This combination of technologies can offer an expansive list of programming suitable for many different device types, while still fitting within practical constraints of the available bandwidth envelope.

MPEG DASH: A Technical Deep Dive and Look at What's Next

Andy Salo
RGB Networks

Abstract

The MPEG DASH standard was ratified in December 2011 and published by the International Organization for Standards (ISO) in April 2012. This paper will review the technical aspects of the new MPEG DASH standard in detail, including: how DASH supports live, on-demand and time-shifted (NDVR) services; how the two primary video formats – ISO-base media file format (IBMFF) and MPEG-2 TS – compare and contrast; how the new standard supports digital rights management (DRM) methods; and how Media Presentation Description (MPD) XML files differ from current adaptive streaming manifests. In addition, the paper will discuss how MPEG DASH is likely to be adopted by the industry, what challenges must still be overcome, and what the implications could be for cable operators and other video service providers (VSPs).

INTRODUCTION

For much of the past decade, it was quite difficult to stream live video to a mobile device. Wide bandwidth variability, unfavorable firewall configurations and lack of network infrastructure support all created major roadblocks to live streaming. Early, more traditional streaming protocols, designed for small packet formats and managed delivery networks, were anything but firewall-friendly. Although HTTP progressive download was developed partially to get audio and video streams past firewalls, it still didn't offer true streaming capabilities.

Now, the advent of adaptive streaming over HTTP technology has changed everything, reshaping video delivery to PCs, laptops, game consoles, tablets, smartphones and other mobile devices, as well as such key home devices as Web-connected TVs and pure and hybrid IP set-top boxes (STBs). As a result, watching video online or on the go is no longer a great novelty, nor is streaming Internet-delivered content to TV screens in the home. Driven by the explosion in video-enabled devices, consumers have swiftly moved through the early-adopter phase of TV Everywhere service, reaching the point where a growing number expect any media to be available on any device over any network connection at any time. Increasingly, consumers also expect the content delivery to meet the same high quality levels they have come to know and love from traditional TV services.

Even though the emergence of the three main adaptive streaming protocols from Adobe, Apple and Microsoft over the past three and a half years has made multiscreen video a reality, significant problems still remain. Each of the three proprietary platforms is a closed system, with its own manifest format, content formats and streaming protocols. So, content creators and equipment vendors must craft several different versions of their products to serve the entire streaming video market, greatly driving up costs and restricting the market's overall development.

In an ambitious bid to solve these nagging problems, MPEG has recently adopted a new standard for multimedia streaming over the Internet. Known as MPEG Dynamic Adaptive Streaming over HTTP, or MPEG DASH, the new industry standard attempts to create a universal delivery format for streaming media by incorporating the best elements of the three main proprietary streaming solutions. In doing so, MPEG DASH aims to provide the long-sought interoperability between different network servers and different consumer electronics devices, thereby fostering a common ecosystem of content and services.

This paper will review the technical aspects of the new MPEG DASH standard in detail, including: how DASH supports live, on-demand and time-shifted (NDVR) services; how the two primary video formats (ISO-base media file format (IBMFF) and MPEG-2 TS) compare and contrast; how the standard supports DRM methods; and how Media Presentation Description (MPD) XML files differ from current adaptive streaming manifests. In addition, the paper will discuss how MPEG DASH is likely to be adopted by the industry, what challenges must still be overcome, and what the implications could be for cable operators and other video service providers (VSPs).

AN ADAPTIVE STREAMING PRIMER

As indicated previously, the delivery of streaming video and audio content to consumer electronics devices has come a long way over the past few years. Thanks to the introduction of adaptive streaming over HTTP, multimedia content can now be delivered more easily than ever before. In particular, adaptive streaming offers two critical features for video content that have made the technology the preferred choice for mobile delivery.

First, adaptive streaming over HTTP breaks down, or segments, video programs into small, easy-to-download chunks. For example, Apple's HTTP Live Streaming (HLS) protocol typically segments video content into 10-second chunks, while Microsoft's Smooth Streaming (MSS) protocol and Adobe's HTTP Dynamic Streaming (HDS) usually break video content into even smaller chunks of five seconds or less.

Second, adaptive streaming encodes the video content at multiple bitrates and resolutions, creating different chunks of different sizes. This is the truly 'adaptive' part of adaptive streaming, as the encoding enables the mobile client to choose between various bitrates and resolutions and then adapt to larger or smaller chunks automatically as network conditions keep changing.

In turn, these two key features of adaptive streaming lead to a number of benefits:

1. Video chunks can be cached by proxies and easily distributed to content delivery networks (CDNs) or HTTP servers, which are simpler and cheaper to operate than the special streaming servers required for 'older' video streaming technologies.

2. Bitrate switching allows clients to adapt dynamically to network conditions.

3. Content providers no longer have to guess which bitrates to encode for end devices.

4. The technology works well with firewalls because the streams are sent over HTTP.

5. Live and video-on-demand (VoD) workflows are almost identical. When a service provider creates a live stream, the chunks can easily be stored for later VoD delivery.

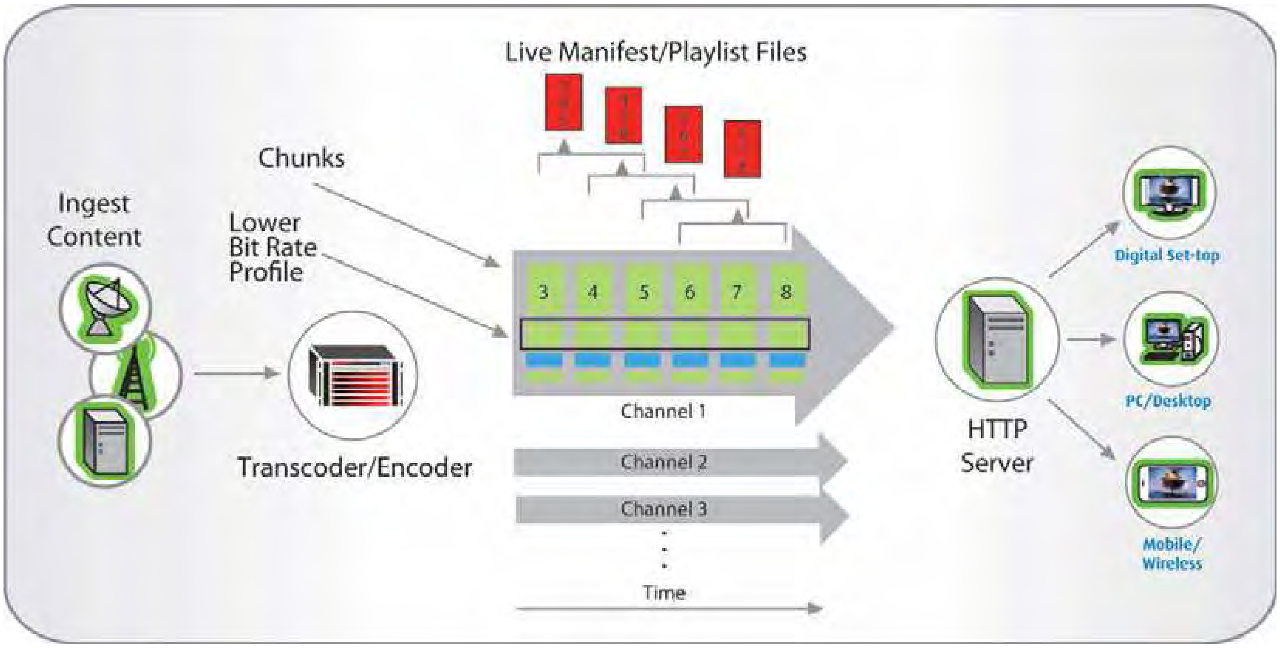


Figure 1: Content Delivery Chain for Live Adaptive Streaming

Sensing the promise of adaptive streaming technology, several major technology players have sought to carve out large shares of the rapidly growing market. Most notably, the list now includes such prominent tech companies as Adobe, Apple and Microsoft.

While the streaming of video using HTTP-delivered fragments goes back many years (and seems lost in the mists of time), Move Networks caught the attention of several media companies with its adaptive HTTP streaming technology in 2007. Move was quickly followed by Microsoft, which entered the market by releasing Smooth Streaming in October 2008 as part of its Silverlight architecture. Earlier that year, Microsoft demonstrated a prototype version of Smooth Streaming by delivering live and on-demand streaming content from such events as the Summer Olympic Games in Beijing and the Democratic National Convention in Denver.

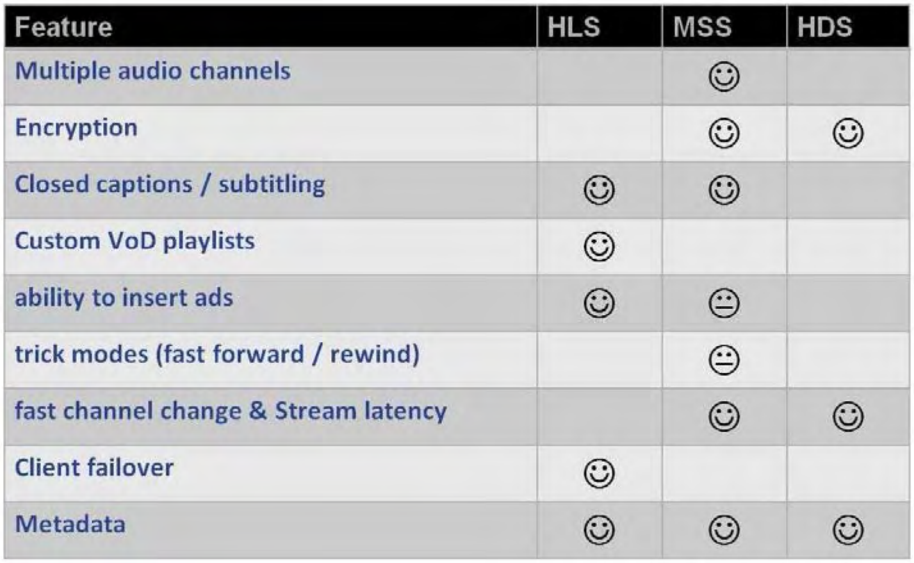
Smooth Streaming has all of the typical characteristics of adaptive streaming. The video content is segmented into small chunks and then delivered over HTTP. Usually, multiple bitrates are encoded so that the client can choose the best video bitrate to deliver an optimal viewing experience based on network conditions.

Apple came next with HLS, originally unveiling it with the introduction of the iPhone 3.0 in mid-2009. Prior to the iPhone 3, no streaming protocols were supported natively on the iPhone, leaving developers to wonder what Apple had in mind for native streaming support. In May 2009, Apple proposed HLS as a standard to the Internet Engineering Task Force (IETF), and the draft is now in its eighth iteration.

HLS works by segmenting video streams into 10-second chunks; the chunks are stored using a standard MPEG-2 transport stream file format. The chunks may be created using several bitrates and resolutions – so-called profiles – allowing a client to switch dynamically between different profiles, depending on network conditions.

Adobe, the last of the Big Three, entered the adaptive streaming market in late 2009 with the announcement of HTTP Dynamic Streaming (HDS). Originally known as “Project Zeri,” HDS was introduced in June 2010. Like MSS and HLS, HDS breaks up video content into small chunks and delivers them over HTTP. Multiple bitrates are encoded so that the client can choose the best video bitrate to deliver an optimal viewing experience based on network conditions.

HDS is closer to Microsoft Smooth Streaming than it is to Apple’s HLS protocol. Primarily, this is because HDS, like MSS, uses a single aggregate file from which MPEG-4 container fragments are extracted and delivered. In contrast, HLS uses individual media chunks rather than one large aggregate file.



Feature	HLS	MSS	HDS
Multiple audio channels		☺	
Encryption		☺	☺
Closed captions / subtitling	☺	☺	
Custom VoD playlists	☺		
ability to insert ads	☺	😐	
trick modes (fast forward / rewind)		😐	
fast channel change & Stream latency		☺	☺
Client failover	☺		
Metadata	☺	☺	☺

Figure 2: Feature Comparison of Three Major Adaptive Streaming Platforms

THE DUELING STREAMING PLATFORM PROBLEM

The three major adaptive streaming protocols – MSS, HLS and HDS – have much in common. Most importantly, all three streaming platforms use HTTP streaming for their underlying delivery method, relying on standard HTTP Web servers instead of special streaming servers. They all use a combination of encoded media files and manifest files that identify the main and alternative streams and their respective URLs for the player. And their respective players all monitor either buffer status or CPU utilization and switch streams as necessary, locating the alternative streams from the URLs specified in the manifest.

The overriding problem with MSS, HLS and HDS is that these three different streaming protocols, while quite similar to each other in many ways, are different enough that they are not technically compatible. Indeed, each of the three proprietary commercial platforms is a closed system with its own type of manifest format, content formats, encryption methods and streaming protocols, making it impossible for them to work together.

Take Microsoft Smooth Streaming and Apple's HLS. Here are three key differences between the two competing platforms:

1. HLS makes use of a regularly updated "moving window" metadata index file that tells the client which chunks are available for download. Smooth Streaming uses time codes in the chunk requests so that the client doesn't have to keep downloading an index file. This leads to a second difference:

2. HLS requires a download of an index file every time a new chunk is available. That makes it desirable to run HLS with longer duration chunks, thereby minimizing the number of index file downloads. So, the recommended chunk duration with HLS is 10 seconds, while it is just two seconds with Smooth Streaming.

3. The "wire format" of the chunks is different. Although both formats use H.264 video encoding and AAC audio encoding, HLS makes use of MPEG-2 Transport Stream files, while Smooth Streaming makes use of "fragmented" ISO MPEG-4 files. The "fragmented" MP4 file is a variant in which not all the data in a regular MP4 file is included in the file. Each of these formats has some advantages and disadvantages. MPEG-2 TS files have a large installed analysis toolset and have pre-defined signaling mechanisms for things like data signals (e.g. specification of ad insertion points). But fragmented MP4 files are very flexible and can easily accommodate all kinds of data, such as decryption information, that MPEG-2 TS files don't have defined slots to carry.

Or take Adobe HDS and Apple's HLS. These two platforms have a number of key differences as well:

1. HLS makes use of a regularly updated "moving window" metadata index (manifest) file that tells the client which chunks are available for download. Adobe HDS uses sequence numbers in the chunk requests so the client doesn't have to keep downloading a manifest file.

2. In addition to the manifest, there is a bootstrap file, which in the live case gives the updated sequence numbers and is equivalent to the repeatedly downloaded HLS playlist.

3. Because HLS requires a download of a manifest file as often as every time a new chunk is available, it is desirable to run HLS with longer duration chunks, thus minimizing the number of manifest file downloads. More recent Apple client versions appear to check how many segments are in the playlist and only re-fetch the manifest when the client runs out of segments. Nevertheless, the recommended chunk duration with HLS is still 10 seconds, while it is usually just two to five seconds with Adobe HDS.

4. The "wire format" of the chunks is different. Both formats use H.264 video encoding and AAC audio encoding. But HLS makes use of MPEG-2 TS files, while Adobe HDS (and Microsoft SS) make use of "fragmented" ISO MPEG-4 files.

Due to such differences, there is no such thing as a universal delivery standard for streaming media today. Likewise, there is no universal encryption standard or player standard. Nor is there any interoperability between the devices and servers of the various vendors. So, content cannot be re-used and creators and equipment makers must develop several different versions of their products to serve the entire streaming video market, greatly driving up costs and restricting the market's overall development.

INTRODUCING MPEG DASH: A STANDARDS-BASED APPROACH

Seeing the need for a universal standard for the delivery of adaptive streaming media, MPEG decided to step into the void three years ago. In April 2009, the organization issued a Request for Proposals for an HTTP streaming standard. By that July, MPEG had received 15 full proposals. In the following two years, MPEG developed the specification with the help of many experts and in collaboration with other standards groups, such as the Third Generation Partnership Project (3GPP) and the Open IPTV Forum (OIPF).

The resulting MPEG standardization of Dynamic Adaptive Streaming over HTTP is now simply known as MPEG DASH.

MPEG DASH is not a system, protocol, presentation, codec, middleware, or client specification. Rather, the new standard is more like a neutral enabler, aimed at providing several formats that foster the efficient and high-quality delivery of streaming media services over the Internet.

As described by document ISO/IEC 23009-1, MPEG DASH can be viewed as an amalgamation of the industry's three prominent adaptive streaming protocols – Adobe HDS, Apple HLS and Microsoft Smooth Streaming. Like those three proprietary platforms, DASH is a video streaming solution where small chunks of video streams/files are requested using HTTP and then spliced together by the client. The client entirely controls the delivery of services.

In other words, MPEG DASH offers a standards-based approach for enabling a host of media services that cable operators and telcos have traditionally offered in broadcast and IPTV environments and extending those capabilities to adaptive bitrate delivery, including live and on-demand content delivery, time-shifted services (NDVR, catch-up TV), and targeted ad insertion. DASH enables these features through a number of inherent capabilities, and perhaps most importantly, through a flexibility of design and implementation. Its capabilities and features include:

- Multiple segment formats (ISO BMFF and MPEG-2 TS)
- Codec independence
- Trick mode functionality
- Profiles: restriction of DASH and system features (claim & permission)
- Content descriptors for protection, accessibility, content rating, and more
- Common encryption (defined by ISO/IEC 23001-7)
- Clock drift control for live content
- Metrics for reporting the client session experience

A Tale of Two Containers – MPEG-2 TS and ISO BMFF

Under the MPEG DASH standard, the media segments can contain any type of media data. However, the standard provides specific guidance and formats for use with two types of segment container formats – MPEG-2 Transport Stream (MPEG-2 TS) and ISO base media file format (ISO BMFF).

MPEG-2 TS is the segment format that HLS currently uses, while ISO BMFF (which is basically the MPEG-4 format) is what Smooth Streaming and HDS currently use.

This mix of the two container formats employed by the three commercial platforms allows for a relatively easy migration of existing adaptive streaming content from the proprietary platforms to MPEG DASH. That's because the media segments can often stay the same; only the index files must be migrated to a different format, which is known as Media Presentation Description.

Media Presentation Description (MPD) – Definition and Overview

At a high level, MPEG DASH works nearly the same way as the three other major adaptive streaming protocols. DASH presents available stream content to the media player in a manifest (or index) file – called the Media Presentation Description (MPD) – and then supports HTTP download of media segments. The MPD is analogous to an HLS m3u8 file, a Smooth Streaming Manifest file or an HDS f4m file. After the MPD is delivered to the client, the content – whether it's video, audio, subtitles or other data – is downloaded to clients over HTTP as a sequence of files that is played back contiguously.

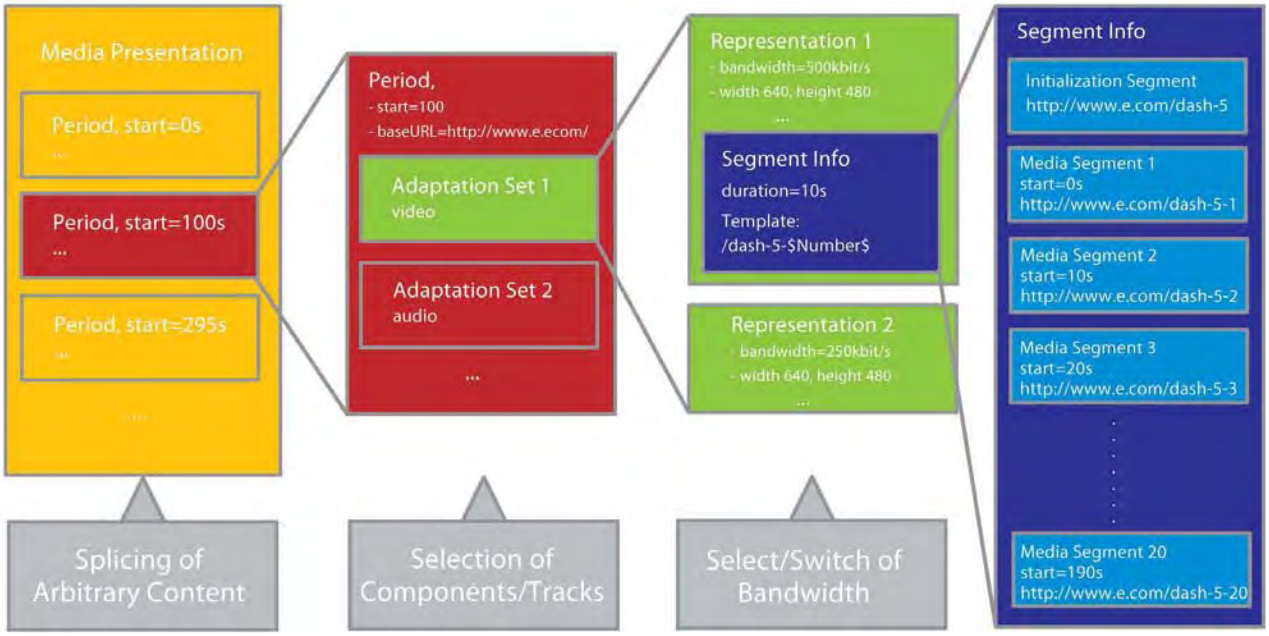


Figure 3: Media Presentation Data Model

(Diagram originally developed by Thomas Stockhammer, Qualcomm)

Like a manifest file in the three commercial platforms, the MPD in MPEG DASH describes the content that is available, including the URL addresses of stream chunks, byte-ranges, different bitrates, resolutions, and content encryption mechanisms. The tasks of choosing which adaptive stream bitrate and resolution to play and switching to different bitrate streams according to network conditions are performed by the client (again, similar to the other adaptive streaming protocols). In fact, DASH does not prescribe any client-specific playback functionality; rather, it just addresses the formatting of the content and associated MPDs.

To see what an MPEG DASH MPD file looks like compared to an HLS m3u8 file, consider the following example. The files contain much of the same information, but they are formatted and presented differently.

Figure 4: Comparison of MPEG DASH MPD and HLS m3u8 Files

Index.m3u8 (top level m3u8)

```
#EXTM3U
#EXT-X-STREAM-INF:PROGRAM- ID=1,BANDWIDTH=291500,RESOLUTION=320x180
stream1.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=610560,RESOLUTION=512x288
stream2.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=2061700,RESOLUTION=1024x576
stream3.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=4659760,RESOLUTION=1280x720
stream4.m3u8
```

Index.mpd

```
<?xml version="1.0" encoding="utf-8"?>
<MPD
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="urn:mpeg:DASH:schema:MPD:2011"
  xsi:schemaLocation="urn:mpeg:DASH:schema:MPD:2011"
  type="static"
  mediaPresentationDuration="PT12M34.041388S"
  minBufferTime="PT10S"
  profiles="urn:mpeg:dash:profile:isoff-live:2011">

  <Period>
    <AdaptationSet
      mimeType="audio/mp4"
      segmentAlignment="0"
      lang="eng">
      <SegmentTemplate
        timescale="10000000"
        media="audio_eng=$Bandwidth$-$Time$.dash"
        initialisation=" audio_eng=$Bandwidth$.dash">
        <SegmentTimeline>
          <S t="667333" d="39473889" />
          <S t="40141222" d="40170555" />

       ...

          <S t="7527647777" d="12766111" />
        </SegmentTimeline>
      </SegmentTemplate>
      <Representation id="audio_eng=96000" bandwidth="96000" codecs="mp4a.40.2"
audioSamplingRate="44100" />
    </AdaptationSet>
    <AdaptationSet
      mimeType="video/mp4"
      segmentAlignment="true"
      startWithSAP="1"
      lang="eng">
      <SegmentTemplate
        timescale="10000000"
        media="video=$Bandwidth$-$Time$.dash"
        initialisation="video=$Bandwidth$.dash">
        <SegmentTimeline>
          <S t="0" d="40040000" r="187" />
          <S t="7527520000" d="11678333" />
        </SegmentTimeline>
      </SegmentTemplate>

      <Representation id="video=299000" bandwidth="299000" codecs="avc1.42C00D"
width="320" height="180" />
      <Representation id="video=480000" bandwidth="480000" codecs="avc1.4D401F"
width="512" height="288" />
 codecs="avc1.4D401F" width="1024" height="576" />
      <Representation id="video=4300000" bandwidth="4300000"
codecs="avc1.640028" width="1280" height="720" />
    </AdaptationSet>
  </Period>
</MPD>
```

MPEG DASH'S PRIME CAPABILITIES – OVERVIEW

As mentioned earlier, MPEG DASH offers a great number of capabilities for adaptive streaming. This section goes into greater detail about many of the prime capabilities.

Codec Independence: Simply put, MPEG DASH is audio/video agnostic. As a result, the standard can work with media files of MPEG-2, MPEG-4, H.264, WebM and various other codecs and does not favor one codec over another. It also supports both multiplexed and unmultiplexed encoded content. More importantly, DASH will support emerging standards, such as HEVC (H.265).

Trick Mode Functionality: MPEG DASH supports VoD trick modes for pausing, seeking, fast forwarding and rewinding content. For instance, the client may pause or stop a Media Presentation.

In this case, the client simply stops requesting Media Segments or parts thereof. To resume, the client sends requests to Media Segments, starting with the next sub-segment after the last requested sub-segment.

DASH's treatment of trick modes could prove to be a major improvement over the way that the three existing streaming protocols handle these on-demand functions now.

Profiles: Restriction of DASH and System Features (Claim & Permission): MPEG DASH defines and allows for the creation of various profiles. A profile is a set of restrictions of media formats, codecs, protection formats, bitrates, resolutions, and other aspects of the content. For example, the DASH spec defines a profile for ISO BMFF basic on-demand.

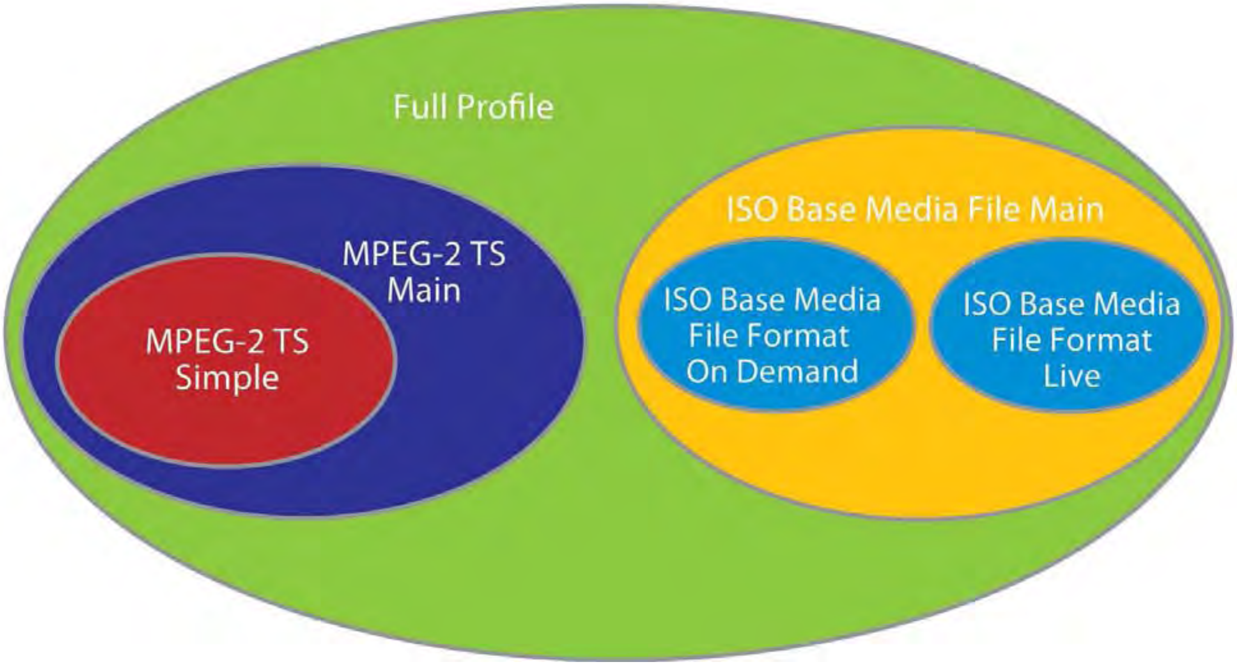

Full Profile
MPEG-2 TS Main
MPEG-2 TS Simple
ISO Base Media File Main
ISO Base Media File Format On Demand
ISO Base Media File Format Live

Figure 5: Describing MPEG DASH Profiles
(Diagram originally developed by Thomas Stockhammer, Qualcomm)

Content Descriptors for Protection, Accessibility, Content Rating: MPEG DASH offers a flexible set of descriptors for the media content that is being streamed. These descriptors spell out such elements as the rating of the content, the role of various components, accessibility features, DRM methods, camera views, frame packing, and the configuration of audio channels, among other things.

Common Encryption (defined by ISO/IEC 23001-7): One of the most important features of MPEG DASH is its use of Common Encryption, which standardizes signaling for what would otherwise be a number of non-interoperable, albeit widely used, encryption methods. Leveraging this standard, content owners or distributors can encrypt their content just once and then stream it to different clients with different DRM license systems. As a result, content owners can distribute their content freely and widely, while service providers can enjoy access to an open, interoperable ecosystem of vendors. In fact, Common Encryption is also used as the underlying standard for Ultraviolet, the Digital Entertainment Content Ecosystem's (DECE's) content authentication system. Common Encryption will be discussed in a bit more detail later in this paper.

Clock Drift Control for Live Content: In MPEG DASH, each media segment can include an associated Coordinated Universal Time (UTC) time, so that a client can control its clock drift and ensure that the encoder and decoder remain closely synchronized. Without this, a time difference between the encoder and decoder could cause the client play-back buffer to starve or overflow, due to different rates of video delivery and playback.

Metrics for Reporting the Client Session Experience: MPEG DASH has a set of well-defined quality metrics for tracking the user's session experience and sending the information back to the server.

MULTIPLE DRM METHODS & COMMON ENCRYPTION

As mentioned earlier, one of MPEG DASH's most important features is its use of Common Encryption, which standardizes signaling for a number of different, widely used encryption methods. Common Encryption (or "CENC") describes methods of standards-based encryption, along with key mapping of content to keys. CENC can be used by different DRM systems or Key Management Servers (KMS) to enable decryption of the same content, even with different vendors' equipment. It works by defining a common format for the encryption-related metadata required to decrypt the protected content. The details of key acquisition and storage, rights mapping, and compliance rules are not specified in the standard and are controlled by the DRM server. For example, DRM servers supporting Common Encryption will identify the decryption key with a key identifier (KID), but will not specify how the DRM server should locate or access the decryption key.

Using this standard, content owners or distributors can encrypt their content just once and then stream it to the various clients with their different DRM license systems. Each client receives the content decryption keys and other required data using its particular DRM system. This information is then transmitted in the MPD, enabling the client to stream the commonly encrypted content from the same server.

As a result, content owners can distribute their content freely and widely without the need for multiple encryptions. At the same time, cable operators and other video service providers can enjoy access to an open, interoperable ecosystem of content producers and equipment vendors.

USE CASES

The MPEG DASH spec supports both simple and advanced use cases of dynamic adaptive streaming. Moreover, the simple use cases can be gradually extended to more complex and advanced cases. In this section, we'll detail three such common use cases:

Live and On-Demand Content Delivery: MPEG DASH supports the delivery of both live and on-demand media content to subscribers through dynamic adaptive HTTP streaming. Like Adobe's HDS, Apple's HLS and Microsoft's Smooth Streaming platforms, DASH encodes the source video or audio content into file segments using a desired format. The segments are subsequently hosted on a regular HTTP server. Clients then play the stream by requesting the segments in a profile from a Web server, downloading them via HTTP.

MPEG DASH's great versatility in supporting both live and on-demand content has other benefits as well. For instance, these same capabilities also enable video service providers to deliver additional time-shifted services, such as network-based DVR (NDVR) and catch-up TV services, as explained below.

Time-Shifted Services (NDVR, catch-up TV, etc.): MPEG DASH supports the flexible delivery of time-shifted services, such as NDVRs and catch-up TV. For the enabling of time-shifted services, VoD assets, rather than live streams, are required. VoD assets formatted for MPEG DASH can be created using a transcoder. Additionally, a device commonly referred to as a Catcher can "catch" a live TV program and create a VoD asset, suitable for streaming after the live event. Because the VoD asset can be streamed in MPEG DASH in the same manner as the live content, the asset can be re-used and monetized by the operator.

Targeted Ad Insertion: Wherever there is video service, there is usually some kind of advertising content to monetize the service. ‘Traditional” ad insertion methods rely on a set of technologies based on the widely used protocols for distributing UDP/IP video: ad servers, ad splicers, and an ecosystem based on zoned ad delivery. But as video delivery transport has evolved via the new set of adaptive HTTP-based delivery protocols from Apple, Microsoft and Adobe, the ad insertion ecosystem has had to evolve to employ new, targeted technologies for insertion and delivery of revenue-generating commercials. The difficulty of inserting ads with the three existing delivery methods is that the protocols don’t support the same ad insertion methods, due to the inherent nature of how the protocols work.

MPEG DASH offers the dramatic potential to help enable adaptive bitrate advertising on many different types of client devices. DASH supports the dynamic insertion of advertising content into multimedia streams. In both live and on-demand use cases, commercials can be inserted either as a period between different multimedia periods or as a segment between different multimedia segments. As in the case with VoD trick modes, this would represent a significant improvement over the way that the three leading streaming protocols currently handle targeted ad insertion.

It is worth emphasizing that DASH supports a network-centric approach to ad insertion, as opposed to a client-centric approach in which the client pre-fetches ads and splices them locally based on interactions with external ad management systems. In DASH, the information about when ads play, which ads play, and how ads are delivered is transmitted through the MPD, which is created and distributed from the network.

PROSPECTS FOR INDUSTRY ADOPTION – CATALYSTS & CHALLENGES

With the development, ratification and introduction of the MPEG DASH platform, MPEG is attempting to rally the technology community behind a universal delivery standard for adaptive streaming media. Many tech companies have already enlisted in the effort, joining the new MPEG DASH Promoters Group to drive the broad adoption of the standard.

Not surprisingly, equipment vendors and content publishers are especially enthusiastic about the new standard. For instance, content publishers savor the opportunity to produce just a single set of media files that could run on all DASH-compatible electronics devices.

The key to MPEG DASH’s success, though, will be the participation of the three major proprietary players – Adobe, Apple, and Microsoft – that now divvy up the adaptive streaming market. While all three companies have contributed to the standard, their levels of support for DASH vary greatly. In particular, Apple’s backing is still in question because of the competitive advantages that its HLS platform stands to lose if DASH becomes the universal standard.

Besides such competitive issues, MPEG DASH faces potential intellectual property rights challenges as well. For example, it is still not clear if DASH will be saddled with royalty payments and, if so, where those royalties might be applied. This section will look at the intellectual property rights and other issues that may yet bedevil the new standard.

Unresolved Intellectual Property Rights Issues: In addition to the competitive issues, there are some unresolved intellectual property rights issues with MPEG DASH. For instance, when companies seek to contribute intellectual property to the MPEG standards effort, the contribution is usually accepted only if the property owner agrees to Reasonable and Non-Discriminatory (RAND) terms. In the case of DASH, though, it is not clear that all of the intellectual property rights (IPR) in the standard are covered by RAND terms.

Non-Interoperable DASH Profiles: Although MPEG DASH may have a single, unified name, it actually consists of a collection of different, non-interoperable profiles. So DASH doesn't solve the problem of different, non-interoperable implementations unless DASH clients support all profiles. This would basically be equivalent to having a client that supports HLS, HDS and Smooth Streaming (which, incidentally, would also address the interoperability problem). Thus, the adoption of DASH doesn't immediately imply a unified, interoperable ecosystem – a DASH world may suffer from the same interoperability issues that HLS, Smooth Streaming and HDS create today.

CONCLUSION

Now that MPEG DASH has been published by the ISO, it seems well on its way to becoming a solid, broadly accepted standard for the streaming media market. Three years in the making, DASH is poised to provide a universal platform for delivering streaming media content to multiple screens. Designed to be very flexible in nature, it promises to enable the re-use of existing technologies (containers, codecs, DRM, etc.), seamless switching between protocols, and perhaps most importantly, a high-quality experience for end users.

Furthermore, most of the tech industry's major players have already lined up firmly behind DASH. The list of prominent supporters includes Akamai, Dolby, Samsung, Thomson, Netflix and, most notably, such leading streaming media providers as Microsoft and Adobe. Apple stands out as one of the few major tech players that haven't fully enlisted in the effort yet. So there's a great deal of hope in the industry that MPEG DASH could actually bring in all of the major players and realize its full market potential.

Yet several critical hurdles remain in the way of DASH's dash to destiny. For one thing, Apple, Adobe and Microsoft must throw their full weight behind the standard and agree to make the switch from their proprietary HLS protocols in the future despite some clear competitive disadvantages of doing so. For another, all industry stakeholders must agree to make their intellectual property contributions to the standard royalty-free.

Neither of these developments will likely happen overnight. So it's not clear yet if MPEG DASH will end up superseding the existing adaptive streaming formats as a true universal industry standard or merely co-existing with one or more of them in a still-fragmented market. As usual, the outcome will depend on what the major vendors decide to do. It will also depend on whether cable operators and other video service providers shift their multiscreen deployments and content offerings to DASH or continue on their current streaming paths. Only time will tell.

Managed IP Video Service: Making the Most of Adaptive Streaming

John Ulm & John Holobinko
Motorola Mobility

Abstract

The paper describes how an operator can leverage adaptive streaming protocols that are used today for unmanaged over-the-top (OTT) content for a complete managed IP video service. The paper describes how this solution is simpler and without some of the challenges imposed by implementing multicast delivery. Motorola's IP video modeling data shows compelling results regarding the relative benefits of adaptive versus multicast.

The conclusions and illustrations presented in this paper will help operators better understand how to: 1) initially deploy managed IP video services via DOCSIS, 2) plan their bandwidth and network resource requirements, 3) support existing video services in IP, and 4) optimize the network resources required as IP video viewership grows from small numbers to ultimately become the predominant means of video delivery in cable networks.

INTRODUCTION

Adaptive streaming is the primary technology for delivering over-the-top (i.e., unmanaged) IP video content to IP devices such as tablets, smartphones and gaming devices through the operator's Data Over Cable Service Interface Specification (DOCSIS) network. Adaptive streaming is the defacto delivery mechanism for OTT services. For managed services however, there is a popular assumption that multicast streaming video should be the principal delivery format to primary screens, not adaptive streaming. However, delivery of managed video in multicast format creates significant complexities for the operator, not the least of which are how to duplicate existing and planned services such as targeted advertising and network-based DVR, amongst others, and managing different segregated service group sizes compared to data services.

This paper presents a proposal to employ a comprehensive *managed* IP video services solution using adaptive streaming protocols with appropriate enhancements. An end-to-end multi-screen IP video architecture is presented, including the role of these adaptive bit rate (ABR) protocols.

The trade-offs of using adaptive streaming versus multicast for delivering managed video services are discussed. One of the other major concerns of operators is the bandwidth that will be required to deliver managed IP video services. Many factors come into play with the introduction of IP video, and our modeling results show that multicast gains may evaporate, so there is no penalty for using unicast-based adaptive protocols.

MANAGED IP VIDEO ARCHITECTURE

Multi-screen IP video delivery requires an end-to-end ecosystem that must encompass data, control and management planes. It must interact with legacy encoding, ad insertion, and content management systems while operating in parallel with traditional linear broadcasting. Operators will migrate towards multi-screen IP video to deliver content to a new generation of consumer devices such as tablets, smartphones and gaming devices; and to enable new cloud based services to attract and retain customers.

[Ulm_CS_2012] described an end-to-end conceptual architecture to support the evolution to IP video delivery. This architecture is segmented into Application, Services & Control and Media Infrastructure layers. Each of these layers is further decomposed into functional blocks.

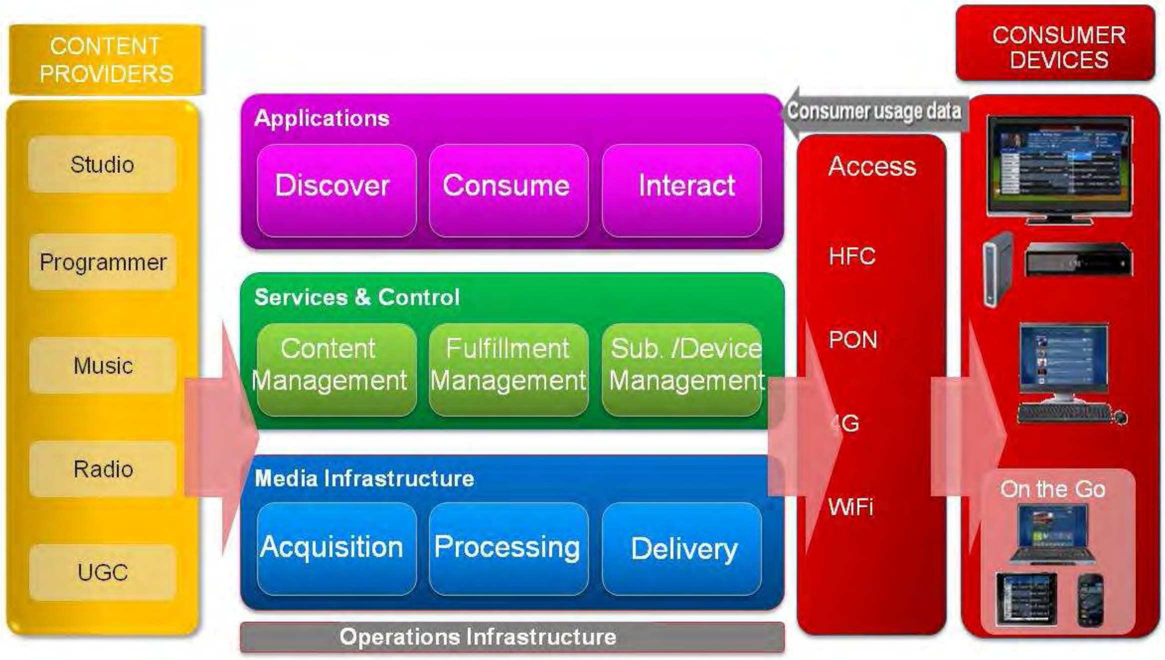


Figure 1: High Level Conceptual Architecture

Figure 1 shows a high-level abstraction of an end-to-end functional architecture for the delivery of IP video from content providers to content consumers. The video service provider must ingest content from multiple content providers, process it appropriately and then transport it over multiple types of access networks to the destination consumer devices.

The representation breaks the functions into three primary layers: Applications layer; Service & Control layer; and Media Infrastructure layer. A fourth functional block called Operations Infrastructure overlays the three primary layers.

Application Layer

The Applications layer provides interaction with the end user and is largely responsible for the user experience. It includes functions that discover content through multiple navigation options such as user interfaces (UI), channel guides, interactive search, recommendation engines and social networking links. It enables the user to consume content by providing applications for video streaming, video on demand (VOD) and network DVR (nDVR) consumption. These applications integrate with the Service & Control layer to authenticate the user, confirm access rights, establish content protection parameters and obtain resources for delivery as required.

The Application layer also provides companion applications which enable user interaction in conjunction with media programs. These may be as simple as allowing interactive chat sessions among viewers watching the same program or enable more complex integration with social media applications. It also enables enhanced monetization with new advanced advertising capabilities such as telescoping ads.

Services & Control Layer

The Services & Control layer is responsible for assigning resources within the network and for enforcing rules on content consumption that ensure compliance from a legal or contractual perspective. It includes functions that manage content work flow through all phases of its lifetime including ingest, transcoding, digital rights management (DRM) and advertising insertion policy. Other functions manage the fulfillment of user requests for content delivery by providing resource and session management, nDVR and VOD management and Emergency Alert System (EAS) and blackout support. Finally, it must manage subscribers and devices to ensure content delivery to authorized consumers in a format compatible with the consuming device.

The Services & Control Layer provides a unified approach for managing entitlements, rights, policies and services for the multitude of devices and DRM domains expected in the emerging adaptive streaming IP video service model. This solution must provide a mapping function between the billing system and the DRM system interfaces, recognizing that leveraging existing billing interfaces provides for a more seamless transition from legacy solutions. Billing should focus on account level transactions – allowing the network and associated DRMs to determine if content viewing is allowed on a specific account or a specific device. A tight integration with compelling DRM solutions is a necessity. By abstracting the complexity of a multi-DRM system, the Service & Control layer efficiently manages entitlements, rights, policies and services for a multitude of devices across a number of DRM domains. These unified provisioning functions will provide an essential building block for end-to-end multi-screen video solutions. For a detailed discussion on this topic, see [Falvo_2011].

Media Infrastructure Layer

The Media Infrastructure layer is responsible for managing video content flow and delivering the media. It includes content ingest, preparations, and delivery to the devices. Functions in this layer acquire content from satellite or terrestrial sources as either program streams or files and encode it for ingest into the system. It processes the content to prepare it for delivery. This includes functions such as transcoding, multiplexing, advertising insertion, EAS, black outs and encryption. Finally, this layer delivers the content to the target device through mechanisms such as Web servers, content delivery networks (CDNs), and streaming servers.

It is in the Media Infrastructure layer where the decision is made on video delivery protocols. For ABR distribution models, this layer includes packaging into appropriate file formats, manifest creation and publishing to a CDN origin server.

The remainder of this paper takes a detailed look at managed content delivery using adaptive bit rate (ABR) protocols.

ABR BENEFITS FOR MANAGED IP VIDEO SERVICE

Using ABR for IP video delivery can be considered a “pull” delivery model in which the end client requests the video data. With ABR, the video content is broken up and stored in a CDN as a series of small files at multiple different bit rates. The end client uses standard HTTP “get” requests to download each file segment into a local buffer from which the content is played out. The client monitors the rate at which downloads are occurring and the available locally buffered content to determine which bit rate to request. If the network is fast, a high quality high bit rate will be selected. If the network is slow, a lower quality, lower bit rate option will be requested. This is an inherently unicast service as there is no coordination between clients (even if they are watching the same content at the same time, two clients would download it independently). A tutorial on ABR for cable may be found in [Ulm_2010]. Below is an in-depth look at many key considerations and benefits in using ABR for a managed service.

CPE: Right Choice for Second/Third Screens

A key driver for migrating to IP video delivery is the ability to deliver services to a wide range of IP devices, in particular personal computers, tablets, smartphones and gaming devices. Operators want to offer these services to remote subscribers who are “off-net” as well as managed IP video services to devices inside their own network. The protocols are applicable to both linear television and on-demand delivery.

ABR protocols are the best choice for these smaller screen devices and off-net operations. They have very simple customer premises equipment (CPE) clients that adapt dynamically to changing internet resource availability. With extremely high churn on CPE devices, it is very important from an operational perspective to support the embedded client on new devices. ABR protocols are becoming the de facto standard for IP video delivery to these devices. With ABR, the operator will not become the long pole in the tent while trying to provide device drivers for the newest gadget of the week.

In-Home delivery of managed IP Video

Since ABR protocols use HTTP, they are extremely well suited for traversing home firewalls. This is in stark contrast to multicast delivery through consumer owned routers. This means that ABR is much better from an operations and support perspective.

The other issue with in-home delivery is that it may span a consumer’s home wireless network with unpredictable latency and throughput. The ABR protocols are also well suited to adapt to this environment.

CDN Considerations

There are some CDN considerations that the operator must review when architecting an IP video delivery system. Traditional VOD systems today use a “push” model where streaming content ‘pushes’ through the system in real time. This approach supports multicast delivery, but requires session management and admission control to secure resources, guaranteed bandwidth from the server to the client, CBR-based video, and dedicated servers.

The server has the added constraints of maintaining correct timing for transmitting content. Any network-induced jitter must be removed by the edge device (edge QAM or set-top box). This approach uses a non-robust transport (e.g. UDP or RTP) which requires added complexity to detect and recover from errors. Because of all of this, a push CDN model cannot exploit general internet CDN technologies for access network delivery.

In an adaptive streaming world, clients “pull” content from the CDN as files or file segments using a reliable HTTP over TCP transport. The client pull approach is CDN friendly and allows operators to re-use HTTP-based Web caching technology that uses standard servers. The CDN caching reduces backbone capacity requirements for both linear and on-demand content. Multicast only reduces backbone traffic for linear content. All of this gives the operator significant cost benefits by leveraging internet technologies. Its state-less architecture also readily scales as needed.

To summarize, a pull CDN model provides the operator with a simpler, more cost-effective system that uses a single IP infrastructure. It leverages internet technologies for performance and resiliency. It supports ABR and enhanced quality of experience (QoE) from a common infrastructure. The operator is able to incorporate public and third party CDN services with its private CDN. Finally, this scales to a global delivery model.

Quality of Experience Considerations

In offering a managed IP video service, QoE is an important consideration for operators. One of the key factors is how the system reacts to congestion. With the high levels of compression in today's video streams, any lost packets can have severe impact on the user's experience. Implementing a multicast- based streaming service puts significant additional burdens on the operator's system. As mentioned earlier, multicast streaming is based on non-robust protocols, so in a heavily congested environment they might lose packets. The operator could choose to over provision the amount of bandwidth needed to prevent these conditions, in which case they are throwing away potential capacity gains from using multicast. The alternatives are to implement some combination of admission control and/or error recovery. An admission control algorithm will be further complicated if variable bit rate (VBR) video delivery is used to maximize bandwidth savings rather than constant bit rate (CBR). An error recovery system introduces new servers into the network and requires custom clients in the consumer devices. Overall, the design, deployment and operation of a multicast-based system are inherently complex.

ABR protocols were developed for Internet delivery with its constantly changing throughput. ABR seamlessly adapts to this varying environment. In a managed network with infrequent periods of congestion, ABR reduces its bit rates during these periods to compensate. The impact on QoE might be comparable to that of running legacy MPEG video through a statistical multiplexer (statmux), which is familiar to operators. ABR also is based on a reliable TCP protocol that has error recovery already built into it, so any packets lost during congestion are automatically retransmitted. Thus, it prevents blocking and other video artifacts that significantly impact QoE. In this case, no network resources need be reserved in advance for the service and ABR reduces or eliminates the potential for blocking. Using adaptive protocols for all IP video delivery helps the operator's overall system become much simpler. More on this topic can be found in [White_2012].

Another QoE consideration is the impact of channel change time. ABR protocols are well suited to fast change times as they can quickly load lower bit rate streams and then switch to higher bit rates as bandwidth is available. Using multicast delivery requires separate additional bandwidth and a proprietary protocol to quick start the video delivery in parallel with the multicast video.

Advanced Services

Another key reason for migrating to IP video services is the ability to offer new advanced services. In particular, this might include highly targeted advertising such as personalized advertisements and telescoping. The system must also support EAS and blackout identical to legacy video services. Using its playlist manipulation, ABR provides the service provider with tremendous capability to re-direct a client on-the-fly with minimal effort and equipment. Supporting these advanced services using multicast delivery becomes problematic.

Miscellaneous Considerations

IP video penetration will occur over a long period of time. This means that the operator's home gateway will continually change during that time as well. Today cable operators have DOCSIS D3.0 devices in the field with 3, 4 or 8 downstream channels. Over the next several years we will see this expand to include 16, 24 and perhaps 32 downstream channels. The operator needs to manage this DOCSIS modem transition. Using ABR and its unicast delivery allows every modem to be in a bonding group suited to its capabilities; multiple bonding groups can then overlap, allowing the cable modem termination system (CMTS) to fully utilize the bandwidth. Multicast delivery runs into multiple problems in a mixed bonding group environment as discussed in [Ulm_2009].

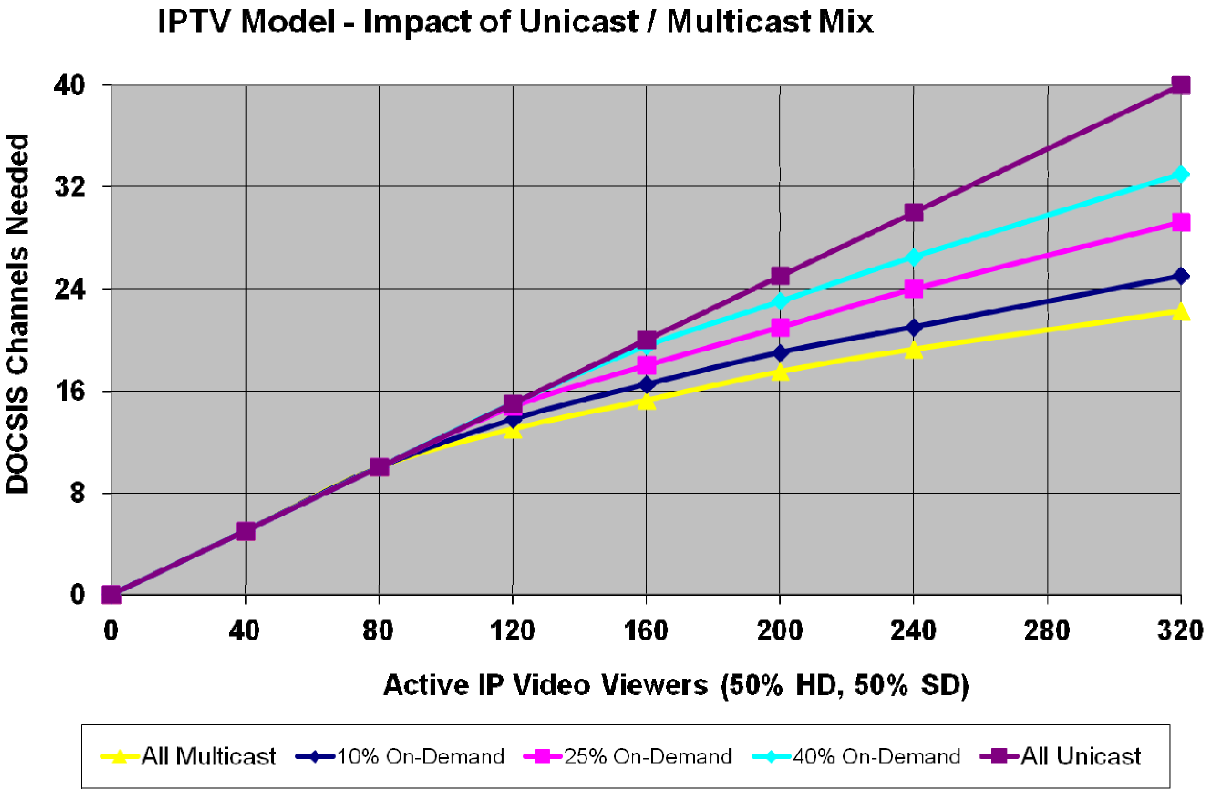


Figure 2: Impact of Unicast / Multicast Mix

ABR BANDWIDTH CONSIDERATIONS

A detailed analysis of bandwidth requirements for ABR compared to multicast was given in [Ulm_CS_2012]. The findings were that, under most conditions, multicast delivery will have little or no bandwidth capacity advantages over ABR unicast delivery. Figure 2 shows some results from that paper.

For early deployments of IP video, the penetration rate will be low. As indicated in this figure, there is no multicast benefit below 120 active viewers. With many operators considering phasing in IP video gradually, the operator also needs to factor in their plans for service group sizes. If the phasing takes 5-7 years, will the operator initiate node splits and cut service group sizes in half during that time? At the same time, increased VOD usage and the introduction of nDVR services might cause a shift from 10% to 25% or even 40% unicast usage. Figure 2 clearly shows what happens when the number of active viewers drops from 320 to 160 or 240 to 120 viewers.

Impact of Multi-Screen Delivery

This analysis was done for a two screen system: 50% of viewers watching high definition (HD) TV content and 50% of viewers watching standard definition (SD) TV content. With multi-screen delivery being a key impetus for IP video services, Motorola extended the IP video capacity modeling to see the effect of multi-screen viewing on capacity requirements.

Below are some sample outputs from the enhanced IP video capacity modeling. This looks at the bandwidth requirements for IP video for two different sized service groups as penetration grows.

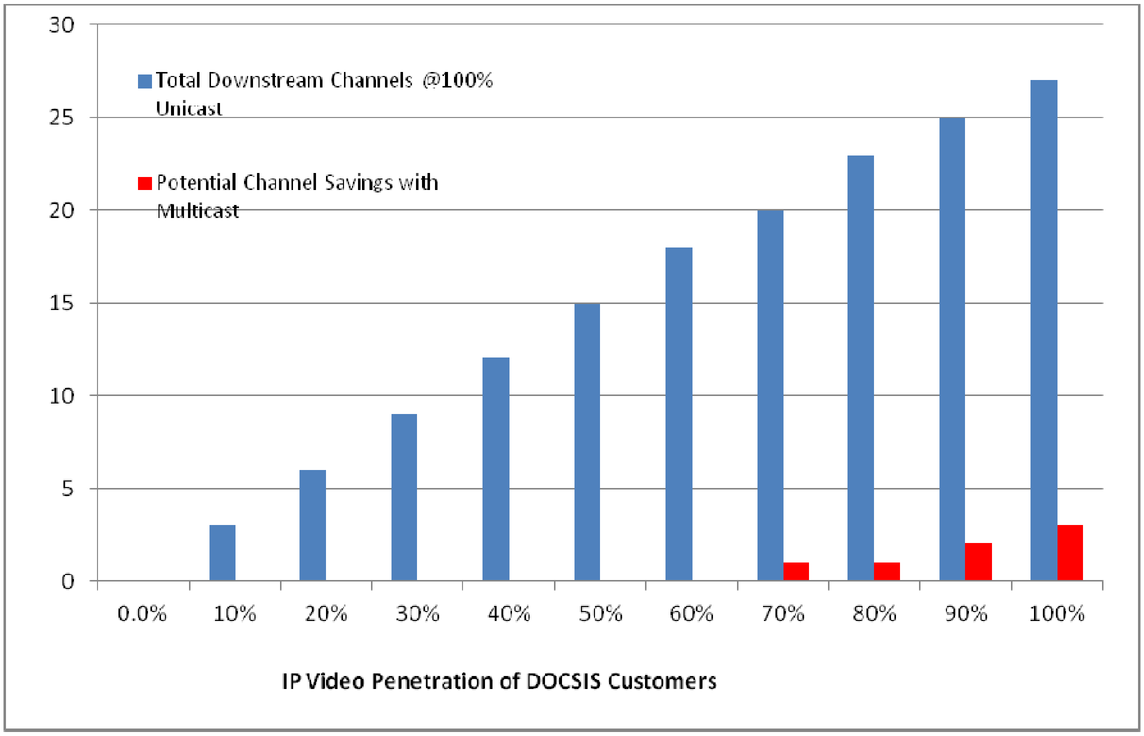


Figure 3: IP Video Bandwidth & Multicast Savings: 320 Active Viewers

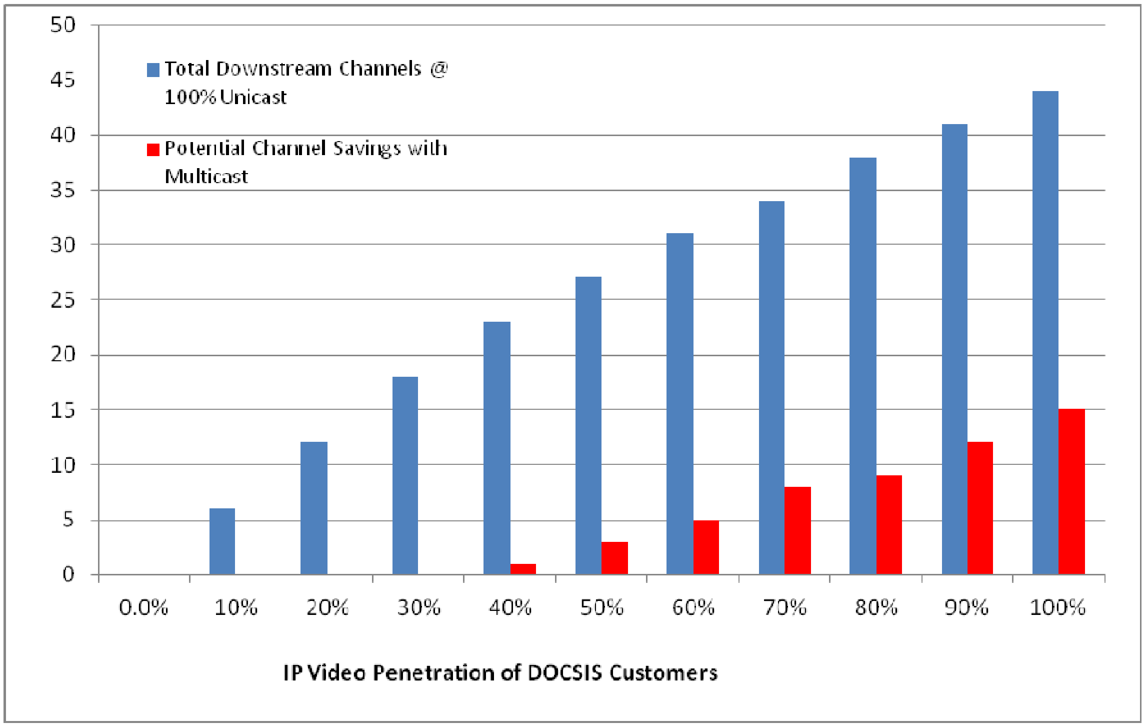


Figure 4: IP Video Bandwidth & Multicast Savings: 640 Active Viewers

In Figure 3, 100% IP video penetration corresponds to 320 active viewers which might represent a 500 homes passed (HP) service group, identical to the analysis above. Figure 4 doubles the service group size to 640 active viewers. In both these examples, viewership is spread across five different screen sizes: 30% HDTV, 30% SDTV; 20% tablets; and 10% each for two smaller screen sizes. It also assumes 25% on-demand usage which is reasonable if nDVR is deployed for the IP devices.

As indicated in Figure 3, the potential multicast gain is non-existent until the operator has reached 70% IP subscriber penetration. Even at 100% penetration, the multicast gain is only 3 channels or ~10% of capacity. This amount is almost negligible in a converged cable access platform (CCAP) environment capable of 64 channels per port.

In Figure 4, the serving group size is doubled. Perhaps the operator combined two fiber nodes to the same CCAP port to get additional multicast gains. Even with this extremely large service group of ~1000 HP, the multicast savings is still less than 20% at 70% IP penetration, yet it requires 34 DOCSIS channels of capacity for the large serving group. The small savings for multicast comes at a significant cost in spectrum used. It also comes in the late stages of the IP video deployment.

QoS in a Multicast Implementation

The purpose of implementing Multicast for delivering managed video content is to save bandwidth. By its very nature, a multicast system only makes sense if fewer channels of spectrum are required than a unicast implementation. Multicast designs are wholly dependent on the assumptions of multicast viewership during peak. At peak viewership, if more programs are being requested than the multicast service group was designed for, blocking occurs resulting in a denial of service. Therefore a prudent design calls for a safety factor in the number of QAMs reserved for the multicast service group. However this flies in the face of the rationale for implementing multicast, which is bandwidth savings.

In contrast, in a unicast implementation, if the bandwidth peak is achieved, the adaptive bit rates are lowered for the viewers in the service group. While video quality may lessen slightly in these cases, there is no denial of service. Therefore, unicast is a better choice for insuring a non-blocking service at peak usage times.

SPECTRUM MIGRATION STRATEGIES

Another very important aspect to IP video migration is finding sufficient spectrum. Some operators have already made more spectrum available by recovering analog TV channels using digital TV terminal adapters (DTA) while other operators have upgraded their hybrid fiber coaxial (HFC) to 1GHz or turned to Switched Digital Video (SDV). This available spectrum is being gobbled up today as more HD content is deployed, VOD requirements continue to increase and high speed data (HSD) services continue to grow at 50% annual rates. So there may still be a need for additional spectrum to ramp up IP video services with a corresponding economic impact.

Early Transition Plans – Hybrid Gateways

One way to significantly reduce spectrum requirements is to convert legacy MPEG-2 linear TV to IP video in a video gateway device that includes a transcoder. This approach requires no new spectrum for linear TV as this video gateway device appears as a set-top box (STB) to the system and uses legacy broadcast content.

The video gateway also has the advantage that it is the single point of entry for video services and allows IP STBs to be deployed elsewhere in the home behind it. These hybrid devices can also operate as IP devices and are pivotal in the transition to an all IP system. Longer term, the transcoding capability and adaptive protocols supported by the gateway may limit the quantity and type of IP devices supported in the home. Eventually the operator will want to support IP devices directly from the "cloud" using their network infrastructure.

A detailed discussion of the home gateway migration is given in [Ulm_CS_2012].

Complete Recovery of Legacy Bandwidth

The previous section on video gateway migration plans helps the operator as they begin the IP video transition. However, the end game is to eventually get to an all-IP system. Legacy MPEG digital TV services may continue to consume 50% to 80% of the available spectrum even after DTA and 1GHz upgrades. Regardless of which path the operator initially took to free up spectrum, eventually they will need to install switched digital video (SDV) to reclaim all of the legacy digital TV bandwidth.

Adding SDV to the mix also increases the need for narrowcast QAM channels. This plays well into a CCAP migration. As the mix between legacy and IP subscribers changes, an operator will need to re-assign SDV bandwidth to IP video bandwidth. This is well suited for CCAP. For a detailed discussion on IP video economics in a CCAP world see [Ulm_NCTA_2012].

Some SDV capacity reclamation modeling results are shown in Figures 5 and 6. Figure 5 shows the total spectrum required for legacy video services as the number of legacy viewers is reduced to zero. It assumes a video service with 180 HD programs (3 per QAM) and 200 SD programs (10 per QAM), so full broadcast requires 80 QAM channels. Figure 6 shows the corresponding SDV narrowcast requirements.

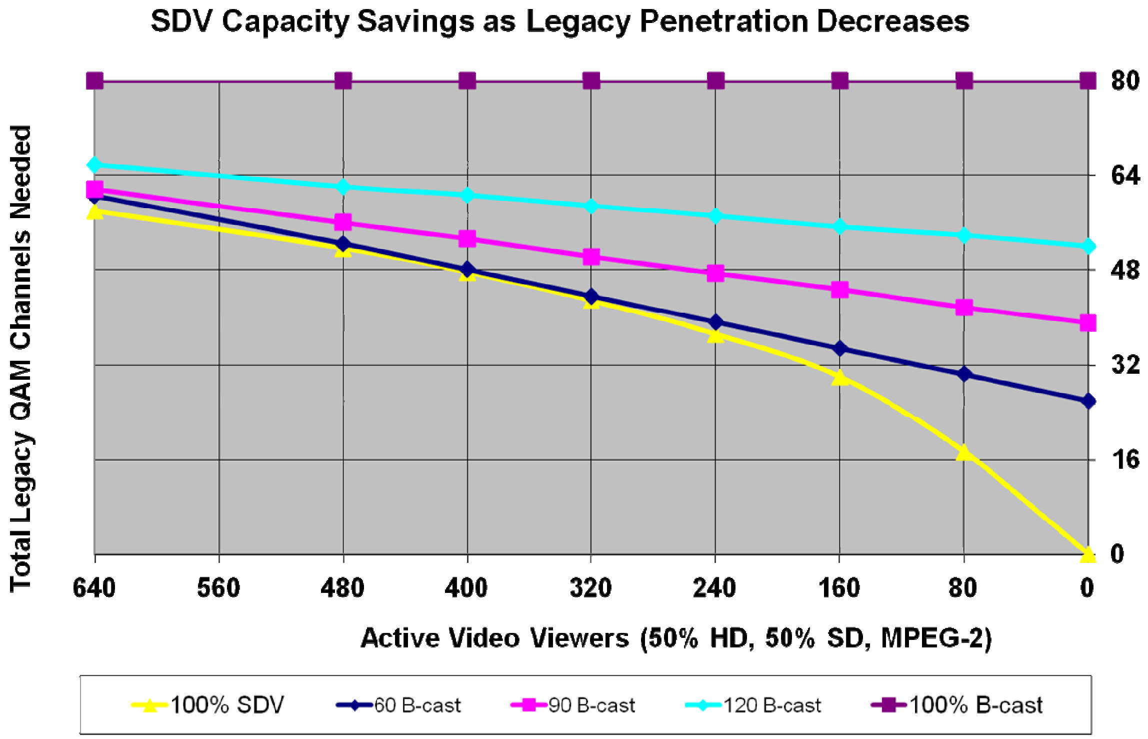


Figure 5: SDV – Total Capacity Savings with Decreasing Penetration

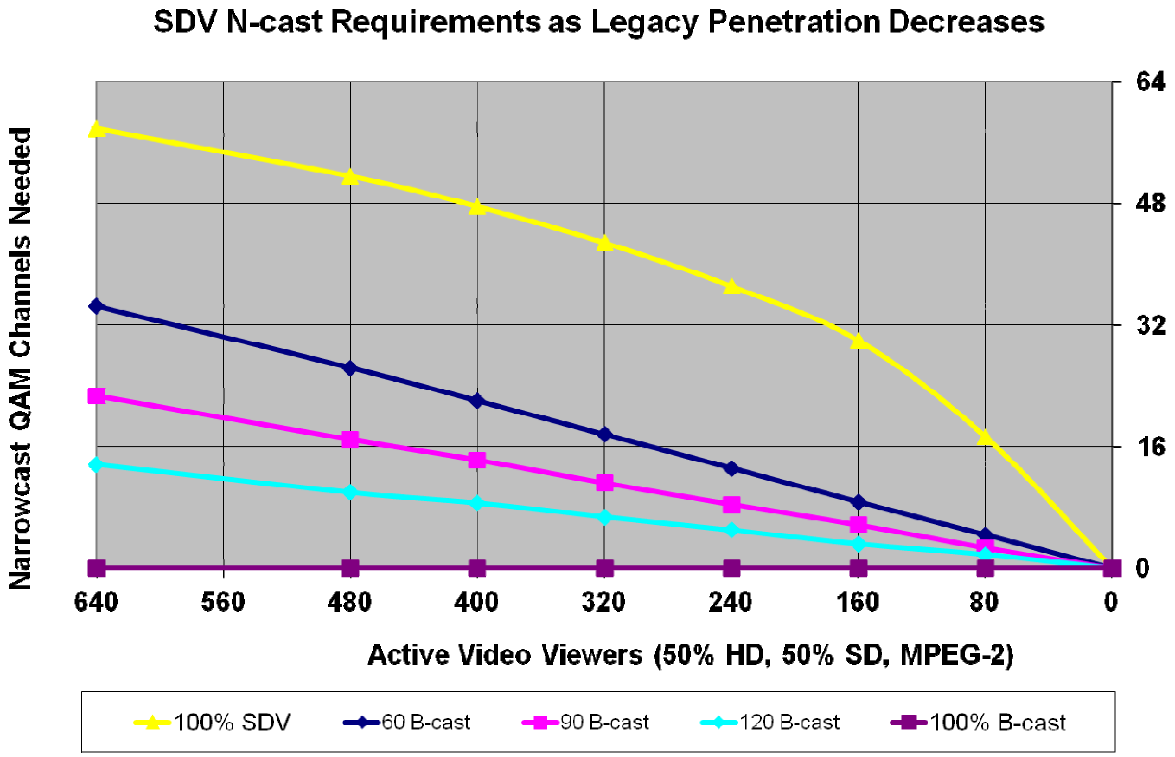


Figure 6: SDV – Total Narrowcast Requirements with Decreasing Penetration

Four scenarios are given varying the amount of switched content up to 100% switched. As shown, 100% switched provides the most bandwidth savings, but requires significantly more narrowcast. The operator has complete flexibility in trading off between spectrum saved and narrowcast QAM requirements. As can be seen in Figure 6, as the number of legacy viewers decreases, there is a corresponding decrease in narrowcast QAM requirements. This allows the operator to repurpose SDV QAM channels as they become freed for DOCSIS channels (HSD or IP video) or additional SDV savings.

It is informative to look at an example where the operator allocates twelve QAM channels for SDV and watch the impact as their legacy viewers are reduced. From Figure 6, the curve representing 120 broadcast programs and 60HD/80SD switched programs crosses 12 QAMs at 560 active viewers. Now looking at Figure 5, this scenario (i.e. 560 viewers, 120 B-cast) requires 64 channels of spectrum, freeing 16 channels (compared to 80 channels for 100% broadcast) for other usage such as IP video growth. As IP video penetration grows, legacy penetration shrinks. The next curve (90 broadcast with 90HD/110SD switched) on Figure 6 crosses 12 QAMs at 320 viewers. Mapping to Figure 5, this scenario (i.e. 320 viewers, 90 B-cast) only requires 50 channels of spectrum, so 30 channels are now available. The next scenario (60 broadcast with 120HD/140SD switched) crosses 12 QAMs around 200 viewers and requires ~36 channels for more savings.

As a result, the SDV spectrum savings are significantly more than multicast gains seen in the previous section. The SDV benefits are also available for small and large service groups. Every operator needs to consider SDV as a crucial part of its IP video migration.

CONCLUSION

Cable service providers will migrate from existing legacy video networks to a full end-to-end IP video system in a number of stages as new services are rolled out. They need to leverage the technology used for these intermediate stages into the final end-to-end system. Therefore, it is critical to have a layered architecture approach as presented in this paper that can isolate the changes between the various components.

Selecting the correct technology is particularly important for the delivery component of the Media Infrastructure layer as it is hardware centric, widely deployed and capital intensive. In particular, this paper focuses on the selection of adaptive protocols as the primary video delivery mechanism and discusses its benefits. ABR enables:

- A wealth of new and constantly changing IP devices
- Easily handles the home environment
- Provides excellent QoE to consumers
- Adapts to congestion without requiring complex admission control or re-try mechanisms
- Leverages internet CDN technology
- Readily supports advanced services including personalized advertising.

The updated IP video capacity modeling results shows the impact of migrating to a multi-screen environment. A 500HP service group may only get 10% multicast gain even once its switched to all IP video delivery.

Understanding the migration plan is a critical piece of the IP video architecture, especially with respect to managing available spectrum. Hybrid video gateways enable the introduction of IP video delivery with minimal impact on an operator's infrastructure. As the system scales, these devices transition to full IP video delivery.

Finally, the operator needs to plan the reclamation of legacy spectrum as they migrate to an all-IP world. This migration will eventually require the use of SDV. The modeling results show that the benefits of SDV are actually greater than the savings from multicast delivery.

In conclusion, the operator needs ABR for its first IP video steps when delivering content to second and third screens; i.e., tablets, smartphones, PCs and gaming devices. Adaptive streaming is the final solution the operator needs once there is an all-IP world with any content, anywhere, anytime, anyplace. We have shown that ABR also handles the transition years and is the only delivery mechanism needed for a managed IP video service.

REFERENCES

[Ulm 2009]	J. Ulm, P. Maurer, “IP Video Guide – Avoiding Pot Holes on the Cable IPTV Highway”, SCTE Cable-Tec Expo, 2009.
[Ulm 2010]	J. Ulm, T. du Breuil, G. Hughes, S. McCarthy, “Adaptive Streaming – New Approaches For Cable IP Video Delivery”, The Cable Show NCTA/SCTE Technical Sessions, spring 2010.
[Falvo 2011]	B. Falvo, D. Clarke, C. Poli, *“Supporting Multi-CAS and DRM Entitlements”*, SCTE Cable-Tec Expo, Nov 2011.
[Ulm CS 2012]	J. Ulm, G. White, *“Architectures & Migration Strategies for Multi-Screen IP Video Delivery”*, SCTE Canadian Summit, March 2012.
[Ulm NCTA 2012]	J. Ulm, G. White, *“The Economics of IP Video in a CCAP World”*, NCTA Technical Sessions, May 2012.
[White 2012]	G. White, J. Ulm, *“Reclaimng Control of the Network from Adaptive Bit Rate Video Clients”*, NCTA Technical Sessions, May 2012.

RECLAIMING CONTROL OF THE NETWORK FROM ADAPTIVE BIT RATE VIDEO CLIENTS

John Ulm & Gerry White
Motorola Mobility

Abstract

This paper provides a brief introduction to adaptive bit rate (ABR) video and discusses why handling this class of traffic well is very important to the cable operator. It then examines the major differences between ABR and the current IP and MPEG video delivery mechanisms and looks at the impact these differences have on the network. Some interesting experimental results observed with real world ABR clients are presented. A number of problems which may develop in the network as ABR clients are deployed are discussed and possible solutions for these proposed. Finally, the paper looks at the cable modem termination system (CMTS) as a potential control point that could be used to mitigate the impact of the ABR clients and regain control of the access network for the operator.

INTRODUCTION

Adaptive bit rate is a delivery method for streaming video over IP. It is based on a series of short HTTP progressive downloads which is applicable to the delivery of both live and on demand content. It relies on HTTP as the transport protocol and performs the media download as a series of very small files. The content is cut into many small segments (chunks) and encoded into the desired formats. A chunk is a small file containing a short video segment (typically 2 to 10 seconds) along with associated audio and other data. Adaptive streaming uses HTTP as the transport for these video chunks. This enables the content to easily traverse firewalls, and the system scales exceptionally well as it leverages traditional HTTP caching mechanisms.

Adaptive streaming was developed for video distribution over the Internet. In order to deal with the unpredictable performance characteristics typical of this environment, ABR includes the ability to switch between different encodings of the same content. This is illustrated in Figure 1. Depending upon available bandwidth, an ABR client can choose the optimum encoding to maximize the user experience.

Each chunk or fragment is its own stand-alone video segment. Inside each chunk is what MPEG refers to as a group of pictures (GOP) or several GOPs. The beginning of each chunk meets the requirements of a random access point, including starting with an I-frame. This allows the player to easily switch between bit rates at each chunk boundary.

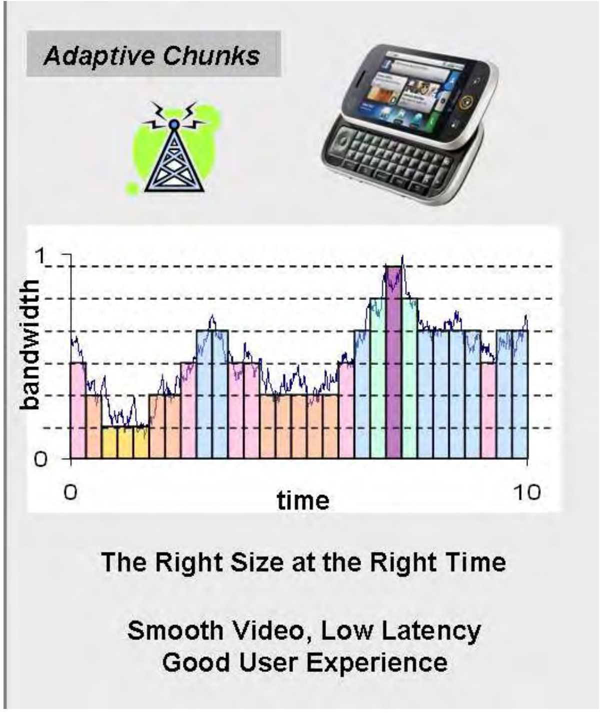


Figure 1 Adaptive Streaming Basics

Central to adaptive streaming is the mechanism for playing back multiple chunks to create a video asset. This is accomplished by creating a playlist that consists of a series of uniform resource identifiers (URIs). Each URI requests a single HTTP chunk. The server stores several chunk sizes for each segment in time. The client predicts the available bandwidth and requests the best chunk size using the appropriate URI. Since the client is controlling when the content is requested, this is seen as a client-pull mechanism, compared to traditional streaming where the server pushes the content. Using URIs to create the playlist enables very simple client devices using web browser-type interfaces. A more in-depth discussion of ABR video delivery can be found in [ADAPT]

IMPORTANCE OF ABR

Second and Third Screens

ABR based video streaming has become the de-facto standard for video delivery to IP devices such as PCs, tablets and smart-phones. ABR clients are typically shipped with (or are available for download to) these devices as soon as they are released. Given the short lifetime of this class of device this is a key enabler, especially compared to the time required to deploy software to traditional cable devices. As mentioned previously, ABR delivery simply requires an HTTP connection with sufficient bandwidth so that it is available both on net and off net. With these advantages, both over-the-top (OTT) and facilities based service providers are leveraging ABR so that essentially all video delivery to second and third screen devices uses this mechanism.

Primary Screen

ABR is also used to deliver a significant quantity of video to television screens in both standard and high definition formats. Over-the-top providers of video service leverage ABR clients installed in platforms such as gaming consoles, Blu-ray players, set-top box-like devices and smart TVs to provide video services to the primary screen. This content rides over the service providers' high speed data (HSD) service and, in many cases, constitutes the bulk of the HSD traffic.

ABR Traffic Load

Studies of Internet traffic patterns [SAND], [VNI] show that video has become the dominant traffic element in the Internet, consuming fifty to sixty percent of downstream bandwidth. Netflix alone constitutes almost thirty-three percent of peak hour downstream traffic in North America. Thus, how well the network supports ABR based IP video is obviously crucial to providing a satisfactory customer experience. In addition, delivery of Internet video to televisions is predicted to grow seventeen-fold by 2015 to represent over sixteen percent of consumer Internet video traffic (up from 7 percent in 2010) [VNI]. Thus, many of the customers will not only be viewing IP video, but will be doing so on a large screen device with expectations of high quality.

In addition to the Internet video explosion, significant amounts of managed service provider video will also migrate to an ABR mechanism, further increasing the percentage of ABR traffic on the network.

Having this much ABR traffic on the network means that it will be a key driver of network costs and with ABR delivering prime entertainment services, how well it is supported will be a key metric for customer satisfaction going forward. Therefore, understanding the issues around delivery of ABR over the DOCSIS network will be crucial for MSO's video service delivery, and for their ongoing profitability.

ABR vs. CURRENT VIDEO DELIVERY

ABR video delivery has a number of very significant differences to both MPEG video delivery and streamed IP video delivered over Real-time Transport Protocol/User Datagram Protocol (RTP/UDP) as used in a Telco TV system such as Microsoft Media Room [MMR]. A number of these differences are discussed below.

Client Control

ABR has been developed to operate over an unmanaged generic IP network in which bandwidth decisions (i.e. choosing the video bit rate to request) are made by the client device based on its interpretation of network conditions. This is fundamentally different from the approaches used for existing MPEG or conventional streamed UDP video delivery, where devices under the direct control of the network operator make the important decisions relating to bandwidth. Thus, in MPEG delivery, the encoding, statistical multiplexing and streaming devices determine the bit rate for a given video stream. These devices are under control of the service provider. Similarly for a UDP streaming solution, the video is encoded and streamed at a selected rate from devices owned by the service provider. In contrast, the behavior of ABR clients is specified by the developer which, in general, will be a third party outside the service provider's control.

Variable Bit Rate

As described previously, an ABR client will select a file chunk with a bit rate which it believes to be most appropriate according to a number of factors including network congestion (as perceived by the client) and the depth of its playout buffer. Thus the load presented to the network can fluctuate dramatically. This is in stark contrast to both MPEG and UDP video streams which are either constant bit rate (CBR) or are clamped variable bit rate (VBR) (i.e., bandwidth can vary up to a maximum bit rate but not beyond it).

A more detailed discussion on the impact on network loading of a number of factors is found in a later section of this paper.

Admission Control

ABR clients join and leave the network as users start and stop applications. From a network perspective, there is no concept of a session with reserved resources or admission control. Again this is the antithesis of MPEG or UDP video in which a control plane operates to request and reserve network resources and determines whether to admit a user. In a controlled network, adding a new user session can be guaranteed not to impact existing users. Once resources are exhausted, any additional session requests will be denied, introducing a probability of blocking into the system. In an ABR model under network congestion, each new session will reduce the bandwidth available to all existing sessions rather than be denied. Thus, users may see a variation in video quality as other ABR clients start and stop. This reduction in quality during peak times is analogous to statistical multiplexing in legacy MPEG video. During peak times, the statmux reduces bit rates across the various video streams to fit within its channel. The ABR system has an advantage in that it will be over a larger channel using DOCSIS bonding.

Congestion Control

With MPEG or UDP streaming video delivery, congestion control is not relevant as the control plane provides admission control to ensure it does not occur. When ABR is used for video delivery, congestion control is a potential issue. The situation is complex in that three levels of congestion control mechanisms are involved operating at different layers in the protocol stack. At the media access control (MAC) level, the CMTS is responsible for scheduling downstream DOCSIS traffic [MULPI]. Operating at the transport level is standard Transmission Control Protocol (TCP) flow control based on window sizes and ACKs, [TCP] and, finally, at the application level the client can select the video bit rate to request. The latter two levels of control (TCP and application) are the responsibility of the ABR clients and as such are outside the control of the network operator. Interaction between these three flow control mechanisms is not well understood at this time and may have unforeseen impacts.

Prisoners Dilemma

As noted above, ABR clients have the responsibility to select the quality (bit rate) of the video they request to download. The algorithms and parameters used by each client to make this decision are outside the control of the network operator. Each client is faced with a decision not unlike the classic “prisoner’s dilemma” [PDIL] in that they can elect to optimize for their own benefit or they can optimize for the common good of all clients on the network (including their own). For example, a very selfish client may never request a lower quality file even during network congestion based on the assumption that other clients will do so, and thus resolve the congestion for them. Commercial pressures to create “better” clients may drive in this direction, but if all clients move to this mode the network will fail. This is not an issue with MPEG or UDP streaming delivery as the network operator has the incentive and necessary controls to offer a quality service to all customers.

Imperfect Knowledge

Clients base their decisions on what to request based on their local knowledge rather than on an overall view of the network conditions. This is in contrast to MPEG or UDP streaming where the network operator provisions the video bit rates based on knowledge of the end-to-end network and expected loads.

The following section on potential problems will address these issues in more depth and attempt to develop some potential solutions.

ABR CLIENT CHARACTERIZATION

As discussed previously, the ABR client plays a critical role in the operation of adaptive protocols. For an operator trying to provide a differentiated quality of experience, it is important to understand how different ABR clients behave under various circumstances.

Motorola research teams took multiple different types of clients into the lab to analyze their behavior. Previous work [Cloonan] discussed results from a simulator. Our goal was to capture live client interaction. Operation during steady state was relatively stable. The interesting observations occurred during startup and when video bit rates were forced to change.

At startup time, clients try to buffer multiple segments as fast as they can. This was particularly obvious for video on demand (VOD) assets where the entire content stream is accessible. Live content tends to have a limited playlist available to the client, preventing large buffer build up. During this startup period, the clients are also calculating the available bandwidth and may decide to switch bit rate. This action may cause some segments to be re-fetched with the new resolution. Overall, the differences between clients seemed fairly subtle for startup.

In our lab environment, the amount of bandwidth available to the ABR client was adjusted. In this manner, the client was induced to switch video bit rates. After reducing available bandwidth, the clients in general made a smooth transition to a lower bit rate. Some clients reacted more quickly than others in down shifting. When the available bandwidth is opened up again, clients started searching for new higher bit rates with the associated buffering of segments, similar to startup. It was in this phase where we saw the most differences between clients. In fact, we saw differences from the same device running different revisions of their protocol.

POTENTIAL PROBLEMS

Based on the above characterization, operators must be aware of some potential problems. As was discussed, there is a burst of additional traffic during startup and when switching to higher bit rates. The system must be capable of handling this additional traffic burst.

Actively managing ABR video traffic may be challenging given that every ABR client may be operating its own disjoint algorithm. This is also compounded since client behavior may change with the download of an updated revision. Bandwidth stability may become a concern if multiple clients become synchronized. For example, the network becomes congested causing a group of clients to lower bit rates. If these clients then sense that bandwidth is available (i.e. it is released due to downshifting by other clients), there may be a surge in traffic that causes congestion, and the cycle repeats.

In general, ABR clients are designed for general Internet usage, so they tend to back off quickly and may be slow to ratchet their bit rates back up. This will create some stability and should prevent the above oscillation, but this may make it challenging to fully utilize the network bandwidth.

There are several fairness concerns that must be taken into consideration. If the current bandwidth utilization is high, then new clients just starting their video may select a lower rate than other clients are currently using. Other forms of unfairness may be introduced when network congestion causes video bit rate changes. Some clients may decide to change while others remain at current bit rates, resulting in disparity between clients.

Another concern, especially for a managed video service, is maintaining a good Quality of Experience (QoE). The more that clients change bit rates, the more potential impact there is to QoE. The system should be designed to minimize unneeded bit rate changes.

For future research, Motorola will expand its investigation to system-level behavior for a large number of disparate ABR clients. It is important that the industry grasps the system dynamics for adaptive protocols.

POTENTIAL SOLUTIONS

In a discussion of potential solutions to problems with ABR video delivery under network congestion, two types of ABR traffic must be considered: managed and best effort. Best effort video traffic is OTT types of service which, in general, would be indistinguishable from general Internet traffic.

Managed traffic would typically be video sourced by the service provider, or by a third party with whom the service provider has negotiated a carriage agreement. How well managed traffic is supported is a significant problem for a service provider as it is, in effect, a branded service for which customers will have a higher expectation.

In general, the following potential solutions apply to a managed IP video service. We will highlight where it also applies to OTT traffic.

Controlled Client

Managed ABR services may be made available only from a specific service provider application downloaded by the user. This removes the issues relating to client misbehavior and enables the operator to predict how the client will handle network congestion events.

It has the disadvantage that the operator must keep the application up to date both in terms of feature parity with other clients and with new devices and operating systems as they are released. It also makes it likely that the user must have multiple applications to access different video sources.

This is not applicable to OTT video from third parties, which will be typically be delivered to either a native client on the device or a client provided by the OTT service.

Session Control

One option to control ABR traffic is to implement a session mechanism similar to those used for more traditional video streaming. In this case a user (or possibly a proxy for the user such as a Fulfillment Manager) requesting a video asset would invoke resource checking and reservation mechanisms in the network control plane. The control plane would reserve access network bandwidth for the video session. Mechanisms such as PacketCable™ Multi-Media (PCMM) [PCMM] are in place today to enable quality of service (QoS) bandwidth reservation over DOCSIS. This is detailed later in the paper.

A problem with this approach is knowing when to start and terminate a session and specifically when to acquire and release the resources. For managed video this could be achieved by using a service provider application as described above. The application would invoke the session setup and teardown as part of the video selection and playing process. Even a controlled application implementation would need a back up mechanism to release resources as the user may simply power off a device or lose connectivity. At the minimum, a "no traffic timeout" would be needed (refer to CMTS section below for more details).

Network Override

In conventional ABR video distribution, the ABR client determines the bit rate of the next file to download from the options in the playlist and retrieves this directly from the content delivery network (CDN). This decision could potentially be overridden from the network in a number of ways.

The playlist file provides the bit rate options specified by the service provider. Normally this selection would be statically provisioned and implemented by the encoding and packaging processes as the video asset was processed. For example, each asset could have files created for 1, 2, 4 and 6 megabits per second (Mbps) and the client allowed to select between these. Modifying the selection options in the playlist file provides a potential mechanism for the network to influence the client operation. Thus in times of congestion, the high bandwidth option could be removed by providing a playlist with only 1 and 2Mbps options. This of course requires run time manipulation of the playlists. A potential problem is the lag from playlist manipulation to actual changes in bit rate selection. Even a short playlist file would probably need to represent video content lasting for a significant time so that this mechanism would have a very slow reaction time to network events. Thus, it would not respond to short term congestion events. However, if the network had well known congestion periods (e.g. 8:00 pm through 10:00 pm) it could be used to reduce congestion during these times. Alternatively, the Session Manager might provide notification when the system is congested. This mechanism would not be applicable to OTT traffic as detecting the playlist files would be problematic, and modifying the third party data is unlikely to be permitted.

CMTS AS CONTROL POINT

For users on an HFC network, IP traffic will always flow through the same CMTS port to reach a user at home. As the shared CMTS to CM link is normally the "narrow pipe" in the video distribution network, this is where congestion would be expected. Therefore the CMTS can potentially provide a useful control point to manage the ABR traffic.

Downstream Scheduling and Queuing

The DOCSIS standard provides very complete QoS functionality which may be useful for managing ABR traffic. DOCSIS QoS is based on the IntServ model of filter and flow specifications [INTS]. If a packet matches an installed filter (i.e. classification) it will be mapped to a specific service flow and then forwarded based on the parameters associated with that flow. Classification is based on matching fields in the packet header such as IP address and Differentiated Services Code Point (DSCP) fields. Thus it could be possible to recognize a managed ABR video packet from a well known source address (e.g. video server) or IP subnet. Alternatively all managed video traffic could use a DSCP marking indicating a preferential forwarding class [DSCP]. Inbound traffic to the network from non-trusted sources such as over-the-top (OTT) video would be subject to DSCP overwrite and set to a base priority such as best effort. The CMTS could then provide preferential treatment for the operator's managed video flows.

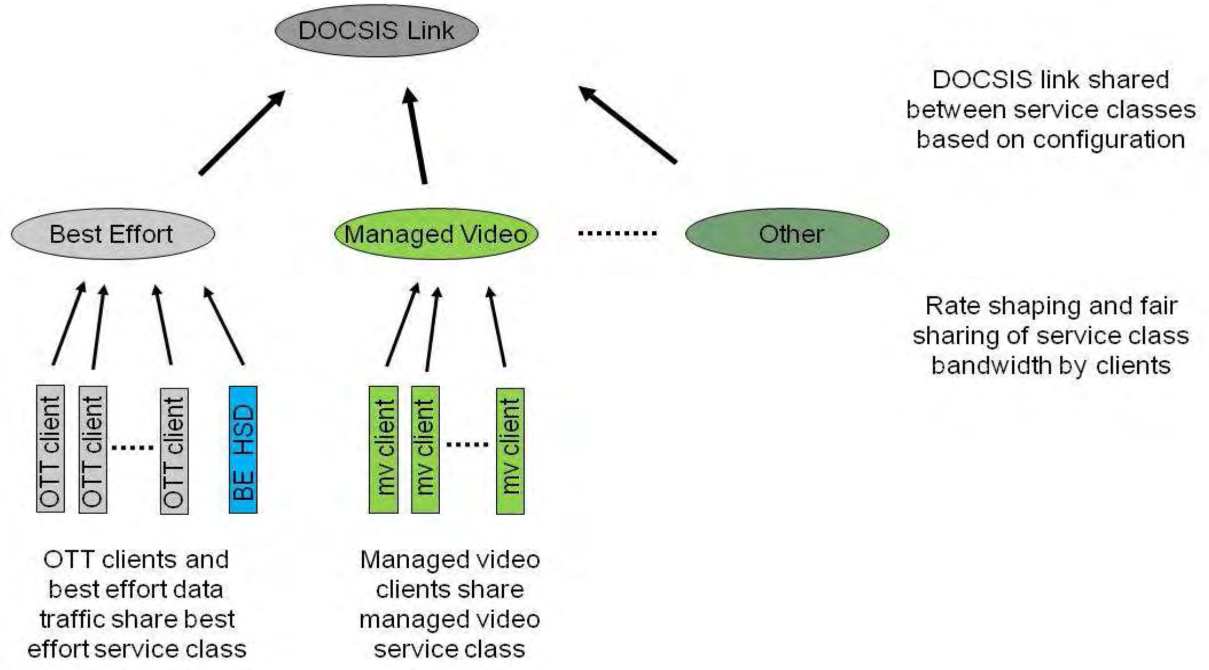

DOCSIS Link
DOCSIS link shared between service classes based on configuration
Best Effort
Managed Video
Other
Rate shaping and fair sharing of service class bandwidth by clients
OTT client
OTT client
OTT client
BE HSD
mv client
mv client
mv client
OTT clients and best effort data traffic share best effort service class
Managed video clients share managed video service class

Figure 2 DOCSIS Link Sharing

If the CMTS supports multi-level scheduling and per-flow queuing as shown in Figure 2, then it can provide fairness between video flows. In this case, each video packet would be mapped to an individual queue (based on the header fields in the packet) within a particular scheduling class such as managed video or best effort traffic. All queues within the same scheduling class share the bandwidth assigned to the class equally so that a single user receives only their fair share and cannot disrupt other video sessions. This mechanism applies to both managed video and OTT ABR video. OTT traffic will be put into the best effort class but will still receive a fair share of the assigned bandwidth for this class. It will, of course, share this with all general Internet traffic. Each scheduling class would be assigned a percentage of the available bandwidth proportionate to its expected load.

Session Control at CMTS

The DOCSIS infrastructure has a mechanism to reserve bandwidth for a flow based on the PacketCable™ Multimedia specification [PCMM]. This provides a potential mechanism to implement resource reservation at the session level. It requires a session establishment and teardown mechanism. In the PCMM model, client applications communicate with an application server (AS) that initiates the QoS requests to the policy server and CMTS. The ABR client application server might be co-located with a session/fulfillment manager, edge server, or user interface (UI) server depending on an operator's control plane infrastructure. Therefore, it would be suited to a managed video service but not OTT. The PCMM / CMTS mechanisms are well understood and include error recovery functions such as the timeout of orphaned sessions.

A potential problem arises in that a video asset may be delivered from one of multiple sources within the CDN. Thus, the filter specification used to identify the packets associated with the session would need to be capable of handling this. This may be as simple as using a known sub-network for the video sources. A more complicated problem is that within the single session, multiple file chunks at different bit rates may be requested due to local events in the client device. The resource reservation for the session could be selected to provide the maximum data rate expected from the client. However, if the client downshifted, this reserved bandwidth would not be used for the managed video but released for use by best effort traffic.

The lab investigations showed that the ABR clients tend to require additional bandwidth during startup and following bit rate increases. The PCMM mechanism can be used to provide a "turbo" mode in which additional bandwidth bursts are allowed for these periods.

CONCLUSION

The impact of ABR traffic on the network is already considerable and is likely to grow significantly as more video is distributed using this mechanism. ABR traffic operates very differently from existing video delivery mechanisms, and in the conventional use case, control over access network bandwidth is essentially abrogated to the device clients. Motorola experiments indicate that these clients vary from device to device and are not necessarily well behaved. Given that they have an incentive to be greedy rather than cooperate for the common good, it seems imperative that the operator finds other mechanisms to control ABR traffic impacts.

A number of options are discussed and the CMTS appears to be a promising location to implement this control. For OTT ABR traffic, the CMTS can provide rate limiting and fair sharing of bandwidth between both ABR clients and other best effort users. This is implemented using existing DOCSIS QoS and CMTS downstream scheduling. For managed ABR traffic, these QoS and scheduling mechanisms may also be used and can also provide segregation of the managed traffic from best effort traffic. With the addition of a session management function in the network, additional control is possible. This enables PCMM control mechanisms to be used to establish service flows for the video streams with defined QoS and reserved bandwidth.

The existing functions provided by the CMTS appear to provide the operator with an excellent control point to impose order on the access network despite the potential for aberrant client behavior.

REFERENCES

[ADAPT]	Adaptive Streaming – New Approaches for Cable IP Video Delivery J. Ulm, T. du Breuil, G. Hughes, S. McCarthy, The Cable Show NCTA/SCTE Technical Sessions spring 2010
[SAND]	Global Internet Phenomena Report Fall 2011; Sandvine
[VNI]	Cisco® Visual Networking Index (VNI) 2011
[MMR]	Microsoft Media Room -www.microsoft.com/mediaroom/
[MULPI]	DOCSIS 3.0 MAC and Upper Layer Protocols Interface Specification www.cablelabs.com
[TCP]	RFC 2581 TCP Congestion Control M. Allman, V. Paxson, W. Stevens
[PDIL]	Kuhn, Steven, "Prisoner's Dilemma", The Stanford Encyclopedia of Philosophy (Spring 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2009/entries/prisoner-dilemma/>.
[INTS]	RFC 1633 Integrated Services in the Internet Architecture: an Overview R. Braden, D. Clark, S. Shenker
[PCMM]	PacketCable™ Multimedia Specification www.cablelabs.com

ABBREVIATIONS AND ACRONYMS

CCAP	Converged Cable Access Platform
CDN	Content Delivery Network
CMTS	DOCSIS Cable Modem Termination System
COTS	Commercial Off The Shelf
CPE	Customer Premise Equipment
DOCSIS	Data over Cable Service Interface Specification
DRM	Digital Rights Management
DVR	Digital Video Recorder
DWDM	Dense Wave Division Multiplexing
EAS	Emergency Alert System
EQAM	Edge QAM device
Gbps	Gigabit per second
HFC	Hybrid Fiber Coaxial system
HSD	High Speed Data; broadband data service
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
MAC	Media Access Control (layer)
Mbps	Megabit per second
MPEG	Moving Picture Experts Group
MPEG-TS	MPEG Transport Stream
nDVR	network (based) Digital Video Recorder
OTT	Over The Top (video)
PHY	Physical (layer)
PMD	Physical Medium Dependent (layer)
PON	Passive Optical Network
RF	Radio Frequency
STB	Set Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VOD	Video On-Demand
WDM	Wave Division Multiplexing

Just-In-Time Packaging vs. CDN Storage

Yuval Fisher
RGB Networks

Abstract

Operators delivering video-on-demand (VoD) to multiple devices using HTTP streaming must select between two options: store assets in multiple formats to be delivered via a content delivery network (CDN), or utilize an on-the-fly, or just-in-time (JIT), packaging to convert VoD assets into the required client format when it's requested. This paper discusses the benefits of JIT packaging and then proposes a model to evaluate the costs associated with each approach, discussing the parameters associated with various use cases. We also discuss the implications of the cost model for more general edge processing, such as just in time transcoding.

INTRODUCTION

HTTP streaming of video based on protocols defined by Apple, Microsoft and Adobe (see [HLS], [MSS], and [HDS]) has led to the development of a new component in the video delivery chain – the packager (sometimes also called a segmenter or fragmentor). This component creates the segmented video files that are delivered over HTTP to clients that then stitch the segments together to form a contiguous video stream. The packager may be integrated into the encoder/transcoder that creates the digital encoding of the video, but often it is a separate component. Separating the components has various advantages, including the ability to capture the output of the encoder/transcoder as a mezzanine format that can be reused for packaging in both live and off-line scenarios.

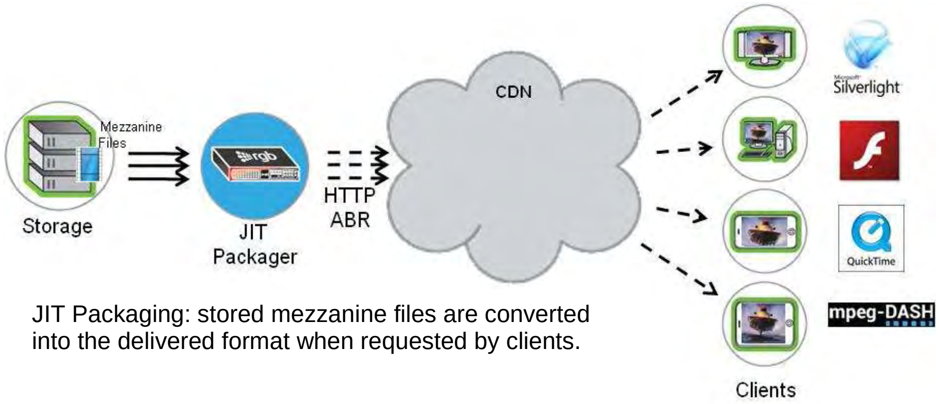
The emerging MPEG DASH (see [DASH]) standard attempts to standardize and unify these protocols under one open specification umbrella; but in the near term, DASH adds more formats that service providers may need to address, since HLS, MSS, and HDS will not disappear immediately, if ever. In fact, DASH has several profiles that have very different underlying delivery formats, so that it may be necessary for packagers to serve not just HLS, MSS and HDS, but an MPEG-2 TS DASH profile and a base media file format DASH profile as well.

In this paper, we focus on one specific use case: just-in-time packaging (JITP), which is applicable for VoD and network digital video recorder (nDVR) applications, including catch-up and restart TV. In all of these applications, each client makes a separate request to view video content (typically from its beginning), so that unlike broadcast video, viewing sessions are independent.

When delivering HTTP streams, two options are possible: either the assets are stored in an HTTP-ready format, so that clients can make HTTP requests for video segments directly from a plain HTTP server. Or, assets can be stored in a canonical (or mezzanine) format which is then converted to HTTP segments as the client makes requests for them – just-in-time. The first option is more disk storage intensive, while the second is more computationally intensive.

Just-in-Time Packaging

In a typical JITP use case, VoD assets are created from live content that is first transcoded into MBR outputs and captured by a "catcher" component that converts the live streams into files stored in a chosen mezzanine format. Alternatively, file assets, rather than live streams, are transcoded into a mezzanine format which uses H.264/AAC for the video/audio codecs and a pre-selected container format. MPEG-2 TS container



JIT Packaging: stored mezzanine files are converted into the delivered format when requested by clients.

format is a natural choice for the mezzanine files, since it can contain much of the signaling present in the original signals in an industry-standard way.

Clients that request a stream from the JIT packager first receive a client-manifest describing the available profiles (bitrates, resolutions, etc). The JIT packager will create the manifest when it is requested the first time; subsequent requests are served from a cached copy. Clients subsequently request specific chunks from the packager which extracts the requested chunks from the mezzanine files and delivers them to the clients. Thus, each client request is served from the JIT packager – the more subscribers that exist, the more JITP capacity is needed.

Selecting a Mezzanine Format

What characteristics should the mezzanine format have? It should:

- be computationally simple to package just-in-time;
- retain metadata in the input streams;
- be a commonly used format with an ecosystem of creation and diagnostic tools.

There are two commonly used mezzanine formats: ISO MPEG file format and MPEG-2 TS files. The first has the advantage that multi-bitrate output can be stored in just one file, as opposed to as many files as profiles, as happens in the MPEG-2 TS case. This makes management of files easier. However, MPEG-2 TS files can provide standardized ways to store many types of commonly used metadata, e.g. SCTE-35 cues for ad insertion points or various forms of closed captions and subtitling, and these are not standardized in the MPEG file case. Moreover, MPEG-2 TS would normally be the format captured in the NDVR use case, and the ecosystem of support tools (e.g. catchers, stream validation tools, stream indexing) is larger in the MPEG-2 TS case. Thus, MPEG-2 TS files make a better mezzanine format than ISO MPEG files in most cases.

WHY USE JITP?

There are a number of reasons why JITP may be a better alternative to pre-positioning assets in all final delivery formats.

Storage Cost Favings

When multiple HTTP streaming formats are used, every asset must be stored in multiple formats, with associated storage costs. This is especially true for network DVR where legal requirements in some regions mandate that separate copies are stored for each customer.

Format future-proofing

The HTTP streaming protocols in use today are still evolving; using JITP of mezzanine-format assets eliminates the need to re-package VoD libraries when these formats change. Changes in formats can be addressed via software updates of the JIT packager, which can then also manage a heterogeneous ecosystem of different format versions (e.g. various flavors of HLS). This is a huge boon to operators who must otherwise decide on a specific version of a format and thus potentially miss features in new format versions or not serve subscribers who haven’t updated their video players.

Single Workflow

Using JITP for VoD with a caching CDN can automatically lead to an efficient distribution of contents in the CDN – that is, the caching of short tail (or commonly viewed) assets in the CDN and the use of JITP for un-cached long tail (rarely viewed) assets. This ensures that new assets automatically migrate into the CDN without requiring a separate offline packaging step in the workflow, as well as a separate, offline determination of which assets are short tail and which are long tail.

Graduated Investment

New VoD service offering using storage rather than JITP would require all assets to be stored in all formats up front, leading to large initial capital expenditure. With JITP, operators can add VoD capacity as the number of subscribers grows with capital expenses that match subscriber growth and revenue.

Unicast Relationship

Because the JIT Packager has a unicast session with the client, it can be used to encrypt VoD sessions uniquely for each client. Moreover, other unicast services, such as targeted ad insertion, can be integrated into the packager. Note that when chunks are encrypted per user, they cannot be cached in the CDN.

COST MODEL

In this section we describe a cost model for comparing storage with JITP. The cost model depends on whether the VoD streams are CDN-cachable or not, as could be the case, for example, if they are encrypted per user. If they are cacheable, the storage in the core used to store the assets, as well as the storage in the tiers of the CDN, can be compared to an equivalent JITP capacity. When the assets are not cacheable, the JITP cost is higher, since both the short and long tail content must be packaged just-in-time.

Cacheable Assets: Storage vs. JITP

A simple cost model (see also [Fisher]) can be created based on a few assumptions. First, we assume that short tail content will be served from the CDN and will not require JITP.

The cost of storing the complete library in multiple formats depends on multiple factors listed in the table below:

	Description	Values
L	Library size (hours)	10K-150K
B	MBR bitrate (Mbps)	10
S	Number of subscribers	100K-10M
P_c	Peak concurrency	5%
P_L	Percentage of long tail requests	10%
C_s	Cost of storage ($/TB)	US $2,000
T	Number of CDN storage tiers	2
F	Number of ABR formats	3

The total cost of storage C_{ts} is then:

$$C_{ts} = C_S \times T \times F \times 3600 \times L \times B \times 10^{-6} \times 1/8$$

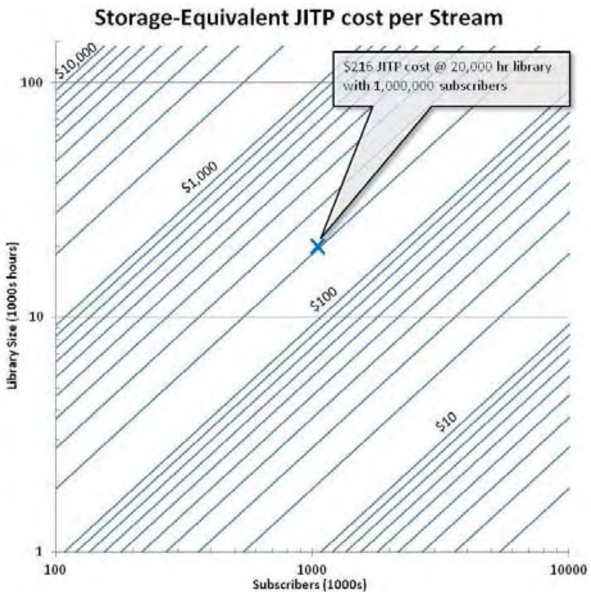
For example, a library of 20,000 hours stored in three formats at the core with two CDN points of presence (or CDN roots or different CDN tiers) would cost $1.08M.

The equivalent cost C_{jitp} of serving a JITP stream rather than using storage is the total storage cost divided by the number of long tail stream requests:

$$C_{jitp} = C_{ts} / (S \times P_c \times P_s)$$

So, in the example above, a million users would have an equivalent JITP cost per stream of $216.

We can look at the parameter space of library sizes and subscriber count to see where JITP provides value. Given that a high-end server can deliver hundreds of simultaneous JITP streams, the graph shows that the range of storage-equivalent JITP cost ranges from low (not even sustaining hardware costs) to very high (where significant savings can be achieved by delivering JITP streams rather than storage). Roughly speaking, the region where JITP leads to cost savings over storage is the upper left triangular half of the graph.



It's worth noting that JITP may incur an additional cost in inbound network traffic, at least when it is centralized. Of course, if JITP is not centralized, then the library must be stored multiple times at the edge, mitigating JITP's value. A complete analysis of every variation is beyond the scope of this paper, but the model described above can be easily modified and used in each situation.

Non-cacheable Assets

When VoD assets are not cacheable, the cost model can still be used by considering 100% of the assets to be long tail. This eliminates the benefit (and cost savings) of caching the short tail in the CDN.

Just-In-Time Transcoding

The cost model does not discuss what type of processing is done in the network – only its cost compared to storage. Since the computational density of transcoding is about two orders of magnitude less than for packaging, the cost graph shows which regions in the subscriber library parameter space are suitable for transcoding as well; this is (roughly) the upper-left triangular portion of the graph that supports processing costs above $1000 per stream.

CONCLUSION

JITP may offer significant cost savings over storage, but its real value may be in other benefits: a simplified workflow, per-subscriber encryption based on unicast delivery, future-proofing against the evolution of formats, and investment and growth in capacity that is commensurate with subscriber growth.

REFERENCES

[Cablevision] 2nd Circuit Court ruling on network DVR http://www.ca2.uscourts.gov/decisions/isysquery/339edb6b-4e83-47b5-8caa-4864e5504e8f/1/doc/07-1480-cv_opn.pdf

[Fisher] Comparing Just-in-Time Packaging with CDN Storage for VoD and nDVR Applications, Proceedings of the Canadien SCTE, March 2012.

[HLS] HTTP Live Streaming, R. Pantos, http://tools.ietf.org/html/draft-pantos-http-live-streaming-06

[MSS] IIS Smooth Streaming Transport Protocol, http://www.iis.net/community/files/media/smoothspecs/%5BMS-SMTH%5D.pdf

[HDS] HTTP Dynamic Streaming on the Adobe Flash Platform, http://www.adobe.com/products/httpdynamicstreaming/pdfs/httpdynamicstreaming_wp_ue.pdf

[DASH] ISO MPEG 23009-1 Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats

Delinearizing Television – An Architectural Look at Bridging MSO Experiences with OTT Experiences

Bhavan Gandhi, Varma Chanderraju, & Jonathan Ruff
Motorola Mobility, Inc.

Abstract

Advent of over the top (OTT) content services by providers such as Netflix, Hulu and Vudu has dramatically altered the media consumption experience and with it the expectations of consumers. OTT services and in some cases cable provider services such as Xfinity TV supplant traditional linear and on-demand offerings. However, despite the availability of all these choices and services, the end-user's media consumption experience is disjoint and detracts from traditional lean-back TV watching. There is an opportunity to build solutions that provide a more cohesive, unified and intuitive user experience for the end-user. This paper describes architectural and system details of a system capable of delivering such an experience.

BACKGROUND

Consumer Experiences Trends

Traditional or linear TV watching, contrary to anecdotal views, is not dead. The ability to access over-the-top (OTT) content has not led to a mass exodus away from television[1]. Looking at absolute numbers, a recent Nielsen study found that people watch an average of 32 hours and 47 minutes per week of traditional TV compared to 27 minutes a week of watching video online[2]. Regardless of absolute numbers, the trend is that people are watching more video than ever before; this includes online, on portable devices, and traditional TV sets as well[3].

It is fair to say that the TV watching experience is evolving; it has been changing from a single-source, single-device experience to one of a multi-source, multi-device experience. In this new world, the incumbent service operator (i.e., MSO) is still the richest single source for linear, broadcast entertainment. Even in this ecosystem, the content delivery and experience infrastructure has been changing to accommodate and support varied devices and content formats.

The increasing number of content sources brings about the more radical experience changes. In some cases, the source is the content provider; other sources include the growing number of internet-based providers of entertainment content such as Netflix and Hulu, who we refer to as OTT operators. The addition of these content sources is fragmenting the user experience and taking it from a lean-back experience to forcing the consumer take an active role in discovering and consuming content from their various sources or subscriptions, while keeping mindful of the devices on which the content is playable. Experience fragmentation is caused by the user having to be cognizant of their subscriptions and the applications from which the content can be discovered and then played. These applications tend to be provider specific, so discovery of content within the provider’s offerings requires being in their application. In the confines of the living room, this experience starts diverging from being lean-back. Even outside of the living room, there is a meaningful need to have a central hub for content discovery and an easy way of consuming the content.

Towards Convergence

The objective of making entertainment content lean back in this changing marketplace is attained primarily by unifying the linear, on-demand, and OTT content discovery process. This should be without regard to the client device that is being used for consuming the content. And, once discovered, content playback should be easy too. Playback capabilities of the specific device should be transparent to the user.

There are a number of client-based applications on various devices that are attempting to achieve unification. Applications such as Fanhattan[4] are trying to unify the content discovery process across OTT content stores such as Netflix, Hulu, Amazon, iTunes, etc., as well as what is generally available linearly. One shortcoming is that linear TV is regional and subscriber dependent; therefore it does not reflect a subscriber's view into available content from their subscription. When it comes to fulfillment, OTT content can be consumed readily if the appropriate clients are supported on the device, however, fulfillment and consumption of broadcast content (through an MSO) is not supported even if the subscriber has a subscription.

Google TV[5] is a client device / application play in the living room to unify the content consumption experience. It is an application platform running Android OS that has the capability of interfacing to the Internet and to the set-top-box. Generally, OTT provider applications (e.g., Netflix) can run on the device to access the repository. There is also an attempt to converge the experience around television, movies, and shows through the Google TV application. This application strives to unify the content discovery from live content and the web; the application allows you to consume the content without regard to the content source (broadcast or internet). If the content is live, the Google TV box tunes the STB to the appropriate channel, and if it is available over the Internet it can be streamed to a client player that supports the appropriate security protocols. The disadvantage of this approach is that the convergence is done at the Google TV client; as such, this experience cannot be replicated across the increasing varying number of devices that are also being used to consume entertainment content.

Server-side (or cloud-based) unification of content discovery and federated content delivery and playback has the potential of bringing to bear the best of the Internet and marrying it with the best of broadcast television. Not only does it unify the content discovery, it also has the flexibility to support applications and experience across a variety of fixed and mobile devices. This is an evolution over unifying content discovery and centralizing content delivery that is espoused by Tranter[6]. We espouse centralizing discovery but federating the fulfillment.

ARCHITECTURAL CONVERGENCE

Architecturally modularizing and separating the content discovery (metadata), control, and data delivery provides flexibility in creating new services. This enables service providers to aggregate content information from multiple sources and to create tools and services that let end-users browse, search, discover, access and control content consumption across their ensemble of devices. This is key to creating a more unified user experience. The end-user is provided a unified content discovery mechanism through an application or guide interface. We have developed a cloud-hosted metadata service that aggregates, normalizes, and correlates content metadata from disparate sources and provides RESTful interfaces to applications. With the prescribed architecture, we allow for secure registration and sign-up to the user's set of subscribed services, whether they are linear broadcast or OTT, and secure access to the user's desired content. One of our primary goals is to achieve a unified discovery experience for the user while simultaneously ensuring that the content distributors or service operators can exercise and enforce their content rights over their media assets.

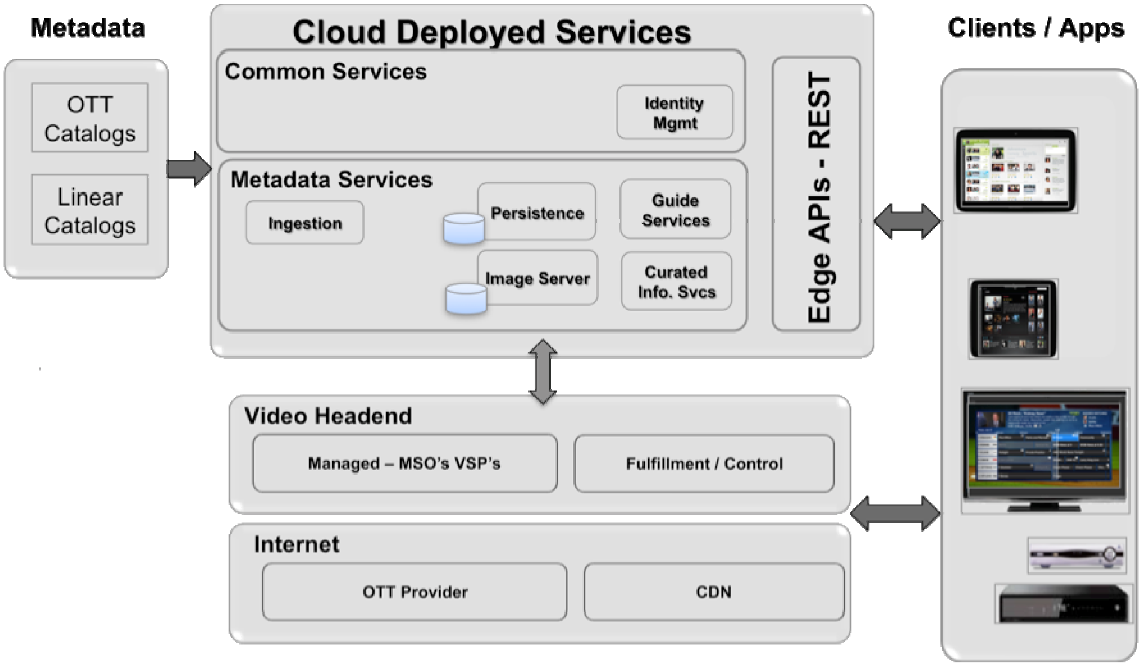


Figure 1. Modular Architecture supporting Converged Experiences

Figure 1 shows a high-level architectural system diagram that disaggregates content discovery from content fulfillment and control. The Cloud Deployed Services is responsible for hosting metadata across the varied sources and providing a unified view for discovery to client-facing applications.
For Internet based OTT content, the URL to play the content may be hosted as part of the Metadata Services or accessed using OTT provided APIs in the client application. OTT content is typically hosted and fulfilled through a content delivery network (CDN).

Taking a similar model with MSO hosted content channels and on-demand content, the client applications need a mechanism to access the content originating from the Video Headend. In our system, we use the Metadata Services as a proxy for passing minimal yet essential control information. Our approach is to publish the URL of the Fulfillment/Control services that the client application can access to discover the video channel (URL). This is modeled after the Internet OTT. By this approach, the data plane and control layers provide flexibility with how the content is delivered and what client devices are supported. The data plane layer is highly flexible and can exercise complete control over video playback on client devices. By minimizing the interdependencies between content discovery, delivery, and ultimately playback, we reduce the likelihood of linear and on-demand information sourced from a provider from getting stale.

In part, the function of an intelligent Fulfillment / Control module is to advertise its location and to enable appropriate content playback. Essentially it acts as a media broker between a user application(s) and supported clients. Once the client application / device accesses the location (URL) of this brokering service and stations that are supported, the client player can then tune to the selected content channels for the supported device type. The inherent assumption is that users are subscribers and the devices are registered with the service operator (and ultimately the Video Headend).

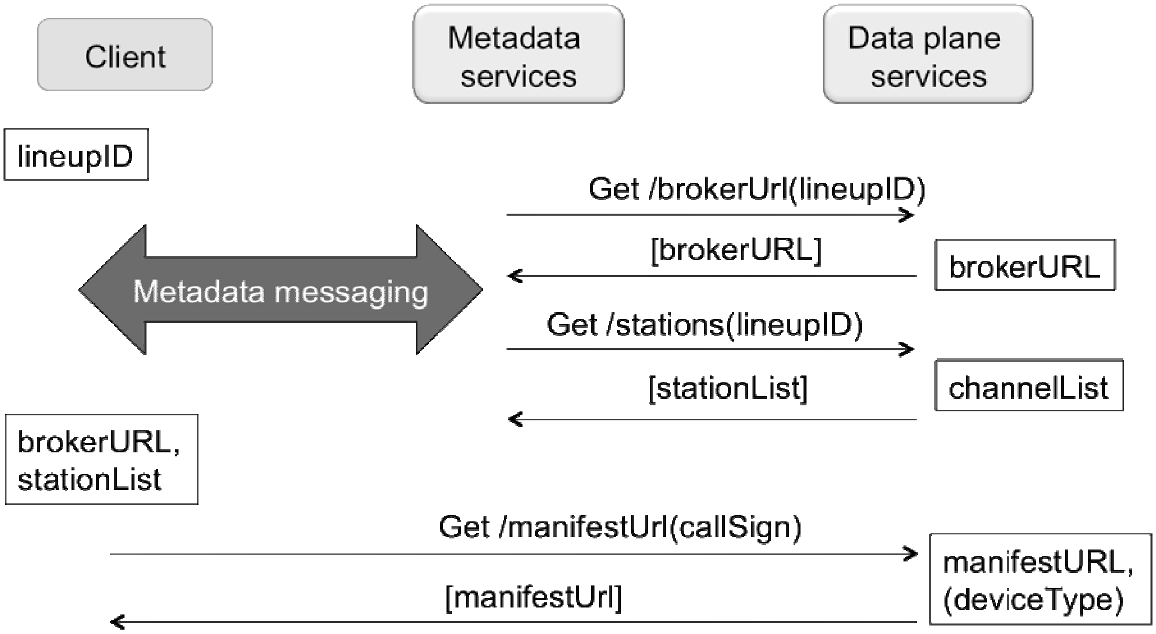


Figure 2. Interaction Diagram between Client, Metadata, and Data Services

Different models can be envisioned depending on whether all the services are hosted and operated by a single operator, or whether there are different operators for the different services.

An example set of interactions is shown in **Figure 2**. The client application first interacts with the Metadata Services to find the address of the appropriate Data plane that is servicing the delivery of the specific lineup (through lineupID). The Metadata services, through interaction with Data Plane services, can find out where the client application should point for getting access to the appropriate linear channels. Given this information (brokerURL, stationList), the application can interact directly with the data plane (Video Headend) to access the linear content (manifestURL) supported on the accessing device. This modularized system provides the advantages of unified content discovery as well as the ability to support an ensemble of devices. Also, since the service providers are ultimately responsible for fulfilling the content to the user / subscriber, they can control access. This allows flexibility in deploying and operating the system.

SOLUTION & DEPLOYMENT ECOSYSTEM

The entire system is built on a proven and robust technology stack that makes use of the latest web services technologies. The system has also been designed to support flexible deployment scenarios.

At a high level the de-linearizing (or converged) television ecosystem is comprised of several subsystems and logical modules that each encapsulate specific functionality. These include:

- Unified Metadata Services subsystem
- Portal & User Interface subsystem
- User Management subsystem
- Device Management subsystem
- Network DVR subsystem that includes scheduling, recording, and archiving
- Fulfillment subsystem
- Dataplane subsystem (Video Headend) that encapsulates transcoders, recorders & streamers

A key element of our solution is the ability to provide a set of unified metadata services to our clients. This allows the clients (internal and external) to rely on a single entity for all content metadata needs irrespective of where the source data is derived from and irrespective of the type of the source metadata (linear, non-linear, broadcast, VOD, OTT etc).

The unified metadata services subsystem is capable of assimilating highly unstructured, inconsistent and incomplete data sourced from dozens of individual metadata sources and turning the data into a consistent, complete and usable set of metadata services that can be consumed by a variety of clients. The unified metadata service exposes feeds and APIs for clients to access linear TV data including scheduling data, lineups, stations and non-linear data such as series, episodes, movies, news programs etc. The unified metadata service subsystem uses a host of techniques including customized data-collection agents (or ingestors) that are continuously tuned to be in sync with ever-changing data publishing formats used by the data-providers that we ingest data from. Examples of data sources include metadata providers such as TMS & Rovi, OTT providers such as Hulu & Netflix. Data-clustering and data-classification techniques are used to normalize and classify related data sourced from multiple metadata providers. The metadata capture and classification has evolved over time and has been enhanced using heuristic learning. In addition a powerful set of editorial tools provide the means and capability to further refine the data through manual oversight. Manual oversight is only required for a small fraction of the data. At the storage layer metadata services use a combination of SQL and NoSQL storage to ingest, classify, store and archive metadata.

Metadata services are deployed in the cloud and expose clean, simple and flexible REST APIs to consumers of the metadata services. JSON is the preferred format for the output of these REST APIs as most clients are browser-based rich-web applications.

Another key element of the solution is the usage of a powerful and flexible framework that is capable of driving the presentation elements on the client (device) side as well as providing an SDK (both on the client side and on the server side) that service operators (or 3rd parties) can use to extend, adapt and customize the look, feel and functionality of the solution. The framework helps abstract different form factors and input paradigms from user-experience developers and assists with ultimately creating uniform experiences across various device types (TVs, Tablets, Smartphones). The choice of client-side technologies (HTML5, JS, CSS) makes it possible to address devices running mobile operating systems such as Android and iOS to STBs running custom Linux images.

The server side components of the solution are deployed in Linux OS, virtualizable, developed on a Java EE platform, conform to the MVC architectural model, make heavy use of the Spring framework and use Hibernate to abstract the storage layer. The server side components are highly modular and communicate with each other mostly using REST APIs. This lends itself to flexible deployment options that can take into account different business and technical requirements that drive the deployment choices of service operators.

CONCLUSION

Being aware that video consumption is evolving and a continued user need for being easily entertained, we have architected and developed a system that allows content from multiple sources to be discovered and consumed on any device. Such a system requires the disaggregation and modularization of the discovery and fulfillment processes. We have used the latest in web technologies to create a flexible eco-system for deploying and operating the system. This allows content to be discovered and consumed from both traditional TV and Internet OTT sources in a unified way.

ACKNOWLEDGEMENTS

Our colleagues Anthony Braskich and Stephen Emeott contributed the specification of the interaction between the Metadata and Data Plane services, including **Figure 2**.

REFERENCES

1. Robertson, A. (Feb. 9, 2012). The Verge. In "*Nielsen: most Americans still pay for traditional TV, but a growing minority go broadband-only*." Retrieved March 22, 2012, from http://www.theverge.com/2012/2/9/2787037/tv-internet-streaming-video-viewing-survey-2012-nielsen

2. Schonfeld, E. (Jan. 8, 2012). Tech Crunch. In*"How People Watch TV Online And Off."* Retrieved March 22, 2012, from http://techcrunch.com/2012/01/08/how-people-watch-tv-online/

3. Indvik, L. (Oct. 20, 2011). Mashable Entertainment. In "*Americans Are Watching More Video Online – and Everywhere Else*." Retrieved March 22, 2012, from http://mashable.com/2011/10/20/nielsen-video-tv-study/

4. (n.d.). Fanhattan. Retrieved March 19, 2012, from http://www.fanhattan.com/

5. (n.d.). Google. Retrieved March 19, 2012, from http://www.google.com/tv

6. Tranter, Steve, 2011. *"Putting the Best of the Web into the Guide,"* SCTE Cable-Tec Expo Proceedings, Atlanta, GA.

BUILDING A WEB SERVICES-BASED CONTROL PLANE FOR NEXT-GENERATION VIDEO EXPERIENCES

Yoav Schreiber, Sunil Mudholkar, Nadav Neufeld
Cisco

Abstract

Next-generation video services require solutions that are aware of real-time data and granular business rules encompassing identity, location, policy, etc., and can make decisions based on that data for a multitude of applications. In conventional video systems, however, this collection of data and business rules resides on disparate elements linked by closed, proprietary connections. This reliance on closed, "siloed" video systems impedes an operator's ability to develop new services and features, or to effectively scale cloud-based video services.

This paper presents a scalable, open-standards approach to orchestrating video services to allow for video control plane extensibility in multi-vendor ecosystems. It describes the architectural foundation for a loosely-coupled, modular video control plane with service provider-grade high availability and scalability. This approach enables video end-points to discover cloud functionality and external systems to expose services and communicate with video endpoints. Drawing on Internet communication methods, the proposed architecture enables a more flexible and scalable video services platform.

INTRODUCTION

An array of market forces is driving demand for new kinds of video services and, ultimately, profound changes to the service provider video systems delivering them. The Cisco Visual Networking Index (VNI) projects that more than 90 percent of consumer IP traffic and two thirds of the world's mobile traffic will be video by 2015.[1] The same study projects 10 billion mobile Internet-connected devices connected by the following year. And, the Cisco Global Cloud Index projects cloud IP traffic to reach 133 exabytes per month by 2015.[2]

These trends point toward a massive shift in consumer viewing habits from traditional, closed TV video systems to an open, cloud-based model. Consumers want the ability to access multiple types of content on multiple types of devices, regardless of the users' location or of the network over which they are connecting (e.g., the managed service provider footprint, an unmanaged Wi-Fi network, a cellular network, etc.). Consumers also seek new kinds of video experiences that integrate conventional video content with cloud services and interactive applications, and extend intuitively across multiple screens.

Service providers are well aware of these industry changes, and many are already moving to expand their video services to new screens and devices beyond the traditional set-top box (STB). However, the traditional service provider video architecture – designed to deliver legacy broadcast and on-demand video content, over a closed network, to a managed STB endpoint – is simply not equipped to support cloud and multi-screen delivery. These next-generation video services demand a level of service orchestration that conventional video platforms do not address. Consider: to deliver a personalized next-generation video experience to a subscriber, the video system must account for:

- Identity
- Device
- Content entitlement

- Location
- Bandwidth availability
- Past user activity
- Social network connectivity
- And much more…

Yet in today's video architectures, this information resides on several disparate, closed systems, including OSS/BSS, content management systems, session resource managers, client software, applications, middleware, etc. In addition, the isolated service "silos" on which conventional service provider control planes rely (i.e., treating managed and unmanaged clients, wired and wireless networks, etc., as entirely separate environments) further impede the service orchestration necessary to deliver a seamless, personalized multi-screen video experience. Conventional video system architectures are also ill-equipped to address the complexity inherent in serving multiple and changing consumer devices connecting to the service, or in optimizing the quality of experience (QoE) based on changing conditions.

Meeting the requirements of modern, cloud-based video delivery will require a new architectural model: a control plane for loosely coupled video systems that is designed specifically to meet the technical requirements of next-generation video services. This paper presents such architecture.

The proposed architecture is a scalable, standards-based approach to orchestrating video services to allow for video control plane extensibility in multi-vendor ecosystems. Drawing on proven Internet communication approaches used by some of the largest web companies in the world, it provides an architectural foundation for a loosely coupled, modular video system with service provider-grade high availability and scalability. Chiefly, this architectural approach:

- Enables video endpoints to discover cloud functionality in a loosely coupled system, and allows external systems to access exposed web services for communication with video endpoints
- Provides a platform to dynamically manage sessions, resources, and workflow in a loosely-coupled system incorporating both managed and unmanaged devices

ARCHITECTURAL ELEMENTS AND KEY CAPABILITIES

The proposed video control plane employs an Internet-based architecture, and as such, represents a significant departure from conventional video systems. Effectively, this approach applies the proven architectural model and design principles used by major web companies like Google and Facebook to contend with massive amounts of data and users, and applies them to video services. This gives service providers a more scalable video services platform, and affords them the same degree of flexibility and speed as web companies when designing, testing, and rolling out new features and applications.

At a high level, the proposed architecture encompasses the following building blocks (Figure 1):

- **Base platform:** The foundation of the video architecture is an open-source operating system, on top of which resides a distributed cache that acts as a shared data store accessible to all loosely coupled elements and workflows in the system.
- **Common messaging infrastructure:** A standards-based, highly-scalable, real-time messaging bus provides the communication framework over which distributed endpoints communicate.

- **Service infrastructure:** The architecture employs a standards-based, hardware-agnostic service infrastructure and workflow engine that enables complex service orchestration with the necessary performance for video services.
- **Session and resource management:** The architecture provides real-time session and resource management capabilities across multiple networks and devices, and is designed to support flexible policy enforcement and dynamic business rules.
- **Applications:** All video applications (i.e., service assurance, device authentication, emergency alert services, etc.), are built upon this platform.
- **Application programming interfaces (APIs) and web services:** the video control plane brokers communication among all elements of the system, including both legacy and cloud-based video applications, via APIs and web services.

Together, these building blocks create a next-generation video control plane that represents a fundamental shift from traditional video systems. Unlike legacy video systems, which rely on tightly coupled client/server communications, the proposed architecture employs a distributed model, similar to that used in web applications. The Internet has solved many of the problems of software resiliency, performance and scale and taking advantage of those is key to building this distributed control plane. In this distributed control plane, functionality and intelligence is allocated to various loosely coupled endpoints (i.e., clients, virtual machines, network elements, etc.), which then advertise, discover, and consume functionality from a shared cache of data.

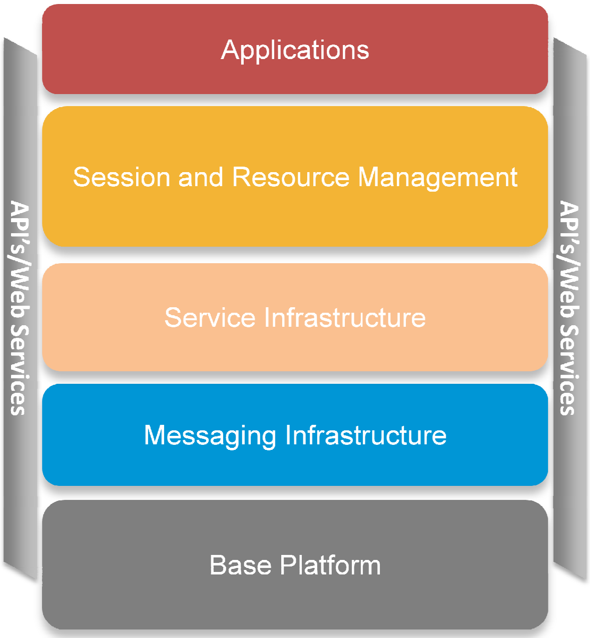


Figure 1. High-level architecture for next-generation video control plane

This data includes all of the essential information that applications need to deliver a video service via the cloud, including presence, state, entitlement, resource availability in the network, etc., all of it updated in real time. The data are stored in a high-speed shared cache, from which they are accessible to distributed applications in real time.

Clients, network elements, and applications can access data stored in the shared cache via a standards-based messaging bus. This common messaging infrastructure connects all elements to the cache via an encrypted, authenticated connection, and facilitates messaging back and forth among the various applications. Cisco has designed the architecture using a widely adopted communication protocol known for its scalability and performance in social presence and instant messaging applications.

Once this core framework is in place – open-source operating system, high-performance data store, and real-time messaging infrastructure – operators can build applications on top of it. These can include everything from straightforward core service functions like device authentication and service assurance, to advanced cloud service offerings such as cloud- or network-based DVR, social TV experiences, and synchronized companion device experiences.

Effectively, the intelligence in the proposed architecture is decentralized – residing in the applications. The proposed video control plane merely acts as a broker for these applications, providing all of the essential information they need to make real-time decisions and deliver cloud-based video services. Together, the shared data cache and common messaging bus functions almost like a web-based news feed: various elements throughout the system publish events or information, and every other element in the system can subscribe to any relevant information. This web services-based communication framework provides inherently more flexibility and scalability than a client/server model, and represents a significant departure from traditional closed video systems, and even some contemporary IP video systems. It should also be noted that the operator need not store every piece of data in the system in a single, centralized real-time cache. Less frequently used data can easily be stored elsewhere in the infrastructure, and remain accessible via the same messaging bus.

Employing this web-based infrastructure for brokering information between applications, the proposed architecture can:

- Orchestrate cloud-based services across multiple devices in real time
- Perform end-to-end session and client/device management for both managed STBs and cloud-connected endpoints
- Provide an interface between multivendor systems and applications
- Achieve service provider-grade scale and availability
- Allow for fully customizable user experience and applications

An Open System, Incorporating Standards and Web Design Approaches

An essential characteristic of the proposed next-generation video control plane is that it is based on open standards to allow for maximum flexibility. As a result, it integrates with legacy systems and with any standards-based third-party technology or application. It is also designed to allow operators to change technology vendors, equipment, systems, etc., as they choose. By avoiding proprietary standards that can age quickly, it provides a more future-ready architecture.

Along the same lines, the proposed architecture is designed to increase service velocity by incorporating web services and design approaches. It is modular and loosely-coupled, allowing for phased introduction of services and technologies. As described later in this paper, the architecture also facilitates service velocity and flexibility through its ability to dynamically manage workflows. The following sections describe this architecture in greater detail.

OPEN, STANDARDS-BASED SOFTWARE PLATFORM

In a legacy QAM-modulated video system, applications and middleware reside on STBs, and video content acquisition and provisioning systems populate back-offices. In this legacy environment, deploying and maintaining software, especially when proprietary, requires a substantial investment of time, skill, and financial resources. As operators transition to cloud distribution based on IP, however, they can take advantage of existing Internet-based standards to achieve greater scalability, and afford greater flexibility and service velocity.

The core technology envisioned in the proposed distributed video control plane has several facets. It is based on an open-source programming language such as Java, it uses web services-based SOA design principles, and it functions as a workflow engine.

Open-Source Technologies

As stated, the proposed architecture is fundamentally an Internet-based approach to video service delivery. As such, it should be based on an open-source or standards-based language such as Java. By using open standards, an operator has its choice of additional open-source libraries and frameworks that simplify the process of building, testing, piloting, and enhancing new services and features.

Java in particular is an excellent fit for a distributed video system. Known for its portability and openness, Java is an apt programming language for an extensible multi-screen video platform with automated workflow, session, content, and other video control plane functionalities. Because Java language code can be represented in the intermediate form Java bytecode, Java can run on a multitude of operating systems that otherwise would require platform-specific machine code. Other reasons for Java's popularity include its efficient memory management and relatively simple object model.

Service-Oriented Architecture

The proposed architecture is designed to operate on a service-oriented architecture (SOA) platform, and should be operable on any standards-based SOA platform. SOA principles allow for loose couplings between clients and servers, and facilitate the development of services independent of the client or underlying platform. As such, SOA design principles help facilitate the systems' ability to share information and functions among multiple distributed applications and system elements in a widespread and flexible manner.

Workflow Capabilities

At the top of the SOA stack is the workflow engine. Users can create workflows by using the workflow engine's graphical editor or by editing XML files directly. Workflows are not hard coded, are not compiled, and do not require any kind of system downtime for modification.

The system supports the real-time execution of multiple workflows, which can be easily extended to add new features rapidly in a controlled manner. These workflows can be characterized as:

- **Atomic:** Once a request begins executing a workflow, it will continue executing that workflow, without interruption.
- **Extensible:** Workflows are defined according to standards and can be modified using standard tools.
- **Flexible:** Workflows are built with a series of nodes that support the basic concepts of sequential operations (IF statement, multi-threading, etc.)

As with the rest of the proposed architecture, the control plane is designed to use a standards-based workflow engine. However, the workflow engine must be fine-tuned for speed to function effectively in a video system. After all, while a lag of a few seconds may be acceptable when initiating a video-on-demand (VOD) session, such a delay in more advanced real-time cloud applications (i.e., pausing or rewinding content in a cloud DVR service) would render the service unusable. Cisco worked to optimize open-source workflow engines to meet these demands, and contributed those gains back to the open-source community.

The communication with external or third-party application procedures also requires a new kind of workflow invocation service. This allows operators to deploy and test features to various overlapping subsets of subscribers, based on criteria such as set-top media access control (MAC) address, account number, service group, or requester IP address range. In the proposed control plane, options for selecting the workflow to execute also include service endpoint and designated market area (DMA) code.

Effectively, this workflow engine provides the tools to create rules, and supports tremendous customization. It also allows for greater flexibility, resiliency, and velocity when rolling out new functions or applications. Traditional systems require operators to shut down the system and come back up when implementing changes, and are unable to execute multiple workflows in parallel. The proposed control plane architecture supports multiple workflows, allowing operators to modify a workflow dynamically, without affecting connections already in use. This enables greater innovation and service velocity by giving operators web-like workflow capabilities, such as A/B testing, where a service provider can use multiple workflows that differ from one another in order to target a specific set of customers, endpoints, or video assets. For example: “Apply workflow (WF) 1 if customer lives in Massachusetts. Apply WF2 if customer is also a high-speed data customer.” The engine can even target individual endpoints, which is useful in beta-testing to familiar customers. Operators can test multiple similar versions of a function or application, and expand or roll them back with relative ease – providing a major boost in their ability to innovate, in less time, at a lower cost. This is typically not possible with a traditional video system.

In the same way, because the video control plane operates via virtualized software instances in a data center rather than tightly coupled hardware systems, the proposed architecture also gives operators greater ability to scale services dynamically. For example, the datacenter can dynamically spin up resources on the East Coast as prime time approaches, and shift those resources to the West Coast as the evening progresses.

VIDEO CONTROL PLANE

The next-generation video control plane is built upon the workflow system described above. It performs three key functions:

- Real-time session management
- Resource management
- Business policy management

Session Management

A next-generation video service must provide a framework for session management across multiple screens, in a variety of video applications. This can include “session-shifting” across devices and networks (i.e., beginning playback on the a TV via the STB, pausing, and then resuming playback later from a smartphone connecting over a cellular network), as well as more advanced applications. One example is a “companion screen” experience that integrates both a managed service (e.g., video content delivered to a managed STB over the service provider’s managed video network) and unmanaged services (e.g., IP data services that complement the STB video service and may be synchronized with it, but are delivered to an unmanaged device such as a tablet or smartphone).

The session management function of the proposed control plane architecture performs the majority of the core functions of a traditional video system back office, but includes support for both QAM and IP environments, and operates according to open, web-services practices.

The traditional approach to video services has been to perform all session processing within a proprietary environment, based on a client/server model with the STB endpoint tightly coupled with back-end servers controlling session and state. In other words, all of the logical software components (resource management, business policy, billing, entitlement, etc.) communicate with each other using closed protocols. The video control plane envisioned here functions differently: The client drives session and state, dictating the format and streaming bit rate required for a specific viewer using a specific device. As discussed, this model is based on the way software works on the web, where diverse applications and devices from multiple vendors share a common delivery language and services framework. By applying this model of session management to video services, operators can gain more flexibility and control.

Self-contained or siloed legacy video back office systems have also had trouble scaling to handle growing volumes of traffic. For a large service provider running a popular VOD service with millions of concurrent users may require dozens of separate session management systems just to handle the load. Additionally, if one part of the system was resource-constrained, the system needed to be expanded as a whole, rather than simply adding an additional node or virtual machine to support that function. Since the distributed control plane envisioned here uses SOA design principles, a shared data cache to store real-time state information, and a messaging infrastructure designed for Internet scale, a single session management system can theoretically scale to serve unlimited clients.

Resource Management

The objective of resource management is to manage video objects in the video distribution system and balance the load among video servers/streamers and networks. To achieve this, the proposed architecture employs resource management tools that intelligently and automatically consider such factors as allocation, video server selection, replication, and cache management to help ensure optimal load balancing among video servers, and to minimize the delay for video requests to be served.

In addition, the resource management function of the proposed architecture relies on the same shared cache as the session management function, and affords the same degree of scalability. The distributed control plane also uses a common resource manager for both legacy and cloud services.

The session and resource management functions of the proposed next-generation video control plane are implemented as logical individual components. Multiple component instances may be deployed throughout the operating environment as virtualized applications in varying degrees.

Flexible Policy Management and Dynamic Business Rules

A traditional TV video distribution system applies a variety of business rules to the delivery of content and services to consumer endpoints. These rules can encompass specific times content can be distributed, specific markets barred from receiving content (for example, blackout rules governing some sports broadcasts), rules barring a device from receiving content without the right content security, etc. For a cloud-based, multi-screen video service, creating explicit rules governing how content can be delivered to every possible IP endpoint in every possible location is simply not practical. The proposed distributed video control plane therefore includes a more flexible business rules engine capable of dictating rules for delivering content through the cloud (for example, allowing streaming of entitled content to any authenticated IP device that supports a particular digital rights management [DRM] system).

The business policy management function of the proposed architecture is powered by a next-generation business rules engine that brings deeper sophistication and intelligence to the process of delivering video services. The business policy management function gives operators a greater level of detail about how the content is going to be consumed. Effectively, it allows operators to store all business rules and policy logic in a centralized script or table, where it can be accessed by other elements in the system. This globally accessible data repository provides the system with vital information on sessions, devices, business rules, etc., and facilitates automated decision-making by the network in applying policy.

This repository is flexible, allowing the operator to define business rules within an XML-based searchable workflow. The workflow can incorporate not just native services associated with the control plane, but allow operators to make “off-board” calls to existing or third-party services. For example, a mobile operator could configure the workflow to make calls to an existing location service, instead of having to recreate that service for the next-generation video control plane. Furthermore, the operator can call entire off-board workflows (not just services), and take advantage of existing third-party business logic instead of having to recreate it. This is yet another example of the value of using an open workflow definition language.

Contrast that with current video systems, which give operators only the most rudimentary knowledge of where the content is being delivered. Essentially, operators know only whether content will be delivered on-net or off-net. With the business policy management tools in the proposed next-generation video control plane, operators can see beyond that, to know exactly what type of device the content is going to and the subscriber consuming it. This allows them to define more finely grained, sophisticated rules for delivering content, giving them an opportunity to generate additional outlet revenue and reduce capital expenditures.

APPLICATIONS AND USE CASES

Once the next-generation video control plane architecture is in place, operators can deploy all applications involved in the video service on top of this framework. This includes essential applications such as device authentication, service assurance, emergency alerts across multiple devices, etc. The proposed architecture is also well-suited to enabling the unique capabilities essential to delivering a cloud-based, multi-screen video service, including the ability to authenticate users and devices among multiple back-end systems (both legacy and IP), and the ability to manage and assure services across multiple devices and networks. The proposed video control plane architecture can also support more advanced applications that take full advantage of cloud capabilities, such as a synchronized companion device experience and a cloud DVR service.

Authentication Among Multiple Back-End Systems

A next-generation video system must communicate with varying types of back-end billing systems, from mainframe to web-based. For cable operators especially, authentication systems are based on a tightly coupled, usually proprietary authentication process between the STB and the back-end billing system, where the STB boots and queries the system for each subscriber’s/STB’s entitlements. These legacy authentication systems will likely remain in place for the foreseeable future, but they present a significant barrier for newer IP media delivery systems and endpoints, which authenticate in very different ways.

Some IP video delivery systems in use today have attempted to bridge this divide. Typically, however, this entails invasive changes to the core of the IP application to support communication with legacy authentication systems. This is to be expected: closed video systems communicate via closed, proprietary mechanisms. Exposing core software elements for authentication (or billing, or other services that require communication with a conventional video back-end system) is typically a significant development project, undertaken at a significant cost. In addition to the potentially onerous costs of this custom integration, this process also impedes an operator’s ability to quickly design and deploy new service offerings that interconnect with legacy billing and authentication systems.

As discussed, the proposed distributed video control plane architecture is an open system. It provides a common infrastructure that can unify legacy authentication and billing systems with newer IP distribution services. In the proposed architecture, this is accomplished via protocol conversion mechanisms that broker this communication and provide the interface to various back-end systems. Rather than incorporating communication with legacy systems into the core of the video control plane, protocol conversion mechanisms deployed at the “edges” of the system communicate with legacy systems, while the core of the distributed services platform remains purely IP-based, and highly scalable. Effectively, this model preserves a kind of stateless, web-aware application core, even as the software communicates with older billing and authentication systems, and frees the video control plane from having to adapt to legacy systems. And, since these protocol conversion mechanisms are open and standards-based, operators need no proprietary intelligence to develop applications to communicate with the IP system.

Service Assurance and Management Across Multiple Device Types

An adaptive bit rate (ABR) streaming capability – the ability to optimize video streams for the specific connecting endpoint based on real-time network conditions – is an essential requirement of a next-generation video system. However, operators cannot rely on a standard, generic ABR functionality. More sophisticated ABR management is required for next-generation services, augmenting basic information about client and network with an additional layer of business rules and priorities, based on a more sophisticated awareness of the network and the subscriber.

Consider a premium cable customer streaming a high-definition program to a TV in her living room via a “smart” TV. Upstairs, her children begin streaming a movie on an iPad. The cable operator would not want an automated ABR function to downgrade the living room TV stream to standard-definition video halfway through the program. The video control plane envisioned here draws on network intelligence to inform ABR decisions, and uses the software control plane to effect quality-of-service (QoS) prioritization and bandwidth reservation in the network. In the scenario described above, the system can draw on operator-defined rules, as well as customizable rules defined by the subscriber, to assure that the large-screen TV in the living room takes priority over a mobile device, and that the primary subscriber takes priority over secondary users.

Synchronized Companion Device Experience

A session-aware control plane for both IP and QAM traffic can be used to deliver complementary viewing services consumed on two devices at the same time, such as viewing a live TV broadcast while using a synchronized application on a tablet. Synchronized companion services can include push applications for polling, alternative content, instant replay and other time-sensitive experiences. A viewer watching “American Idol,” for example, could receive bios of contestants pushed to the companion screen when contestants come on stage, and a real-time voting application to vote for the winner during the show.

The proposed video control plane architecture is an ideal platform for deploying these types of synchronized multi-screen applications. The real-time messaging infrastructure allows the operator to synchronize the web applications, companion screen, and live TV stream to enable these and other real-time interactive multi-screen applications. The same messaging infrastructure can also support social applications, such as the ability to allow a subscriber watching a TV show to identify and chat with friends who are watching the same show. These and other synchronized multi-screen applications can benefit service providers by differentiating their services, enhancing subscriber loyalty, and driving customers to higher subscription tiers – and all are enabled by the proposed video control plane architecture.

Cloud DVR Service

Many operators are now seeking to move content storage back into the network, rather than on DVR appliances in the subscriber’s home. These nDVR or cloud DVR services offer operational advantages over traditional DVR service offerings – most notably, the ability to reduce capital expenditures on costly DVR appliances. For a cloud DVR service to function effectively, however, operators need to manage those cloud-based resources and enforce business logic across unmanaged devices to deliver a seamless transition of experience for the subscriber. A cloud-based DVR service requires a system with extremely high performance that can scale sessions well beyond normal VOD utilization behavior, with hundreds of millions of assets.

The proposed video control plane architecture with its next-generation, high-performance workflow engine meets all of these requirements. In addition, because the architecture employs a web-services based control plane, it simplifies the process of extending the DVR service to other IP devices. These capabilities allow service providers to roll out differentiated, revenue-generating services like cloud DVR itself, but also allow for additional revenue-generating services. For example, operators can invoke a workflow that detects when a subscriber’s DVR storage is nearly full, and pushes a message out to the user’s companion device to ask if the subscriber would like to purchase additional cloud DVR storage space.

CONCLUSION

For many service providers, the shift from traditional QAM-based video architectures to an open, cloud-based distribution model is a five- to 10-year transitional exercise. But the video services landscape is rapidly evolving, and this transition already is well underway. Video service providers are faced with the need to adopt cloud distribution capabilities even as they serve existing customers over legacy infrastructure.

A key to navigating this transition is handling video session, resource, policy, and other functions from a common software control plane. The proposed next-generation video control plane architecture can manage this transition, offering a range of overlapping and mutually reinforceable benefits. These include:

- **Service velocity:** Transitioning to a video control plane based on open programming languages and web services offers a competitive advantage. By providing web services management capabilities for video services, the platform helps enable rapid testing and deployment of new features and applications to a large number of end devices and targeted sets of customers, while allowing operators to continue taking advantage of legacy systems in a graceful manner.
- **Flexibility:** A core value of SOA-based systems, flexibility is exemplified in the proposed control plane architecture and its real-time execution of multiple workflows. Operators have the ability to develop and test all manner of features and applications, in both targeted and large-scale deployments. The ability to dynamically manage multiple workflows in parallel also allows operators to roll out changes and upgrades with no downtime.
- **Scalability:** While legacy systems could expand to handle growing volumes of traffic only with difficulty and cost, the proposed architecture scales simply by adding another node or virtual machine to support any given function. Based on Internet design and communication principles, it provides a platform for video services at massive scale, capable of supporting theoretically unlimited clients.

Ultimately, the proposed video control plane fills a critical gap in the ongoing evolution of service provider networks into a cloud-based delivery framework. Legacy transport technology has many years of life remaining. This platform allows providers to continue supporting those services, while adapting to emergent market realities and using the most efficient cloud-oriented software architecture, techniques, and methods to deliver compelling new subscriber experiences.

REFERENCES

[1] Cisco. (2011). *Cisco Visual Networking Index: Forecast and Methodology, 2010-2015.* Cisco. Retrieved April 1, 2012, from Cisco.com: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

[2] Cisco. (2011). *Cisco Global Cloud Index: Forecast and Methodology, 2010-2015.* Cisco. Retrieved April 1, 2012, from Cisco.com: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

TELEVISION 3:0 - THE MERITS AND TECHNICAL IMPLICATIONS OF CONTROLLED NETWORK AND CLIENT CACHING

Edmond Shapiro, VP Project Delivery Americas
NDS, Ltd.

Abstract

The cable industry has long debated the merits of using general purpose devices to access cached information in the network (commonly referred to as "cloud storage") as opposed to using cached information stored locally (on a device or within a home network), and in what combinations. In the past these trade-offs have involved the location of video on demand (VOD) and digital video recorder (DVR) storage. Today technical design decisions have become even more complex as engineers grapple with the growing number of caching permutations that will facilitate the deployment of Television 3.0, the next generation of IP based advanced digital cable services.

This paper analyzes network design considerations that cable engineers should consider when architecting Television 3.0, the next generation of IPTV applications using an array of cacheable information that includes: Application logic (cached JavaScript), Presentation logic (Remote and Local User Interface], Content (Adaptive Bit Rate (ABR) and Progressive Download (PDL) files as well as Metadata and Network Interfaces.

By highlighting the importance of an abstraction layer herein referred to as the Television 3.0 Common Service Framework, this paper explores hybrid architectures that permit network operators to dynamically cache information at multiple locations within a network – to dynamically adapt deployed services from one type of device to the next and from one region to the next – constantly evolving as new devices and network resources are made available in a rapidly changing technological environment. Smart software design builds an agile, future-proof foundation to increase deployment velocity of advanced services and enhance the operator's brand through improved system performance and better user experiences. Smart software design also avoids the many pitfalls of the past that have afflicted cable operators – from outdated devices and vendor lock-in, to degrading performance and feature bloat, to network-wide equipment upgrades in support of new services.

Specific applications and services highlighted in this paper include:

- *Content protection solutions – Conditional Access System (CAS), Digital Rights management (DRM)*
- *Content Delivery Networks – global, regional and federated*
- *Content recording and playback – DVR scheduling, resource allocation*
- *Ad insertion –graphic and video*
- *Multi-screen service delivery*

BACKGROUND

Television 1.0

The television application is a communication technology for the sharing of moving images with a group of people: the “mass media”. The television transport network is more efficient when it can deliver the same information to more than one person at a time, as was originally the design of the analog terrestrial, cable and satellite networks.

The *Moving Pictures Expert Group* (MPEG) standards enabled the broadcasting of digital television services for the first time, within this paper referred to as the *Television 1.0 service*. As with analog, the Television 1.0 ecosystem was designed to efficiently transport broadcast digital information to a mass of people.

Digital television rapidly increased the amount of content or number of channels that a consumer could view at any moment in time and therefore the Electronic Program Guide (EPG) application was invented to improve content discovery.

With few exceptions (parental controls for one), the EPG user experience remains the same for every viewer. The underlying content changes (Linear and VOD events), but the EPG screens remain constant.

Web 1.0

Though the Internet is defined as the inter-connect of many different private and public IP networks for the purposes of sharing information, in the minds of most consumers, the Internet has become synonymous with the World-Wide Web (WWW) or just Web.

When the Web was invented, the Web user experience was as impersonal as the Television 1.0 experience. The Web was made up of pages written in a HyperText Markup Language (HTML), transferred from one computer to the next using the HyperText Transfer Protocol (HTTP). Each web browser eventually saw the exact same screens (same experience as Television 1.0).

Differences in user experience from one user to the next arose from the distance travelled by the HTML file. If a file crossed too many networks, it might be slowed down or even stopped altogether. The Web doesn’t care whether a file is located on a computer in the same city or on the other side of the world. The response time for every Web request is unpredictable because there are no guarantees of reliable transport between the web server and the clients. User experiences on the web are considered “best effort” because of this unpredictability.

The first implementation of these standards came to be known historically as “Web 1.0”. As the Hyper*Text* name implies, the first HTML files were simply text files carrying textual information. As long as text was the only content being broadcast on the web, the size of the text file made little difference to the user experience. Best effort was simply good enough, even when accessing files across very slow networks.

However, user experience designers quickly grew frustrated at the impersonal Web 1.0 experience and began to deploy “richer” graphical user interfaces including non-text based files, so called “binary files”, such as music, photo and video files, which led to an explosion in the size and number of files managed by web designers (see Figure 1).

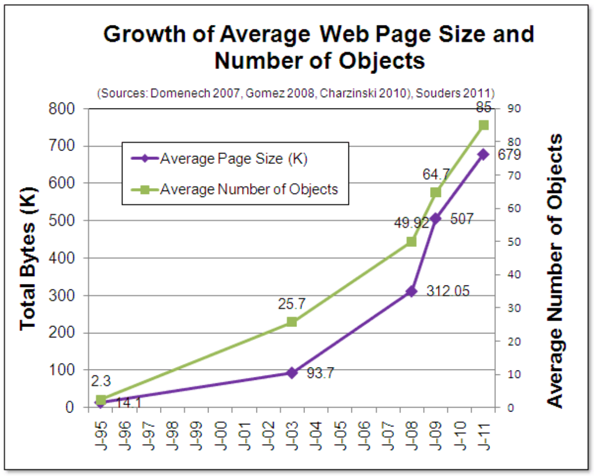


Figure 1: Growth of the Average Web Page

The best effort mode of the Internet was unable to adapt to the demands of the Television 1.0 application developers. Broadcasting (or multicasting) IPTV over the Internet has been notoriously difficult to achieve. Every network between the content distributor and the consumer has to agree to pre-allocate enough bandwidth to carry the fixed bandwidth required by the IPTV service. Only a private network, managed by a single network operator, has ever effectively scaled network capacity to implement this application (e.g. AT&T UVerse).

Web 2.0

User experience designers soon adapted and began to dynamically generate HTML files exposing a richer more personalized user experience for each user and device type. This personalization of web services became known as “Web 2.0”, or the social web, and has been exemplified by the success of web service providers such as Facebook and Twitter.

The successful Web 2.0 service providers learned to work around the “best-effort” design of the Internet by overlaying a virtual network on top of it, called a Content Delivery Network (CDN). The CDN was used to rapidly distribute and store (or cache) the many media files to as many edge networks as possible, as close as possible to the mass of the consumers, thereby avoiding network overload and reducing the number of network hops between the consumer and the media files, and thus reducing the unpredictability of the Web user experience.

CDNs take advantage of specialized algorithms to redirect HTML hyperlink requests to the nearest cache location of the requested media file. These algorithms are constantly and dynamically evaluating network boundaries to avoid bottlenecks and to determine optimal routes.

Over time this CDN virtual network adapted to many different uses as the various types of Internet connected devices exploded. Where once a web application could assume that all web browsers were located on a Personal Computer with a single screen size and a local cache, now mobile devices with smaller screens and no cache had to be accommodated. By delivering different size graphic files and deciding upon remote or local caches, the CDN was able to optimize content delivery to each type of device.

An equally important Web 2.0 development was the widespread adoption of a standardized programming language called JavaScript and the Extensible Markup Language (XML) that enabled web application developers to selectively retrieve parts of the HTML file based on local context or inputs (e.g. user actions, cookies, etc.), as opposed to retrieving the entire HTML file.

Television 2.0

With the growing capacity of CDNs to deliver rich media to Web 2.0 users, demand grew for the delivery of streaming media services such as Radio and Television. As traditional IPTV could not be predictably delivered over the Internet, new CDN friendly techniques were required.

Adaptive Bit Rate (ABR) technology arose from this challenge. Where traditional IPTV content was pushed at a fixed rate, ABR content is pulled at a number of different bitrates that can be influenced by the CDN as well as the local application context.

Instead of broadcasting a common monolithic media file to every consumption device, ABR technology delivers an index file (or manifest file) instead. This abstraction allows for different versions of the media file (size, resolution) to be cached and consumed at any time or place. Where the CDN is able to locate a faster network or when a client connection improves, then the user's viewing experience improves accordingly.

ABR technology has become the foundation for a generation of *Television 2.0* services. These Television 2.0 services have enabled service providers to deliver television to any type of device, and no longer just to televisions. ABR is particularly well suited for unmanaged or minimally managed networks, especially home WiFi networks.

Different ABR standards have competed in the marketplace, being driven by the commercial interests of major CE device manufacturers (e.g. Apple, Microsoft). MPEG recently standardized the *Dynamic Adaptive Streaming over HTTP* or DASH specification which incorporates a number of these approaches.

<u>Cable's Challenge</u>

Cable network operators have watched as popular Over-The-Top (OTT) service providers have taken advantage of ABR technologies to deliver high-quality (even HD quality) services over their fixed and mobile IP networks.

ABR technology is so efficient that it will squeeze nearly every bit of available bandwidth from the access network, limiting alternative services that the cable operator might wish to provide over that same network.

In some countries "net neutrality" regulations restrict cable operators from differentiating or prioritizing any type of broadband service. Cable operators are forced to implement service neutral bandwidth and data caps to control the growth of these OTT services.

The impersonal EPG of legacy Television 1.0 services cannot compare to the personalized and social media experience of many Television 2.0 applications. This is ironic for a cable operator given that the legacy broadcasting Television 1.0 network is generally more efficient at distributing television services.

To compete, cable operators have deployed their own multi-screen Television 2.0 services. However, these same net-neutrality regulations are being tested as subscriber usage limits may be ignored when using the cable operator's own Television 2.0 delivery networks.

From a technical perspective, there is little difference between a private IP sub network and the legacy Quadrature Amplitude Modulation (QAM) as both are based on the same MPEG networks that cable operators used to deliver legacy broadcast and narrowcast television services. All of these services must coexist on the same access network or "last mile".

As ABR encoding (or transcoding) can occur at any control point in the content delivery network, regulators will find it increasingly difficult to make these net neutrality distinctions.

A consumer who chooses to stream ABR content on their own has the option to purchase a CE device, such as Sling Media, to transcode and transmit set-top box content to any other device. A cable operator who delivers the same ABR service from the network will not only save the consumer the purchase price of such a device, but will reduce energy costs for all consumers by centralizing this functionality in the “cloud”.

Television 3.0

OTT services delivered by global CDNs to any network on Earth are ideal for content creators and broadcasters who wish to expand viewing audiences, but only if the access networks are capable of delivering their service.

Broadband service providers who control the access networks will compete on the capabilities of their network infrastructure and will be judged by consumers on the user experience of these OTT services.

Though these various service providers may compete, there is also an incentive for them to cooperate. New *Television 3.0 services* will utilize these same CDN and ABR technologies but in a more network aware fashion.

Where Television 2.0 services utilized CDN overlays and local device context to optimize the user experience, Television 3.0 services will go further, benefiting from a deep and intimate knowledge of the underlying IP networks, and leveraging a *Common Service Framework* to abstract the user experience.

Caching and abstraction techniques led to the advances in Web 2.0 and Television 2.0 content delivery. Extending these techniques with the addition of new communication and collaboration tools will be the foundation for the common service framework of Television 3.0.

For example, publish and subscribe techniques will make it possible for an access network provider to expose to a Television 3.0 service provider the caching resources of the Edge CDN closest to the subscriber. Television 3.0 service providers who design their applications to account for these network variations will inevitably produce a better user experience.

Today most home networks are minimally managed by professionals. This makes the home network the last frontier for a managed IPTV service. Television 3.0 service providers will expand their management of the subscriber’s home networks as well, such that the Edge CDN may very well be within the subscriber’s home.

Television 3.0 services can be extended into the subscriber’s home network only if a home gateway is capable of supporting these services. For instance, a managed gateway supplied by the cable operator could assure the bandwidth needed to supply a 3D service to every 3D capable device within the home network.

Such a technique will make possible other advanced services as well. For instance, managing the home network for the subscriber would enable plug and play video-conferencing, home security and energy management services to co-exist with television on most devices.

A Common Service Framework that is network aware will incorporate all of these advanced services into the service provider’s branded user experience across any device and on any network. Television 3.0 features that are available only in the home network will be disabled on the road. Other features that are only available with a higher subscription tier will be managed in a common fashion across all devices.

The Television 3.0 service benefits from network awareness but should not be dependent upon it. All of the components of the Common Service Framework must be optimized for use within and outside of a managed IP network. For instance, this means that a DRM client application should provide robust enough security to validate a user and their consumption device anywhere in the world, and not just inside of their home or within their cable network.

Finally, as the Internet itself adapts to these new paradigms (e.g. Software Defined Networking), the Television 3.0 service will fully enable consumers everywhere to experience the true "TV Anywhere" service that they desire.

LEVERAGING ADVANCED CACHING TECHNOLOGY

Usage Context

The term "application framework" describes a software structure for developing software applications within a specific operating system or environment. The responsibility of the application framework is to provide "context" to the users of the framework, which are a set of software components called clients or "applications".

In a Television 1.0 service, the application framework that provides a common set of services for accessing the television transport stream is commonly referred to as "Middleware".

Middleware enables a common set of applications, such as the service provider's Electronic Program Guide (EPG), to share set-top box hardware resources, without being tied to a specific set-top box manufacturer. *Digital Video Broadcasting Multimedia Home Platform (*DVB MHP) and CableLabs *OpenCable Application Platform* (OCAP) are two standardized sets of middleware functions or APIs.

The Television 1.0 service has a relatively fixed or static usage context. The users of a middleware framework are expected to reside in a fixed location on a single set-top box attached to a single television. Dynamic events are limited to remote control or front panel user inputs or network control messages that are typically managed by a conditional access function.

The Television 2.0 service has an equally fixed or static usage context as well. OTT services are typically manipulated at the source (in the network) to conform with the requirements of a specific CE (Consumer Electronic) device manufacturer's chosen application framework, typically Microsoft (Windows), Apple (iOS) or Google (Android).

The Television 3.0 service has a much richer usage context, as it is designed to flexibly adapt to a more complex set of environmental variables. A Television 3.0 service may be executing on a fixed device such as a set-top box, or on a mobile device such as a Tablet.

On a fixed device, the Television 3.0 service must adapt to the same usage context as a traditional set-top box, for instance, by supporting a front-panel interface. However, the Television 3.0 service might also support the geospatial feedback that it acquires from the mobile hand-held device.

All Television 3.0 applications whether fixed or otherwise will respond to the same network control messages including subscriber entitlement or service feature changes. The service and content protection function of the Common Service Framework (historically referred to as Digital Rights Management or DRM) must constantly validate the usage context of the media consuming application, ensuring that the content distributor's commercial agreements are respected and that content will not be used for any purposes other than the intended ones.

Within such a complex operating environment, prioritization of usage context becomes critical. For example, in a telephony enabled device, such as a smartphone, the application framework may need to determine for each event whether the television application or the phone application takes precedence. The same Television 3.0 services running on a smart TV, smartphone and tablet will react differently to each type of event, and may in fact be programmed by the user to react differently.

Though service providers have the option of developing a different set of Television 3.0 applications, each optimized for the specific CE manufacturer's application framework and unique usage context, this will inevitably be seen as a costly and infinitely expanding endeavor, as all of these applications will need to be supported and maintained indefinitely.

Alternatively, Television 3.0 service providers will draw upon the experience of Web 2.0 service providers by abstracting their services through the use of the Common Service Framework.

The Television 3.0 service provider will balance the goal of a complete abstraction layer that minimizes device specific development, with the desire to leverage unique device specific capabilities (e.g. larger or higher resolution screens, better memory management, unique man-machine interfaces, hardware security hooks or greater portability).

The HTML5 standard, currently in development, is expected to facilitate the deployment of a common set of applications and services across compatible devices. HTML5 includes richer graphical capabilities and more complex JavaScript application logic. As device specific application frameworks standardize on the JavaScript standard interfaces, Television 3.0 service providers will deploy more features in a common fashion, and thereby reduce their dependency on device specific or downloadable application frameworks.

As an example, HTML5 utilizes JavaScript to abstract the usage context of the local device. Through the use of JavaScript to access the device's local storage, effectively extending the virtual CDN network into the device, Television 3.0 application developers will be able to cache service information that improves the predictability of the user experience, potentially approaching the reliability that consumers experienced in Television 1.0 applications (at least within managed networks).

Built-in DRM functions may already exist in many CE device specific application frameworks. However, in order to achieve a common set of security functions across every device type, the service provider will inevitably require a global set of content and service protection functions to be included within the Common Service Framework. These security functions may leverage device specific security capabilities or usage context, but must never be completely dependent upon them.

For example, to ensure that there exists a trust hierarchy, the Common Service Framework might leverage any hardware based personalization or security functions that are exposed by the device manufacturer (for example Unique Device IDs). Alternatively, DRM clients may be integrated with the device's application framework, such that a DRM application may be downloaded to provide dynamic security hooks that may be leveraged by the trust functions of the Common Service Framework.

Context Control

For a service provider who is designing a Television 3.0 service across fixed and mobile devices using a Common Service Framework, a key decision is whether that service should take advantage of network specific features or whether it should be agnostic to the underlying transport technologies.

If for example a cable operator deploys a network agnostic (Television 2.0) application on a tablet alongside a traditional set-top box application (Television 1.0) over the same access network, then the network resources required to deliver the same quality of service to both devices may end up being twice the resources that would have been required if both devices had implemented a common Television 3.0 application that utilized a network aware Common Service Framework.

Most broadband service providers have already adapted to the demands of Television 2.0 service providers by scaling their access networks to enable ABR to coexist alongside legacy analog and digital cable television services. To avoid the overhead cost of simulcasting ABR video over the same access network that already delivers similar content in a traditional digital video format, the network operator must allow the Television 3.0 Common Service Framework to interface with the legacy network controller systems.

To permit a Television 3.0 service to interoperate with the legacy digital cable plant requires a complex interoperability design that adapts existing Television 1.0 infrastructure to the Common Service Framework. This includes the ability to leverage legacy System Information (SI) and Conditional Access (CA) services already being transported alongside the broadcast digital video service.

The access network operators and service providers must agree on a context control interface to communicate the availability of hybrid or legacy services, and to enable the Television 3.0 service provider to control the access network transcode or transcrypt resources that would be required to convert content to the required consumption format. This includes transferring the service protection metadata of the legacy conditional access system to the Common Service Framework for use by the Television 3.0 applications.

The advantage of deploying a network aware Television 3.0 service is that it can be made to be more scalable by reducing demands on the access network resources. The disadvantage is that interoperability costs may be greater for the network aware Television 3.0 service (see Figure 2).

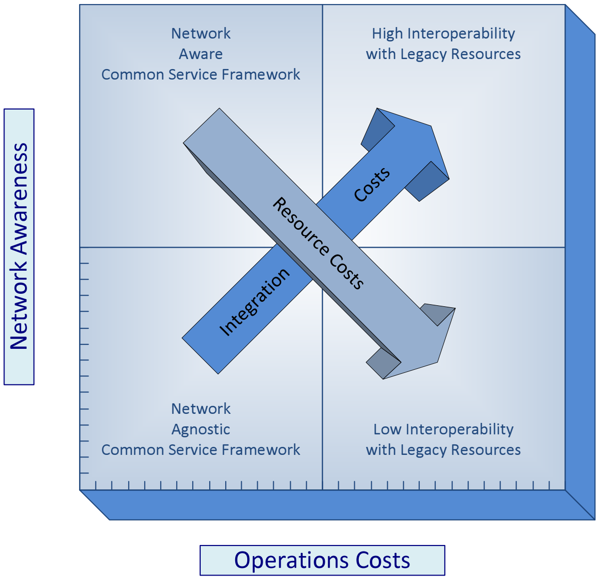


Figure 2: Cost of Network Awareness

Including legacy network awareness into the Television 3.0 Common Service Framework should be built-in from the start as it will become ever more difficult to retrofit fielded applications. Legacy network awareness need not be deployed into the field on day one. These features may exist in the

Common Service Framework, but may be turned off in the first deployments, thereby reducing and/or delaying the legacy interoperability costs.

Inevitably a hybrid gateway device will be required to support Television 3.0 service interoperability with the legacy television services. New Internet TV compatible transport standards such as MPEG DASH may never be feasibly deployed on all existing fielded hardware. Until a service provider forces every subscriber to replace their incompatible legacy equipment, a hybrid device will be useful (and cost-effective) in translating between the legacy transport and the new Internet standard transports.

A hybrid gateway device might be installed within the home or outside of it, but regardless of where it is situated it will serve the same function. The hybrid gateway translates between the legacy service and transport controls, for instance by leveraging MPEG2 transport streams (TS), system information (SI) and conditional access (CA) information to acquire and terminate the legacy broadcast service, and then transcoding, transcrypting and translating these services into the Television 3.0 service exposed to the newer multi-screen applications.

Note: A more detailed discussion of media gateway termination technology is available at: Architecting the Media Gateway for the Cable Home

Context Aware Services

As mentioned above, an application may require that the Television 3.0 service will distinguish between conflicting user priorities based on context.

As an example, when viewing television on a mobile phone, the user may choose to pause their viewing in order to answer the phone call. In a pure streaming model, the application would determine availability of pausing by validating whether a “catchup” version of the program is available for bookmarking and later streaming from the paused location.

But if the same mobile application happens to be situated within the user’s home network then a DVR capable application might already be recording a legacy broadcast version of the program to a local storage device, altogether eliminating the need for additional network “catchup” streaming resources. Further, if the Television 3.0 Common Service Framework were capable (e.g. DRM content controls permitted), then the application might be able to stage the content for offline as well as online viewing, enabling the DVR recorded content to be viewed in a park or on a plane.

If the subscriber exposes local storage in the home for the purposes of viewing television content, then the same Common Service Framework (in a managed network) might use Progressive Download techniques (PDL) to persist as many formats of the content in the home as are required by that home’s devices, avoiding the future need for in-home devices to go back to the network for viewing. The same PDL technique may be used to pre-position personalized advertising content.

Another example of a context-aware service framework is the ability to limit or expose service provider resources based on an application’s user privileges. For instance, limiting the quality of the television content might be dependent on a user’s data quota. The service provider might for instance permit the user to limit their household’s access to HD quality content when a certain threshold of usage is met every month, or for a specific household device or user.

Such authorizations may in fact be federated in the Television 3.0 model. For instance, additional personalized metadata about a specific television event might be available to a subscriber only if they subscribe to multiple service providers. As an example, the Common Service Framework may permit the CDN to be managed by one service provider, but the content discovery may be managed by different ones, only sharing a common content identifier. For instance, subscribers to Common Sense media or Rotten Tomatoes might have additional descriptive information about the current movie that the user is accessing from their cable subscription.

To deploy a Television 3.0 Common Service Framework capable of unlimited TV viewing anywhere, as has been described in this paper, the Common Service Framework must be capable of implementing a very robust contextual control interface over the content as well as the content delivery network, whether connected to a network or offline. The context control interface logic itself must be cacheable along with any associated context control metadata including related content and service information required for discovery, protection and transport of the content, so as to permit offline as well as online viewing. The ABR manifest in DASH may be used to implement such a context control interface. The XML manifest file standardized by DASH may be accessed from a stream or from a cache. The DASH manifest file may be adapted and/or extended on the fly by any intermediate control point or it may be kept in its original pristine form, untouched as any associated content is transported through the content distribution network.

For instance a service provider might choose to regionalize, localize and/or personalize the original broadcast manifest file as well as any associated sidecar files. Sidecar files may be used to extend the original manifest or index file, for instance by describing network specific abstractions (and might be required in cases where the original manifest file is write-protected). Examples of content personalization include frame accurate insertion of an overlay graphic, an alternate video replacement, or any other form of advanced advertisement, as well as the inclusion of a user or a group’s specific bookmarks.

At each point that the manifest file is transferred from one sub network to another over the entire content distribution network, the manifest or associated sidecar files will be subject to controlled manipulation as required by the context specific needs of the service provider or the end user.

As an example, in the case that a service provider is leveraging a local storage device to permit offline and online viewing of their controlled content, the service provider might pre-stage personalized content or metadata to be used in place of the broadcast content or at other pre-defined interstitial points. Such a Television 3.0 application would not only allow for the display of personalized advertisements, but would indeed allow for any type of personalized content – the same movie might be available in a specific subset of the five parental advisory formats for each user in the household.

Through these ABR synchronization techniques, the Television 3.0 application may be able to access new types of contextual metadata, for instance enabling users to skip through episodes in a series – or articles in a video news journal. The user might even personalize their application to prioritize content based on their location (for instance emphasizing movies shot in Paris over British comedies and then vice-versa as location changes).

Additional Television 3.0 contextual services will be made possible by the development of a robust context control interface. For instance, content discovery and recommendation searches may be persisted and prioritized by users to enable automatic organization of future programs (or versions of programs) in a much more personalized fashion.

The Television 3.0 content will adapt to the use of contextual metadata. Television series will include metadata to allow viewers to automatically catch-up to favorite plots. Movie directors will include metadata to allow viewers to experience their stories differently, depending on personal desires (e.g. family-friendly fare, racy endings, and mood-sensitive plot lines).

Consumers most likely will agree to reductions in privacy in return for more personalized Television 3.0 content – that will be delivered along with more personalized advertisements (which may actually be of interest to consumers). Contextual control of playback will assure advertisers that the consumers have actually viewed their information, and will enable consumers more instant gratification (e.g. immediate purchase of the actress's dress).

The concept of DVR scheduled and recorded content will evaporate over time as all content will be available at any time and in any place. Instead of recorded content, users instead will refer to "My Content Library" in order to distinguish between personally interesting content and everything else. DVR schedulers will evolve into personal recommendation and content discovery tools. All content the user ever viewed will be available to them, but only the recent content most likely to be viewed next will be displayed within personal recommendations list.

Network Aware Services

A few years from now, the content delivery networks of today will be considered as outdated as the Web 1.0 applications of a decade ago.

A service aware contextual gateway application might be deployed at every sub-network interface point on the content delivery network. The content delivery network control application itself might be virtualized and contextualized in the same fashion as the television applications described above. Today network switches and routers are fundamentally constrained by the Open Systems Interconnection (OSI) model to a very limited visibility of application needs within a specific OSI level.

Using the same kind of abstraction model described above for television content distribution, IP packet distribution can be equally freed from the constraints of the existing network control models.

New forms of "network aware services" will be enabled to adapt more easily to the physical constraints of the underlying sub-network. An example of this is a residential gateway that dynamically routes consecutive video packets across both a home wireless and home wire-line network (e.g. MoCA, 80211.AA) depending on temporal noise characteristics and error correction on each physical transport medium.

The poster art representing a movie to be displayed in a television application might be dynamically adjusted not only by the device screen size, but by the capacity of the underlying IP network on which it was transferred to the device.

As each network gateway application is empowered to make service aware optimization decisions, the network controller will coordinate and mediate conflicting needs of applications, service providers and access network operators.

CONCLUSION

In a world where every consumer desires to have their subscription services on any device at any time, service providers must learn to live with an endless variety of devices and an infinite number of services, delivered over any type of network, both offline and online. In the short-term, every service provider must be able to deliver their services in managed, unmanaged and hybrid environments equally well.

The rapid ascent of audio-video services delivered with ABR technologies and the rapid adoption of the MPEG DASH standards exemplify this trend. ABR manifest and associated content and metadata files may be cached and manipulated at every point in the content delivery path to assure consumers access to television services whether in the home or on the go.

Use of similar next generation caching techniques will be extended throughout service delivery platforms to assure that every operator service benefits from similar scalability paradigms, including user interfaces and collaborative communication features.

Just as Television 2.0 took advantage of techniques developed for the social web to optimize delivery of television over unmanaged networks, Television 3.0 applications will adapt those techniques to the needs of network operators who require consistent managed and branded television service to be delivered to any subscriber at any time and any place.

Intelligent Caching In An ABR Multi-Format CDN World

Patrick Wright-Riley, Brian Tarbox
Motorola Mobility, Inc.

Abstract

In their infancy, content libraries contained a few thousand pieces of content and most vendors put a copy of everything everywhere. As the contents grew to tens of thousands of titles, Central Libraries were added and Least Recently Used (LRU), then intelligent caching, was employed. As content libraries have grown by orders of magnitude and now adding Adaptive Bit Rate / multi-format copies to the mix, some suggest intelligent caching is no longer possible. Motorola asserts that intelligent caching is both possible and even more critical today than ever. Intelligent caching still plays a valuable role in the ABR Multi-Format world.

INTRODUCTION

In the last ten years the industry has experienced at least three distinct generations of thinking on the approach to placement and duplication of content. We define the first generation as a time when content libraries were small enough that each Video On Demand (VOD) system maintained its own copy of each piece of content. These libraries were stored on spinning media and were served either directly via disc arrays or DRAM. These libraries tended to contain a few thousand titles of standard definition content. Caching in these systems was something that happened in the disk driver or the VOD server's memory backplane.

Generation two can be characterized by the slow introduction of high definition content and libraries of tens of thousands titles. This increase drove the capital expenditure equation high enough to discourage the placement of all content at every site. Thus, the Central Library approach containing the "Gold Copy" along with smaller edge libraries that maintained copies of the commonly viewed content being watched by subscribers within their domain was introduced. Many VOD systems were constructed with the 80/20 rule where it was assumed that 80% of the subscribers viewed the same 20% of content. Given this assumption, distributed edge libraries used a simple Least Recently Used (LRU) caching algorithm to determine which 20% of the content from the edge library was essential to maintain. As it turned out, this content distribution model did not produce the content storage and reduction in network congestion operators expected. This dilemma led to the development of an alternative approach called intelligent caching. Intelligent caching (IC) incorporated additional information about content viewing behaviors beyond what LRU could provide. From there IC became the norm for caching at the edge. However not so far down the road, the explosion in SD and more HD content storage requirements combined with a growing number of smart devices and tablets, Adaptive Bit Rate/Multi-Rate was destined to become part of the picture.

In the third generation, content libraries jumped again to hundreds of thousands of titles, with HD now dominating the content scene. Today this content is now chunked and replicated into several bit rates and wrapped in several formats. Thus, hundreds of thousands of titles can easily become millions or billions of file chunks linked by manifest files.

Conventional wisdom suggested that reaching these levels of processing and file management rendered intelligent caching obsolete. It's suggested LRU caching within the content delivery network (CDN) is both the best that can be done and is enough. Given that intelligent caching increased the efficiency of edge content retention such that 98% of the content was properly retained, it seems reasonable to explore if those benefits can be retained in an ABR, multi-bit rate environment.

PROBLEM DEFINITION

What is Caching and What Drives It?

Caching is a predictive activity. When caching, the system uses data about past viewing behaviors to make assumptions about future viewing behavior associated with a given channel. The assumption typically results in the allocation of a scarce resource before it is actually needed. The "hit rate" of the cache is the percent of time that the assumption turns out to be correct. The impact of the hit rate is based on the comparative cost of permanently reserving the resource against the cost of allocating the resource on-the-fly. To achieve this second, more efficient method, caching is most powerful, especially given that in terms of network usage and related congestion, when the cost of real-time allocation is high.

In order to quantify the value of caching we have to look at the differential cost of resource allocation. Disks are substantially slower than memory and networks are substantially slower than disks. Some VOD vendors made a business out of this differential by attempting to build systems where the entire active portion of the content library lived in memory, or DRAM. This was a successful strategy until the growth of the library outpaced the growth in memory chip size. The battle then moved from DRAM vs. Disk to Disk vs. Library where the relative cost of late allocation was even higher.

Basic Caching

Caching algorithms are characterized by the predictive algorithms employed within the CDN. Caches are assumed to be filled with content at all times. Thus, the critical decision is actually which content item to remove from the edge storage. The most basic type of algorithm supporting this capability is the Least Recently Used or LRU. This algorithm maintains a usage timestamp for entries in the cache and when content removal is required will eject the item with the oldest usage time and replace it with new content. Such systems update the usage timestamp whenever there is a "hit" on the item. These algorithms can be compared to the psychological principal of "Win-Stay, Lose-Shift" where a successful outcome will cause a subject to make the same choice again and an unsuccessful outcome will cause the subject to make a different decision.

This leads to the need to understand the content viewing behavior attempting to be predicted. In the case of content viewership there are two approaches: attempting to predict the future behavior of a given viewer or the future behavior of a group of viewers. By tracking content watched by a particular viewer, inferences can be drawn regarding the potential viewing of that content by the set of all other viewers. However, when using an LRU algorithm to do so, the system simplifies the analysis to a single parameter—time last used—and may not be fully representative of the likelihood of future viewing by a group. Thus, although LRU has some value, it is greatly limited when compared to more intelligent, multi-parameter caching algorithms.

Comparison of LRU with Garbage Collection

Java is the world's most popular computer language and its performance is largely dictated by the behavior of its memory reclamation or garbage collection system. The Java computer language's Garbage Collection (GC) system is one of the world's most studied caching systems. Valuable insights may be gleaned by comparing GC with various other caching algorithms.

One of the primary drawbacks of a simple LRU approach is that it understates or ignores the effect of what GC calls infant mortality of reference. Many objects have a usage model of initial creation followed by limited use, ending with no further activity. In a computer program a variable might be declared, used in a single computation and then discarded. Similarly, in a television viewing experience a user might tune to a channel, watch for a few seconds and then move on. In this scenario the content would have a very high LRU score. In essence, the naïve algorithm employed by LRU would preserve the item in cache in spite of its low actual usage. This confirms the low predictive strength of LRU. This is important when we consider that most CDN "intelligent caching" systems are based on LRU approaches.

To achieve greater predictive power, an algorithm must incorporate a more sophisticated object lifecycle model; an object being a piece of content or a chunk of content carrying specific bite rate and format characteristics. Such a lifecycle model is typically generational. The GC partitions the cache into three generations: 1) Eden, 2) Tenured, and 3) Permanent. When objects are first created they live in Eden. The system periodically scans the memory list looking for items to eject. Items in Eden that are not ejected after two passes, meaning they still have active references to them, are promoted to Tenured. Items living in Tenure that survive more passes are promoted to Permanent and thus remain much longer in cache. There are actually two types of GC passes: full and partial. Partial collections are run quite frequently, have relatively little impact on system throughput and do not examine the Permanent cache. Full collections on the other hand are comparatively rare, can often affect system throughput and do look at the Permanent cache. So, content that exists in the Permanent cache are only occasionally examined for ejection.

Segmented LRU cache uses a similar (though limited) system. There are two LRU lists. Items initially live on the first list and after a second "hit" get promoted to the second list. While this is certainly better than a simple LRU, there is still a world of difference between noticing a second hit and true intelligent caching.

Advantage of Intelligent Caching

Intelligent Caching is a term we reserve for systems incorporating a more sophisticated object usage model. Such a model must acknowledge the realities of content viewing such as channel surfing, free content preview, time of day and day of week viewership patterns and other patterns of apparent viewership that may or may not represent true viewing of content.

The bottom line is that content hits, initial or passive, are not predictive or representative of actual viewership until the aggregate viewing time has exceeded a certain quantum of time. Once the aggregated viewing time of the content has passed a threshold (which may be dynamic and involve multiple analytic parameters) then statistical inferences may be made about the future likelihood of additional views. This is the basis for intelligent caching algorithms and where their value lies above LRU algorithms.

Factors That MAY Diminish Predictability

If in a multi-bit rate, multi-format world where content is delivered over a CDN, one could argue that caching, in its entirety, is unnecessary. There are factors that are commonly cited as evidence that caching is no longer possible or valuable. These may or may not eliminate the usefulness of all caching algorithms, but they certainly provide a challenge to the usefulness of some caching techniques. It is helpful to remember that the caching algorithm attempts to extrapolate from the past exposure or access of content the future possibility of that content being viewed again, possibly by another viewer. The problem in understanding and valuing cache is that "the same content" may now exist in multiple copies, in different formats and bit rates, with chunks spread across multiple edge streaming servers. (See section Content Affinity for more details about chunk distribution).

Multiple Formats

As we enter into the second half of 2012 the video format battle is raging on. Apple's HLS format appears to be dominant, but the Microsoft and Adobe formats are still contenders. Although it is unclear what position these latter companies are taking with respect to future support, they cannot be discounted. At the same time the DASH specification is evolving and may, over time, acquire significant share. While many hope to support fewer than four formats, that time is not yet here (and may never arrive).

There are at least two ways to address the question of how multiple formats effect cache predictability.

One way is to ask the question, "Does the fact that a piece of content was viewed in a particular format "enough" provide any evidence that it will be viewed again in the future…in that same format and/or in other formats?" Since the multi-format ABR world is so new it's hard to anticipate future usage patterns. To the extent that we can extrapolate from existing usage patterns it seems safe to assert that content reaching a threshold of use is in general more likely to receive future plays than content that has not reached the threshold. It is also reasonable, though untested, that reaching the threshold on a particular piece of content in one format is at least weak evidence for the future popularity of that content under a different format. To state the opposite one would have to assert that popularity in one format provided zero evidence of possible future popularity under another format which is unreasonable.

A second approach to this problem is to think about multistage packaging and common formats. As has been discussed in other papers, there is an ongoing debate of the merits of packaging in various locations. Some lobby the benefits of Central Packaging. Others point out the potential benefits of customization from Edge Packaging. An interesting hybrid approach is to perform an initial round of chunking and manifest creation in the center, followed by a real-time component that transwraps content and performs unique, targeted manifest generation. From a caching point of view this approach defers the combinatorics of multiple formats until well downstream of the CDN. A cache element located "upstream" of this real-time transwrapper might see just a single format, thus diminishing its value.

Multiple Bit Rates

The key to adaptive bit rate streaming is the availability of multiple representations of each piece of content. This can be seen from at least two points of view. On one hand the same content might well be viewed at a different bit rate on a phone, a tablet and a big-screen LCD based simply on the capability of the various devices. In this slice of the world each stream might have a different bit rate, but does not necessarily change its bit rate during the presentation. In the other slice, each client responds to the ever-changing load on the network by asking for smaller content when the network is slow and larger content when the network is fast. This is the grand assumption behind most ABR streaming.

The problem is that it invokes the dilemma of the commons: when there is a shared and limited resource, the greater good is often different from the individual good. When the network is congested, every viewer will fully support the idea that everyone else should limit their bandwidth such that "I" can continue streaming the highest quality experience. And everyone else feels the same way. This can be controlled if the client software is controlled by the infrastructure providers in that their client software can enforce the self-limiting behavior. On the other hand, does anyone doubt that clients will be made available that attempt to game the system to consume more than "their fair share" when the network is congested? We assert that it remains to be seen just how many distinct bit rates are actually active for a given content. So, while at first blush ABR might multiply the number of different copies of each piece of content by a factor of six to ten per format, the actual number may be significantly lower, perhaps three or four per format.

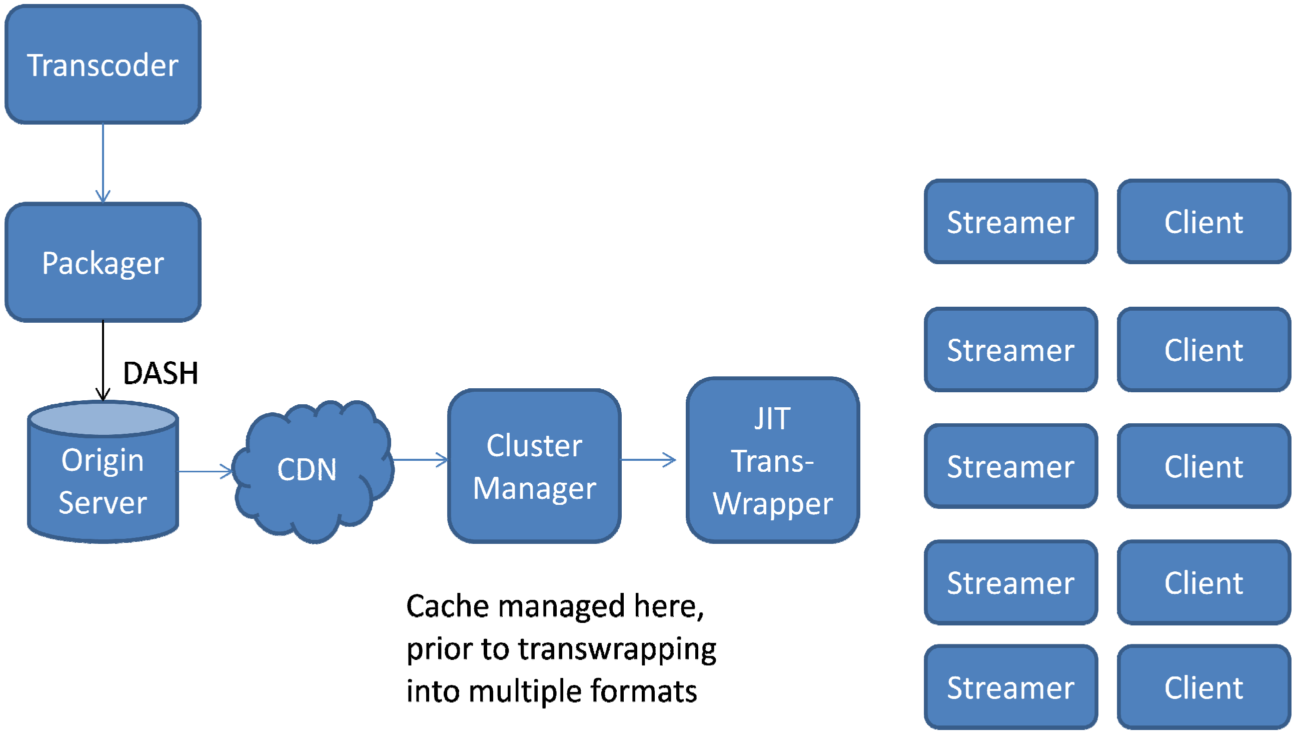


Figure 1: Possible Caching Prior to Transwrapping

<u>NDVR – Unique Copy</u>

Unique copy basically eliminates the ability to do caching at all. For those unfamiliar with the concept, a legal ruling has declared that if some number of viewers record the same content, the NDVR system must store a unique and distinct copy of that content for each of those viewers. In the systems, operators are explicitly forbidden to store a single copy and manage viewer access to that copy. So the only opportunity for re-use of segments or manifests would be if an individual user watched a recorded show multiple times—probably not sufficient to take advantage of caching.

While this might be seen as ending any discussion of caching, keep in mind that Unique Copy presents problems for many aspects of the system. It is anticipated that some vendors may push the envelope of mixed common copy / unique copy systems, especially outside North America. In this scenario, caching may have a larger role to play.

Personalization

Personalization is the process of converting a general video stream into one tailored for a particular viewer or group of viewers. Two main categories exist here; ad insertion and blackout (both are discussed in more detail in the paper "*Complexity Considerations for Centralized Packaging vs. Remote Packaging*" being presented at this conference.) In each case a stream that logically could be used to satisfy many stream requests is turned into one that is usable for a subset of those requests. To the degree that this personalization happens upstream of the caching system it will naturally render the caching system useless.

Factors That May Enhance Predictability

While many types of systems suffer from added scale, caching algorithms actually tend to work better in larger environment, if simply because there is more data to use for decision making and there is more content to provide a better opportunity to employ caching to enhance performance. There will undoubtedly be many different sized deployments of video systems, now and in the future. CDN-enabled, multi-format, multi-bit rate systems will be overwhelmingly biased towards the larger of these deployments; the cost of the complexity associated with such CDN systems precludes them from the smaller tier two and tier three deployments.

This then leads to the next important question which is, "Where does the caching engine live in the CDN architecture?" If it lives on the edge server, then it is limited to the total number of streams supported by that server. Many edge servers are relatively small devices supporting only a few thousand streams. The chances of getting meaningful hit rates in such a small environment are correspondingly low. On the other hand, if the caching engine lives near the edge, but in the CDN it might well be able to see dozens or hundreds of the edge servers. This scale changes everything. The chances of getting several play requests for a given content out of several hundred thousand streams is quite reasonable.

The Role of Content Affinity in CDN Caching

Most diagrams of ABR streaming show the client talking directly to an edge Packager or the CDN; the role of any edge server is not discussed. Motorola believes this is a mistake and causes large opportunities for caching via the use of Content Affinity to possibly be overlooked. If the diagrams do show an edge streamer they tend to show only a single one. In almost all cases any reasonably sized deployment will involve dozens or hundreds of edge streamers since each such device typically only supports a few thousand streams at a time.

Technical papers that have included a multiplicity of edge streamers have tended to view them as interchangeable, even on the per stream basis. It has been asserted that each chunk request from a client might be serviced from a different edge streamer, assuming that every edge streamer has the same chunks. This is then described as a resilient stateless design that can trivially survive the loss of one or more edge streamers. Some of that is true, but at a cost. The cost is that by making server selection stateless we remove the possibility of using knowledge from previous states to improve our caching.

Content Affinity is the process whereby all streams for the same content are directed to the same edge streamer. This can result in enormous savings in both disk space and network bandwidth utilization. If all streams for Spiderman, as an example, go to the same streamer, there is a far greater opportunity for fragment re-use than if the streams for Spiderman are distributed randomly to several dozen streamers.

If we accept the gains that can be realized from Content Affinity then we must look to see which deployment models give the best chance of using Affinity to our advantage. Figure 2 shows one such configuration.

The client makes its initial request to a Cluster Manager (CM) which is a control plane application that maintains the knowledge of which edge streamer has which content. The CM selects a streamer and issues an HTTP redirect message to that device. The client re-issues the request to the streamer which either services it directly if possible or defers to the Edge Packager to create the manifest, if necessary.

Note that Content Affinity is a separate concept from caching and the CM contains no storage of manifests or content chunks. The CM simply directs streaming requests in such a way as to increase the likelihood that the target Edge Streamer will already contain the required chunks for a stream.

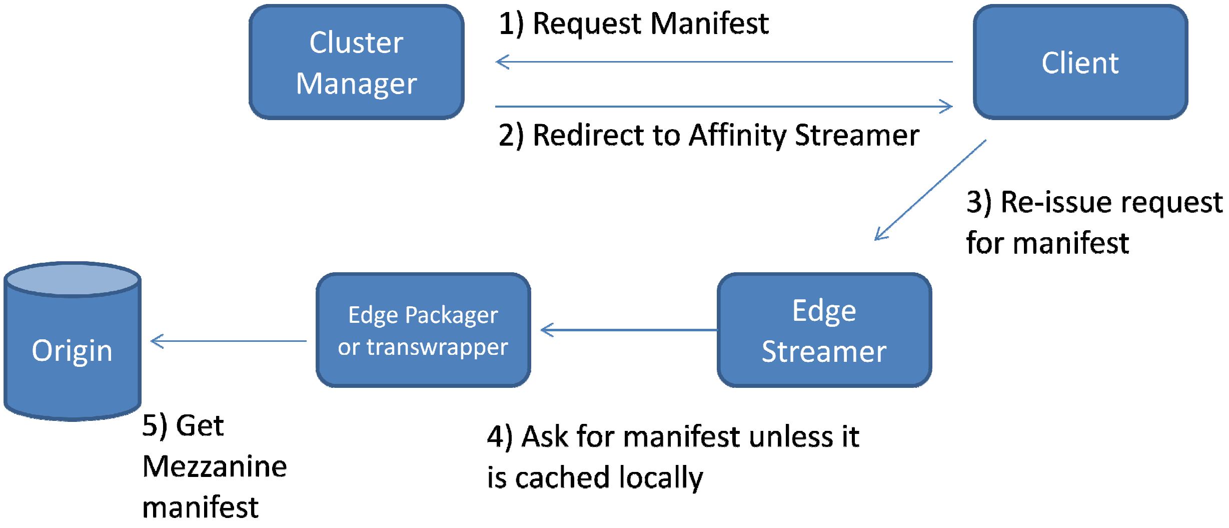

Cluster Manager
1) Request Manifest
Client
2) Redirect to Affinity Streamer
3) Re-issue request for manifest
Origin
Edge Packager or transwrapper
Edge Streamer
5) Get Mezzanine manifest
4) Ask for manifest unless it is cached locally

Figure 2: Content Affinity Deployment

Comparing Intelligent Caching with LRU Caching in a CDN

LRU-based caching in a CDN uses no intelligence about the content, its placement, or its usage. The algorithm simply notices which chunks were the least recently used and discards them when it determines that it needs to create space for new chunks. There is a single ordered list of the chunks logically maintained at the edge of the CDN. This single list covers all chunks sent to all edge streamers. It also makes no use of the fact that chunks may actually be related, i.e., being part of larger piece of content. This can be a benefit as well as a drawback.

If a viewer is channel surfing and briefly visits 20 different channels for 5 seconds each, then the system will likely generate the highest time-last-used values for those chunks and so they will remain stored in cache over other content that should be kept instead. A more intelligent system would never have promoted those chunks as they are clearly of transitory usage. On the positive side, since the system views each chunk individually it would not use those minor play times to promote later chunks from the same pieces of content.

An intelligent caching system would tend to treat such channel surfing as below the threshold for promotion within the cache and would thus not eject other, more popular content.

To put it another way, consider the case where three clients sampled a piece of content but ultimately were watching a different content and a fourth client sampled and then watched the content the others sampled. Putting aside the bit rate and format questions for a moment, we should objectively conclude that the program being watched by three viewers was more popular than the other content and should bias any limited resources such as caching towards the more popular program. The CDN/LRU-based cache cannot do that as it uses a strictly time-last-used algorithm rather than a hit counting-based algorithm. The Content Affinity-based system, on the other hand, allows for the direction of the common content to a common edge streamer and the one-off content to a different edge streamer. This automatically increases the locality of usage of each piece of content to a given pump and thus increases the hit rate of the particular pump's cache.

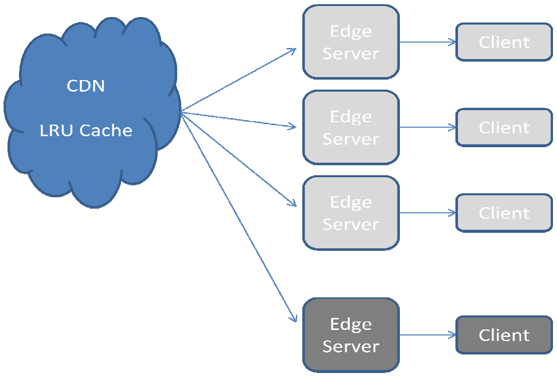


Figure 3: LRU-based Caching of Popular/UnPopular Content

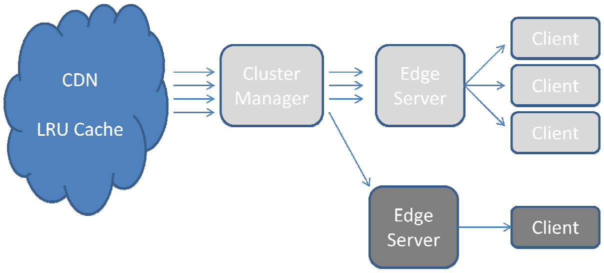


Figure 4: Affinity-based Caching of Popular/UnPopular Content

CONCLUSION

Historically, Intelligent Caching has been shown to provide significant reductions in the need for potentially expensive content storage. This benefit should not be discounted lightly. We have described several of the challenges facing intelligent caching in a multi-format ABR streaming environment. Some of these challenges such as the legal requirement for unique copy NDVR may prove insurmountable. We have, however, shown several opportunities that may allow the use of intelligent caching in other domains to have significant benefits over LRU caching. In particular, we have shown that the affects of Content Affinity can be profoundly and positively affected by efficient, intelligent caching algorithms.

COMPLEXITY CONSIDERATIONS FOR CENTRALIZED PACKAGING VS. REMOTE PACKAGING

Brian Tarbox
Motorola Mobility

Robert Mack
Motorola Mobility

Abstract

Adaptive streaming protocols will be a critical component for operators offering IP video services. One of the key functions in Adaptive streaming is a "packaging" function that creates playlists/manifests, segments the video into chunks, and "wraps" the chunks to make them suitable for one of several protocols. There is an ongoing debate as to the merits of where to perform adaptive stream packaging within a service provider's content delivery network (CDN). Various analyses have considered centralized, distributed, and edge packaging architectures. These analyses primarily considered the CDN bandwidth and storage savings that could be attributed to distributed/edge packaging architectures versus the operational complexity that would likely result. In addition, these evaluations focused more on video on demand (VOD) rather than linear content distribution.

For many Service Providers the ability to centralize all transcoding and packaging operations is appealing, particularly if they own the CDN and are therefore less concerned with the per-bit content distribution costs. For other Service Providers, particularly those that want to augment their existing service with streaming capabilities, but are sensitive to these costs and the costs associated with standing up large centralized video processing centers, the ability to customize content at the edge may make more sense. So, for example, a Tier 2 or Tier 3 operator may want to augment his offerings out of an existing Regional Headend.

In addition, edge packaging may offer options that can reduce the complexity associated with providing certain desired system functions. For example, considering regional ad insertion and blackout, edge packagers can incorporate simple functions that emulate similar legacy system capabilities which minimize the impact to a service provider's network and operations.

Finally, this paper will also explore some of the unique functional capabilities that packagers can offer in support of centralized or regionalized architectures, including intelligent access network capacity management, playlist obfuscation for ad insertion, regionalized blackout, and support for both legacy and advanced advertising in adaptive environments. This will enable operators to fully understand the trade-offs of implementing various packaging architectures and make the right choices when rolling out IP video services.

INTRODUCTION

Previous papers examining where to perform packaging focused on network bandwidth and capital expenditures as the variables to measure. This paper examines other factors that must be explored in order to enable sound decision making in systems architecture design. These factors focus on the processing complexity required to output segments and manifests under various expected conditions. These complexity considerations may make the support for certain desired capabilities problematic. In particular, this paper examines the ability of the various packager configurations to support regionalized and targeted ad-insertion as well as to support blackout processing.

BACKGROUND

Packager Locations Within the Distribution Network

Until recently most discussions about packager locations listed center and edge as the options. In 2012, however, the concept of a hybrid center/edge packager has gained traction. Each will be described briefly.

In the center packager, the packaging function configuration is embedded within the transcoder, or the packager is connected to the output of the transcoder which, in turn, is connected to the origin server. All video stream processing, including fragmentation, ad-insertion, manifest creation, etc. is performed prior to any client request for the content. All fragments and manifests that might be required or requested are deposited onto the origin server.

In the edge packager configuration, the transcoder outputs its fragments (e.g., MP4 or FMP4) and optionally a mezzanine manifest file onto the origin server. One or more packagers are located “south” of the CDN. The client connects directly to the packager or to an optional edge server which redirects to the packager. The packager then requests fragments and the optional mezzanine manifest file from the origin server, transwrapping on-the-fly into the appropriate format for the requesting clients (e.g., HLS, HDS, HSS).

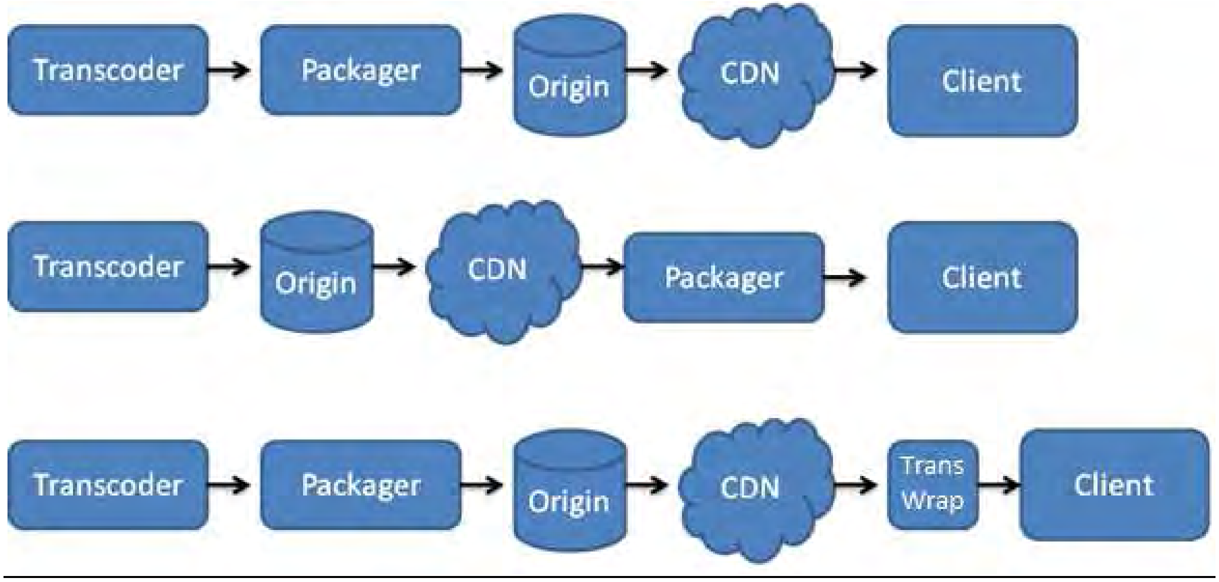


Figure 1: Three Styles of Packaging

The hybrid solution is similar to the center configuration except that it adds a transwrapper component located between the client and the CDN. This is a relatively new style of configuration and so many options are possible. The “north side” packager might transwrap to one of the four standard formats (HLS, HDS, IIS, DASH), leaving the transwrapper to re-wrap only if a given request was for a different format. The transwrapper component might also perform session or region specific operations such as ad-insertion or blackout.

Regionalization Perspective

Some service providers may want to augment their existing home cable/data service with streaming capabilities for in-home IP devices. They may not being trying to stand up an OTT type of service to off-net subscribers, rather they want to offer current subscribers the ability to use their portable devices or want to deploy IP STBs with streaming capabilities. Furthermore, they plan to perform multi-rate transcoding within their existing Central Headends and will deliver the multi-rate transcoded transport streams or mezzanie files to the Regional Headends via their IP distribution networks or potentially via CDN. Just as in their legacy systems, regional customization takes place at the Regional Headends. This is where an Edge Packager could be used to perform regional ad insertion or even blackout processing.

Tier 2 and 3 service providers may also wish to offer streaming services and are sensitive to CDN content distribution costs. Similarly, they may wish to receive multirate transcoded content directly from content providers or resellers and need to perform regional customization (e.g. regional ad insertion).

FUNCTIONAL CAPABILITIES SUPPORTED BY PACKAGERS

Any discussion of ad insertion must occur in the context that viewers do not, in general, want to watch ads. They will go to substantial lengths to avoid ads, with smartphone and tablet users running clients that have been specially designed to defeat ad presentation systems. If one builds a manifest that looks like the following, you can bet that someone will find a way to avoid watching fragment three.

```
 <fragment time=1, length=10,
uri=LOTR.mp4>
 <fragment time=11, length=20,
uri=LOTR.mp4>
 <fragment time=21, length=80,
uri=http:/www.ad-decision-system.com>
 <fragment time=81, length=10,
uri=LOTR.mp4>
```

Figure 2: Easy to Defeat Manifest

Center Packaging Ad Insertion

Central packaging implies that all of the work to create fragments and manifests is completed prior to any request for content. The transcoder outputs to the packager which outputs to the origin server and that “transaction” is complete. The transaction from the client to the origin server (through one or more CDNs) is a completely separate transaction. It may not be apparent, but this is true regardless if the content is VOD, linear, or network digital video recorder (nDVR). In both VOD and nDVR there is a gap between the recording/packaging of content and its eventual playback. Even in the linear case, however, the packager is essentially filling a bucket (the origin server) while the client is emptying that bucket. This becomes more clear if the linear case is expanded to consider StartOver TV. StartOver is basically linear with a limited ability to jump back to the start of a program. In adaptive bit rate (ABR) linear TV, there is what amounts to a 30-second jitter buffer to deal with packaging and manifest creation; StartOver just expands that buffer to 30 minutes.

This implies that all manifest information for regional or targeted advertisements must be built prior to content request. This can be accomplished in two ways. First, one could create multiple manifests, one for each ad region. The session manager then needs to point the client to the region-specific manifest. It does not appear that this approach is feasible for targeted advertizing. The second approach is to build a single manifest such that the ad decision manager (ADM) or some other component is invoked at playback time.

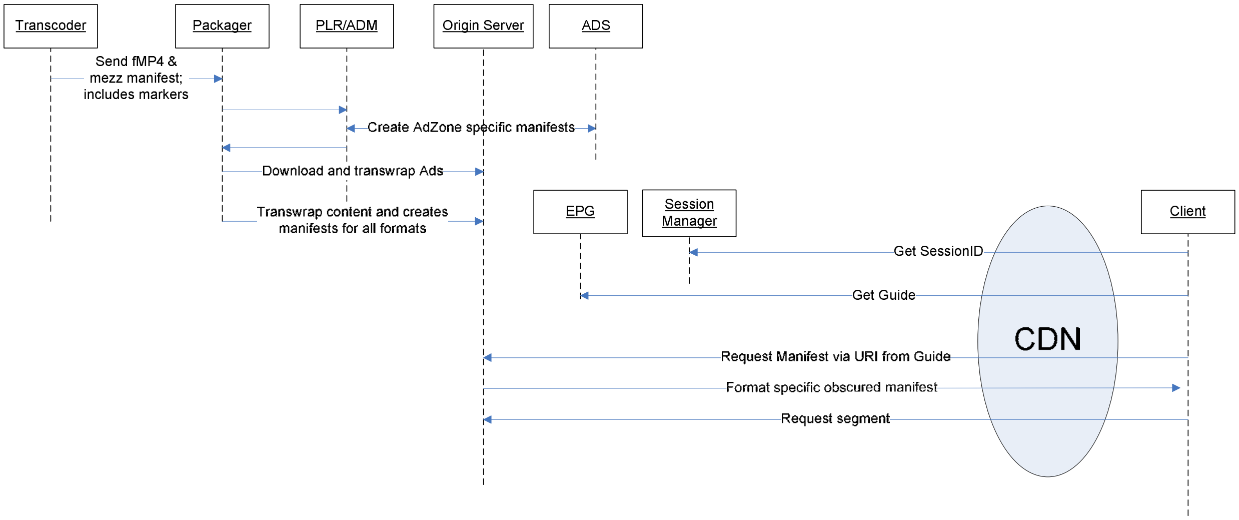

Transcoder
Packager
PLR/ADM
Origin Server
ADS
Send fMP4 & mezz manifest; includes markers
Create AdZone specific manifests
Download and transwrap Ads
Transwrap content and creates manifests for all formats
EPG
Session Manager
Client
Get SessionID
Get Guide
CDN
Request Manifest via URI from Guide
Format specific obscured manifest
Request segment

Figure 3: Center Packager for Ad Insertion

For the second approach to work, the manifest entry corresponding to an ad (or set of ad segments) must be a uniform resource identifier (URI) that can be processed by the origin server or by the ADM itself. One could imagine generating a URI pointing to the origin server that was encoded in such a way that the origin server could detect it, perform some processing on it, and issue an HTTP redirect to the ADM with some localization parameters supplied. Alternately, the URI could point to the ADM directly with some guarantee that the parameters required for localization would be directly supplied by the client. In this last case the URI cannot effectively be obscured. In addition, the client may be able to spoof the parameter(s) used to direct the system towards the targeted ads.

Edge Packaging Ad Insertion

Edge packaging defers the creation of manifests and fragments until the content is requested. This means that a session ID (targeted ads) and/or region ID (regionalized ads) are known at packaging time. Given this information, the packager can use its Play List Rebuilder (PLR) function to query the ADS during manifest creation. It can also cache those results. Since it does not have to pre-build the manifests, it can build regionalized manifests only as needed. This is a savings since it is unlikely that every program that is recorded will, in fact, be requested from all possible ad regions. Granted, manifest creation is not expensive, but managing an explosion of files that might never be used can add complexity to the overall solution.

Since manifests are only created as needed, it becomes possible to create targeted or per-user/per-session manifests. Depending on the ADM being used and the campaign that is in force, it may be desirable to create viewing experiences tailored for a single user at a particular time. While such a manifest cannot be cached since it is a onetime use artifact, at least it can be created. In the standard central packaging configuration, targeted manifests are simply not possible. It should also be observed that while the manifest for a targeted viewing cannot reasonably be cached, the content and ad fragments associated with that viewing may well be cachable.

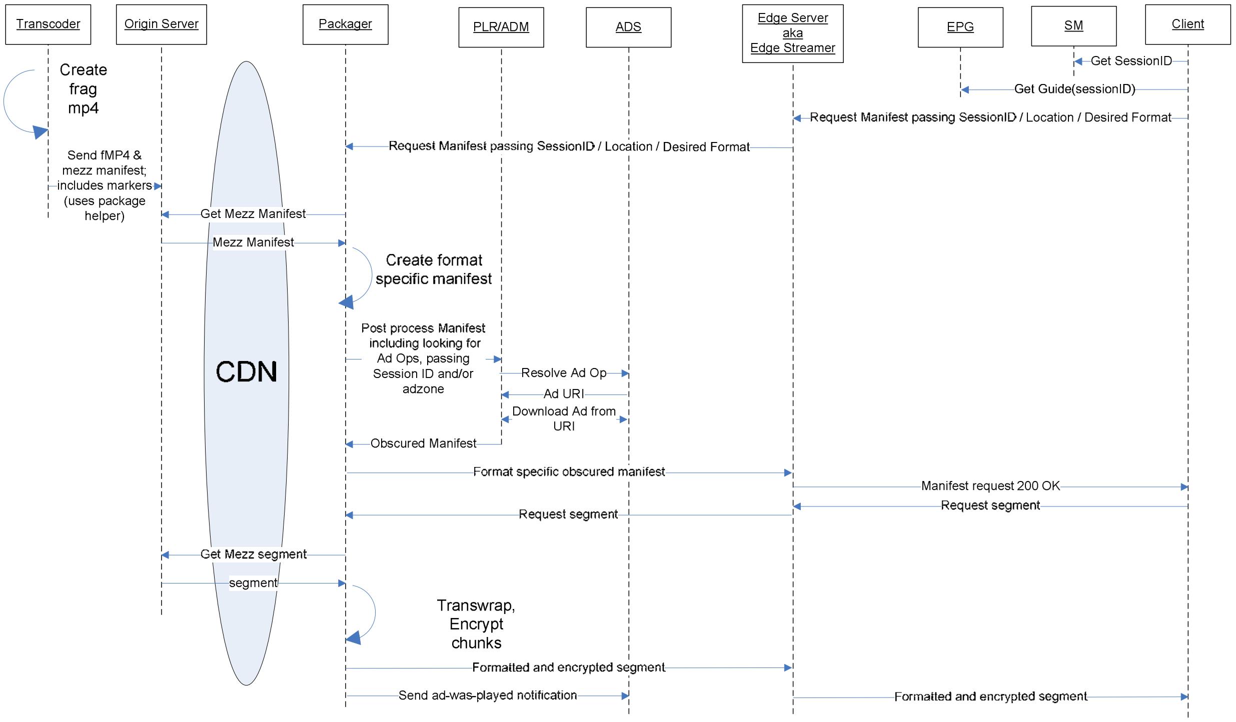

Transcoder
Origin Server
Packager
PLR/ADM
ADS
Edge Server aka Edge Streamer
EPG
SM
Client
Create frag mp4
Send fMP4 & mezz manifest; includes markers (uses package helper)
CDN
Get SessionID
Get Guide(sessionID)
Request Manifest passing SessionID / Location / Desired Format
Request Manifest passing SessionID / Location / Desired Format
Get Mezz Manifest
Mezz Manifest
Create format specific manifest
Post process Manifest including looking for Ad Ops, passing Session ID and/or adzone
Resolve Ad Op
Ad URI
Download Ad from URI
Obscured Manifest
Format specific obscured manifest
Manifest request 200 OK
Request segment
Request segment
Get Mezz segment
segment
Transwrap, Encrypt chunks
Formatted and encrypted segment
Send ad-was-played notification
Formatted and encrypted segment

Figure 4: Edge Packaging Ad Insertion

Viewership Management

Another significant problem with Central Packaging is related to fulfillment or viewership management. As seen in Figures 3 and 5, there is no obvious way to indicate when an ad is actually played by the client. Keep in mind that the origin server is acting as a simple web server and the distinction between content fragments and ad fragments has been obscured. This appears to leave us unable to inform the ADS when the fragments corresponding to a particular ad were ever actually viewed by a client. This would seem to imply the need to put some level of intelligence in the origin server, yet the proper level of that intelligence is elusive. If a pattern to the manifest entries is declared such that the fulfillment observer could determine which fragments were actually ads, one would have to assume that the clients could detect the pattern.

Alternately one could imagine creating a back channel from the origin server back to the packager providing fragment-requested information for all fragments. The packager presumably knows which fragments are ads and could be the component to send the fulfillment message to the ADS That begins to blur the lines of function design for the packager and also makes assumptions about packager knowledge.

To avoid this one could add a component to the control plane path such that whenever a fragment was fetched by a client this new component could detect when an ad fragment was requested and inform the Ad Management Service. This control plane component would have to deal with the fact that an ad fragment might be cached by the CDN and not actually fetched from the Origin Server.

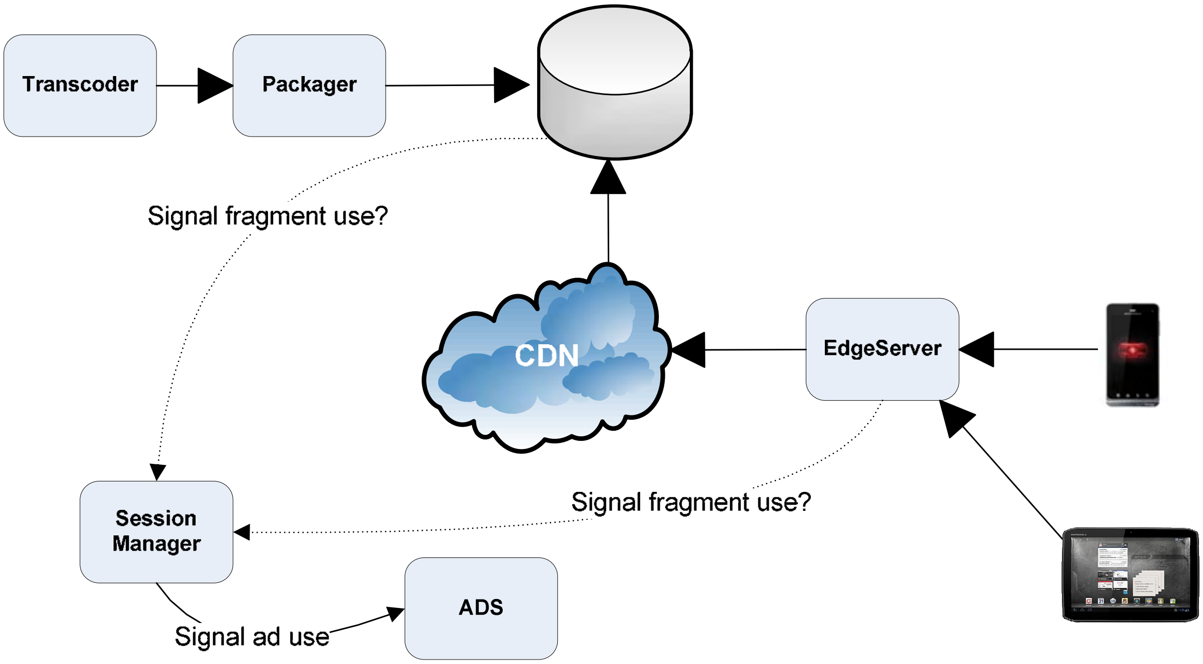


Figure 5: Center Packager Viewership Monitoring

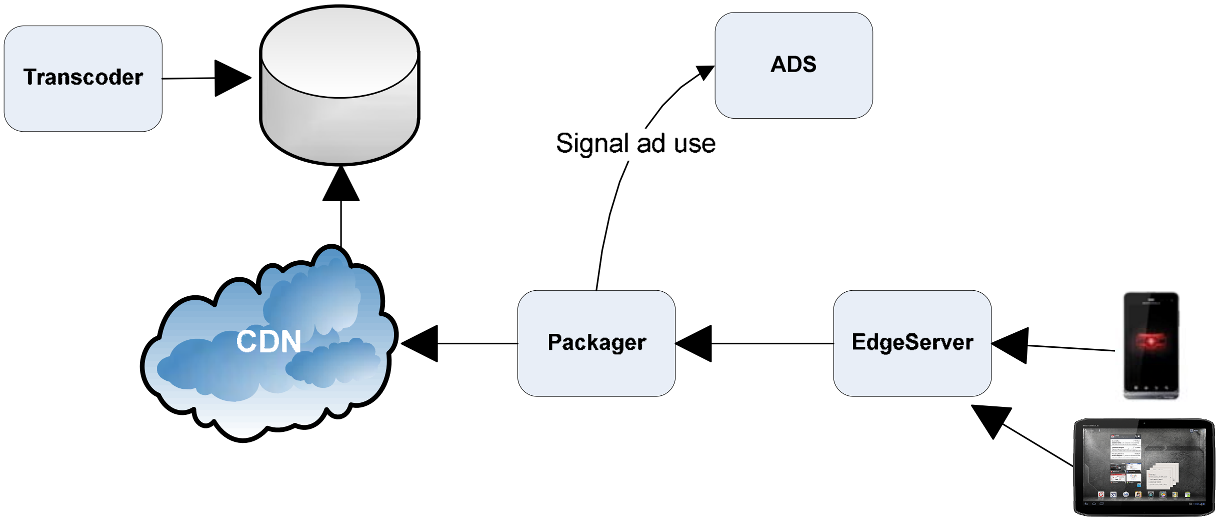


Figure 6: Edge Packaging Viewership Monitoring

Such a centralized control plane component would also seem to become a bottleneck and/or single point of failure as it would appear to need to be involved in every fragment request from every client.

Yet another approach would be to split the packager function into separate transwrapping and manifest generation components. One could then position the manifest generation component logically between the client and the origin server. This hybrid option will be discussed in more detail later

Each of these problem becomes simpler in the Edge Packaging deployment shown in Figure 6. As can be seen from the diagram the packager is involved in all fragment requests and can directly inform the ADS about fragment downloads. This is especially convenient since the Packager component is already well positioned to have knowledge of which fragments are and are not advertisements.

Ad Zone Dynamism

Late binding of manifest creation also allows for dynamism in the set of ad zones. Ad zones are only applied to manifest creation at session creation time. While this may not be a large difference for linear or nDVR applications, it can be for VOD content. VOD contents are ingested once and then may exist in the system for weeks or months. Some classic content might stay in the library for the life of the system. It is certainly imaginable that an operator might want to change the configuration of their ad zones during this time window. It's not clear that there is a way to modify ad zones in a center packaging configuration.

Time Based Ad Selection

Edge packaging also provides more options for ad selection. For example, suppose the system wants to generate a targeted ad based on the playback time of the content. An example would be a show-teaser ad suggesting that the user watch the "next" program. Except in the case of linear content, "next" will mean something different at playback time than it did at record time. If a user records a show on Monday at 7:00 PM they may receive teaser ads for the 8:00 PM Monday show on the same channel. Playing that content back on Tuesday results in a teaser ad that is, at best, useless and, at worst, defeats the viewers quality of experience. Receiving promotions for a show that you missed and cannot, in fact, watch could frustrate viewers. Therefore, the ad decision that will be reflected in the manifest should be made at playback time, not at record time.

Hybrid Packaging

A hybrid approach to packaging should be explored to round out the possibilities. As alluded to earlier, this approach involves using a central packager to perform content and ad fragment chunking. All such fragments are loaded onto the origin server, along with one or more undifferentiated manifest files.

Between the origin server and the client is another component, the transwrapper. This component or software service may be co-located with the origin server or may be a deployed on separate hardware. The intention is that the URI supplied to the client for obtaining the manifest should resolve to the transwrapper component.

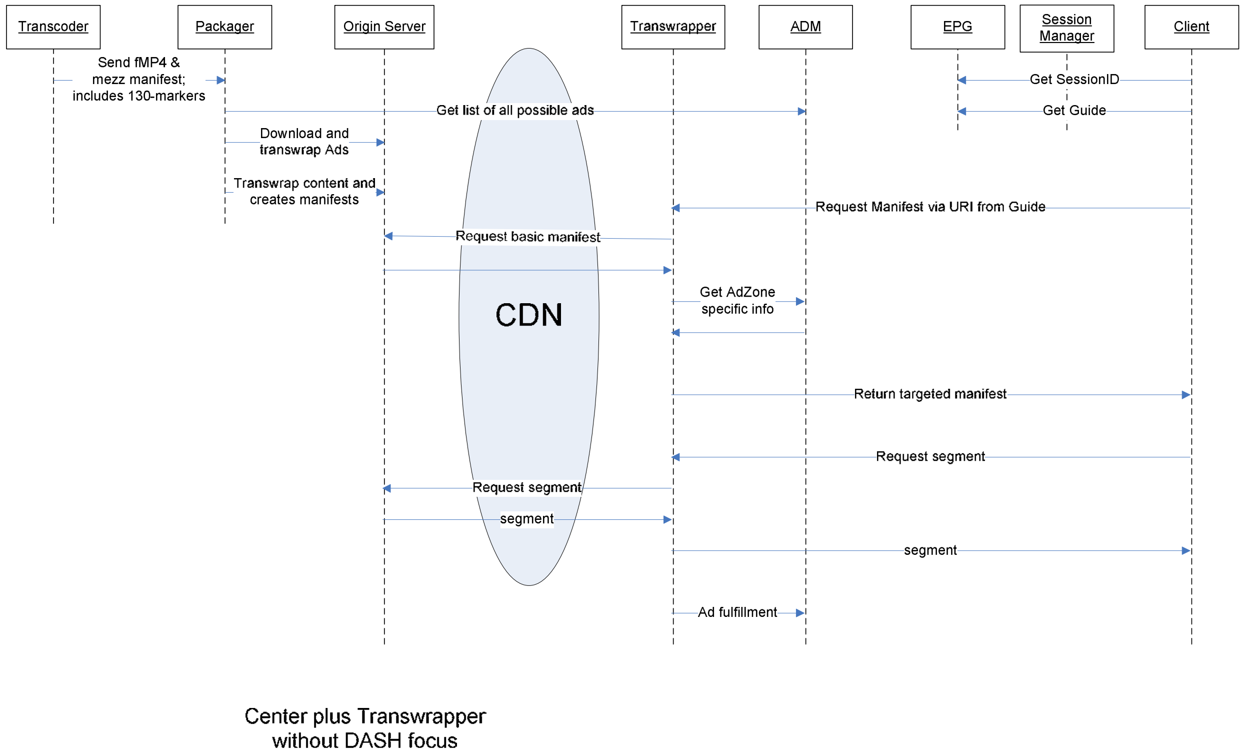


Figure 7: Hybrid Packaging

The transwrapper uses information in the client request to a) transwrap as needed to the client's requested format, and b) perform ad personalization. To do this, the client must supply a regionID for regionalized ads or a sessionID for personalized ads or equivalent data that can be resolved into an appropriate ad zone. Based on that derived ad zone, the transwrapper assembles a manifest tailored to clients within that ad zone.

At first glance, this seems like a reasonable compromise that achieves many goals. From another point of view, however, the system now has both center and edge packaging components. In other words, it's not clear that Hybrid packaging has any advantage over Edge Packaging.

CENTRAL & REGIONAL BLACKOUT APPROACHES

It is anticipated that blackout control will be a required function in multi-screen environments, wherein IP set-top boxes, in-home portable devices, and portable devices outside the home will have to be restricted from receiving content based upon their location (or the subscriber's home location) during a service substitution event. Today, blackout is enforced through the content provider's uplink control system. Normally, a retune command is inserted at the uplink and targeted to individual integrated receiver/decoders (IRDs) known to be operating within a specific region. When a specific IRD observes a retune command addressed to it, it mutes the video stream or replaces it with an alternate service for the duration of the blackout. Similar functionality can be provided in multi-screen systems by manipulating playlist/manifest files during a blackout event.

In a centralized architecture, where packaging and manifest creation is performed, a new manifest or sequence of manifest files needs to be created during the blackout event for clients within the affected region. The new sequence of manifest files will direct those affected clients to tune to alternative content for the duration of the blackout event. Through an element known as a Blackout Manager, unique regional manifest files are generated for all the blackout regions under its control. The Blackout Manager must have knowledge of the CDN topology and specifically the mapping of each edge cache to the specific geographic region it services. It is required to continuously monitor for blackout events by processing IRD retune messages for its regions. When a blackout is in effect for a given region, it requests manifest updates from the playlist rebuilding function, which are subsequently published to the CDN. The updated manifests reference new URLs that point to alternate content during a blackout for one or more affected regions.

Figure 8 illustrates the system. The blackout manager requests that the playlist rebuilder generate unique manifests for each of the three regions under its control, namely, Pittsburgh, Philadelphia, and State College. The manifest (M) is retrieved from the packager and the SportsNetwork.ServiceProvider.net/pitt, SportsNetwork.ServiceProvider.net/philly, and SportsNetwork.ServiceProvider.net/StateCollege manifests are created and published to the CDN. The content identified within the different regional manifests can be identical until such time as a blackout event is required to be enforced. At that time, the manifest file for that area is modified by replacing the blacked out content URLs with URLs for the content to be substituted.

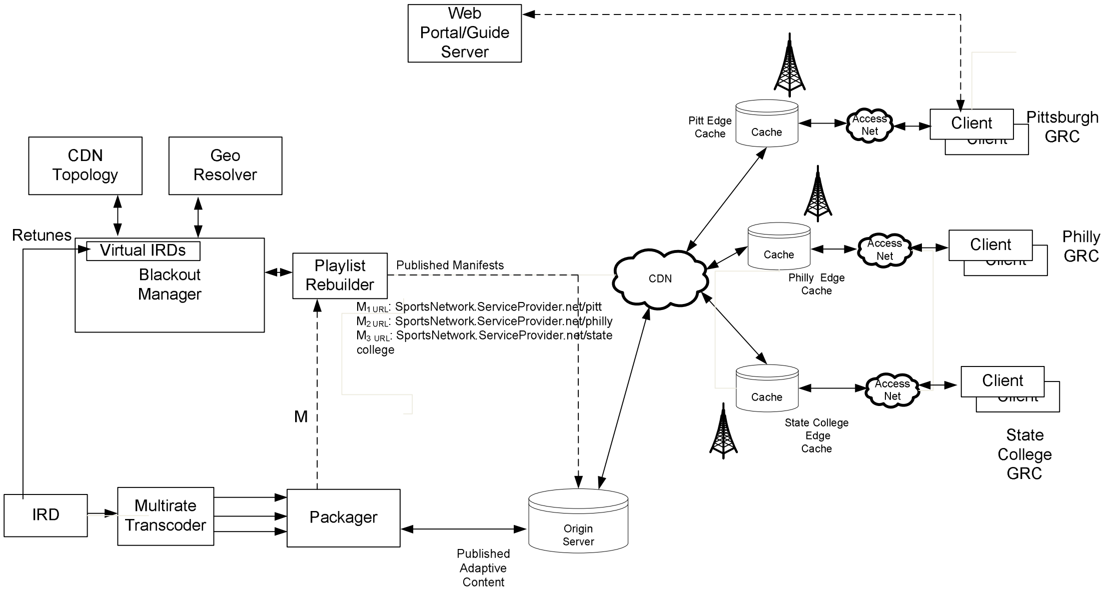


Figure 8: Centralized Blackout Management

A particular client within a given blackout region can retrieve the right manifest for that area through a number of techniques listed here:

1) Client GeoLocation: Client geolocates itself using embedded GPS technology or geo position services available on the network. The client reports its location to an upstream control plane element—for example, a session manager—which, in turn, identifies the appropriate URL for the manifest associated with that region. Alternatively, the client can construct an HTTP request that includes location metadata, which results in the return of a location-specific manifest.

2) Control Plane GeoLocation: The client is geolocated by control plane elements within the network. For example, a session manager that the client communicates with could use a geoLocation service to resolve the client's source IP to a location within the network. The session manager, in turn, identifies the appropriate URL for the manifest associated with that region to the client.

3) Edge Network GeoLocation: Edge network elements append location metadata into a client's HTTP request relying on the network's knowledge of where HTTP requests entered the network.

Each of these options has advantages and disadvantages as described below. For the first option, it is necessary to have clients that can perform geo location processing. Without this capability, they would be unable to request a manifest for their particular GRC and would likely receive the most restrictive manifest (blackout area), even if they were not located within a blacked out area. It's also not hard to imagine the development of downloadable applications that will allow clients to spoof their actual location within an HTTP request for content.

In the second option, control plane elements are in the critical path of determining the client's location at all times. This is particularly difficult if the client is mobile and is crossing different blackout zones. Each time the client enters a new zone it must be detected by the control plane elements so that a new manifest for that zone can be delivered.

The third option is the most ideal way. Here, a function exists within edge distribution network elements (e.g., CDN cache) that can append location-specific metadata into the client's HTTP request for the manifest file. For example, if Service_Provider is the subscriber's service provider, and if the subscriber is trying to acquire the SportsNetwork broadcast, the guide/navigation function would provide the SportsNetwork URL, SportsNetwork.ServiceProvider.net/index. The location- specific metadata would be inserted within the access network or at the boundary (edge cache) between the access network and CDN ingress point, so as to accurately identify the physical location of the client. This would result in a modified URL, SportsNetwork.ServiceProvider.net/pitt/index. This is a simple and reliable method that even works if mobile clients cross GRCs dynamically. As the client moves in and out of different access points, the network elements at the edges of the network add location-specific metadata that becomes part of the request, resulting in the return of the appropriate manifest. Of course this approach requires this functionality to be incorporated into access networks or CDNs in a standardized fashion.

An alternative that models today's blackout system solutions uses the packager to enforce blackouts within a given region. As illustrated in Figure 9, edge packagers are located within the different blackout regions.

Each live packager is configured to assume a virtual IRD identity. The IRD retune messages received from the satellite downlink are carried within metadata that is distributed to all the packagers in the regions. Each packager filters for retune messages destined for its VIRD identity. A live (edge) packager that observes a retune message addressed to it will update the manifest it is creating with URLs that point to alternate content during the duration of the blackout.

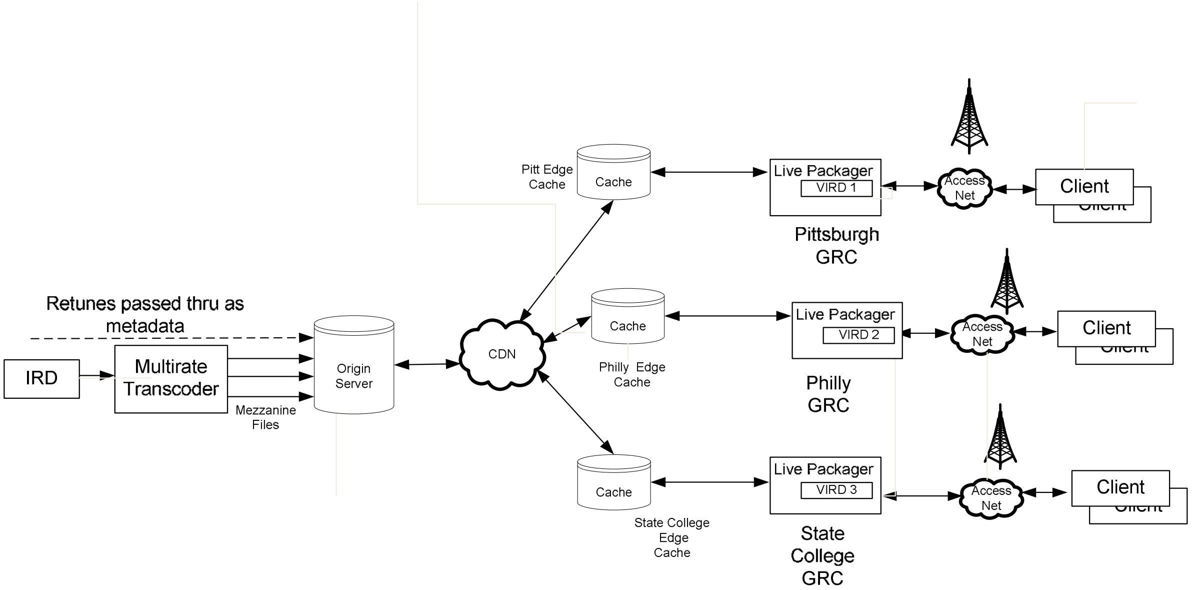


Figure 9: Edge Blackout Architecture

The following is a summary of the advantages and disadvantages provided by the centralized and edge/regional blackout solution options:

Centralized

Advantages

- All operations are managed centrally.
- All content processing equipment can be co-located.

Disadvantages

- A blackout management function is required that understands the GRCs it manages and CDN topology.
- A scalable playlist rebuilding function is required.
- Network changes may be required to append location information to HTTP requests or higher level managers may be required to point the client to the appropriate manifest URLs

Edge

Advantages

- There is no need for a centralized blackout management function.
- There is no need for a centralized, highly scalable playlist rebuilder function.

Disadvantages

- Requires deployment of edge packagers.
- During BO, content replacement has to be facilitated at the edge packagers.

CONCLUSIONS

The choice of where to locate the packaging function is a complex one with implications far beyond capital expenditure decisions such as how much hardware to purchase and where to locate it. Previous discussions of central versus edge packaging have focused on the costs of network infrastructure. There are however many other important factors that should be considered. The packager has been thought of as a simple component that provides chunking and creates manifests. As has been described in this paper the truth is that many of the high value features of the overall system are highly dependent on the packager configuration. Ad insertion and blackout control are two examples of such high value features.

As an operator, do you want to support generalized ads, regional ads, or targeted ads? How concerned are you about the race between ad providers and ad-defeating clients?

Are you focused exclusively on one of several possible subsets of the viewing experience (VOD, linear, nDVR, ABR, multi-screen)? Depending on the subset, the problem space changes. One may want to consider a phased approach or look at the entire problem when planning a deployment.

Of course, the choice may not be between center or edge packaging. It may well be between center and Hybrid packaging, and edge packaging. One could well argue that simply using edge packaging is the simpler solution.

VIRTUAL ENVIRONMENT FOR NETWORKING TESTING AND DESIGN

Judy Beningson, Colby Barth, Brendan Hayes
Juniper Networks, Inc.

Abstract

This paper describes the use of virtual environments for the testing, design and modeling of networks. This paper will also explain the architecture and technology behind these virtual networking environments, and will highlight two real world use cases. The paper will also cover the benefits and limitations for cloud-based network modeling and testing to help operators determine the best uses.

INTRODUCTION

Operators who own and run IP transport networks understand that testing new protocols, design changes and/or modeling service introductions can be challenging. Most operators have access to a test lab for such purposes, but these labs have limitations in terms of scale and flexibility. Even the largest test labs do not approximate the size of an actual production network; smaller operators' labs may be non-existent or so small that any realistic control plane scalability testing is simply not feasible.

Due to size, budget availability and space limitations of current physical test labs, it can be difficult to test or design for the same level of scale as an operational network. Additional challenges result from the requirement for physical "racking and stacking". To test different topologies or configurations typically means making changes to physical connections and systems, which can be time-consuming and in some cases can have an impact on the number of test iterations.

Physical labs are also costly to both acquire and maintain. There is typically some level of capital outlay required for new projects, and once equipment is purchased, there are recurring costs associated with power, space, cooling and maintenance.

While physical labs are absolutely a critical part of any operator's test and design toolkit, because of the aforementioned limitations in terms of scalability, flexibility and costs many have considered the possibility of moving some testing and design exercises into the software realm. In fact, there exists several commercial and open-source software-based network simulation tools (e.g., GNS3, Olive), but these introduce another set of challenges and limitations. Generally these solutions are not officially supported by the major network equipment manufacturers, so features, protocol behavior and capabilities vary between what is available in software and what one will see on an actual network. For example, some of the router simulation software options lack forwarding capabilities. Other offline modeling tools can show results that diverge from actual world behavior. While these software solutions certainly have their place, to be able to test and design with confidence, one needs to conduct tests with the actual code that will run in your physical network.

To help fill the gap between physical test labs (realistic but limited scale and flexibility) and traditional software simulation solutions (flexible but limited realism), networking equipment vendors such as Juniper Networks are now offering cloud-based services that enable operators to create and run networks in a virtual environment. These environments enable users to create and operate virtual networks consisting of fully functioning router/switch "stacks" of network equipment operating systems. Some solutions also include virtual machines of the test equipment you would expect to see in a physical lab.

These cloud-based environments have the benefit of using virtual resources—so they are immensely flexible and scalable—and are also fully supported by network equipment vendors. This latter point ensures feature parity across multiple versions of router OSes and protocol consistency across both the virtual environment and physical gear.

Use Case	Virtual environment solution
Scalability	✓
Protocol interop	✓
OSS/BSS integration	✓
What-if scenarios	✓
Alternate Network architectures	✓
Training/Education	✓
Hardware testing	✗
Forwarding performance	✗

Table 1: Virtual Testing Environment use-cases

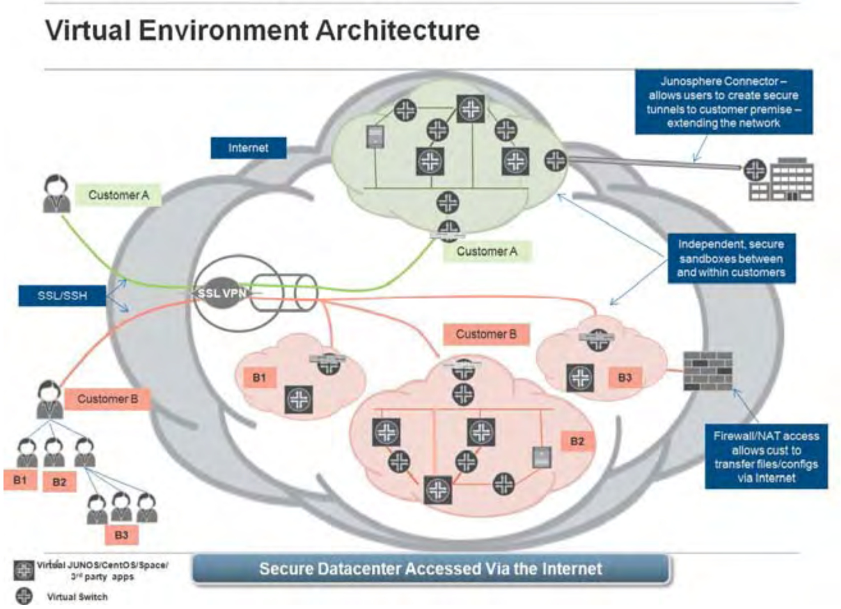
Within a virtual environment, operators can essentially replicate their production network and conduct test and design exercises with a level of scale and realism not otherwise possible, along with many other use cases. Refer to table 1. However, because it is a virtual environment, some tests are simply not possible. In this paper, we will outline the technology behind these virtual environments; examine some real-world use case examples; and discuss the benefits and limitations of such solutions.

VIRTUAL ENVIRONMENT

The network virtualization environment used for the tests described in this paper is a Juniper solution (marketed under the name Junosphere), and it is essentially used to create networks in virtual, rather than physical space. These virtual networks can be used for design, test and training exercises without the need for physical gear while providing a true instance of a router operating system (in this case, Junos) along with an emulated data-plane.

The key components of a virtualized networking system are:

- A secure, multi-tenant Data Center, optimized for high-speed networking between servers and network-attached storage
- A virtual machine (VM) management


Virtual Environment Architecture
Junosphere Connector – allows users to create secure tunnels to customer premise – extending the network
Internet
Customer A
Customer A
SSL/SSH
SSL VPN
Independent, secure sandboxes between and within customers
Customer B
Customer B
B1
B2
B3
B1
B2
B3
Firewall/NAT access allows cust to transfer files/configs via Internet
Virtual JUNOS/CentOS/Space/ 3rd party apps
Virtual Switch
Secure Datacenter Accessed Via the Internet

layer customized for creation of network topologies

- A series of VM images, that users can load on demand
- A graphical user interface which allows users to save and store custom topologies as well as control permissions and access to the service

Each of these components is covered in more detail below.

Data Center

Because the virtual environment will be used to create and operate networks, the demands on it are quite different from most cloud environments or services, which traditionally are priced and offered based on compute power and/or storage. It would be very difficult to simulate a network in these environments, so it was necessary to build out an entirely new, next generation, cloud Data Center for the foundation of this virtual networking environment. The data center is a combination of Intel-based servers and network-attached-storage, with all Ethernet ports connected together via Juniper EX Ethernet switches. DC file upload and download protection is provided via high-end firewalls, and end-user topology access is secured via the SSL VPN gateway software. The cloud is located in a high-availability colocation facility that provides rack space, cooling, redundant power and high-speed, redundant Internet access. DC uptime is designed to be 24x7, 365 days per year, with service maintenance windows roughly occurring monthly. Finally, a publicly accessible URL completes the access.

Virtual Machine Manager

The real brain of the solution is the Virtual Machine Manager (VMM) software that handles the virtual machine creation and deletion as well as the unique job of VM inter-working. A purpose-built cloud for this virtual networking environment was required because we are building customer-specified networks of VMs, and not just leasing workload CPU cycles and/or access to storage.

The VMM used is a Juniper-developed KVM/QEMU-based solution that provides the ability to scale according to the size of the computing platform, offering support of complex network topologies as well as hosting a mixture of Junos, Unix and other 3rd-party VMs. VMM takes in via its API an execution script that, in conjunction with the Virtual Distributed Ethernet (VDE) switches, provide emulated Ethernet segments to which virtual machines are able to interconnect. VMs within a user's space are able to communicate over these emulated segments, the interfaces operating in the same way that a Layer2/Layer3 interfaces on a regular physical device would. VMM, thus, creates a "VMM topology" per customer which is a unique instantiation of the VDE Switch process, the number of VMs, and the type of VMs. The spaces are "secure"; VMs from User A are unable to communicate with those of User B.

Virtual Machine Images

During the instantiation of the VM by the VMM software, a personality (image file) is loaded onto the VM. This personality decides the operation of the VM. Within the virtual environment discussed in this paper, the available image files included:

- VJX1000 – a virtual version of a Juniper router/switch – that supports current releases of the Junos operating system. It is a "real" operating system, with an emulated forwarding plane capable of supporting all routing (MPLS, VPLS, v4, v6, multicast) and firewalling (stateful firewall) features. The virtual machine is able to operate as a regular Juniper device, without the need for hardware to be present.

- Junos Space - a network management application platform that can be used to provision, monitor, and manage Juniper devices
- Centos – a Unix host image for customers to add custom applications or host configurations
- Partner images from leading design and test vendors such as:
  - Cariden Technologies (MATE) [1]
  - Packet Design Insight Manager [2]
  - Spirent Virtual Test Center [3]
  - Mu Dynamics Studio [4]

This paper describes specific experiences, and therefore the images above are restricted to what was available within the existing virtual environment. It is possible that virtual machine image files representing other vendors or technologies could be incorporated into a similar virtual networking environment.

User Interface

The user interacts with the virtual network via a web-based user interface (UI) that lets users access the environment from any browser-equipped laptop or tablet. The UI is an application built as a multi-tenant provisioning tool for account, capacity and library management. It provides the GUI-based control of resources, allowing users to schedule their access times, store their topology files, and build their unique networks on-demand.

IN THE WILD

As previously mentioned, a virtual environment can provide significant value when trying to evaluate new technology and/or test specific large-scale protocol scenarios for a network. A physical lab environment is essential for router/switch hardware testing and validation but in almost all cases cannot provide the topologic resources to determine how a technology or protocol with act on an actual network.

In the next two sub-sections, we will discuss two scenarios where a virtual environment is used to validate network operation in the presence of new technology. For each use-case we will briefly describe the problem and/or challenge followed by a description of how virtual networking resources were used to solve the problem.

Use-case #1: Large Scale Core Network Scaling

In this example, an operator is trying to validate several simultaneous technologies to enable a more efficient method of scaling their core network. This represented a fundamental architectural shift that required a much more detailed test environment than could be provided by a set of off-line modeling tools and a few routers in a lab. The goal was two fold:

- Introduction of a MPLS “optimized” packet forwarding paradigm through the use of BGP labeled-unicast sub-address-family [5]
- Introduction of a multi-plane core architecture and the Aggregation/Edge connectivity

The network and technology migration is illustrated in the figure below.

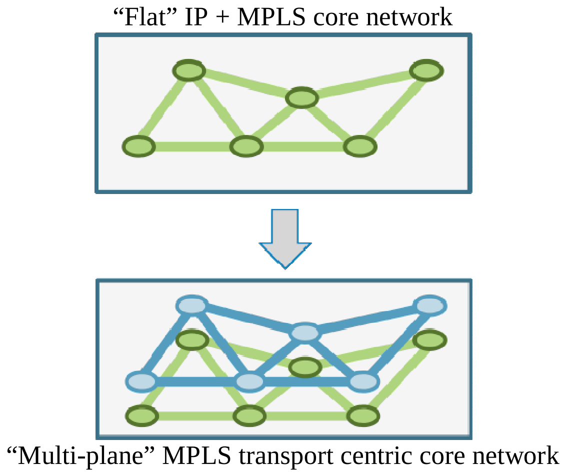

"Flat" IP + MPLS core network
"Multi-plane" MPLS transport centric core network

Figure 2: Network Architecture validation

The challenge the operator faced was how to conceptualize and visualize the target network, test the required protocol modifications, test the introduction of new protocols, and subsequently validate the forwarding properties in the network.

It was essential to be able to validate the changes on a mirror image of the current core network which consisted of a number (10's) of PoPs geographically dispersed across the U.S. in order to ensure the correct routing policy changes, interaction of additional protocols, and validate the protocol architecture.

In addition to generally validating the modified network architecture, the operator now had a working virtual model of the target network in order to train their operations teams, practice and validate change-order methods and procedures as well a working documented target network.

Use-case #2: Protocol Scaling Characterization

In this use case, an operator wanted to very specifically characterize the memory and forwarding impact on their routing infrastructure if they enabled a new protocol extension. The protocol extension was a Border Gateway Protocol (BGP) extension called Add-path [6]. We will briefly describe BGP Add-path in the next few paragraphs before getting into the specific operator example.

BGP has implicit withdraw semantics on each of its peering sessions, where an advertisement for a given prefix replaces any previously announcement of that prefix. If the prefix completely goes away, then it's explicitly withdrawn. BGP scaling techniques such as route-reflector and confederations are widely used in networks of all shapes and sizes. These techniques result in information hiding—for example, available backup routes are hidden. This may be good for scaling, but can problematic in other ways. BGP Add-path addresses some of these inefficiencies.

There are a number of reasons to enable BGP Add-path.

- Faster convergence, robustness and graceful shutdown schemes that require backup paths. This is because route reflectors eliminate backup paths.
- Stability and correctness schemes that require additional paths. For example fixes for MED oscillation or MED misrouting
- Multipath schemes that require multiple next hops
- And, implicit withdraw alone is potentially a problem for some types of inter-AS backup schemes

As you can see, much like the previous use-case, the operator was faced with multiple challenges:

- Would BGP Add-path provide the expected functionality?
- How would the additional BGP paths affect the routing resources of their network?

- Do they leverage the current BGP design or could further optimizations be realized?

It was essential for the operator to build a virtual representation of their current International core network to baseline BGP behavior and resource utilization. Another requirement was the need to be able to access and import, as closely as possible, their current peering locations in order to replicate the current BGP table "attributes".

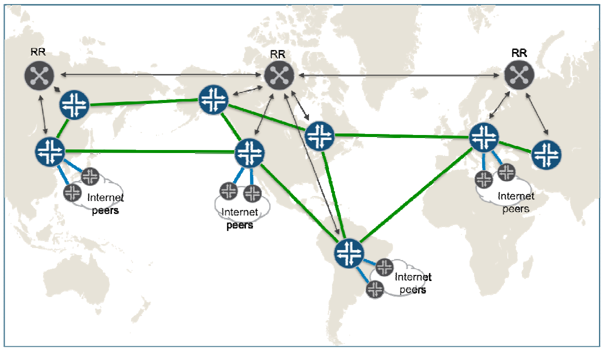


Figure 3: International Core network with regional route-reflectors (RR) for BGP scaling

The resulting virtual network representation allowed the operator to not only characterize their current design, validate BGP add-path and understand specific add-path configuration requirements but also developed multiple future architectural scenarios where indeed BGP Add-path not only delivered the required functionality but could also result in reducing the network resources required to scale BGP.

CONCLUSIONS

Virtual networking environments are a new development that leverage the technologies and concepts popular in cloud computing, and apply them in new ways to solve a fundamental problem for network operators. While virtualized environments will never be a complete replacement for hardware testing, they can provide the resources that allow operators to perform large-scale topology design or testing exercises that would not otherwise be possible. In this paper, we have outlined the technologies behind a specific virtual networking environment implementation, and several use cases, but these technologies and use cases can vary beyond what was discussed within the scope of this paper. In any form, virtual networking environments can be a powerful addition to an operator's design and testing toolkit.

FURTHER READING

QEMU/KVM references/publications

http://www.linux-kvm.org/page/Main_Page
http://wiki.qemu.org/Main_Page

Network virtualization references:

Flexible Cloud Environment for Network Studies:
http://edusigcomm.info.ucl.ac.be/Workshop2011/20110311002

BGP Route Reflection:

http://www.ietf.org/rfc/rfc2796.txt

REFERENCES

[1] http://www.cariden.com/
[2] www.spirent.com
[3] www.packetdesign.com
[4] http://www.mudynamics.com/
[5] http://tools.ietf.org/html/rfc3107
[6] http://datatracker.ietf.org/doc/draft-ietf-idr-add-paths-guidelines/

STRATEGIC CAPITAL - A FORMALISM FOR INVESTING IN TECHNOLOGY

Marty Davidson
Society of Cable Telecommunications Engineers

Abstract

Spending decisions in cable today are complex. Long gone are the days of prioritizing OPEX over capital or purchasing via simple volume related discounts. Capital is now under an intense microscope. This paper presents a way to strategically and logically determine the optimal purchase price that will minimize the total cost of ownership, identify ways to drive efficiency into a workforce by identifying the proper division of labor and it will make way for the possibility of technological innovation through a 'creative destruction' process that will enable long-term growth.

INTRODUCTION

Currently, telecommunications service providers face stiff competition with new entrants every day and must search for solutions to the challenges and difficulties of growing revenue as well as margins. They must do this while dealing with the continued high fixed cost of doing business and the multitude of seemingly simultaneous priorities. Additional pressures exist due to operators being evaluated on a free cash flow basis. Under the existing economic climate, more often than not, this pressure is mis-prioritized and translates to demands for lower priced Customer Premises Equipment, or CPE. When this happens, a caustic force is unleashed that actually increases total costs and negates the scientific possibility of technological innovation. Due to the many ramifications of such a decision, the development of an evaluation schema is required.

This paper provides a formalism for a new way to think about how features in equipment that have the potential to translate into lower costs over time can be objectively and agnostically assessed. After this valuation is completed, decisions that optimize performance and lower OPEX can be made at the time of purchase. A specific example used is the consideration of strategic technical investment in CPE diagnostic elements that optimize operational costs by identifying applicable processes and the possibilities for the proper division of labor. It is shown via this formalism that through this type of upfront investment, service providers will reliably identify and improve not only their fiscal position, but also the quality of customer experience and will be well armed for the ever-evolving subscriber/revenue battle. Lower operational costs via these types of strategic technical investments in CPE will be shown to have additional advantages that can be evaluated using the formalism to determine how they would aid in improving capital efficiency and the ability of a cable operator to react even more quickly to new service needs and market forces. Finally, the formalism will provide a mechanism for operators to determine which new features are critical enough in long term cost-benefits to warrant standardization so that all equipment supports the features. A key goal of this formalism is to implement the type of industrial efficiency and quality envisioned by the likes of Frederick W. Taylor and W. Edwards Deming by specifically coupling equipment procurement decisions into a longer-term process of continued technology improvement to enhance the competitive position of cable operators. But another goal is to provide a mechanism for the type of disruptive process of transformation or 'creative destruction' via new equipment and service capabilities that accompanies the kind of radical and rapid innovation that is the force that sustains long-term economic growth.

FINANCIAL PRIORITIZATION

Opinions vary as to where, when and how our current economic climate started, be it deflation, deleveraging, debt accumulation, etc. associated with the housing & financial bubbles. Initially, the economic downturn actually benefited service providers as consumers limited their expenses for activities like going to the movies. The desire for entertainment was still strong so subscribers turned more and more often to home entertainment services. While the concern over the potential for a significant age of deflation was being ignored by the masses, some companies began to feel the real impact to their top and bottom lines. Telecommunication service providers seemed to initially weather the storm, however, as the economy kept declining and lagging, its impact to these providers began. To the credit of the industry, bold changes began happening, but not all the changes were for the betterment of the business in the long term. One example of this is when operators reduced expenses but cut not only the fat, but also the muscle and sometimes into the bone. In the short term, when these changes were looked at in a silo they appeared to be very reasonable; however, when you couple such decisions with being evaluated on a free cash flow basis, some very dangerous things happen. Purchasing organizations are incented to drive prices lower and lower, which in itself is the right intent. The danger is when decisions on capital purchases are based purely on purchase price. When this happens without taking into consideration the 'hidden' costs in operations, the total cost of ownership can far outweigh any purchase price savings. Additionally, technological innovation is stymied and the possibility of the 'creative destruction' process for sustained fiscal growth vanishes. Joseph Schumpeter popularized the idea of 'creative destruction' based on the economic theories of Karl Marx and he believed innovation shifted the powers in a market place by the introduction of new competitors and that 'creative destruction' described the dynamics of industrial change.

In order not to limit a new age of industry pioneers, a methodology is needed to holistically evaluate purchasing decisions that will lead to the most strategic investments in capital possible. A formalism is presented here that identifies a new parameter called Optimal Purchase Price, which takes into account a wide array of considerations one could use when negotiating equipment purchases, whether that be with a vendor or with the purchasing department within their own company. This prescription for strategic capital purchases leverages a Total Cost of Ownership, or TCO, approach and is not a Cost Benefit Analysis. Performance differences between pieces of equipment should be evaluated relative to the importance to the purchaser. This formalism looks at capital investments from concept to test to deployment to operational integration to trouble resolution to future proofing.

OPTIMAL PURCHASE PRICE

To begin to define the Optimal Purchase Price or OPP, a base upon which can be built is required. That base is the traditional, actual purchase price that an operator would pay for a given piece of equipment. While this paper does consider equipment throughout the network, from the national distribution centers through the backbone, headends, hubs and HFC plant, the predominate evaluation comes from Customer Premises Equipment, or CPE.

Traditionally, the purchase price is evaluated on a Return on Investment, or ROI, basis. ROI is a function of base purchase price, BPP, average revenue per unit, ARPU, and average expense per unit, AEPU. Essentially it is the time period that operational cash flow takes to recover the capital purchase, usually expressed in a number of months.

$$\mathrm{ROI} = \frac{\mathrm{BPP}}{\mathrm{ARPU} - \mathrm{AEPU}} \qquad (1)$$

For the purpose of this formalism, a normalized payback can be considered. One characterization of this is seen in Figure 3.1.

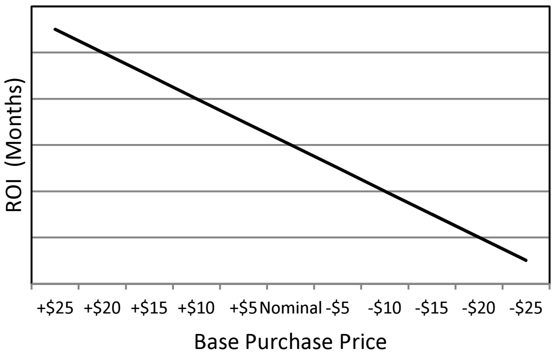


Figure 1

This curve is linear but it certainly has multiple Purchase Price Factors, or PPFs, which can influence it non-linearly such as PPF_1 which accounts for equipment volume discounting or other price influencing factors. Another adjustment that can be made on the base optimal purchase price is differential pricing for multiple organizations or PPF_2. One example of this is sometimes referred to as most favored nation pricing. For a given operator PPF_2 is ignored, but is a valuable tool for a comprehensive analysis across multiple perspectives.

Once the base OPP is established, the incremental components of total cost factors must be defined and evaluated.

Pre-Deployment Test Cost

Before the operational cost impacts of a deployed device can be considered, an evaluation of Pre-deployment Test Costs or PTC must be made. These apriori considerations include:

- Software, firmware and hardware related costs that come from issues that are identified in lab or field trial evaluations and require new versions prior to deployment. Each of these costs has a related scale factor based on the likelihood of needing multiple revisions. Software typically requires 10-20 times more revisions than hardware or firmware.
- Lab testing costs which encompass lab setup, test, evaluation, post analysis, tear down and personnel costs, whether performed internally or externally to a given operator.
- Field trial expenses including training, planning, trial management, field and customer care resources, increases in calls and truck rolls as well as tangential components to account for costs due to customer dissatisfaction and poor press.

$$\mathrm{PTC} = \sum_{i,j,k}^{n} (\mathrm{SF}_{\mathrm{SW}} * \mathrm{Sw}_i + \mathrm{SF}_{\mathrm{FW}} * \mathrm{Fw}_j + \mathrm{SF}_{\mathrm{HW}} * \mathrm{Hw}_k) \ (2)$$

Each component has built into it the number of resources in the lab, field and management of the project, the associated costs for these resources and the time it takes to resolve the issues that have been identified.

Cost of Deployment

Once a piece of equipment has made it through the lab and field trial hurdles, deployment begins. Operators use multiple strategies for deploying new hardware, firmware and/or software. Deployments could start from a few friendly users to a small market with limited deployment, all the way up to a national or company-wide roll out. There have been numerous situations where small deployments did not identify operational issues until an appropriate level of scale was met. As such a fiscal evaluation of deploying new technology must be used. The Cost of Deployment, or COD, is proposed and is a major component of OPP.

The most influential factor in COD is the increase in trouble rate. This increase has been shown to add a significant cost to doing business. When a piece of equipment from a new supplier is introduced into the field, the customer-reported trouble rate can increase, CRT_i, as much as 30 percentage3032733840 points. There are numerous cost drivers when this happens such as: increased calls into customer care, CC_c, increased truck rolls, TR_c, both valid and in error (traditionally 10-15% of all trouble calls into customer care translate into a truck roll in error) and resources on the team that manages the tickets being worked, TM_c.

$$\mathrm{COD} = f(\mathrm{CRT}_{\mathrm{i},}, \mathrm{CRT}_{\mathrm{t}}) * (\mathrm{TR}_{\mathrm{c}} + \mathrm{CC}_{\mathrm{c}} + \mathrm{TM}_{\mathrm{c}}) \quad (3)$$

Pick a dollar figure for a call into care, a truck roll and a hourly labor rate and you will see how significant this parameter can be. But that is just the beginning as this is a problem that just keeps on giving. There is a major influence on all of these expense increases, namely, the time it takes to get the customer-reported trouble rate back down to normal levels.

As seen in Figure 2 below, getting back to the normal trouble rate can take 18 months and with new technology or product introduction this can be even longer.

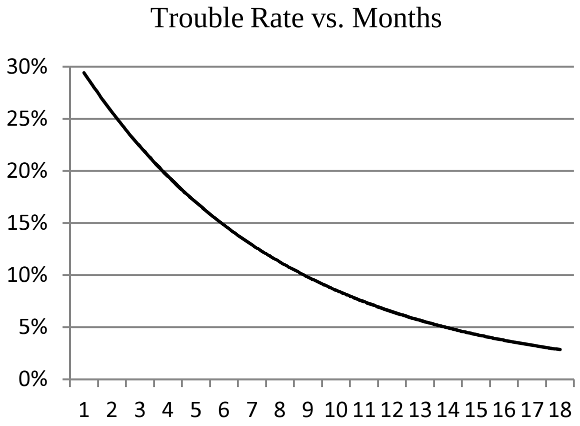


Figure 2

Equipment Combo Factors & Locale Weight

As mentioned in the introduction of OPP, the major focus of this paper is on CPE even though there are other network equipment influences (NEF) built into the formulation. Equipment Combination Factors, or ECF, take into consideration what services can be enabled on a given piece of CPE and what actual services a customer is paying for on that CPE, this is referred to as CPE_F. For example the lowest ECF components are stand-alone set top boxes and cable modems. Just above that are home gateways, WiFi enabled modems and eMTAs. Additional weighting is applied to devices that carry critical services like lifeline voice and home security, CPE_w. This is reflected in the matrix operation to determine ECF.

$$\mathrm{ECF} = [\mathrm{CPE}_{\mathrm{F}}] * [\mathrm{CPE}_{\mathrm{W}}] + f(\mathrm{NEF}) \quad (4)$$

The location of the equipment being deployed also has an impact on the overall total cost that needs to be considered and is reflected in this analysis as ELF. Factors considered in ELF include every locale where equipment could be deployed (EDL) from the home through the HFC network to the backbone and into national data centers. The degree of influence that errors associated with new deployment have on the customer population is weighted appropriately (EWF). This weighting function is proportional to the number of subs potentially impacted by it and a characteristic function of the device itself.

$$\text{EWF} = f(\text{device}) * \sum_{i=1}^{n}(\text{subs})_i \quad (5)$$

The functional combination of these two elements provides the overall equipment locale weight, which can be seen in Figure 3.

$$\text{ELW} = f(\text{EDL, EWF}) \quad (6)$$

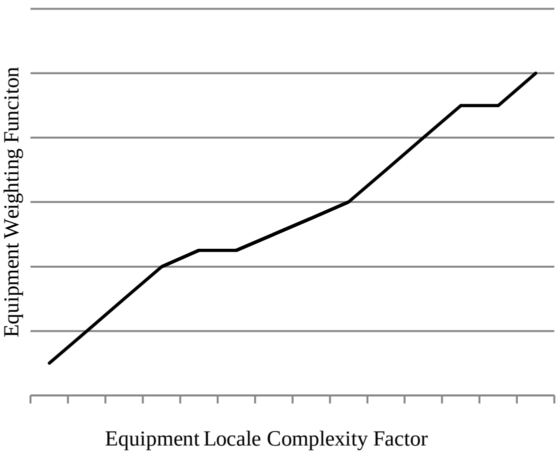


Figure 3

Optimal Ease of Use

Whenever a new piece of equipment or software is introduced into service, there are some differentiators between products that can have an impact on real operational costs. Training is the first element in the calculation of Optimal Ease of Use, or OEU. Training material must first be developed. These could be as small as talking points posted to a call center knowledge base or as involved as a multi-day, hands on session with a live instructor. Once training is developed, the degree of complexity, which can be correlated to time off the job, varies as described. But it is not only the length of training that is of concern, it is the complexity associated with it and the probability that repeat training would be needed. Representing the training aspects of this analysis, Training Development & Deployment, or TDD is used.

OEU is also influenced by the degree of difficulty or ease with which a user can debug and solve a problem on a given device. This is reflected in the Total Time Usage Factor, or TTU.

Standards are so well embedded into our daily life that the average worker rarely, if ever, considers the impact of standards. The Standards Product Factor, or SPF, is a factor that lends itself to the ease of integration, training, etc. when compared with non-standards based products. Standards based products allow for efficiencies to be realized and this can lead to a division of labor which can re-purpose resources to more important and complex challenges.

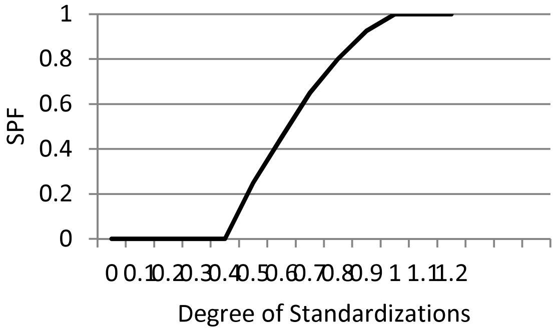


Figure 4

SPF can be looked at as an inverse function so that if a product is standards based, it will help lower the total cost of ownership. This leads to the formulation of OEU.

$$\text{OEU} = f(\text{TTF}) * f(\text{TTU}) * f(\text{SPF}) \quad (7)$$

There are multiple other considerations that could be included in the OEU calculation such as: how much a technician likes a particular product and thus an internally created efficiency of how it helps improve his or her daily work duties; the support provided by a particular vendor; or the creativity and innovation instilled in an employee inspired by the technology and associated ease of use.

Customer Type Factor

The customer must not be forgotten in this analysis, so the introduction of an OPP parameter for the customer is necessary. CTF, or the Customer Type Factor is a complex, non-uniform variable that is heterogeneous in nature. If only one service was provided to a customer and each customer had the same propensity for calling when things didn't work correctly, assessing the CTF would be a much simpler effort, as opposed to the ever growing number and complexity of products a customer may have, as well as the level of, or lack thereof integration that exists. CTF is a function of the products or services a customer has, their likelihood to call into customer care based on a characteristic distribution, the number of different revisions of software, firmware and/or hardware and the types of equipment and level of integration of such devices.

$$\mathrm{CTF} = f(\mathrm{products}) * \mathcal{L}(\alpha|\mathrm{x}) * f(\mathrm{revisions}) * f(\mathrm{integration}) \quad (8)$$

An additional component that could be considered in CTF, but is not reflected here, is if an operator were to prioritize service for their most valued subscribers.

Technical Advancement Advantage

Every piece of equipment has its merits and its opportunities for improvement. As rapidly as technology evolves, as well as the associated operations and customer expectations, a relationship between a given piece of equipment and the technological advantages that it provides is proposed as the Technical Advancement Advantage, or TAA. TAA is the calculated as:

$$\mathrm{TAA} = (\mathrm{FPF} + \mathrm{HPF}) * f(\mathrm{CPD}) \quad (9)$$

Both FPF, the Future Proofing Factor and HPF, the Historical Performance Factor functions are characterized similarly as described by following which is then normalized.

$\sigma_i \geq 1$	i^{th} device 100%
$0.3 < \sigma_i < 1$	i^{th} device $\sqrt{\sigma_i - 0.3}$
$\sigma_i \leq 0.3$	i^{th} device = 0

FPF is essentially the ability of a given piece of equipment to extend its operational usefulness. An alternative, inverse way to think about this would be the less changes required over the life of products from a technological operations perspective. HPF is a confidence value in a vendor who is trusted and has demonstrated past performance of delivering what has been requested. The higher value in both of these factors correlates to a positive impact on TAA and overall OPP.

FPF & HPF Weighting Factors

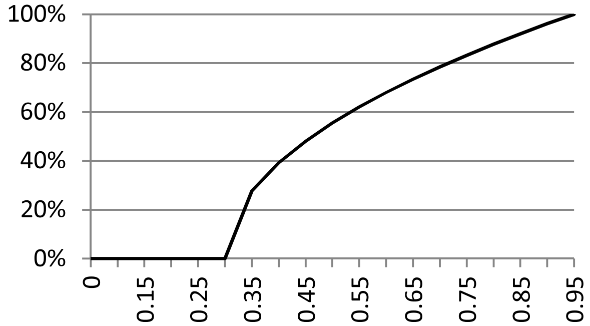


Figure 5

CPD, or Customer Platform Diagnostics, is another proposed functional scaling variable that highlights a piece of equipment's overall diagnostic ability to cross platforms and reduce overall time to repair. One representation of this function is that it gives an increasing, exponential positive benefit to the overall calculation because significant improvements in this area are challenging to come by to say the least.

Each of the attributes of this advanced technology element can be very individualistic and multiple other relationships could be used.

Smart Energy Adjustment

Economic times have made it more difficult for operators find costs savings in their business, but one of the more recent areas of focus is on energy use. Power bills are still a major component of cable expenses and both space and existing power are becoming rare resources. In order to factor energy into the equation of capital purchases, a Smart Energy Adjustment, or SEA, is needed. Presented here are four areas for consideration.

The first component of SEA is the Energy Efficiency Factor, or ε_F, which is a calculation of how efficient a given piece of equipment is. Proposed here is a ratio measure of average throughput and total power used, e.g. bits/watt.

$$\varepsilon_F = \frac{\bar{\mathrm{X}}_\mathrm{n}}{\mathrm{P}_\mathrm{T}} \quad (11)$$

Other elements that need to be included in the smart energy calculation are: density (a function of throughput and physical area), size (a function of how much space a given device occupies, particularly critical in centralized equipment locales) and diagnostic ability. For the purpose of this formalism they are reflected as:

$$\varepsilon_D = f(\tau, A) \quad (12)$$
$$\varepsilon_S = f(S) \quad (13)$$
$$\varepsilon_{PD} = f(t, I) \quad (14)$$

The power diagnostic factor, ε_{PD}, is an intriguing area that could have significant impacts on energy consumption and power availability. The ability here is for a device to understand the historical current (or voltage or power) use and correlate it to potential failure modes, essentially looking at energy as a proactive indicator of overall service availability and reliability. This is major focus related to the fiscal health of the industry and an opportune area for further research.

Diagnostic Capability Determinant

One of the most crucial areas of focus in this formalism is the value of investing in technology, particularly in CPE, that can include diagnostic elements that lead to the optimization of operational costs by identifying customer impacting issues throughout the lifecycle of the customer. This identification, as detailed below, can ultimately lead to process efficiencies and thus even greater savings and enable the possibility of even more technical innovation. Characterization of this capability is done through the definition of the Diagnostic Capability Determinant, or DCD. There are four drivers of DCD, the first of which is Pre-Customer Realization, or PCR. PCR outlines the ability of diagnostics to identify a service related issue before a customer would notice it. Ideally, the best scenario would be if an event could be identified before it happens, however, there are events that will always be impossible to prevent. The time variable in the PCR equation accounts for this situation and is reflected in the overall calculation as the duration of an event multiplied by a function of the percent of time that identified instance occurs times the frequency of occurrence. The calculation is shown with a summation because there may be multiple devices with identical alarms that are worked independently by the work force.

Additionally, unique issues can occur simultaneously. The summation is across the total of all of these events.

$$\mathrm{PCR_i} = \sum_{\mathrm{i}}^{\mathrm{n}} \Delta \mathrm{t_i} * f(\frac{\mathrm{I_i}}{\mathrm{I_T}} * \frac{\mathrm{F}_i}{\mathrm{F_T}}) \quad (15)$$

As mentioned, customer-impacting events are impossible to prevent, but PCR identifies how well the diagnostic capability works in a pro-active fashion. When an event actually impacts a customer it is critical that we identify a DCD component that measures how well the embedded software can identify and distinguish an issue and provide information to the service provider to remedy the situation, which is suggested here as the Diagnostic Activity Factor, or DAF. DAF is a nonlinear function that heavily weights quicker resolution of troubles, as is seen in Figure 6.

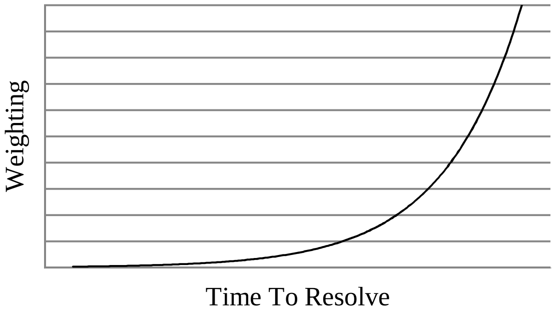


Figure 6

Even with a high DAF, the cause of the problem may not be known. For example, the problem or occurrence may be identified and the problem resolved quickly, which is the initial priority in operations, but the underlying cause was not determined. The Post Issue ID, or PID, addresses the intrinsic value in knowing what caused the problem. PID is a measure of how specific diagnostics are in their ability to identify the actual cause of the problem. Many times using posteriori data can help put new alarm parameters or thresholds in place or identify new process steps or errors in an existing processes.

Figure 7 articulates a scalar multiple that can be used in the DCD calculation. Notice that if no or minimal post problem identification exists, the value for PID is zero. Above that, a three tier value is proposed. These values should be evaluated based on the particular type of equipment in use and the services it supports. Another consideration is how many actual devices are or would be deployed.

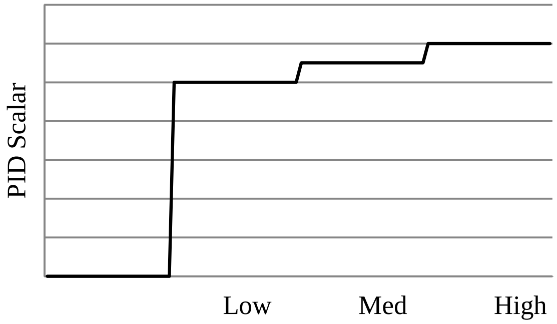


Figure 7

The last consideration in DCD is a proposed parameter that reflects the accuracy of diagnostic recommendations. In operating a network there are many times when data being presented point to a particular issue but when further due diligence is performed, the identified issue is inaccurate. The value of such a parameter is individualized on how important that is to a given user. Here we simply call it Error ID Avoidance, or EID and is a function of user ranked importance, ϕ.

$$\mathrm{EID} = f(\phi) \quad (16)$$

Combining the fore mentioned components of DCD leads to the following calculation. PID and EID are important factors but their influence is adjusted appropriately when compared to DAF or PCR.

$$\mathrm{DCD} = \mathrm{DAF} * (\mathrm{PCR} + \frac{\mathrm{PID+EID}}{\mathrm{PCR}}) \quad (18)$$

Workforce Effectiveness Principle

The last element of OPP is a parameter called the Workforce Effectiveness Principle, or WEP, which is composed of four parts. The first three parts are directly correlated to the technician using a given device and the fourth is a new concept addressing the possibility of quantifying the ability to distribute labor in the most efficient way.

Two of the WEP components measure a technician’s ability to work with a given piece of equipment. Both are straightforward in the sense that their intent is to assess the technician’s interactions during installation and troubleshooting. They are called Installation Ease, or IE, and Troubleshooting Ease, or TE. An ideal approach for these factors would be to perform a time and motion study, using the techniques to identify business efficiency through Frederick Winslow Taylor’s Time Study work combined with work of Frank and Lillian Gilbreth on Motion Study. This will provide a historical baseline and then a static, multi-tier variable based on a suggested difficulty factor here called, δ, can be determined.

$$\mathrm{IE} = f(\delta_{IE}) \quad (19)$$
$$\mathrm{TE} = f(\delta_{TE}) \quad (20)$$

A less scientific measure, but perhaps even more valuable consideration in WEP is the technician’s confidence in working with a given piece of equipment, which here is called the Technician Confidence of Use, or TCU. Anyone who has managed a workforce of technicians can readily articulate the benefits of a confident and enthused team. TCU is a proposed measure to capture just that.

Distribution of Labor, or DOL, is the main driver in WEP and on a macro scale can have the most significant impact on operational expenses. The reason that DOL is so significant is that it looks at the current workforce operations and processes through a ‘Scientific Management‘ lens that Frederick Winslow Taylor proposed and used in the Efficiency Movement. DOL attempts to evaluate the most efficient ways to accomplish the tasks at hand by using advanced diagnostics that will enable the problems to be worked in a more efficient manner. As such WEP is defined as:

$$\mathrm{WEP} = f(\mathrm{DOL}) + \frac{\mathrm{IE+TE}}{||DOL||} + f(\mathrm{TCU}) \quad (21)$$

Due to the inherent complexity for this key component of OPP and because of its many interrelated degrees of freedom, computational algorithmic analysis is required.

Summarizing Optimal Purchase Price

There are a dozen main contributors that have been used to describe OPP. Each of these components has its own level of complexity and interrelatedness to the others. A structured model is required using computational algorithms with bounded, varying randomized inputs for the many individualized computations proposed in this formalism. A Monte Carlo type analysis is suggested that is specifically targeted at reducing the overall total cost of ownership, including resource reallocation efficiencies. The largest challenge of such a model will be integrating those components that are more “soft”, less deterministic and highly dependent on the individual or company prioritization of such elements. One such example is how worker satisfaction is valued via parameters such as TCU, which was described earlier as part of WEP.

Once this is done, a new ROI can be evaluated taking OPP into account and the overall value of a product can be effectively evaluated, both from the operator and vendor perspectives. Next the varying lifecycles of a product should be considered. This may be done by creating different ROI analyses, e.g. via the Monte Carlo analysis mentioned above. If, within a given set of parameters, the ROI exceeds the expected lifecycle, one would need to iterate the formalism to identify areas of cost that could be removed from the business and thus creating a lower OPP with a potential a higher TCO. These realized savings can be thought of as 'insurance' against unforeseen costs. This approach is particularly applicable in the current business climate given how short life cycles can be.

INDUSTRIAL QUALITY & EFFICIENCY

There are many possible ways to divide where the work should be done vs. where it is being done. There are three primary tasks investigated here: issue identification, fix implementation and resolution confirmation. The most fundamental view of efficiency in this discussion is purely how much more efficient a worker can be doing the same tasks as were done previously. This worthwhile endeavor is the same approach outlined by W. Edwards Deming in his approach to Total Quality Management. Workers and management alike should continually assess their duties to look for opportunities for improvement. Detailed chronicling of this work is imperative to drive consistency into the services that are provided to end customers, i.e. standardization. Once the work is documented, methods and procedures can be built into training materials and subsequently the training being given, bringing efficiencies to the training resources as well. Once standardized, many tasks can then be modeled and implemented in software tools to remove menial labor tasks. When this happens, the proverbial flood gates open and one can investigate how to divide the workforce and reallocate the work in the most efficient manner. One example of this would be how the three primary tasks mentioned above could be divided. Taken in order, issue identification could be implemented in software and the verification of issues could be done in a centralized work group. This work group would have fewer total resources than a distributed model as they would be able to fill in the otherwise distributed, individual lows of work with the volume that comes from the total distributed load. Additionally, the speed of identification would also increase. The fix implementation process would also improve. Not all work could be centralized, but any fix that could be done remotely could move into a centralized group and similar efficiencies could be realized.

Finally, much like the issue identification scenario, issue resolution confirmation could be moved to a centralized work group, once again creating a way for the most efficient manner possible.

TECHNOLOGICAL INNOVATION

Some believe that the source of the Western idea of 'creative destruction' is the Hindu god Shiva, who was thought to be the destroyer and creator simultaneously. However, as mentioned earlier, Joseph Schumpeter is credited with introducing the term 'creative destruction' in his famous book, *Capitalism, Socialism and Democracy*. It was in this book where he described how innovation can cause the disruptive process of transformation.

So how does 'creative destruction' apply to this formalism? First, an example is necessary to baseline the concept. In the retail market, previously, many small, older, local companies historically offered retail consumer products. The distributed nature of this model left little opportunity for expense reductions. Then came the technological innovations that Wal-Mart introduced, including new ideas such as personnel, marketing and especially inventory management. While these innovations destroyed businesses like Montgomery Ward and Woolworths, it created a whole new set of technology that spawned other businesses and innovations. Another example is the destruction/creation cycle of 8-track to cassette to compact disc to MP3.

If the proper division of labor outlined in this paper is considered, work is moved from a distributed workforce to a centralized one requiring fewer resources overall. This destroys the structure of the legacy labor pool, but creates an increased level of customer service, reduces operational expenses and opens the door to a whole new set of technological innovations that could never have been imagined before this change, i.e. a disruptive innovation that helps create new business opportunities for existing and new vendors. Along with these new business opportunities comes the scientific possibility for further innovation, aka, 'creative destruction' which leads to rapid and sometimes radical change that yields economic growth over the long-term.

There are other tangible benefits that are realized from the suggestions in this paper, such as increased customer service and loyalty, i.e. reduced churn, marketing advantages as in brand strength, reduced advertising costs and thus lower costs of acquisition. Additionally, an interesting benefit is how capital may be used more effectively and spent in places where the greatest benefits are. Reduced cycles for new product introduction may also be realized as would softer advantages like internal and external public relations.

CONCLUSION

Solely choosing purchase price for equipment based on traditional volume discounts or on a 'lowest cost wins' basis is not sufficient. Today's high fixed cost of doing business, simultaneous priorities, ever increasing level of competition and return on investment expectations, demand a new approach.

Determining the optimum purchase price for equipment can be identified through the combination of the many factors described earlier. These factors require operators to take a look at the capital costs with a new total, comprehensive perspective. This approach requires that personnel in operations, engineering, marketing, finance and purchasing work together in evaluating the total cost vs. assessing it in silos with competing priorities.

Evaluating products in a manner described in this formalism provides: the possibility for operational performance optimization; product and operational standardization; lower short term costs; distinct competitive advantages; increased customer service levels; reduced product deployment times; the proper division of labor; and the radical and rapid innovation required to sustain long-term economic growth.

It is imperative that operators take into consideration the kind of upfront investment and implement a capital strategy that takes into account the vast and complex array of variables that contribute to the overall fiscal health of their business and set themselves up for the next generations of success.

References

- Schumpeter, Joseph A. (1994) [1942]. “Capitalism, Socialism and Democracy”. London: Routledge.
- Reinert, Hugo; Reinert, Erik S. (2006). "Creative Destruction in Economics: Nietzsche, Sombart, Schumpeter”
- Mitcham, Carl and Adam, Briggle “Management” in Mitcham (2005)
- Zandin, K. (2001), Industrial Engineering Handbook, 5th edition, McGraw-Hill, New York, NY.
- Deming, W. Edwards (1986). Out of the Crisis. MIT Press.
- In-Home Support Services, SCTE Home Networking Primer Series
- Innovator's Guide to Growth - Putting Disruptive Innovation to Work. Anthony, Scott D.; Johnson, Mark W.; Sinfield, Joseph V.; Altman, Elizabeth J. (2008). Harvard Business School Press
- Trends Over Time in Server Energy Use, Performance, and Costs: Implications for Cloud Computing and In-house Data Centers. SCTE SEMI Primer Series
- Network Management Fundamentals, Alexander Clemm, 2006, Cisco Press
- DOCSIS Checklist for PacketCable™ Reliability in the Outside Plant - Downloadable DOCSIS Checklist for the PacketCable™ Reliability. From SCTE Outside Plant Seminar
- How to Identify and Build Disruptive New Businesses, MIT Sloan Management Review Spring 2002
- Monitoring and Managing a DOCSIS™ 3.0 Network. SCTE DOCSIS Primer Series
- The W. Edwards Deming Institute, Fostering Understanding of The Deming System of Profound Knowledge

ENGINEERING ECONOMICS – DOCSIS 3.0 CHANNEL BONDING FOR IMPROVED NETWORK ECONOMICS

Amit Garg, James Moon
Comcast Corporation

Abstract

DOCSIS 3.0 (D3) enables channel bonding, i.e. multiple downstream (DS) and multiple upstream (US) RF carriers that can be combined to provide a wideband service.

In this paper, we demonstrate that channel bonding not only provides multi-system operators (MSOs) with the opportunity to offer faster speeds to their customers, but also provides an opportunity to reduce capital required to meet the growing traffic demand. Combinatorial models were used to assess the opportunity for such load balancing gains. Later, empirical data was used to measure the load balancing gains achieved on a plant with D3 cable modem termination systems (CMTSs). Eventually, results from a trial with paying subscribers demonstrated the impact of providing D3 modems to select customers.

The paper demonstrates that instantaneous load balancing achieved through channel-bonding provides carriers with substantial improvement in engineering economics.

A NEED FOR SPEED

The Internet eco-system is flourishing; subscribers love the ease and convenience of broadband access, while content providers have embraced this new platform to provide an ever-increasing plethora of data intensive services – video email, video chat, video-conferencing, music, streaming video, cloud storage and cloud computing to name but a few.

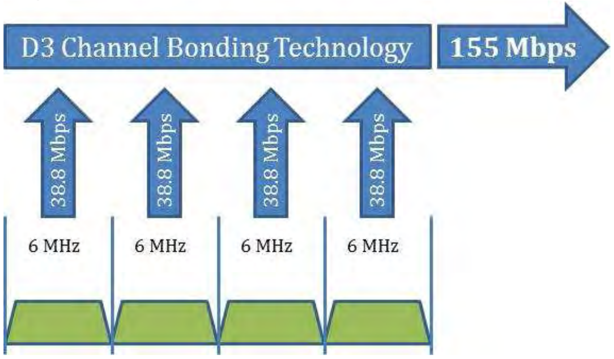
Over the last 15 years the Internet has grown from a novelty to a necessity. Be it communications, travel plans, information, education, news or entertainment, individuals are very likely to use the Internet. Over the same period, internet access has undergone a massive shift, from dial-up modems providing 14.4 kbps to always-on broadband access at DS speeds in excess of 100 Mbps as consumers have embraced faster and faster broadband speeds. MSOs have been leading the way in providing broadband access – by embracing DOCSIS as a way to provide broadband services to their customers. Until a few years ago, in the absence of cable's competitive broadband services, the only way to get a 1.5Mbps service was to pay upwards of $1,000 per month for a T1 from the local or competitive telephone company. Today, the most common broadband packages, with DS speeds in excess of 6Mbps, start at around $40 to $50 per month.

In the early days of DOCSIS, MSOs in North America provided broadband service by using a single 6 MHz channel for DS, and another 3.2 MHz carrier to provide US service. Improvements in modulation eventually enabled ~38 Mbps DS and ~ 10 Mbps US with each of these carriers. And, until recently, a single 38Mbps DS carrier was shared across a group of subscribers (service group) to provide customers with economical access to broadband speeds, while ensuring that customers received a desirable experience.

The development of the DOCSIS 3.0 standard changed that. D3 enabled multiple channels to be bonded into a single service group and was a direct result of subscribers' appetite for faster and faster speeds. To provide speeds in excess of 38 Mbps bonded RF channels become an absolute necessity as the service speeds exceed the offered line rate.

Today, MSOs in North America are typically bonding 4 to 8 DS channels and are beginning to bond 2 to 3 US channels. Channel bonding has enabled DS speed offers in excess of 100 Mbps, while demonstrating DS speeds of up to 1 Gbps. US speeds of 20 Mbps have been offered to customers; US speeds of up to 100 Mbps have been tested. This growing ecosystem has resulted both in an increase in number of subscribers using broadband, as well as increased demand per subscriber. In recent years, demand per subscriber has been growing at ~45 to 50% CAGR. For an MSO, this translates into the need to double the capacity of their high speed data (HSD) networks every 18-24 months.

Figure 1



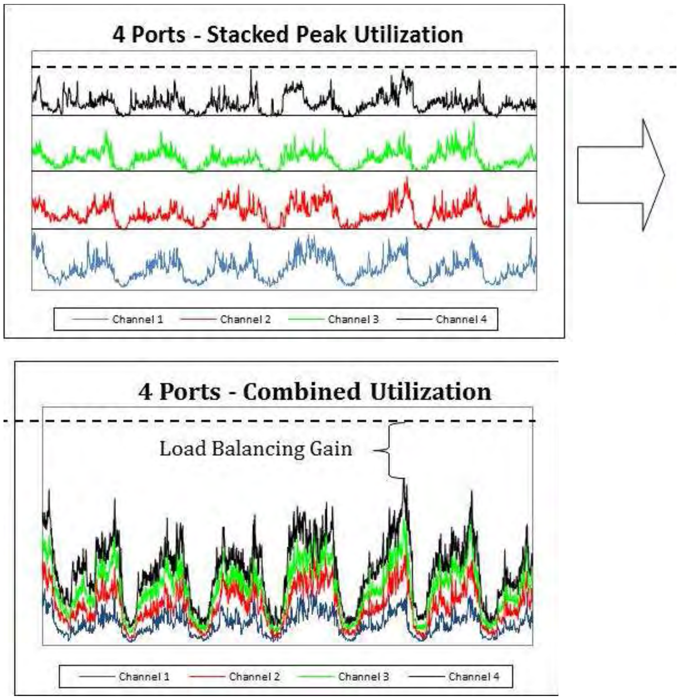
WHAT OTHER ADVANTAGES MIGHT ARISE OUT OF CHANGES IN ARCHITECTURE?

Once the D3 rollout began, there was an increased interest in understanding what other benefits might be derived from the bonded channels – something akin to increased operational efficiency of trunks as explained with Erlang math – fatter pipes are more efficient. Specifically, did channel bonding enable any statistical multiplexing or load balancing gains?

LOAD BALANCING GAINS

Load Balancing enables better use of network bandwidth by managing the network to the Peak of the bonded group and not by managing each port to its own peak. In our case, while observing the utilization of individual DS channels, it was noted that the peaks for multiple channels rarely occur at the same instance. For our purposes, the diagram below illustrates how we viewed the opportunity for statistical load balancing gains.

Figure 2



The top chart has peaks stacked, one upon the other; the bottom has traffic layered. The latter provides stat-mux gain over the former.

Early on, it was very clear to us that channel bonding could unlock some fairly significant network efficiencies as we increase the number of channels included in each SG. Our work helps determine the ranges of those gains and efforts that might be needed to capture those gains.

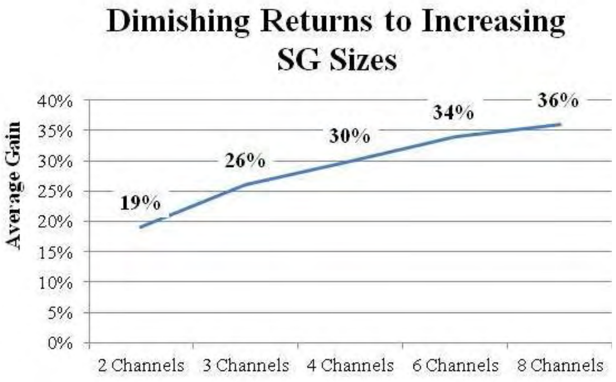
COMBINATORIAL ANALYSES TO APPROXIMATE STATISTICAL MULTIPLEXING GAINS

Prior to deployment of actual D3 networks, attempts were made to quantify the magnitude of hypothetical statistical multiplexing gains that could be possible. To that end, combinatorial models[1] were used to combine pools of existing DS channels into hypothetical service groups of 2, 3, 4, 6 and 8 channel combinations.

1. 5-minute channel utilization records were used.
2. All ports were combined into 2,3,4,8 channel bonded-groups.
3. Peak utilization for the period for each DS channel in the hypothetical SG used in the calculation was noted.
4. A SG peak for the combination was calculated by layering each of the 5-minute values for the channels that made up the hypothetical SG and finding the SG peak value. [Value A]
5. Gains were calculated by dividing the calculated SG peak by the sum of the individual peaks of the channels comprising the SG [Value B] and subtracting 1.
6. Distributions of these potential gains i.e. [(Value B-Value A)/Value B] are summarized in Figures 3 and 4.

Later, larger samples that included more channel/SG combinations from multiple markets were developed to evaluate the gains. Evidence indicates diminishing returns. 2 channels provide 19% gain, 3 channels provide 26% gain, or an incremental 7% points over 2 channel. 4 channels provide 30% gain, or an incremental 4% points over 3 channels, etc.:

Figure 3

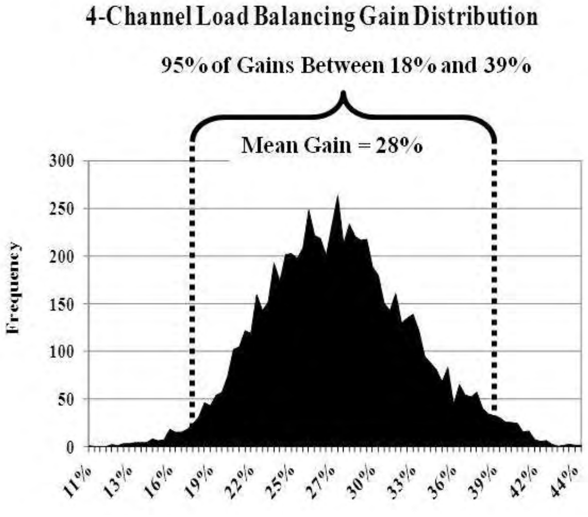


THERE IS A DISTRIBUTION WITH RESPECT TO LOAD BALANCING GAINS

While we were able to calculate the average gains from 2, 4, 6, and 8 port combinations, a quick glance at the distribution chart in Fig. 4 illustrates the fact that the gains are not uniform, but are normally distributed.

Our initial work focused on a single CMTS with 22 ports. We grouped the 22 ports into 7,315 4-Port SG combinations ($_{22}C_4$) and calculated the gain for each.

Figure 4



In this set, we found that the "worst" combination provided a gain of 11%, while the "best" was nearly 45%.

[1] Models implemented in Matlab. Neha Gadkari performed simulations in support of this analysis.

For our purposes, it appeared that network efficiency gains of approximately 25 to 30% could be realized for the (then) typical DS SG deployment of 3 or 4 channels.

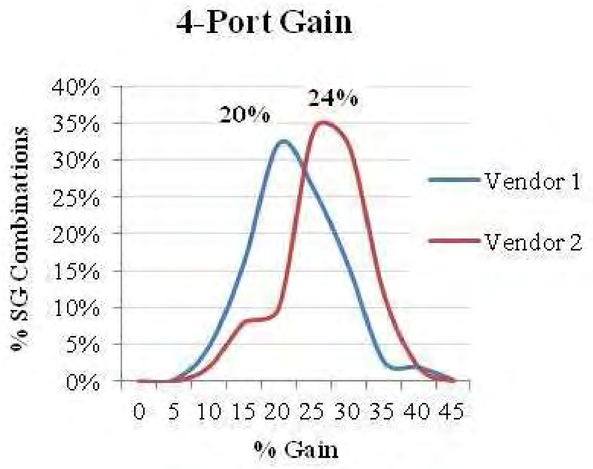
PRODUCTION D3 SERVICE GROUPS EVALUATED FOR REMAINING STAT MUX GAINS

As empirical data became available on production D3 service groups, additional combinatorial analyses were performed to determine if there was any load balancing opportunity remaining as the CMTS vendors had implemented load-sharing algorithms to balance traffic on SGs where majority of cable modems were still not D3. Our data set for this portion of the analysis consisted of over 300 4-channel service groups.

The data showed that most of the 25-30% load balancing gain was still on the table. For one vendor, the average opportunity was~ 20%, while it was closer to ~24% for the other vendor. This led us to conclude that significant gains could be achieved only through instantaneous statistical load balancing.

While vendors raced to develop various load-sharing algorithms to help balance demand across multiple RF channels, it was clear that without the deployment of significant numbers of D3-enabled devices that these significant statistical multiplexing gains would prove elusive as D3-devices enable instantaneous load balancing.

Figure 5



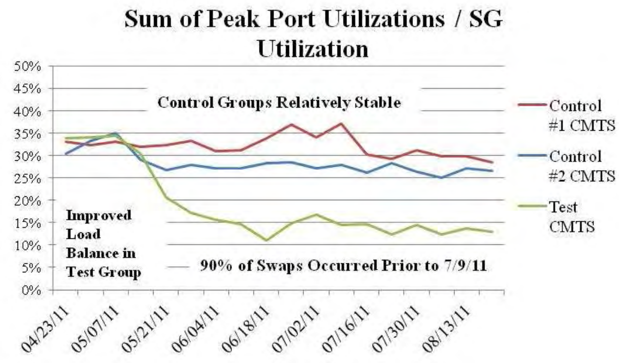
TESTING THE HYPOTHESIS

To further our understanding of what additional stat mux gains could be attained, we conducted an experiment where we provided D3-enabled gear to a large number of subscribers on a CMTS in one Comcast market. Two additional CMTSs in the same market were used as controls.

Over a period of approximately 2 months, select customers were provided new D3 CPE or modem and self-install kits. In addition, all new additions in the market (test as well as control CMTSs) were provided D3-CPE to prevent new users from inadvertently influencing results.

Measurement of the available gains on the SGs of the test CMTS as well as those on the controls indicated that there were about 30 to 35% gains available at the beginning of the study. As targeted modems on the Test CMTS were swapped by our customers and as new customers were supplied D3-enabled gear, we found that the deployment of the D3 gear was generating the desired effect – that load balancing gains were being generated (see Fig. 6). That is, the sum of the peaks of the individual channels that made up the SG and the actual SG peak were converging.

Figure 6



IMPLICATIONS FOR NETWORK EFFICIENCIES

Given a sufficient penetration of D3 gear that most, if not all of the hypothetical gains can be realized. A one-time gain of 20-30% in network capacity offers meaningful returns and can be exploited by MSOs to improve the bottom line as load balancing gains provide savings for years to come.

A simple model illustrates the annual network impacts of a 20%, 1-time gain (see Fig. 7.) In year 1, a 20% impact is recorded. Capital expenditures in that year plummet 64% vs. the business as usual (BAU) view.

Figure 7

	Demand Index - BAU	Demand Index with 20% Gain	Incremental - BAU	Incremental with 20% Gain	% Reduction in Annual Incremental Demand
Year 0	1.00	1.00			
Year 1	1.45	1.16	0.45	0.16	-64%
Year 2	2.10	1.68	0.65	0.52	-20%
Year 3	3.05	2.44	0.95	0.76	-20%
Year 4	4.42	3.54	1.37	1.10	-20%
Year 5	6.41	5.13	1.99	1.59	-20%
Year 6	9.29	7.44	2.88	2.31	-20%

However, the gain is the gift that keeps on giving; with annual expenditures continuing to track 20% below BAU figures.

CONCLUSIONS

While channel bonding has enabled MSOs to offer DS speeds in excess of 100 Mbps and US speeds in excess of 20 Mbps, it offers significant engineering economics. With 4 or 8 bonded channels, MSOs can expect 25-30% gain in network efficiencies through instantaneous load balancing. As more cable modems are upgraded to D3 over the next few years MSOs will benefit from these engineering efficiencies in their capital outlay for years to come.

The Economics of IP video in a CCAP World

John Ulm & Gerry White
Motorola Mobility

Abstract

The paper outlines an IP video architecture and determines the relative cost contributions from the major components. For current equipment, the DOCSIS® downstream channel is shown to be the major contribution to infrastructure cost. As next generation Converged Cable Access Platform (CCAP) systems are deployed this will fall enabling a cost effective IP video platform to be realized. At this point other cost contributors become more significant. CDN and nDVR trade off options are discussed. Finally the paper looks at spectrum migration options to release the bandwidth needed to deliver IP video service.

INTRODUCTION

Delivery of IP video will be a major factor driving cable infrastructure during the next few years. Studies of Internet traffic patterns [SAND], [VNI] show that video has become the dominant traffic element in the Internet consuming 50 to 60% of downstream bandwidth. Cable's "Over The Top" (OTT) competitors account for much of this traffic, with Netflix alone constituting almost 33% of peak hour downstream traffic in North America.

To remain competitive cable operators need to deliver IP video to the rapidly expanding tablet, PC, smart-phone and gaming device market. They must leverage the same cost effective technologies as OTT competitors for this and for video delivery to the primary TV. Thus service providers must deliver two forms of IP video: unmanaged OTT off-net and managed video services on-net. This has caused the industry to become very focused on the implications of offering IP video over a DOCSIS® (Data Over Cable Service Interface Specification) channel.

Over the years, the relatively high cost of a DOCSIS channel has impacted potential solutions for IP video. In the past this spawned multiple alternate proposals. Bypass architectures such as DOCSIS IPTV Bypass Architecture [DIBA] were proposed as alternatives and bandwidth saving mechanisms such as multicast and variable bit rate (VBR) technologies investigated. These have become somewhat redundant with the recent surge of adaptive bit rate (ABR) protocols among consumer devices. This unicast delivery mechanism based on HTTP has become the defacto standard for IP video services to this class of devices. In fact, ABR may be used for all IP video traffic including primary screen [CS_2012].

Thus a critical question for operators is how to deliver unicast based IP video cost effectively. It is important to understand the cost implications for DOCSIS downstream channels in the future.

IP VIDEO ARCHITECTURE

To understand the economic impact of migrating to IP video, the system must be separated into key elements. Components of a Managed IP video Architecture are detailed in [CS_2012] and [Ulm_NCTA_2012].

Figure 1 shows a high level abstraction of an end to end functional architecture for delivering IP video from content providers to content consumers. The video service provider must ingest content from multiple content providers, process it appropriately and then transport it over multiple types of access networks to the destination consumer devices.

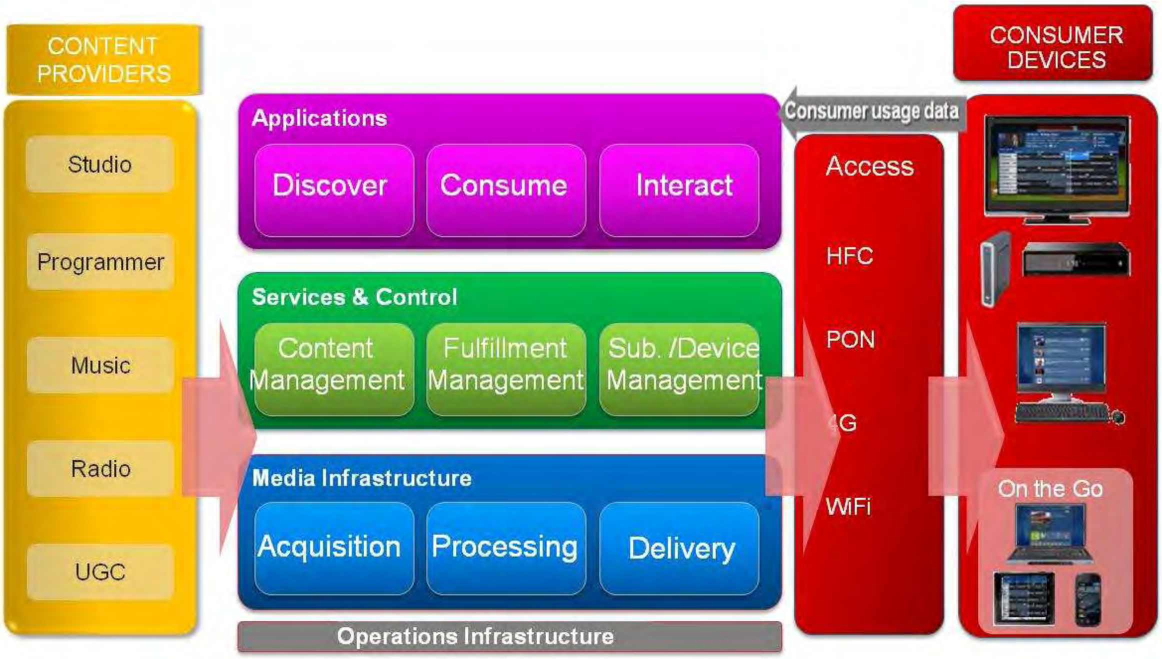


Figure 1 IP video Functional Model

This functional model is used to develop a high level breakdown of the costs for IP video delivery and to compare the relative contribution of each component. This will enable operators to understand the impact of the major cost drivers and make intelligent system trade-offs in their IP video architecture.

MAJOR COMPONENTS AND COST IMPLICATIONS

Content Providers

The number of content sources is increasing. Traditional streamed linear television broadcasts from studios and programmers may be received over satellite or terrestrial links in MPEG-2 and MPEG-4 formats. User-generated content and other Web based multimedia sources must also be supported, but will more typically be delivered as file-based assets.

Costs associated with ingesting content from content providers scale based on the number of program sources ingested and the cost of the material. Once purchased and ingested, these programs are shared across all subscribers. The cost per subscriber is not materially impacted by changes within the delivery infrastructure and thus these costs are not considered further in this paper.

Consumer Devices

One of the principal drivers towards a service provider IP video infrastructure is to be able to support generic IP-based consumer devices such as smart-phones, tablets and gaming devices. The range of consumer devices appears to be almost limitless in terms of screen sizes and resolution, network data rates, processing power, mobility, media format support and DRM support.

Most of these consumer devices are owned directly by the consumer. The one exception to consumer-owned devices might be the IP set-top box. For this analysis, it is assumed the operator will have some leasing revenue associated with the IP STB so it is not considered as part of the infrastructure costs. There are some cost tradeoffs in the use of home gateways and hybrid video gateways which will be considered later in the paper.

Access Networks

A primary reason to move to an IP video infrastructure is that it can be access network independent in contrast to existing MPEG/RF video infrastructure. For the purpose of this investigation only the hybrid fiber coaxial (HFC) access network will be considered.

Core IP Network

The components of the IP video architecture interconnect via the same generic IP core network used for all video and high speed data service delivery. The costs of the core network are amortized over multiple services and all subscribers. Thus the cost contribution to IP video service on a per subscriber basis is relatively low.

Application Layer

The Application layer provides interaction with the end user and is largely responsible for the user experience. It includes functions that: 1) discover content through multiple navigation options such as user interfaces (UI), channel guides, interactive search, recommendation engines and social networking links; 2) consume content by providing applications for video streaming, video on demand (VOD) and network DVR (nDVR) consumption; and 3) provide companion applications which enable user interaction in conjunction with media programs such as interactive chat sessions.

Applications are typically implemented in software running on servers in the data center with a thin client application on the consumer device. The applications may be provided by the service provider, the device provider or a third party. Costs associated with the applications layer are thus shared between these entities. On a per subscriber basis these are relatively small as they are amortized over a large number of subscribers.

Services & Control Layer

The Services & Control Layer is responsible for assigning resources within the network and for enforcing rules on content consumption that ensure compliance from a legal or contractual perspective. It includes functions that manage: content work flow from ingest through to delivery; the resources needed to ensure content is delivered to users when requested; and subscribers and devices to ensure that content is delivered to authorized consumers in the required format.

The Services & Control Layer is implemented as a set of software applications running on servers in the service provider network. These applications are typically licensed on a per subscriber or per session basis. Thus costs are a combination of hardware platform and software licensing. The basic control components required include: workflow and session management, DRM control, and resource management.

Media Infrastructure Layer

The Media Infrastructure Layer is responsible for video content delivery from the content provider to the consuming devices over the access network. This includes acquiring content from satellite or terrestrial sources (as either program streams or files), encoding it for ingest into the system and processing the content to prepare it for delivery. This is where the heavy video processing occurs and functions such as trans-coding, multiplexing, advertising insertion, encryption and publishing to a content delivery network (CDN) are found. This layer must also deliver the content to the target device through mechanisms such as web servers, CDNs, and streaming servers.

Costs for content reception and encoding scale on a per content stream basis. The content is shared across many subscribers so that the per subscriber cost is low. Packaging costs may scale based on content streams for a pre processing model or on subscribers if a just-in-time model is used. The choice between these is based upon a trade-off between packaging, storage and transport costs [PACK]. CDN costs do scale on a per subscriber basis.

RELATIVE END TO END INFRASTRUCTURE COST

The relative end to end cost of delivering IP video to a subscriber includes contributions from all of the components mentioned above and each component can have a wide range of variability. The Application Layer and Services & Control Layer products tend to be software on standard server platforms in a data center where costs are shared over a very large number of subscribers. The Media Infrastructure Layer is the component that contains the specialized hardware products and is where most of the operator investment occurs. Rather than attempt a detailed investigation of all of these components, the focus of the paper is on how changes in the network access pieces of the Media Infrastructure Layer impact the cost model.

Cable modem termination system (CMTS) ports to date have been deployed to provide high-speed data (HSD) and voice services. CMTS costs on a per subscriber basis have been relatively low. This cost point has been possible because HSD services could be heavily over-subscribed. IP video has a very different service model and cannot be over-subscribed to the same extent. A single CMTS channel can support anywhere from a half dozen high definition (HD) to a couple dozen smaller active video streams depending on the encoding rate used. The ratio of high definition to standard definition content now becomes very important. At historical CMTS pricing points, this translates to an order of magnitude of $100's per IP video stream.

Each of the components above is highly configurable which can result in wide variations in the end to end cost analysis. To understand the relative costs of these components required a nominal use case based on data from actual products and bid responses. This is shown in Figure 2 on a cost per active video user basis. For this example, the CMTS cost is roughly ten times the costs ascribed to the other major components. It is clearly the most significant cost driver for IP video and will be the primary focus of the rest of the paper.

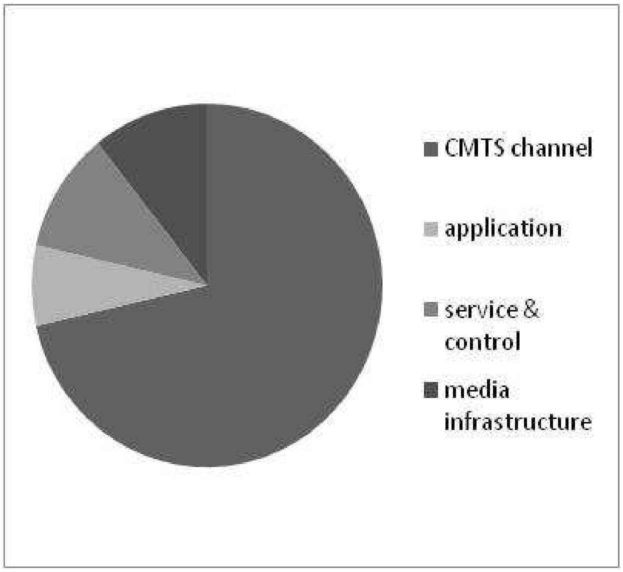


Figure 2 Per video user cost contributors

IP VIDEO & DOCSIS CHANNEL COSTS

CMTS Costs – Historical Perspective

DOCSIS is now 15 years old, having first been established in March 1997. Over that time, it has continued to evolve. In the early days, the cost per downstream channel was above $10,000. Early implementations had fixed downstream to upstream ratios (e.g. 2x8), so if more downstream bandwidth was needed, the system was burdened with the cost of more upstreams whether or not they were needed.

In addition to the fixed ratio, these early CMTS's were focused on offering a robust voice service for the operators. This introduced significant costs as these CMTS became carrier grade incurring the associated redundancy overheads.

Thanks to Moore's Law, these costs were reduced over time. Two architectural changes accelerated this trend. First, the DOCSIS 3.0 specification (D3.0) was developed and released. This laid the groundwork to enable multiple bonded channels per downstream port. At the same time, CMTS architectures shifted to decoupled architectures where upstream and downstreams could scale independently of each other. Some vendors chose a modular CMTS (M-CMTS) path for this while others implemented decoupled architectures within their Integrated CMTS (I-CMTS). As D3.0 was deployed, this helped to accelerate the reduction in cost per downstream channel as multiple channels were now implemented per port and the upstream burden per downstream channel was reduced.

So where are we today? Based upon recent research from Infonetics (Q4 CY2011), the revenue per downstream (channel) will decline in calendar year (CY) 2012 to approximately $1,600. After several years of significant reductions following the introduction of D3.0, the industry is starting to see price declines level out. Infonetics has forecasted that CY12 will see a 10% drop over CY11 which is substantially less than the previous two years.

As we move forward with unicast based IP video, it is very important to understand the cost implications for DOCSIS downstream channels going forward.

CCAP Disrupts DOCSIS Density & Pricing

Recent industry and CableLabs® efforts have defined a new specification called CCAP that is a high density combination of CMTS and edge QAM (EQAM) in a single unit. Current CMTS products may only support 4 or 8 channels per downstream port. The initial version of CCAP is defined to support 64 narrowcast channels per port, with a flexible channel mix between DOCSIS and EQAM. Future CCAP products may support 128 or even 160 channels per port, enough to fill the entire 1GHz downstream spectrum. Clearly, CCAP causes a disruptive shift in downstream densities, increasing by a factor of sixteen! With these densities, there will be a corresponding decrease in the cost per downstream channel. For IP video deployment, it is very important to understand how CCAP will affect access network costs.

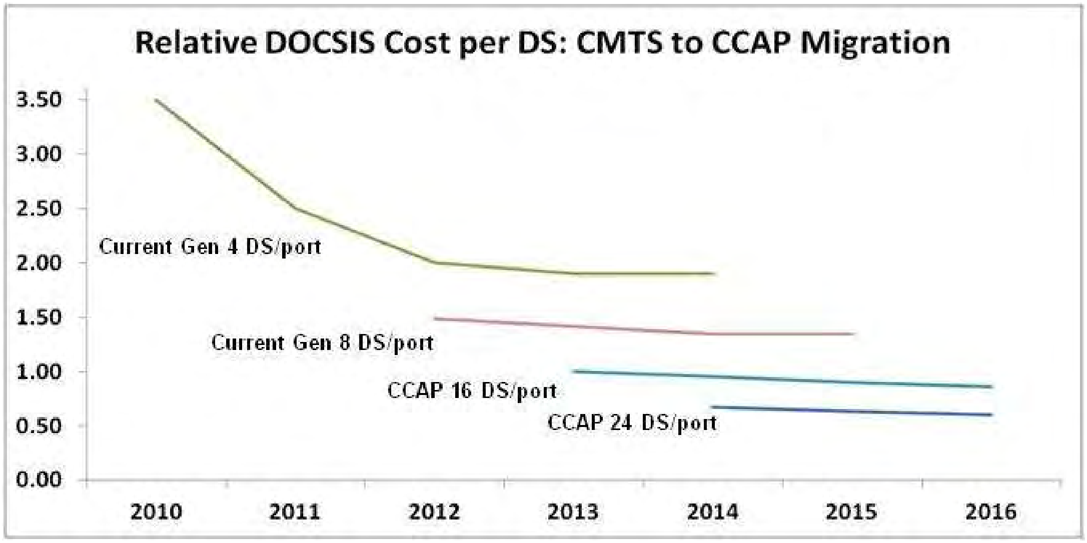


Figure 3 DOCSIS Downstream Cost

Initially, operators will only need a fraction of the CCAP capacity. Even if they wanted to deploy more channels, the spectrum required is a very scarce resource. For an operator to buy the full CCAP capabilities but only use a fraction of its capacity (e.g. 16 downstream channels) would cause a significant spike in the cost of downstream channels. CCAP would not be cost effective compared to current CMTS platforms. Therefore, vendors will need to license channels, similar to what is done today for high-density EQAM products. This allows CCAP products to be deployed while offering competitive downstream channel costs; vendors then defer revenue to a later time once additional channels are licensed and operators gain the benefit of deploying systems with longer lifetimes. Figure 3 above depicts the downstream channel cost trends over time for current CMTS with 4 and 8 downstream channels per port; then speculates where CCAP with 16 and 24 downstream channels per port might be positioned relative to current CMTS pricing.

To further explore this, Motorola developed an economic model for CCAP deployments around licensing algorithms. As discussed previously, a model where the full CCAP costs are paid up front will be difficult to justify on a cost per channel basis. On the other extreme, selling CCAP channels at the average price per channel based on a fully deployed product is also problematic. The system must be designed to support the full working load. If only a small number of channels are licensed to start, then vendors will lose money on initial deployments with no guarantees of future revenue. This would inhibit product development.

The ideal model required a licensing algorithm that would reflect the expected channel deployment. As referenced in [Howald], downstream capacity can be expected to continue at the 40-60% annual rate. Based on this along with an assumed starting point of 16 downstream channels per port, Table 1 shows how the downstream channel deployment is modeled.

Year	Total Downstreams	Incremental Downstreams
2013	16	-
2014	24	+8
2015	32	+8
2016	48	+16
2017	64	+16
2018	96	+32
2019	128	+32

Table 1 Downstream Growth

Note that this is reasonably close to the 50% growth per year that is often quoted.

Another factor that must be taken into consideration is that operators have a limited budget to spend in a given year. Infonetics forecasts show that CMTS revenue is only expected to grow 5% annually over the coming years while overall capacity above is growing at 50%. This implies that the CCAP downstream cost per channel must drop year over year (YOY) as larger number of channels are introduced in later years.

The results from our economic model are shown in Figure 4 and Figure 5 below. The baseline was 16 downstreams (DS) per port for the initial year and the average cost per downstream channel is shown for the sequence described in Table 1. Figure 4 shows the ratio with 16 DS being the 1.0 baseline. Figure 5 is interesting in that it plots the same data with a log scale. Even though Figure 4 shows each sequence getting progressively closer together, Figure 5 highlights that there is a roughly fixed percentage decrease YOY.

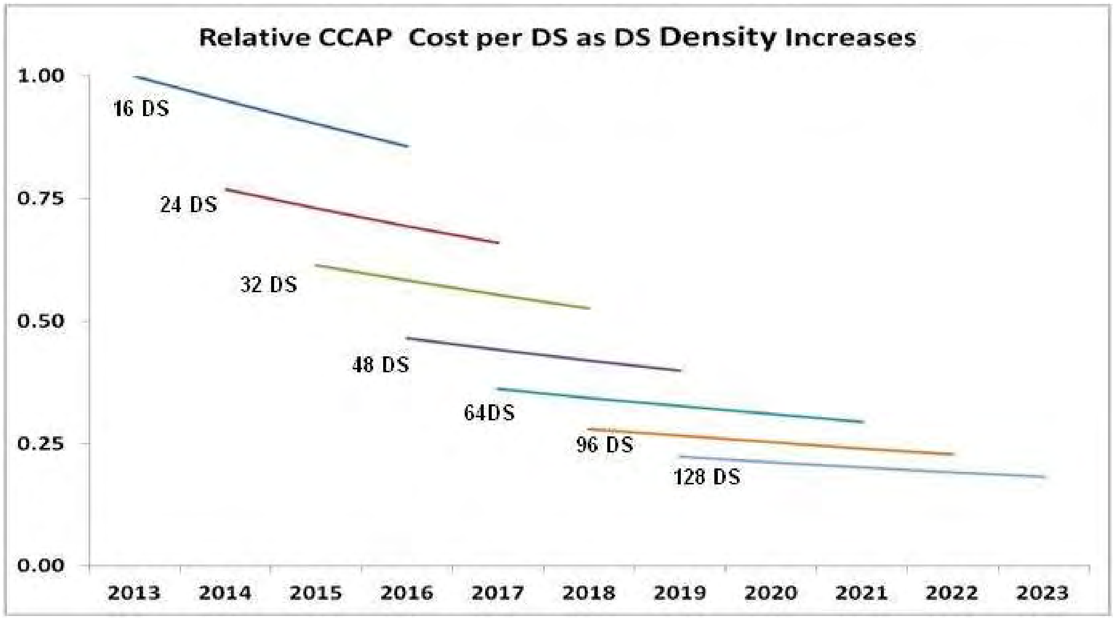


Figure 4 Cost Per DS at Higher Density

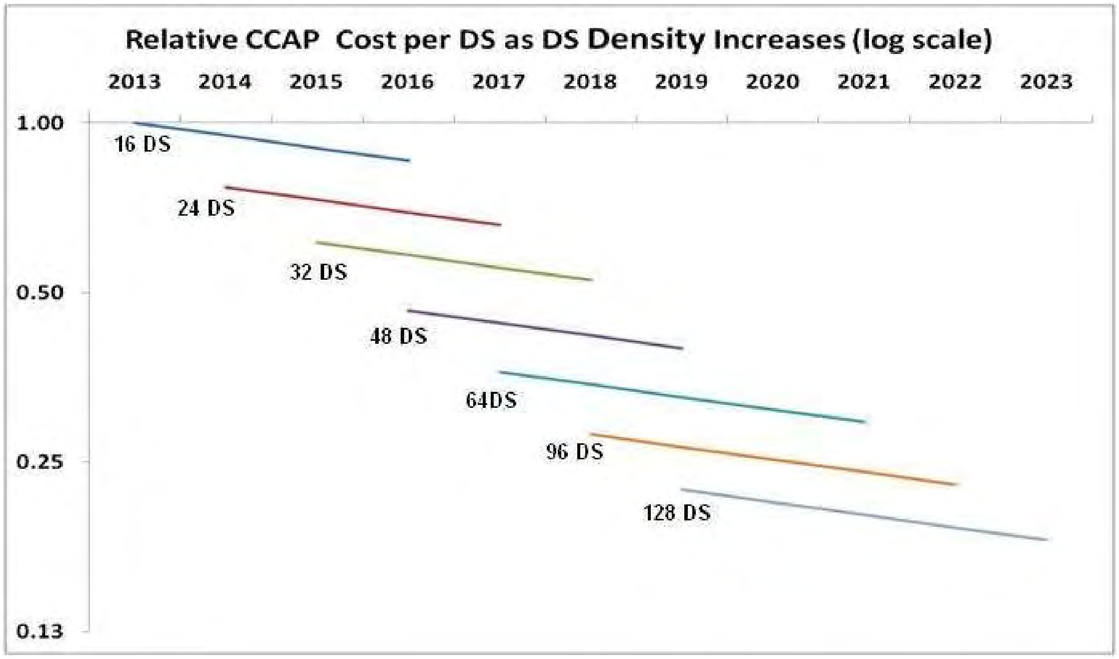


Figure 5 Cost Per DS at Higher density (Log Scale)

A licensing model like this is beneficial to both customers and vendors, assuming the initial starting point of 16 downstreams is sufficient to the vendor for initial installation and the YOY decrease in costs per downstream channel is sufficient to enable the operator to incrementally add channels in ever larger amounts within their budget.

Disclaimer: the above analysis is hypothetical and not based on any real products. It shows some possibilities for licensing algorithms that may be beneficial to vendors and customers. Every vendor may implement their own licensing algorithm and market conditions may cause these licensing algorithms to change over time.

As seen in Figure 4 and Figure 5, the economics around IP video deployment will vary over time. Costs will be higher initially but volumes will be lower. As IP video penetration ramps up, DOCSIS channel costs start to drop substantially.

Another important aspect is that IP video deployment is an incremental addition onto an existing DOCSIS HSD infrastructure. Therefore, it is critical to understand the incremental costs for downstream channels, not just the average costs which were previously discussed. This can be best explained by an example. Let's start with 16 downstreams as a baseline cost. Now suppose once there are 32 downstreams, the average cost per channel is 75% of the baseline cost per channel. In reality, the first 16 channels cost 100% and the incrementally added 16 channels were just 50% of baseline, giving a weighted average of 75%. So the incremental cost of 50% is the number that should be used for IP video economic analysis.

Taking the analysis further, CCAP leverages high-density EQAM technology. In the extreme, the incremental addition of a downstream channel could approach that of a high density EQAM product. Infonetics research shows that the average QAM cost was $163 in CY11 and forecasts that it will drop to $86 by CY16. Note that these are the average cost per QAM.

From our previous analysis, the incremental cost per channel could be substantially less. So it would not be a reach to suggest that the incremental cost per QAM several years from now may reach $40 per channel. This is an interesting number as the industry will approach $1 per Mbps for downstream bandwidth.

Working with this number for IP video economics, an IP video HD stream @ 5Mbps would therefore cost $5 to transport. Note that a few years ago this may have been $200-$400 using older CMTS downstreams. This radically changes the IP video economics. An updated chart with relative infrastructure costs is shown in Figure 6 below and shows the DOCSIS component has fallen from being the major cost contributor to become comparable to the other elements in the total cost. At this point other components become just as significant to the overall cost model.

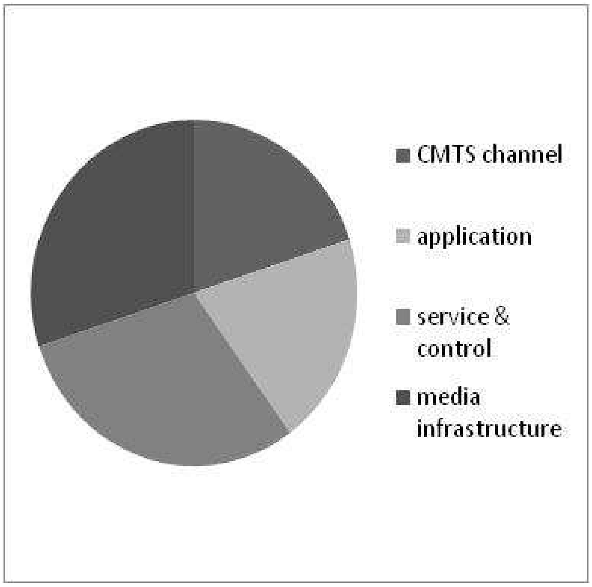


Figure 6 Post CCAP Cost Contributors

OTHER COST CONSIDERATIONS

CDN Options

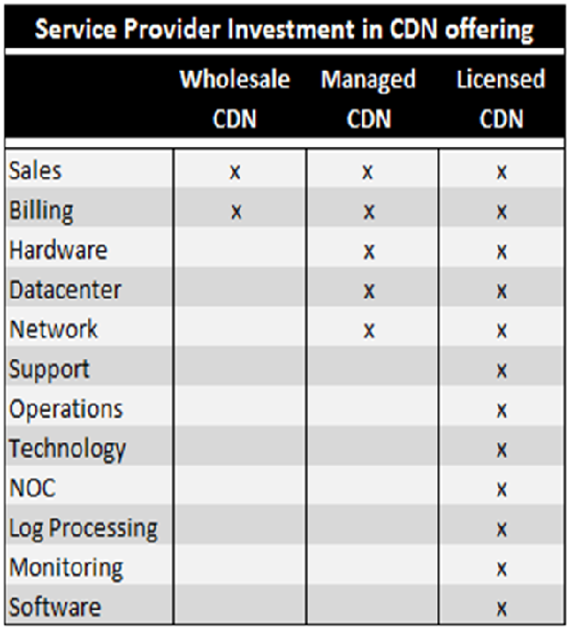
As operators migrate to IP video services using ABR, they will be able to leverage internet CDN technology for video delivery. There are a wide range of options to achieve this with a corresponding range of costs.

Initially many operators may purchase CDN services from one of the worldwide CDN providers. Eventually an operator may enter into a wholesale relationship with that CDN provider in order to resell CDN capacity directly to content providers and web site servers. This may allow the operator to extend their brand to the CDN services as well.

A possible next step in the CDN progression would be to install a managed CDN. In this step CDN nodes are added inside the service provider network but are still managed by the CDN provider. This allows the service provider to deliver content internally on their own nodes and network while still leveraging global access through the CDN partner company. The service provider minimizes operational expenses (OPEX) since the CDN partner still manages the internal CDN.

Finally, the service provider can install a licensed CDN. Equipment and software are deployed on the service provider's network and the provider assumes responsibility for operations and support. At this stage, the service provider can participate in a federated CDN exchange with other CDNs to deliver content outside their own CDN.

Table 2 shows the various functions associated with each of the three approaches. From a cost perspective, the wholesale approach requires the least amount of up-front investment but it is also the most expensive on a per-bit-delivered basis. Each step then requires more investment from both a capital expenditure (CAPEX) and OPEX perspective, but continues to result in lower costs for delivering each bit of content.



Service Provider Investment in CDN offering

	Wholesale CDN	Managed CDN	Licensed CDN
Sales	x	x	x
Billing	x	x	x
Hardware		x	x
Datacenter		x	x
Network		x	x
Support			x
Operations			x
Technology			x
NOC			x
Log Processing			x
Monitoring			x
Software			x

Table 2 Service Provider CDN Options

Transcoder and Storage Trade-offs

For linear television service, there is traditionally no storage costs associated with it. The content is encoded/transcoded, prepared and delivered to the consumer. With the new world of IP devices arriving, operators will want to go beyond simple linear television service to these devices and offer the ability to time shift. Consumers have become accustomed to their DVR for the television screen and will demand the same service for their IP devices. This will create a need for network based or "cloud" DVR services (nDVR).

Some current legal rulings based on existing content contracts require that nDVR content have a unique copy for each subscriber that records it. Other services offered today with re-negotiated content agreements allow single copy storage provided the fast-forward feature is disabled. The relative cost impact of nDVR is affected dramatically by the ratio between these.

Multi-rate ABR also exasperates the problem since a unique copy of a piece of content must now be stored in multiple bit rate formats. An example of this cost impact is shown in Figure 7.

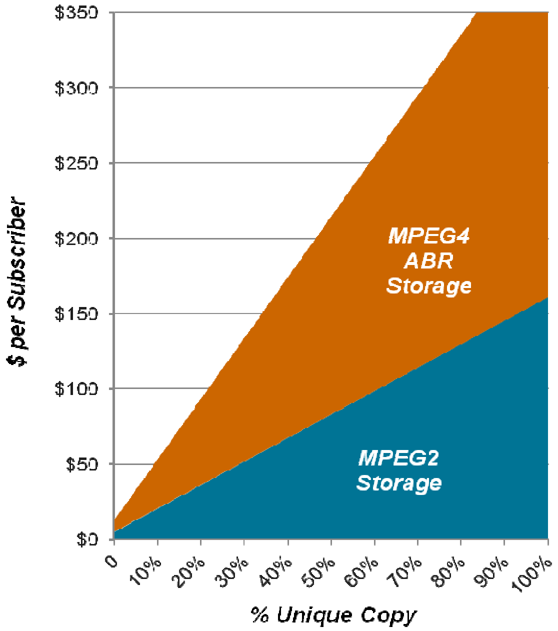


Figure 7 Storage Costs – nDVR

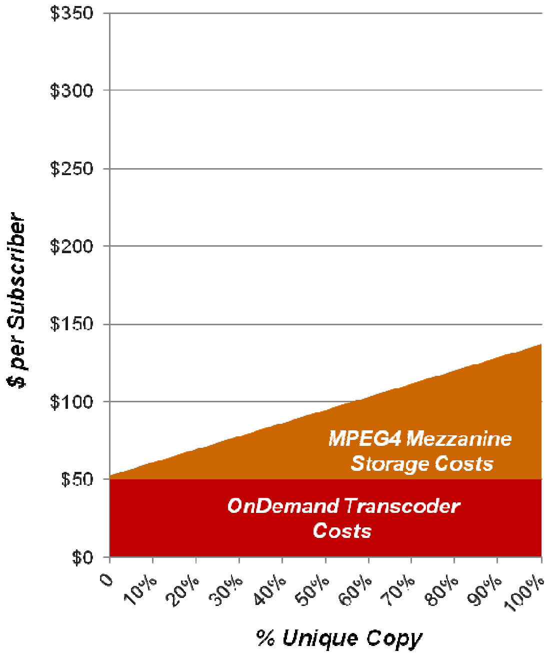


Figure 8 Transcoder vs. Storage

An alternative approach is to store a limited number of mezzanine formats in the nDVR storage and then transcode the content to the appropriate ABR bit rate on the fly when it is being viewed. Figure 8 shows an example of how costs may be impacted.

This creates a tradeoff between storage costs and transcoders costs that is constantly shifting. Many factors go into this analysis and the on-demand transcoder costs can vary significantly. This is an area where a disruptive change in transcoder costs could significantly change the landscape.

SPECTRUM MIGRATION STRATEGIES

Another very important aspect to IP video migration is finding sufficient spectrum. Some operators have already made more spectrum available by recovering analog TV channels using digital TV terminal adapters (DTA) while other operators have upgraded their HFC to 1GHz or used switched digital video (SDV). This available spectrum is being gobbled up today as more HD content is deployed, VOD requirements continue to increase and HSD services continue to grow at 50% annual rates. So there may still be a need for additional spectrum to ramp up IP video services with a corresponding economic impact.

Early Transition Plans

One way to significantly reduce spectrum requirements is to convert legacy MPEG-2 linear TV to IP video in a home gateway device. To support ABR devices in the home requires a transcoder in the home gateway device. Simple stand-alone devices are available today that accomplish this. This is an excellent approach for early deployments as it has almost no impact on infrastructure costs for rolling out linear IP video services. It also requires no new spectrum as this home gateway device appears as an STB to the system.

The next step in this migration is to introduce hybrid video gateways that also incorporate transcoding technology. These perform the same IP video conversion for linear TV described above for delivery to multi-screen IP devices. The video gateway also has the advantage that it is the single point of entry for video services and allows IP STBs to be deployed elsewhere in the home behind it. These devices can also operate as IP devices and are pivotal in the transition to an all IP system. As above, it can have a minimal impact on infrastructure costs to start and allows the operator to grow its IP video infrastructure at their own rate.

A detailed discussion of the home gateway migration is given in [CS_2012].

Complete Recovery of Legacy Bandwidth

The previous discussion on home gateway migration plans helps the operator begin the IP video transition. However, the end game is to eventually get to an all-IP system. Legacy MPEG digital TV services may continue to consume 50% to 80% of the available spectrum. Regardless of which path the operator took to free up spectrum, eventually they will need to install switched digital video (SDV) to reclaim all of the legacy bandwidth.

Adding SDV to the mix also increases the need for narrowcast QAM channels. This plays well into the previous CCAP analysis in this paper. Also, as the mix between legacy and IP subscribers change, an operator will need to re-assign SDV bandwidth to IP video bandwidth. This is also well suited for CCAP. A more detailed analysis of the SDV migration is in [Ulm_NCTA_2012].

CONCLUSION

Operators must deploy unicast ABR video to remain competitive. The infrastructure costs of providing this service are currently dominated by the cost of the downstream DOCSIS channels needed. With the development and deployment of high density CCAP platforms the cost per downstream is expected to fall dramatically, enabling the operator to deploy sufficient channels to meet demand while remaining within budget.

In the early days of CCAP deployment, not all channels will be used creating a potential disconnect between the capacity of the platform and the cost per channel deployed. The paper offers a framework for licensing which should be mutually acceptable to vendors and operators to circumvent this hurdle.

With the DOCSIS channel cost reduced significantly other cost components become more significant. ABR video is conveniently and cost effectively delivered via a standard internet CDN. A range of options to implement this are available from complete outsourcing to in house each offering different trade-offs in OPEX and CAPEX.

As nDVR is deployed into the ABR infrastructure another set of trade-offs will be required. For each recorded asset the multiple bit rate versions required can either be created at record time or created at play out from a recorded mezzanine format. In this case the trade off is between storage capacity and real time transcoding costs.

Operators will need to find the downstream bandwidth required for IP video delivery. Several options are available to do this. Home gateways may be used for early deployments in parallel with legacy MPEG-2 video. As the move to all IP video progresses the amount of MPEG-2 channels will decrease so that they can be economically delivered using SDV. CCAP is well suited for this.

The operator has multiple choices to make but will be able to deploy the technology required to remain competitive in an IP video environment.

REFERENCES

[SAND]	Global Internet Phenomena Report Fall 2011; Sandvine
[VNI]	Cisco® Visual Networking Index (VNI) 2011
[DIBA]	M. Patrick, J. Joyce, “*DIBA – DOCSIS IPTV Bypass Architecture*”, SCTE Conference on Emerging Technology, 2007
[CS 2012]	J. Ulm, G. White, *“Architectures & Migration Strategies for Multi-Screen IP Video Delivery”*, SCTE Canadian Summit, March 2012.
[Ulm NCTA 2012]	J. Ulm, J. Holobinko, *“Managed IP Video Service: Making the Most of Adpative Streaming”*, NCTA Technical Sessions, May 2012.
[PACK]	Unified Content Packaging Architectures for Managed Video Content Delivery, Santosh Krishnan, Weidong Mao, SCTE Cable-Tec Expo 2011
Howald 2011]	Dr. Robert Howald, “*Looking to the Future: Service Growth, HFC Capacity, and Network Migration”*, 2011 SCTE Cable-Tec Expo Capacity Management Seminar,, Atlanta, Ga, November 14, 2011
Howald 2010]	Dr. Robert Howald, “*Boundaries of Consumption for the Infinite Content World”*, 2010 SCTE Cable-Tec Expo, sponsored by the , New Orleans, LA, October 20-22, 2010
[Howald 2012]	Howald, Ulm, *“Delivering Media Mania: HFC Evolution Planning”*, SCTE Canadian Summit, March 27-28, 2012, Toronto,

ABBREVIATIONS AND ACRONYMS

CCAP	Converged Cable Access Platform
CDN	Content Delivery Network
CMTS	DOCSIS Cable Modem Termination System
COTS	Commercial Off The Shelf
CPE	Customer Premise Equipment
DOCSIS	Data over Cable Service Interface Specification
DRM	Digital Rights Management
DVR	Digital Video Recorder
EAS	Emergency Alert System
EQAM	Edge QAM device
Gbps	Gigabit per second
HFC	Hybrid Fiber Coaxial system
HSD	High Speed Data; broadband data service
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
nDVR	network (based) Digital Video Recorder
OTT	Over The Top (video)
STB	Set Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VOD	Video On-Demand

Author Index

BITS, BIG SCREENS, AND BIOLOGY

Dr. Robert L Howald and Dr. Sean McCarthy
Motorola Mobility

Abstract

High definition television (HDTV) has dramatically improved the consumer viewing experience. As such, despite its hunger for precious bandwidth, increased HD programming continues to be a key industry objective. However, as evidenced by the exhibits and technology on display at the Consumer Electronics Show (CES) this past January, today's HD, is just a step in the progression of consumer video. In addition, today's video processing and delivery is also undergoing significant change. In this paper, we will explore new generations of HD video and the innovative enabling technologies that will support them. We then roll-up the components and project their impact to network architecture.

Specifically, we will consider advanced formats, beyond just emerging 1080p60 (blu-ray) HD. Recognizing the expectation of a very long HFC lifespan, we will quantify how QFHD (aka 4k) and even proposed "Super Hi-Vision," or UHDTV, stack up for consumer services. We will assess practical and human factors, including those associated with HD-capable second screens, such as tablets. We will quantify physiological variables to the optimization of the video experience, such as personal through immersive screen sizes, viewing environment, and high frame-rate television.

On the encoding side, we discuss H.265 High Efficiency Video Coding (HEVC) against its own "50%" objective. And, just as we considered human variables associated with the user experience, we can take advantage of human biology to deliver the highest perceived quality using the smallest number of bytes. Using new signal processing models of the human visual system (HVS), the ultimate arbiter of video quality, a unique combination of bandwidth efficiency and high perceived video quality can be achieved. This technique, called Perceptual Video Processing (PVP) will be detailed, and its impact on video quality and bandwidth quantified.

In summary, we will evaluate long-term network prospects, capturing the potential trajectory of video services, innovative encoding techniques, emerging use cases and delivery, and shifting traffic aggregates. In so doing, an enduring network migration plan supporting multiple generations of video and service evolution can be projected.

INTRODUCTION

Decades of broadband growth and an ever-increasing range of video services has given operators a sound historical basis upon which to base future growth trends, which is critical for business planning. Service growth and subscriber satisfaction with the portfolio of media delivered to them provides new revenue opportunities. To meet these demands, key decisions must be made for upgrading hubs, homes and the access networks. The prevailing MSO approach has been a very successful pay-as-you grow approach, capitalizing on technologies as they mature and as consumer demands require. This has worked extremely well because of the latent HFC capacity, which incrementally was mined as necessary by extending fiber, adding RF spectrum, incorporating WDM optical technologies, and delivering digital and switched services.

As IP traffic has grown aggressively, video quality has also moved ahead, albeit at a more gradual pace. The appetite for HD is being fed at this stage of the evolution, but the HD lifecycle itself has only just begun. As cable systems deliver 720p and 1080i formats, the ability to support and deliver 1080p quality already exists in the CE and gaming worlds. Flat panel televisions continue to become larger, more capable, and lower cost. Their size already is breaching the boundary of where a “normal” viewing distance would benefit from a yet higher quality video signal. 2k and 4k (Quad Full HD or QFHD) formats have entered the conversation and the demonstration rooms. These formats are being explored and seemingly will inevitably lead to a new service offering. Beyond QFHD is the Ultra-High Def (UHDTV, 4x QFHD)) format, or Super Hi-Vision, invented by NHK in Japan in the mid-1990’s. At that time, it was foreseen by NHK to be a consumer format in the 2030 time frame.

EVOLUTION OF VIDEO SERVICES

Spatial Resolution

With the advent in particular of HDTV, development of QUAD HD (2x in each dimension) and UHDTV, the video and CE industries have a strong understanding of the relationship among resolution required, screen size, and viewing distance.

Just as visual acuity is measured and referenced to object sizes at defined distances, the display size and placement relative to the viewer is a key piece of the resolution requirement equation. Figure 1 is a straightforward way to see how these factors interact [40] based on recommendations provided by multiple professional organizations, home theatres experts, and retail manufacturers. Generally, for a fixed resolution (linear trajectories on the plot), a larger screen size is best viewed further away. For a fixed screen size, higher resolutions are best viewed by sitting closer to allow for the full benefit of the increased detail on the display. Finally, for a fixed distance from the display and the higher the format resolution, the larger the screen size should be.

As a simple example, a 50 inch screen, if viewed more than 20 ft away or greater, will begin to lose the benefit of HD at 720p, and provide an experience more akin to Standard Definition 480p. Sitting too close, such as 5 ft away on a 100” 1080p screen, threatens quality due to distinguishing of pixels. This chart thus also explains the increased pixel count of UHDTV based on a 100” display recommendation and wider viewing angle (closer).

The guidelines come from different organizations and retailers, and while they tend to cluster around similar recommendations, they are not in complete agreement. This is generally due to the varied perspectives of the organizations, such as, for example, what sells more TVs. The range of recommendations varies from about 1.5x-2.5x of display size for viewing HD content, with the lower end corresponding to 1080 resolution.

The recommendations are also correlated to an assumption about visual acuity as it relates to the ability to resolve the image detail. They are also associated with viewing angle considerations. For example, the recommended optimum fields of view are given as about 30° (SMPTE) or 40° (THX) in the horizontal plane. In the vertical plane, simpler guidelines are designed around avoiding neck strain, and so describe maintaining at least a 15° vertical field of view. The maximum recommended, beyond which neck strain is a risk, is a 35° viewing angle.

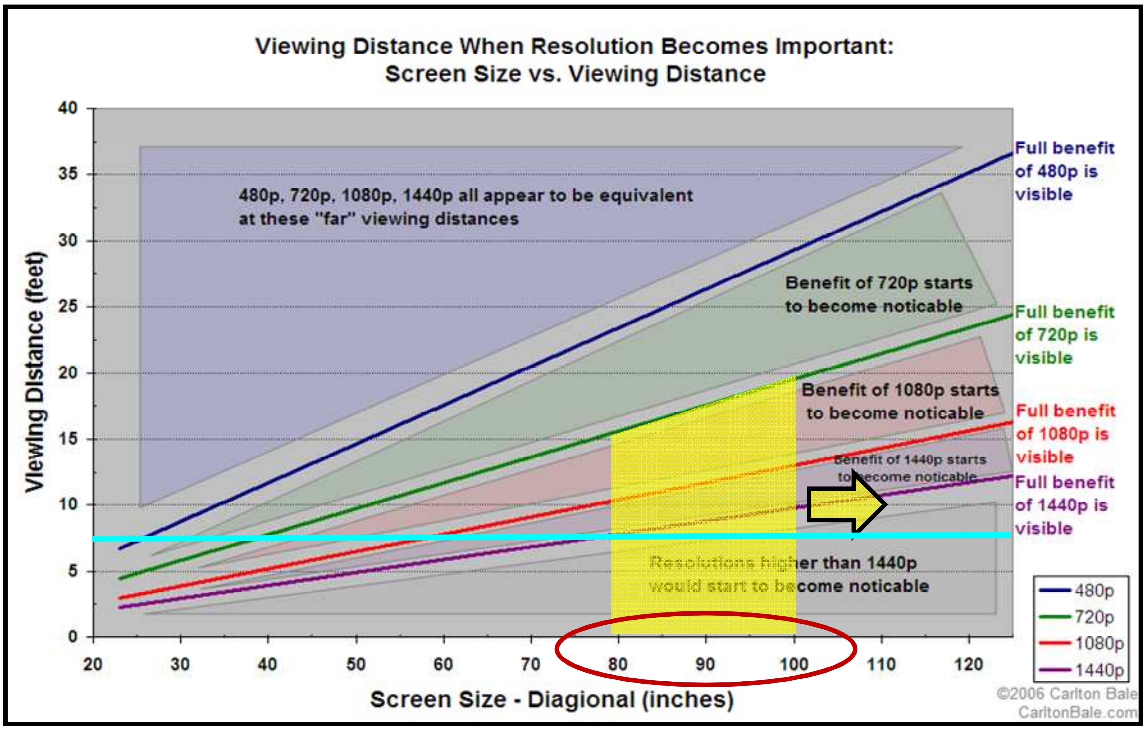


Figure 1 – Screen Size, Viewing Distance, and Spatial Resolution

Let's ponder modern display capabilities. Consider the bottom right corner of Figure 1, shaded yellow. A typical viewing distance in the home today is about 7.5-9 feet, which certainly has been driven in part by historical screen sizes. It is not surprising for anyone who has visited a big box retailer recently that flat panel screens are available now at ever-increasing sizes, such as those shown in the shaded yellow range of Figure 1. At 7.5 feet distance (light blue line), "only" a 55" screen could show perceptible benefits for resolutions better than 1080p (light blue line crosses red line). At 80", flat panels have fully breached the 2560x1440 resolution threshold, sometimes referred to as Extreme HD (4x 720p HD resolution) in the gaming world. The next stop beyond this is QFHD at 3840x2160p. Based on this figure, there is potential viewing value for this format screen size and larger.

Note that UHDTV, was viewed as a 100 inch screen, but also viewed at only about 1 meter (3.3 feet). The intent was to generate the feeling of immersion. Studies by NHK concluded that feelings of discomfort often associated with immersive viewing such as IMAX level off with screen size at a certain point. In the case of UHDTV, the angle at which this occurs is for 80 inch screens. Therefore, a screen fully 100 inches is not expected to present an increased probability of discomfort, but yet yields the level of immersion and video quality desired in the experience.

Now consider Figure 2. Not only do larger primary screens translate into the need for better spatial resolution, our secondary screens also have gotten larger, simultaneously more portable, *and* capable of high quality video such as HD. The explosion of tablets has put an entire new generation of high-quality video capable screens literally at our fingertips for deployment virtually anywhere relative to our viewing perspective.

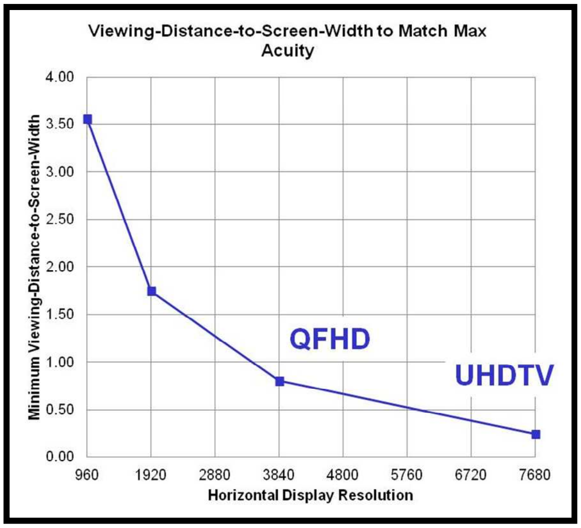


Figure 2 – Screen Size, Distance, and Resolution – Mobile Viewing

In the figure above, it is easy to see how the 1920x1080 resolution can be improved upon for reasonable viewing conditions. For a 10" Motorola Xoom tablet, for example, if the screen is about 17" away, its spatial resolution can be perceptibly improved with a higher resolution format. It is not hard to envision this scenario, for example, on an airplane or with a child in the backseat of a car.

The case for full QFHD or UHDTV on the 10" tablet would be difficult to make based on this figure without some other accompany variables. Nonetheless, clearly screen sizes and portability in this case are combining to change the paradigm of mobile viewing environments far, far away from the legacy of QVGA resolution at 15 frames per second.

Dynamic Resolution

The term "dynamic resolution" refers to the ability to resolve spatial detail of objects in motion. The 30 Hz (interlaced), 50 Hz, and 60 Hz frame rates have origins in AC line rates, and thus are not scientifically tied to video observation and testing. They simply exceeded what was known at the time about 40 Hz rates causing undesirable flicker. Most early analysis on frame rate was to ensure that motion appeared realistic (seamless), as opposed to a sequence of still shots. There was less focus on eliminating other artifacts of motion, such as smearing effects. Yet, as spatial resolution has improved, temporal resolution has not. Interlaced video itself is a nod to the imbalance of motion representation – exchanging spatial resolution for a higher rate of image repetition to better represent motion than a progressive scanning system of the same bandwidth.

The above frame rates have since become embedded in tools and equipment of the production, post-processing, and display industries, and so, with respect to frame rate, we are hostages to the embedded infrastructure and scale of change that would be required to do anything else. As such, HDTV standards today are based on the 60 Hz interlaced or progressive frame rate. It has been suggested [1] that with larger and brighter displays of higher resolution, the frame rates in place based on practical implementation limitations of the 1930s era ought to be reconsidered, of course while recognizing a need to maintain some level of backward compatibility.

As displays become larger and of higher resolution and contrast, the challenges to effectively displaying motion increases, because the edges to which movement is ascribed are now sharper. What is optimal? There is not a firm answer to this question. The human visual system streams video continuously in a physiological sense, so the question is around the processing engine in the brain. Various sources describe tests where frame rates of 100-300 fps show perceived improvements compared to 60 fps [1, 41]. The difficulty of performing this type of testing – content and equipment – limits how much has been learned. There are potentially positive encoding implications to these higher frame rates. Intuitively, more rapidly arriving frames ought to be consistent with better coding efficiency, as it is likely that there is less variation frame-to-frame.

We will not consider any changes to frame rate beyond interlaced/30 to progressive/60. But this is a variable to keep an eye on as larger screens and live sports viewing collide.

Formats and Bandwidth Implications

High Definition has had a major impact on the industry in multiple ways. On the positive side, the Quality of Experience (QoE) delivered to the consumer is tremendously improved. HD has enabled cable operators to strengthen the service offering considerably. And, like the DVR, HD has very much the "once you have it, you never go back" stickiness to it.

Conversely, while HD services certainly act to increase revenue, they also create a significant new bandwidth burden for the operator. Whereas 10-12 standard definition (SD) programs can fit within a single 6 MHz QAM bandwidth, this number drops to 2-3 HD programs in a 6 MHz QAM. This loss in efficiency is compounded by the fact that HD today represents a simulcast situation – programs delivered in HD are usually also transmitted in the SD line-up. For all of the subsequent analysis, we will base MPEG-2 SD and HD program counts per QAM on averages of 10 SD/QAM and 2.5 HD/QAM. Obviously there cannot be a fractional number of programs n a QAM slot. The 2.5 assumes that for MPEG-2 encoded HD, an operator may chose 2 or 3 in a 6 MHz slot based on content type, and the QAMs are equally split with both.

Perhaps most worrisome with respect to bandwidth is that current services are basically HD 1.0. Only the first generation of formats are deployed – 1280x720p and 1920x1080i. The improvement over SD is so vast that it is easy to wonder what could possibly be the benefit of even higher resolution. However, as we showed in Figure 1, it is relatively straightforward to show how the continued advancement of display technology at lower and lower costs, in particular consumer flat panels, leads to reasonable viewing environments where resolution beyond 1080-based systems would be perceptible. In addition to the flat screen scenario, similar analysis in Figure 2 showed similar conclusions for"2nd screen" tablet viewing. All modern tablets support HD quality viewing. Coupled with realistic use cases that are likely to include close viewing distances, higher resolution scenarios may add value here as well.

We will consider the effects of two next generation video formats on the HFC architecture's ability to support them – QFHD and UHDTV. QFHD has had prototype displays being shown since approximately 2006 and has entered the conversation as the big box retailers now routinely display 80" screen sizes. Analyst projections have placed QFHD in the 2020 time frame for deployment timeframes. A comparison of these two formats against standard HD, and other formats, is shown in Figure 3 [19].

Note that QFHD works out to 4x the pixel count as 1080 HD, and UDHTV works out to 16x the pixel count. In each case, there is the possibility of higher bit depth (10-bit vs. 8-bit) as well, which translates into more bits and bandwidth. We will assume this is taken advantage of in the latter case only. As a result, we arrive at the following set of potential scaling factors, without any assumptions about possible latent compression efficiencies on top of conventional gains projections for new display formats.

SD to:

1080i – 4x
1080p – 8x
QFHD – 32x
UHDTV – 160x

It is of course premature to know precisely what compression gain may be available for advanced formats, since these enhanced formats are in their infancy. For now, we will rely on the resolutions to correlate with bandwidth, with the 8-bit to 10-bit pixel depth for UHDTV and the frame rate for p60 (doubling the information rate) as the only other variations quantified.

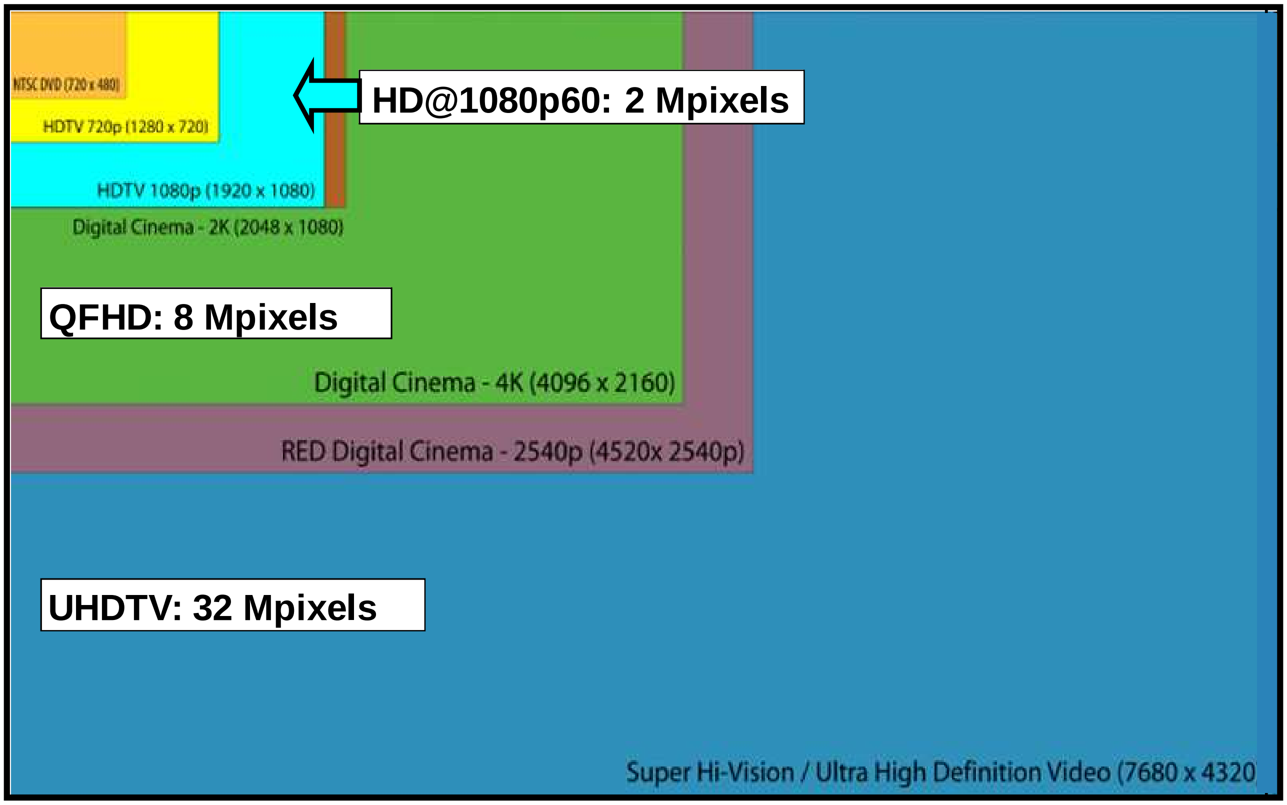

NTSC DVD (720 x 480)
HDTV 720p (1280 x 720)
HDTV 1080p (1920 x 1080)
HD@1080p60: 2 Mpixels
Digital Cinema - 2K (2048 x 1080)
QFHD: 8 Mpixels
Digital Cinema - 4K (4096 x 2160)
RED Digital Cinema - 2540p (4520x 2540p)
UHDTV: 32 Mpixels
Super Hi-Vision / Ultra High Definition Video (7680 x 4320

Figure 3 – Beyond High Definition Formats Comparison

VIDEO COMPRESSION – STILL ON THE MOVE

It may not seem like long ago, but it is nearly 10 years since the Advanced Video Coding (AVC) [27, 30] international standard was completed in 2003. AVC – also known as H.264 and as MPEG-4 part 10 – has been a remarkable success. It has enabled IPTV and HDTV to take hold and grow commercially. It has enabled Blu-Ray video quality at home. And it has been powering new models of delivering digital video over the internet. AVC and its equally successful predecessor, MPEG-2, are expected to continue to play an important role in the digital video economy for many more years, but they'll be soon joined by a new entrant to the international standard portfolio -- High-Efficiency Video Coding (HEVC) [6, 10, 18, 22, 31, 32].

In many ways, HEVC is a close cousin to AVC. Both are of the same genus of hybrid block-based compression algorithms that incorporate spatial and temporal prediction, frequency-domain transforms, data reduction through quantization, and context-sensitive entropy encoding. Where HEVC stands out is in the wealth and sophistication of its coding tools, and in its superior compression efficiency.

Figure 4 captures the state of the set of core MPEG compression standards in the context of their lifecycle.

Efficiency

First and foremost for any compression standard is the simple question of how much more efficient it will be at compressing video streams. HEVC aims to double the compression efficiency of its AVC predecessor. AVC itself doubled compression efficiency compared to MPEG-2. That means that a consumer quality HDTV program delivered using 16 Mbps today with MPEG-2 (like a cable TV QAM channel supporting 2-3 HD channels) would need only about 4 Mbps using HEVC. As we will see in subsequent analysis, it also means that we might reasonably expect to be able to deliver Super HD (4kx2k) over the bandwidth we use today for regular HDTV, enabling yet another generation of enhanced video services.

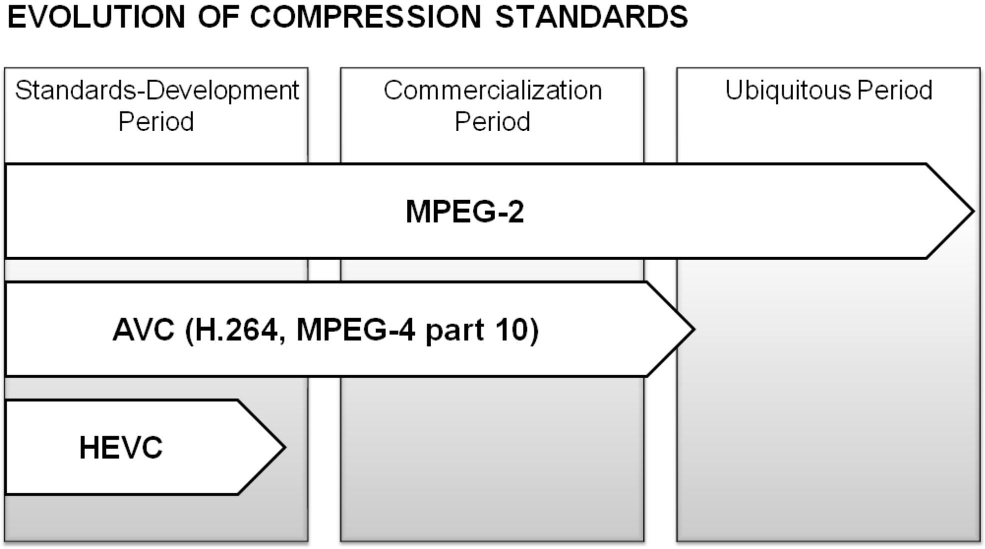

EVOLUTION OF COMPRESSION STANDARDS
Standards-Development Period
Commercialization Period
Ubiquitous Period
MPEG-2
AVC (H.264, MPEG-4 part 10)
HEVC

Figure 4 – State of Current Video Compression Standards

At the onset of the HEVC development process, the ITU-T and MPEG issued a joint call for proposals [33]. Twenty-seven proposals were received and tested in the most extensive subjective testing of its kind to date. Scrutiny of the proposals entailed 134 test sessions involving 850 human test subjects who filled out 6000 scoring sheets resulting in 300,000 quality scores. The conclusion [34] was that the best proposals yielded 50% bit rate savings compared to AVC at the same visual quality. The potential for another 50% gain launched the Joint Collaborative Team for Video Coding (JCT-VC), and HEVC development formally got underway.

In late 2011, JCT-VC reported another series of compression-efficiency tests [35] using objective rather than subjective methods. Those test showed that HEVC had not yet hit the 50% mark with scientific certitude, but was very close and had excellent prospects for additional gain. Table 1 shows the results from objective tests comparing HEVC to AVC High Profile. The tests were conducted using various constraints to examine the efficiency of HEVC for several important potential use cases: broadcast such as over cable, satellite, and IPTV that need random-access features to support fast channel change and trick modes, low-delay applications such as video conferencing, and intra-only compression that uses only spatial prediction within each frame of video to support applications such as contribution-quality video.

The bit rate savings listed in Table 1 represent the point at which HEVC and AVC High Profile produce the same peak-signal-to-noise ratio (PSNR). Though PSNR can be a sometimes inaccurate metric of subjective video quality [25, 39], the data in Table 1 are consistent with the earlier extensive subjective testing [35] and are thus expected to be valid predictors. The data of Table 1 represent overall average performance of the various HEVC use cases for a wide range of resolutions from 416x260 to 2560x1600 [13]. It is clear from Table 1 that HEVC substantially outperforms AVC High Profile.

Other results from the JCT-VC report on objective tests are displayed in Table 2. These results provide insight into how HDTV might differ from mobile devices with regard to HEVC efficiency.

Table 1 - Compression Efficiency of HEVC compared to H.264/MPEG4 part 10 AVC

NOTE: Relative Compression Efficiency is calculated as 1/(1 - Bit Rate Savings)

Example Use Case	Encoding Constraint	Bit Rate Savings	Relative Compression Efficiency
Broadcast	Random Access	39%	164%
Video Conferencing	Low-Delay	44%	179%
Contribution	All-Intra	25%	133%

.

Table 2 – Current Compression Efficiency of HEVC for HDTV and Smartphone

Display	Width	Height	Bit Rate Savings Compared to AVC High Profile	Relative Compression Efficiency
HDTV	1920	1080	44%	179%
Smartphone	832	480	34%	152%

Table 2 points out that HEVC's gains for HDTV resolutions are greater than for smartphone resolutions. They are also greater than the average over all random-access results shown in Table 1. These results hint that HEVC may become relatively *more* efficient for emerging resolutions beyond HDTV, such as 4K (4096 x 2048) and Ultra HD (7680x4320). If such proves to be the case, market forces might help accelerate deployment of HEVC as a way for operators and display manufacturer to offer new beyond-HD options to consumers.

It is important to note that both MPEG-2 and AVC improved significantly as they moved from committee to market. Even today, MPEG-2 and AVC continue to become more efficient as competition pushes suppliers to find new ways of improving quality and squeezing bits. The same dynamic is expected with HEVC. It should experience additional improvements, rapidly, when it emerges from the standardization process, followed by long-term, continuous honing through commercial competition. It is common in industry circles to project that HEVC will achieve its targeted doubling in compression efficiency – it is simply a matter of time.

For purposes of our subsequent analysis of HFC capacity and services, we will assume that HEVC will indeed ably achieve its 50% goal when commercially available.

Under the Hood

Some of the AVC efficiency gains were the result of new coding techniques such as context-adaptive binary arithmetic entropy coding (CABAC). Yet a large part of the gains came from making existing tools more flexible. Compared to MPEG-2, for example, AVC provided more block sizes for motion compensation, finer-grained motion prediction, more reference pictures, and other such refinements.

HEVC also gains its performance edge by using newer versions of existing tools. One of the most significant enhancements is that the concept of a macroblock has morphed into the more powerful concepts of Coding Units (CU), Prediction Units (PU), and Transform Units (TU).

Coding Units are square regions that can be nested within other Coding Units in a hierarchical quad-tree like manner to form an irregular checkerboard. The advantage is that smaller Coding Units can capture small localized detail while larger Coding Units cover broader more uniform regions like sky. The result is that each region in a picture needs to be neither over-divided nor under-divided. Avoidance of excessive segmentation saves bits by reducing the overhead of signaling partitioning details. Judicious subdivision saves bits because the details within each terminal Coding Unit can be predicted more accurately.

Prediction Units extend the "just-the-right-size" coding philosophy. Prediction Units are rectangular subdivisions of Coding units that are used to increase homogeneity – and thus predictability – within Coding Units. If a particular Coding Unit encompasses a region of grass and a region of tree bark, for example, an encoder might attempt to arrange the boundary between Prediction Units so it matches the grass-bark boundary as closely as possible. Together, Coding Units and Prediction Units create a quilt of more homogeneous patches that are easier to compress than regions of heterogeneous textures.

Transform Units are also subdivisions of Coding Units. The objective is to position and size Transform Units such that a picture is subdivided into mosaic of self-similar patches when viewed from the frequency domain. One of the dominant visual artifacts in MPEG-2 and AVC is the distortion that sometimes occurs near sharp edges and around text. This artifact is a result of performing a transform and quantization across radically different textures on either side of the edge. In HEVC, Coding, Prediction, and Transform Units work together to more precisely decouple the textures flanking the edge thereby reducing spillover and avoiding the visible defect.

Other coding tools also get a makeover in HEVC. Intra prediction supports many more directional modes to discriminate the angular orientation of lines, edges, and textures more exactly. Inter prediction has improved interpolation filters to yield higher quality motion vectors. And there are less costly ways of sending motion vector information to the decoder. HEVC also gains at least one new kind of loop filter targeted at improving both objective and subjective visual quality.

Not all the enhancements in HEVC are incremental. HEVC will be capable of delivering high-quality video to every conceivable device from the size of a thumbnail to a wall-filling 8k x 4k display in wide-gamut color palette that rivals the natural world. That is an opportunity for unparalleled consumer experiences.

Commercialization -- Profiles & Levels

Compression standards of the caliber of HEVC are complex amalgams of sophisticated algorithms and protocols. In the past, specific subsets of capabilities and features of MPEG-2 and AVC were organized into Profiles with Levels to aid commercial adoption and facilitate interoperability between vendors. It would be unsurprising if HEVC also adopted a family of Profiles, but at the moment the HEVC Committee Draft [37] specifies only Main Profile, which roughly corresponds to AVC High Profile.

Within the HEVC Main Profile, the Committee Draft does specify a number of Levels. Each Level corresponds to a maximum picture size (in terms of number of samples) and maximum pixel rate for the luma component. From these constraints, it is possible to indicate the minimum Level that would correspond to various consumer devices, as we do in Table 3 for smartphones; HDTV on tablets and flat panels at home; and next-generation beyond-HDTV displays.

Table 3 - How HEVC Main Profile Levels Might Correspond to Various Displays

	Example Format	Width	Height	Frame Rate	Minimum Level
Smartphones	QCIF	176	144	15	1
	CIF	352	288	30	2
	480p	854	480	30	3
	QHD	960	544	60	3.1
HDTV	720p	1280	720	60	4.1
	1080p	1920	1088	30	4.1
				60	4.2
Beyond HDTV	4K	4096	2160	30	4.2
				60	5.1
	Ultra HD	7680	4320	30	6
				60	6.1

Note that most smartphones and sub-HD resolutions would be supported starting at Levels 1 through 3, depending on the picture size and frame rate. Note that any Level above the minimum Level could be used. HD resolutions would be supported starting at Level 4. Beyond-HD resolutions would require at least Level 5 & 6 with one interesting exception. Super HD 4k x 2k resolution at 30 frames per second shares Level 4.2 with 1080p 60 frames per second. It may turn out that operators will be able to leverage Level 4.2 in the future to provide consumers with both 1080p60 sports content and Super HD 4k film content (24 frames per second).

Next Steps

The process of earnest creation of HEVC began in 2010 with a Call for Proposals (CfP). There have now been nine JCT-VC meetings in which approximately 200 attendees per meeting created and debated over 2000 input documents. In **February 2012**, JCT-VC issued a complete draft of the HEVC standard called the Committee Draft [37] which will be refined over the coming months. The Committee Draft also serves as a starting point from which to explore development of commercial HEVC products. The Final Draft International Standard is scheduled to be made available in January 2013 for formal ratification.

HEVC is well on its way. And, as we shall see in the next section, it will be an essential component of future advanced video services for cable operators, based on what we are able to project today for service mix, spectral constraints, and likely migration strategies.

TRAFFIC AND SPECTRUM

Dynamics of the Shift to IP Video

While video resolution affecting bandwidth requirements presents an enormous capacity challenge, it is not the only variable driving spectrum use. In addition to bandwidth growth of the video itself, the nature of the traffic aggregate being delivered is changing as well. There are many variables in play, virtually all of which are driving towards increasing unicast delivery of video content:

- More content choice
- Time-shifting
- Trick play expectations

- Network DVR (nDVR),
- Video capable IP device proliferation (tablets and smartphones)
- Shrinking service groups

And, of course, over-the-top (OTT) viewing from web-based content providers is already unicast delivery.

As a result of these shifts, the gains typically afforded by multicast capability, or bandwidth reclamation gains associated commonly with SDV architectures, begin to evaporate. Consider Figure 5 [23]. On the right edge of the curve, we can see by comparing the DOCSIS channel count required for delivery of unicast compared to multicast that for a large group of active users and predominantly linear content, there is significant, exploitable gain. This converts to important bandwidth savings. This has been the lesson of SDV widely deployed in HFC networks today. However, these deployment advantages are based upon the content choice and the size of service groups of the time. Today, as node splits occur, the growing use of a variety of IP clients consuming video, increased choice etc., the operating point on the curve shifts to the left.

The crosshairs in the figure (60% penetration x 60% peak busy hour viewing on a 500 hhp node) represents a reasonable operating point in a system outfitted with 200 HD and 200 SD programs available as switched services. There is clearly much less gain at this point, suggesting only a modest savings in exchange for the complexity of multicast. Some optimization steps may be taken to most efficiently allocate spectrum, but with an eye toward simplicity of architecture as well. This approach is shown in Figure 6 [23]. This diagram illustrates the concept of broadcasting the very popular content to take advantage of programming where simultaneous viewing is likely to occur regularly, optimizing use of bandwidth while maintaining simplicity in the architecture. Analysis in [23] suggests that the vast majority of gain, around 80-90%, occurs in approximately the first 20 programs.

Thus, a combination of broadcast and unicast may be the end result of an IP Video system weighing the tradeoffs of efficiency and complexity. The modest loss of efficiency of “all unicast” in the figure is recovered through the use of a small tier of broadcast services. And, as service groups continue to shrink, there will be virtually no bandwidth efficiencies lost at all. This is illustrated in Figure 5, for example, for the 80 active IPTV viewers.

Next Generation Video Formats: Parallel Characteristics to IP Video

The dynamics commonly associated with 2nd screen viewing may also come to pass in the next generation primary screen video world. There is a large permutation of video formats for mobile viewing, being usurped today by high quality formats. The likely similar dynamic to emerge for primary screens is simply that new formats will get introduced well before other formats are retired. Historically, this would suggest a need to simulcast formats to ensure all customers have their video needs served based on what formats they can support on the TV sets in their home. With more formats arriving, and an overall accelerated pace of change, this could create a bandwidth Armageddon given the nature of the advanced formats relative to bandwidth consumption. However, as we shift into the IP Video world today built around 2nd screen compatibilities, we are developing and deploying tools for discovery and delivery of a large permutation matrix of formats and protocols based on the different capabilities and interfaces of IP client devices.

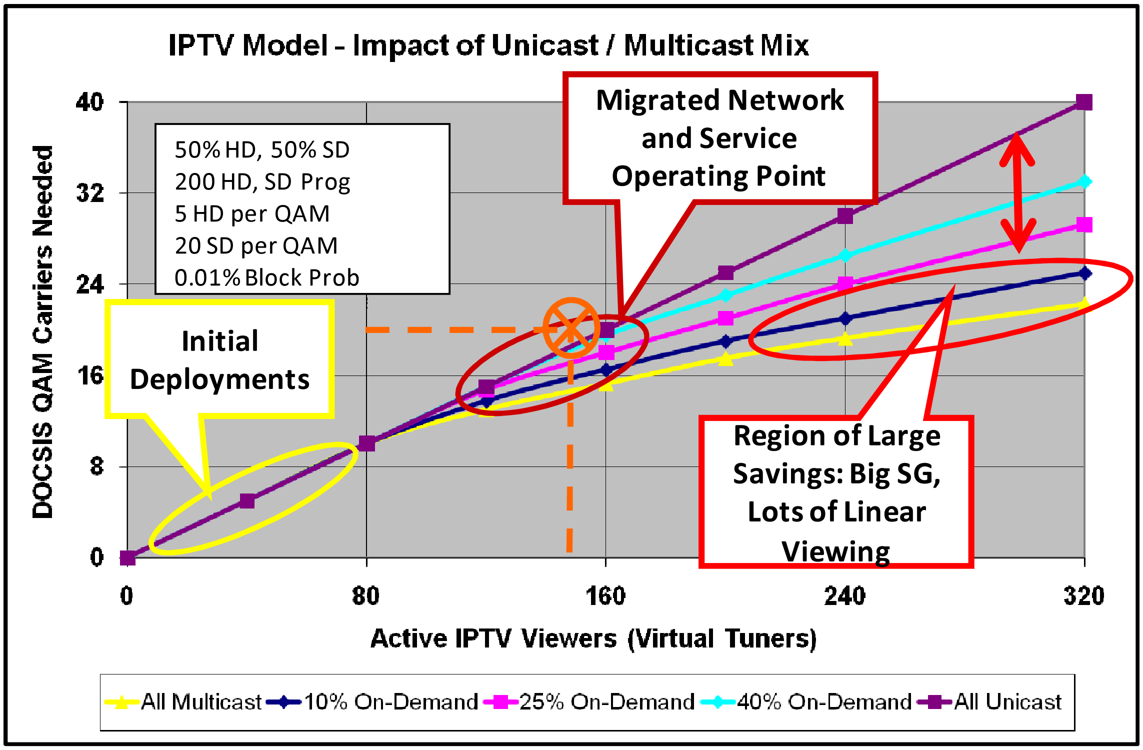


Figure 5 – IP Video Shifts the Spectrum Allocation Methodology

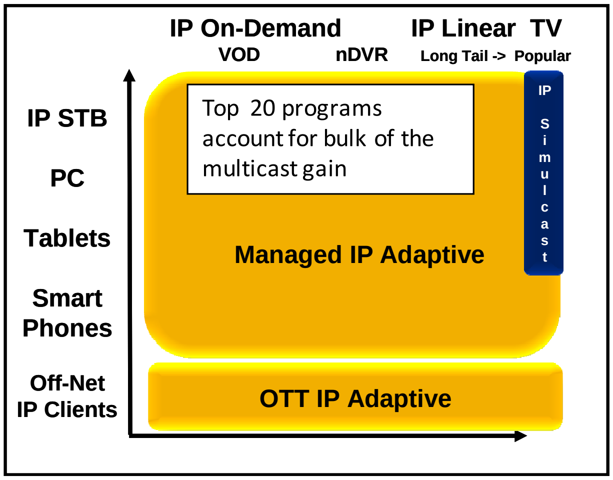


Figure 6 – Optimizing IP Video Delivery

This same dynamic could occur in the future with new high-resolution formats and smart TVs, with the only difference being that the process will take place with respect to discovering and adaptation to *primary* screen capabilities. The intelligence required is being built today to serve those 2nd and 3rd IP screens. By the time, for example, QFHD is a video format scaling in volume, the migration characteristics driving traffic to nearly all unicast will have taken place. As such, primary screen format discovery will be timely for keeping simulcast requirements at bay.

The model that we will assume as we assess the network implications is one of a small set of broadcast (conservatively quantifying with 40 total broadcast programs), with all other traffic as unicast. We will assume that the remaining traffic for video – the video unicast – is inherently captured in the traffic projections as part of 50% CAGR on the downstream. It may, in fact, be precisely what the CAGR engine of growth *is* for IP traffic over the next decade. A contrasting view would be to project HSD growth at 50% CAGR, but add to this video traffic aggregates representative of video service rates of an aggregate [13].

NETWORK IMPLICATIONS

It is quite simple to illustrate a network capacity problem in the face of increasing video quality and resolution, which directly translates into more bandwidth required. In Figure 7 we find the intersection of traffic growth, video services, and time in order to help guide MSO decision timelines. The trajectories moving upward from left to right show a commonly assumed Compound Annual Growth Rate (CAGR) of 50% over a period of ten years offset with two breakpoints over the course of the decade where a (perfect) node split takes place.

While the HFC available capacity in the downstream is over 5 Gbps when considering the highest order modulation profile currently utilized (256-QAM – the yellow horizontal threshold) it is of course not all available to support IP traffic today. The vast majority of today's spectrum is set aside for video services. Figure 7 charts the growth of IP services, but also quantifies the setting aside of spectrum used for video services. These video services that are the moving target that we are looking to quantify here. The bandwidth set aside for video services is subtracted from the 870 MHz capacity to identify the threshold for when the IP traffic would exceed the available spectrum to support it. These thresholds are the horizontal lines on Figure 7.

Four thresholds are shown bounding the available capacity over the course of 10 years. The first, baseline case (red) identifies the available spectrum for data services growth if the video service offering is made up of 60 Analog carriers, 300 SD programs (30 QAM slots), 50 HD programs (20 QAM slots), and 8 VOD slots. The math for this distribution of broadcast and VOD is quite simple: 60+30+20+8 = 118 slots consumed for video services, leaving 18 slots for DOCSIS.

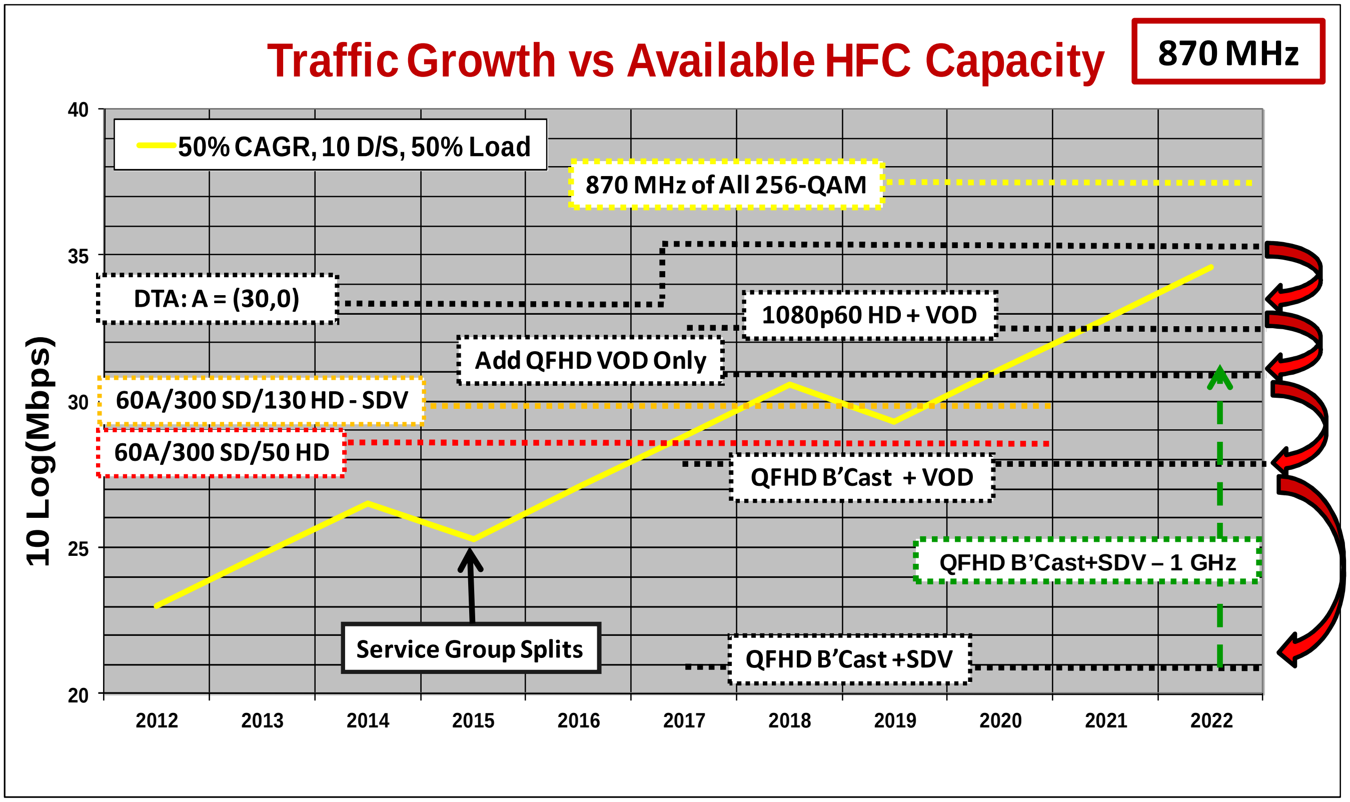


Figure 7 – New Resolutions Project to Massive Spectrum Management Concerns

Under an assumption that today's downstream DOCSIS carriers consume 200 Mbps of capacity (50% peak busy hour usage of 10 deployed downstream slots), then this video service architecture supports IP traffic growth through the year 2016, assuming there is one service group split along the way.

The orange threshold identifies the available headroom for IP growth if Switched Digital Video (SDV) is deployed, and the SDV achieves 3:1 gains for both SD and HD. Also, the HD program count is increased to 130 (modeled after a specific operator example objective). The broadcast tier in this case is limited to 60 Analog carriers, and the top 40 most popular channels offered, which are broadcast in both HD and SD. All other programs are on the SDV tier (about 20 SDV slots). The benefits of SDV are clear in Figure 7. Despite more than doubling of the bandwidth-intensive HD programming, we nonetheless gain new capacity for IP growth.

As powerful as SDV is for reclaiming spectrum, it is only reclaiming QAM spectrum, which is already inherently efficient in delivering digital video. There are further, large spectrum gains available by instead reclaiming spectrum from the Analog carriers through the use of digital terminal adaptors (DTAs). In Figure 7, the implementation is a phased approach. Phase 1 is a reduction of Analog slots from 60 to 30 – the black threshold that extends through 2017. In 2017, it is suggested that Phase 2 kicks in, whereby all Analog carriers are removed. This is the second black threshold, where now well over 3 Gbps has been freed up as capacity for IP growth. This chart and analysis process also identifies the flexibility available in downstream spectrum management. There are many knobs and levers associated with decisions on service mix and use of tools available for bandwidth growth.

Of course, the core issue as new video services evolve is that a 10-year plan demands consideration of these bandwidth-hungry next generation video possibilities. Ten years of tools and projections are encouraging, but the projection is based on video services and technology as we know them today. The plan can quickly implode by considering the capacity when including the integration of new generations of HD.

Four phases of next generation video evolution are identified by the red arrows on the right side of Figure 7. First, consider simply that all of the HD becomes 1080p60 HD – broadcast, SDV, and VOD. It is assumed this format does not require a simulcast phase (existing STBs and HDTVs support the format if it is available to them). The drop in available capacity (the first red arrow on the right hand side of Figure 7) reflects about a lost year of lifespan, all other assumptions the same.

Next consider that a Quad Full HD (QFHD) format is made available on VOD as an introduction to this format in its early days, as capable televisions become available to early adopters. The current VOD allocation remains (1080p60) in this case, so this advanced VOD service is completely additive in terms of spectrum. It is assumed that this format is deployed only using MPEG-4 compression. Nonetheless, as revealed with the second red arrow, we see a larger step downward in available capacity, which now is just over 1 Gbps. Roughly another year of lifespan is lost, all other assumptions the same. Furthermore, this would drive the timing of the second node split for downstream in 2018 if a QFHD VOD tier were to become viable in that time frame.

Now consider the third step, whereby QFHD was used for Broadcast HD and VOD, but not SDV. Note we have not included a simulcast of standard HD, even though it is TBD at this point whether a QFHD format can be "down-resolutioned" to standard HD. Certainly this is not the case in today's televisions or STBs feeding televisions, but it is likely to be a consideration in future iterations. The 4x scaling of standard HD is of course, in part, to make it more likely that systems can take advantage of current processing in the video chain through the simple integer scaling factor of pixels. Not delving into the details of how this might play out, we quantify the impact of a change in the broadcast and VOD to QFHD. The effect identified by the third red arrow is to drop network capacity down to about 600 Mbps, and clearly this is eating into any hope for supporting long-term IP traffic growth.

Lastly, now consider that the SDV tier is converted, but the VOD is not. An example of why this might be practical is that as the IP migration takes place, it might be determined that the legacy VOD infrastructure is not permitted to grow with new MPEG-2 TS based investment. These investments would be made instead in the IP domain, with VOD being one of the first phases of the video services migration to IP. In this case of Broadcast and SDV supporting QFHD as opposed to standard HD, we clearly see, in the form of the lowest black threshold on the chart at about 21 dB (just over 100 Mbps of capacity available for IP traffic, or three DOCSIS channels), the hopeless situation for next generation video without some new ideas and evolutionary approaches to be supported over the HFC network.

To point out a measure of hope that hints at some of the consideration we will account for later in the paper, the upward pointed green arrow shows where this situation would instead fall if there was 1 GHz worth of spectrum to work with. The spectrum freed up by 1 GHz of HFC compared to 870 MHz is about 22 slots, which works out to almost 900 Mbps using 256-QAM. New spectrum is but one tool we will evaluate as a means to enabling the migration of next generation video services

Note also that we have as yet not even attempted to factor in any capacity effects associated with Ultra-High Definition Television (UHDTV) as a potential format.

LONG-TERM VARIABLES: GOOD NEWS – BAD NEWS

We observed in Figure 7 that there was an obvious problem brewing under the assumptions made based on considering HFC architectures, services, and technology, as we know each today.

In Table 4, we begin to make the case for why the situation may not be as dire as these projections. On the left hand side of Table 4, "Losses," we quantify in the decibel language of the projection analysis the potential bandwidth penalty of the new formats, quantified in the row identified based entirely on the resolution difference. Again, it may be determined in practice that the encoding process more favorably compresses the formats than is portrayed in Table 4, but for now we will simply rely on encoding efficiency gains consistent with the average savings attributed to H.264 and H.265 using today's HD format. In each case, this amounts to 50% savings, based on early evaluations of H.265 and our prior discussion on HEVC.

Table 4 – Video & Network Variables: Losses and Gains

Losses	dB	**Gains**	dB
1080p60	3.00	H.264	3.00
QFHD	6.00	H.265	3.00
UHDTV	6.00	Split	3.00
10-bit	0.97	N+0	9.21
Frame Rate	0.00	Mod Profile	0.97
Total	15.97	VBR/D3	1.55
		Total	20.73
Difference	**4.76**	(All)	
	7.76	(HD to QFHD)	

The conservative assumption for 1080p60 is that it is 2x the bandwidth required of 1080i30. For the purposes of this study, as is generally done in practice as well, we will not distinguish between bit rates of 1080i and 720p although the former is roughly 12% more bits of transport rate.

We consider a 10-bit depth of field for UHDTV, but no additional overhead associated with changes to subsampling. We also do not consider any additional frame rate impacts on transport bandwidth. While scan rates of television rates have increased, and, as discussed, studies [1] reveal that frame rates higher than 60 Hz are perceptible by humans, there appears to be no move afoot to standardize in the market place on anything higher. Additionally, UHDTV is standardized around a 60 Hz frame rate. While research noted above has shown perceptibility by human of up to 300 Hz, it would not be fair to impart new bandwidth associated with new frame rates at this stage, even if they are to take shape. It is intuitively likely that higher frame rates lend themselves to more similarities between adjacent frames. We leave the variable in the chart because we believe that in time, formats will begin to experiment with higher frame rate delivery. We should keep this variable in the back of our minds as a possible wildcard.

The "Losses" when added together in the worst case of UHDTV as the final phase is 15.97 dB, which we will round to 16 dB for discussion purposes.

Now, let's take a look at the "Gains" in Table 4. Some of these we have already observed as "HFC as we know it" gains in our projection chart. We identified service group splits by the traffic growth breakpoints in the chart, which recognized the virtual doubling of bandwidth (ideally) in a typical node split. The average bandwidth allocated per subscriber in the split service group is now twice as much.

We also capture the service group split function here identified as N+0. In this case, we are recognizing that rather than perform further business-as-usual node splits after another round of this expensive activity, an "ultimate" split is executed instead, where the fiber is driven deepest – to the last active. The impact to the average bandwidth made available per subscriber is much greater in this case, with the homes passed per N+0 node assumed to be 40. Note that we identify one split prior to N+0 in Table 4. In the actual timeline-based model we will capture the move to N+0 as an extra split (two total) prior to the migration to N+0. We captured the decibel effect (3 dB) within the N+0 adjustment in the table to match what we will show on the subsequent projection analysis.

Lastly, we applied the benefits of MPEG-4 encoding in introducing QFHD – obviously better than MPEG-2, but also clearly not enough itself to compensate the bandwidth growth. This is intuitively obvious enough, seeing as the MPEG-4 gains do not offset the resolution increase in pixel count. However, it should be pointed out that the 1080p60 case shown in the trajectory of Figure 7 may indeed be offset by the introduction of MPEG-4 to deliver that service. In fact, it is reasonable to consider that 1080p60 as a service does not become a video service offering *until* MPEG-4 is available.

We showed in the Figure 7 a hopeful sign in the form of a different total available spectrum – 1 GHz vs. 870 MHz. However, because our starting assumption of 870 MHz may be optimistic or pessimistic, and because the spectrum expansion discussion is a wide-ranging one, we will address the physical bandwidth component in a subsequent discussion dedicated to spectrum.

We identify three other "Gain" variables – the subsequent generation of encoding, H.265, the use of IP Video delivery using bonded DOCSIS channels, and the opportunity to be more bandwidth efficient in an evolved (i.e. N+0) HFC architecture

High Efficiency Video Coding (HEVC, H.265)

As described, HEVC is in the heavy lifting phase of development and standardization, has as an objective a 50% better bandwidth efficiency of video transport, all while also yielding a higher quality. It appears to be on the track to achieve these targets.

The time-to-market for encoding standards and time-to-scale of advanced video formats follow roughly similar temporal cycles in terms of years. They are not necessarily in phase, but in both cases long evolution cycles have been the norm. As shown in Figure 8, it has been to the case in the past that the encoding gains served to continually drive down the rate of video (all SD for a time), even as slow as the pace of encoding development was. This singular fact explains the rise of over-the-top video. Data speeds raced ahead while video rates continuously dropped, crossing paths about seven years ago.

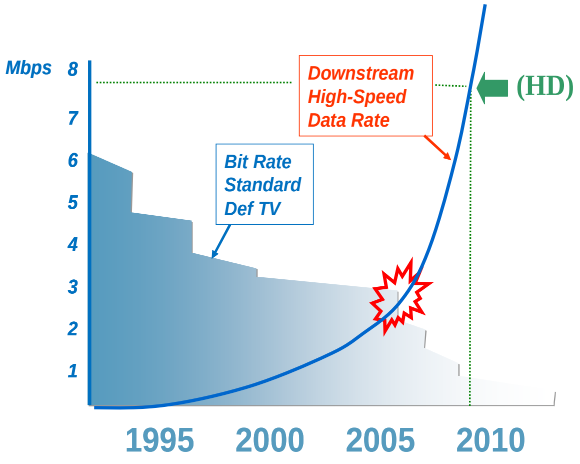


Figure 8 – Compression Meant Video Rates Only Decreased for Many Years

Now, however, demand for more HD has exploded, and display technology advanced significantly as well. It appears that the continuously accelerating pace of technology development will mean that higher quality, better resolution video will proceed faster than the process of standardizing encoding techniques. There is no accelerant to such a process, and arguably the increasingly competitive technology environment could lead to a slower standardization process, with service providers caught in between.

As indicated, early evaluation of H.265 and the conclusions drawn around this work described previously suggests that it will indeed achieve its target objective of 50% savings in average video bandwidth.

IP Video

Legacy architectures are based on simple traffic management techniques that allot an average of 3.75 Mbps per standard definition video stream to fit 10 such streams within a 40 Mbps single-carrier downstream QAM pipe. The heavy lifting of bit rate allocation is done at the MPEG level, whereby video complexities are estimated, and a fixed number of bits in the pipe are allocated to the ten streams under the constraint not to exceed 37.5 Mbps total. The same process plays out over High Definition slots, but in this case only two or three HD streams are part of the multiplex.

The introduction of DOCSIS 3.0 adds channel bonding to the toolkit, which, with the addition of MPEG-4 encoding, increases the stream count by over and order of magnitude relative to the transport pipe size. The net effect is the ability to use law of large number statistics for both SD and HD to the favorable advantage of less average bandwidth. So many independent streams competing for so much more pipe capacity results in a self-averaging effect that yields more efficient use of an N-bonded channel set when compared to MPEG-2 based video over N single channel QAM slots.

Self-averaging suggests that variable bit rate (VBR) streams can be used, recognizing the peaks and valleys will be handled inherently by the statistics (actually a capped VBR). Several prior analyses [16] of DOCSIS-based delivery, taking advantage of favorable statistics of wide channels to better handle the peaks and valleys of video traffic, shows that capped variable bit rate transmission yields a bandwidth savings that can be exploited. We use a 70% scaling as the bandwidth required for VBR-based channel bonded DOCSIS video in comparison to CBR-based single carrier QAM transport.

Fiber Deep Migration

"Business as Usual" HFC migration has been shown to be well-suited to about a decade of video and data traffic growth, without any new or special tools or techniques to accomplish this lifespan [13]. As discussed, use of node splitting in the HFC architecture reaches its ultimate phase when the last active becomes a fiber optic node. This architecture goes by various names – Passive Coax, Fiber-to-the-Last-Active (FTLA), or N+0. Regardless of the name, the architectural implications have two core components: small serving groups - on the order of 20-40 – and the opportunity to exploit new coaxial bandwidth becomes much more straightforward (30 assumed). The lifespan provided by BAU splits will not only make N+0 more cost effective due to RF efficiencies, but it will also leave operators within a stone's throw of FTTP should the need arise as an end state.

An important "side" benefit of an N+0 architecture is that the quality of the RF channel improves dramatically without the noise and distortion contributions of the RF cascade. The result is a higher SNR HFC link in the forward path. Because of this, we then consider more bandwidth efficient modulation formats. In Table 4, we have assumed that 1024-QAM will be readily accessible in such architectures, and in particular if new FEC is also implemented.

Finally, the removal of all RF amplifiers in the plant leaves only taps, passives, and cabling between node and subscriber, a much simpler scenario for flexible and expanded use of new coaxial bandwidth. Prior analysis [12] has shown how 10 Gbps (GEPON) and higher downstream capacities become conceivable in this architecture.

As fiber penetrates deeper into the HFC architecture, ultimately perhaps landing at N+0, the possibility of exploiting more bandwidth efficient modulation profiles exists, especially if the forward error correction (FEC) is updated from J.83 to modern techniques with substantially more coding gain. Here, we assume 1024-QAM supplants 256-QAM, for 25% added efficiency [15].

The dB Balance Sheet

Adding up the "Gain" side of Table 4, we find a total of about 20.7 dB, vs. 16 dB of "Loss." The encouraging information here is that this implies that, in principle, we can convert our current HD lineup fully to UHDTV and this would still be supported over the HFC network. All else equal, HFC lifespan would not be compromised in the face of IP traffic growth – the trajectory thresholds would not drop. This is so because the net of the gains and losses is a positive 4.7 dB. Thus, the thresholds would actually rise. Better yet, if the only format we bother concerning ourselves for business planning purposes is QFHD, then we have another 3 dB or headroom in our net gain.

The flaw in this good news story, of course, is that by the time we are considering QFHD, the IP CAGR is already threatening video service thresholds. We are at or near the end of the ten year window of migration. We are looking to extend HFC lifespan *beyond* this decade to the next while introducing these advanced video services. The excess gain can be viewed as available overhead for a simulcast transition. Based on Table 4, there is 4.8-7.8 dB to work with as part of enabling the possibility. While the services our transitioning, the IPV evolution is taking place, and the network is undergoing BAU migration, there are some "Not Business As Usual (NBAU)" evolutions expected to be taking place as well related to spectrum and architecture.

We will use Table 4 and these NBAU evolution factors to extend the projection through another decade, and draw conclusions on the intersection of video evolution, traffic growth, capacity, and the role of CAGR.

NEW SPECTRUM CONSIDERATIONS

Now let's consider the spectral aspects that were discussed in the last section, but not quantified in Table 4.

Figure 9 illustrates the anticipated spectrum migration of the HFC architecture long-term. A key driver discussed in great depth in [13, 14] is the necessity of operators to do something to address the limited upstream for the future. There are no easy answers to new upstream spectrum, and this figure describes the most effective approach and best performing from a modulation efficiency and flexibility standpoint, and which also yields the most efficient use of spectrum long term. The later is perhaps *the* key long-term primary objective for HFC spectrum evolution.

Because of the reasons outlined in [13] and [14], we foresee a phased approach to spectrum migration, consistent with the way operators incrementally deal with infrastructure changes in the context of dealing with legacy services and subscribers. The end state of the spectrum migration is shown in the bottom illustration of Figure 9, where some level of asymmetry consistent with what supports the downstream/upstream traffic ratio, will remain. No matter where the Frequency Domain Diplex (FDD) architecture lands in terms of diplex split, it is most assuredly going to yield a downstream capable of over 10 Gbps, and an upstream capable of over 1 Gbps.

While Figure 9 represents the most likely evolution scenario, other versions may come to pass. However, for any implementation, it is virtually guaranteed that the 10 Gbps/1 Gbps targets, at least, will be achieved. We will use this certainty in our projections in determining the ability of the evolved HFC architecture to deliver next generation video service in the face of continued growth in high-speed data services.

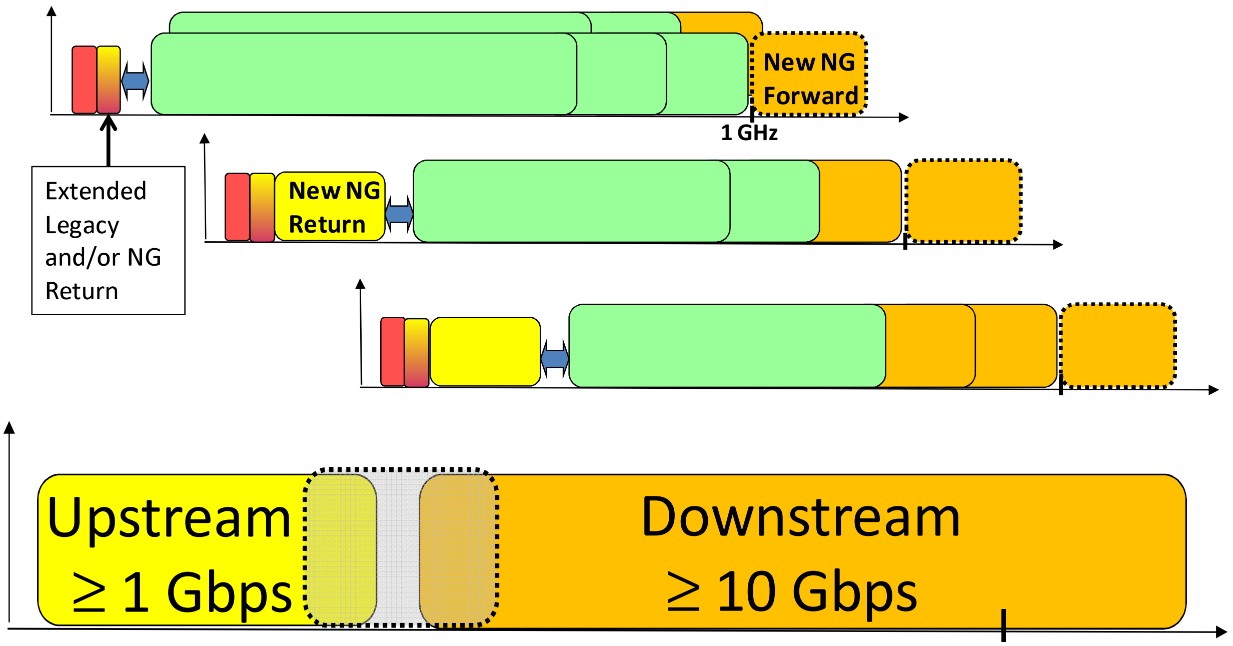


Figure 9 – Probable Evolution of the Cable Spectrum

PUTTING IT ALL TOGETHER

We now revert back to our original problem of capacity growth, and extended timeline of Figure 7 to account for the introduction of new generation of video formats. Beginning with Figure 10, we take into account all of these factors of video bandwidth growth and capacity preservation, placed in the context of HFC lifespan.

Video Service Delivery Assumptions

As we discuss video formats such as QFHD and UHDTV, it is reasonable to assume that HEVC has a key role, that fiber deep migration has continued to take place and is quite far down the path, and that the IP Video transition is in full swing, and possibly even complete. It is also reasonable to suggest that *unless* these evolutions take place, it is not practical to consider new tiers of advanced video services. Under this assumption, QFHD and UHDTV only become service in the cable network over IP, and only when HEVC is available in products for deployment.

The transition model is, of course, critical, as every new format introduces a period of simulcast if a service represents a broadcast. Conversely, in a full IP transition and a fully unicast architecture, the resolution and format become part of control plane and discovery. There is no wasted simulcast bandwidth, just any bandwidth penalty paid if the migration of video service delivery from the “legacy” efficiencies of broadcast to a dominantly unicast architecture is not properly managed (see Figure 5 and 6).

The Intersection of Video Services and Traffic Growth

Now let’s consider Figure 10. Figure 10 is a modified Figure 7, extended through the end of the next decade, managed with an N+0 migration, and accounting for various capacity enhancing techniques discussed above.

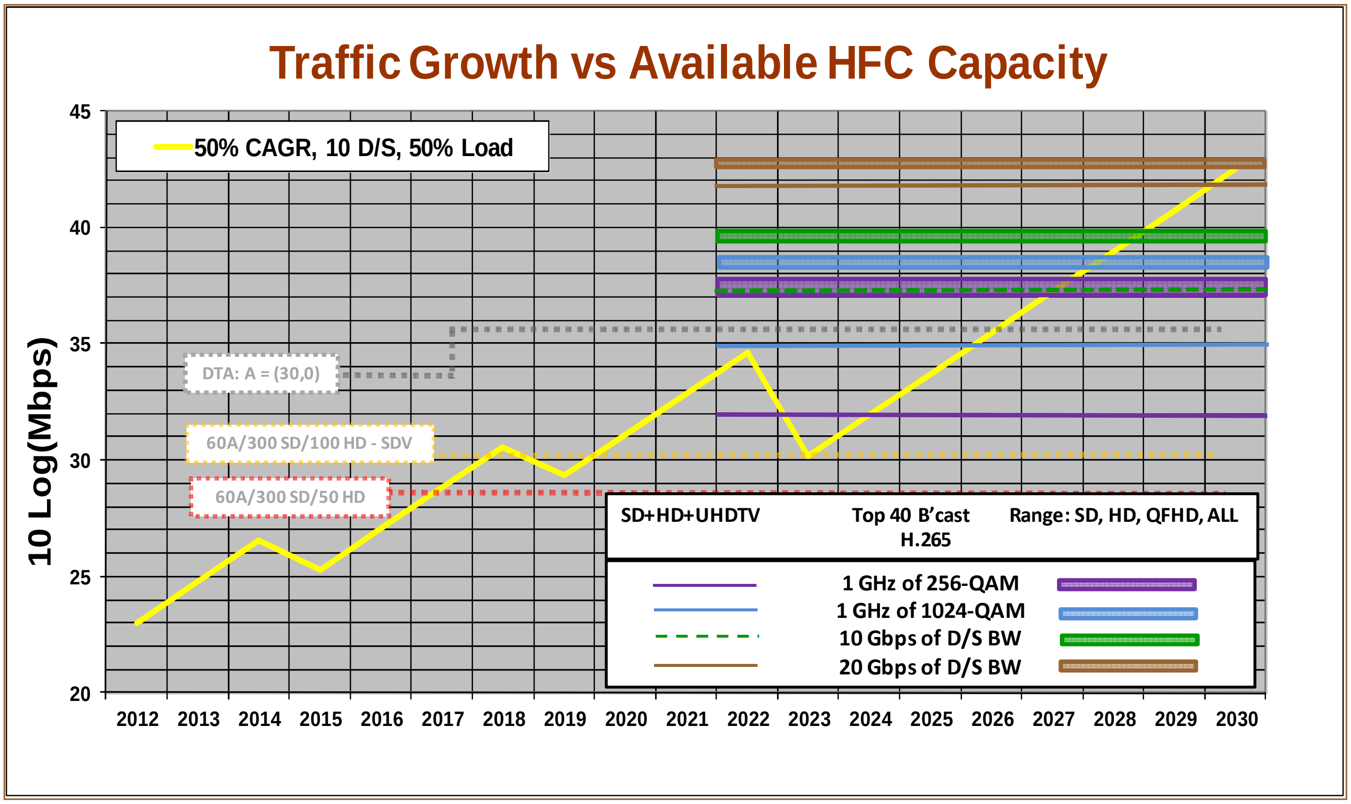


Figure 10 – Next Gen Video, Traffic Growth and HFC Capacity Limitations

The CAGR description is no different than Figure 7, it only goes on for longer, and sees a steeper breakpoint in 2022, representing the final phase migration to N+0. The Figure 7 thresholds are shown, in faded form, for reference against the cases to follow.

The legend at the bottom right is described as follows.

In all cases, we are talking about thresholds set by having a *static IP broadcast of the Top 40* channels. We are therefore taking advantage of IP video efficiencies, only as we know them today and previously identified in Table 4. Recall, we indicated that for a 200/200 lineup of SD/HD, then the top 20 programs would amount to 80-90% of the multicast gain in a switched IP system capable of multicast. The conclusion from that analysis was that a simplified, near-optimal architecture might instead be a mix of full broadcast and unicast, recognizing that all dimensions of network and service evolution are towards more unicast. From that, we have conservatively used a Top 40 program broadcast, which essentially would account for all of the multicast gain. At 40, it will likely come at the expense of some inefficiency of spectrum use versus multicast, but we prefer to err on the side of setting aside more spectrum for the purpose of a conservative analysis.

Also, because we are ultimately after the second-decade phase of HD evolution, we implement the next phase of compression evolution, HEVC, in calculating the long-term thresholds.

Four cases of available capacity are identified:

1) 1 GHz of spectrum carrying all 256-QAM, or 6.32 Gbps (purple)
2) 1 GHz of spectrum carrying all 1024-QAM, or 7.9 Gbps (blue)

3) A 10 Gbps downstream, in light of our prior conversation about the evolution of cable spectrum and key objectives (green)
4) A 20 Gbps downstream – enabled only through an N+0 architecture with a further extended use of coaxial bandwidth, requiring additional plant evolution of the passive architecture, including tap changes (brown)

Four cases of video formats are also analyzed. However, three of them fall close to one another in net capacity impact, and are lumped together in a "range" identified by a *rectangle* of the associated color on Figure 10. The fourth, most burdensome case is, not surprisingly, that which includes the introduction of UHDTV under the bandwidth assumptions we have identified previously – 160x the bandwidth requirement of the SD resolution and format. These UHDTV cases are identified by *lines* of the associated color – note that the green, 10 Gbps line is dashed, only because it overlaps the rectangular threshold range of the 256-QAM case.

The three cases in proximity whereby a rectangle is used to identify the threshold range are (in each case a simulcast of the Top 40):

1) SD + 1080p60
2) SD + 1080p60 + QFHD
3) SD + QFHD only

The latter, for example, makes sense if we consider that the integer relationship of formats (4x scaling of pixels) makes for the potential that next generation QFHD screens are also capable of displaying a "down-res" to 1080p60, or the STB/CPE function is capable of performing this function for the television. The range of remaining capacity in these three cases seems intuitively very close, and in fact is always within about 1 dB. This is a product of three things:

- Large capacity made available by all-QAM to 1 GHz, at least
- HEVC whittling down SD and 1080p60 rates by a factor of one-quarter
- The nature of the chart, based on nonlinear CAGR, is decibel units which tend to compress large numbers, which is illustrated by recognizing we are quantifying the impact of 18 years of aggressive compounding of traffic.

Let's examine what Figure 10 reveals.

First, consider UHDTV as a format that is mid-to-late next decade in scale at the earliest. It is not realistically able to be supported by HFC, at least under the assumptions we have used here. Even the most favorable of evolution deployments shown here – 20 Gbps of downstream capacity – suggests that persistent CAGR coupled with this broadcast video service runs out of room before the end of the decade. The vast majority of the bandwidth is the UHDTV itself, so eliminating the simulcast component is negligible to this conclusion.

By contrast, if we look at the QFHD scenarios, and view this as a format eligible at the end of this decade, then even the least capable case of 1 GHz of 256-QAM bandwidth offers nearly a decade of support for this scenario, with a range reaching exactly to the end of the next decade (2030) before a threshold breach of HFC capacity. This bodes well for the ability of tools available – just as we understand them today – to manage through an aggressive combination of video service evolution and persistent CAGR of IP traffic. It remains to be seen if this form of the evolved HFC network is the most cost-effective approach to enabling this service mix. But, it is surely comforting to know that the possibility exists to support such services with a 2012 understanding of technology, recognizing in addition the long time window of observation in which to adapt strategy and technology accordingly.

Note, of course, that since we have used 10 Gbps and 20 Gbps and not QAM calculations, these threshold apply equally to any access network that would set aside IP bandwidth for 40 channels as described herein. However, since other architectures may be full multicast, a broadcast adjustment (removing this lost capacity) might be in order for an accurate comparison. This is quite easy to accommodate by noting that 10 Gbps is simply 40 dB on Figure 10, while 20 Gbps is 3 dB higher at 43 dB. It is clear that there is very little difference in lifespan implied between these thresholds and those with broadcast allocations in this stratosphere of bit rates and continuance of CAGRs.

Settling of CAGR

In Figure 11, we show a modified case, whereby the assumption is made beyond this first decade that CAGR *decreases* to 32%. We chose this settling of CAGR at 32%, such that the net CAGR for the period through the end of the next decade is an 18-year average CAGR of 40%.

The logic behind this assumption is that we have seen this aggressive march forward of CAGR driven primarily by over-the-top (OTT) video services. In the model developed here, we are already allocating spectrum for most popular video services, and thus using video also as a driver for CAGR could be considered double counting, at least in part (the most-watched part) of this phenomenon. In addition, the vast history of CAGR growth has been around *catching up* with our ability to download and/or consume media – audio, then video. Once these media consumption appetites are satisfied, then it is possible that a CAGR settling will take place, with limits set by behaviors and eyeball counts [11]. Of course, it may simply be replaced by as-yet-to-be-determined non-media consumption applications, or altogether different kinds of media consumption that is bandwidth-busting, such as volume displays. That, however, seems beyond even the extended time frames we are evaluating here.

The above reasoning was completely qualitative, and it may in fact turn out that aggressive 50% CAGR persists indefinitely, or possibly increases. Nonetheless, because of the ramifications of long-term CAGR variation, we thought it useful to show this perspective, and that an 18-year average of 40% CAGR was a reasonable amount of settling to consider. Note that only at the year 2030 exactly would the 40% average and the 50%-32% model meet. The trajectories along the way getting to those points will, of course vary.

Now let’s evaluate what Figure 11 below says about video services evolution, capacity, and time.

First, observe now that *every* QFHD case indicates a lifespan of the network *through* the end of the next decade, even the 1 GHz, 256-QAM only case. This is a very powerful statement about the impact a settled CAGR may have on the support of advanced video services. It is also a reminder about the dramatic mathematical and planning implications of 18 years of compounding.

For UHDTV, there still does not appear to me much hope for a lasting solution to broadcast support, under what seems like the reasonable assumption that it does not make a large-scale service appearance until 2025 or beyond. The best case scenario in Figure 11 only suggests that 20 Gbps of network capacity covers the UHDTV scenario plus traffic growth into 2032-2033, which is then very shortly after it would have been introduced.

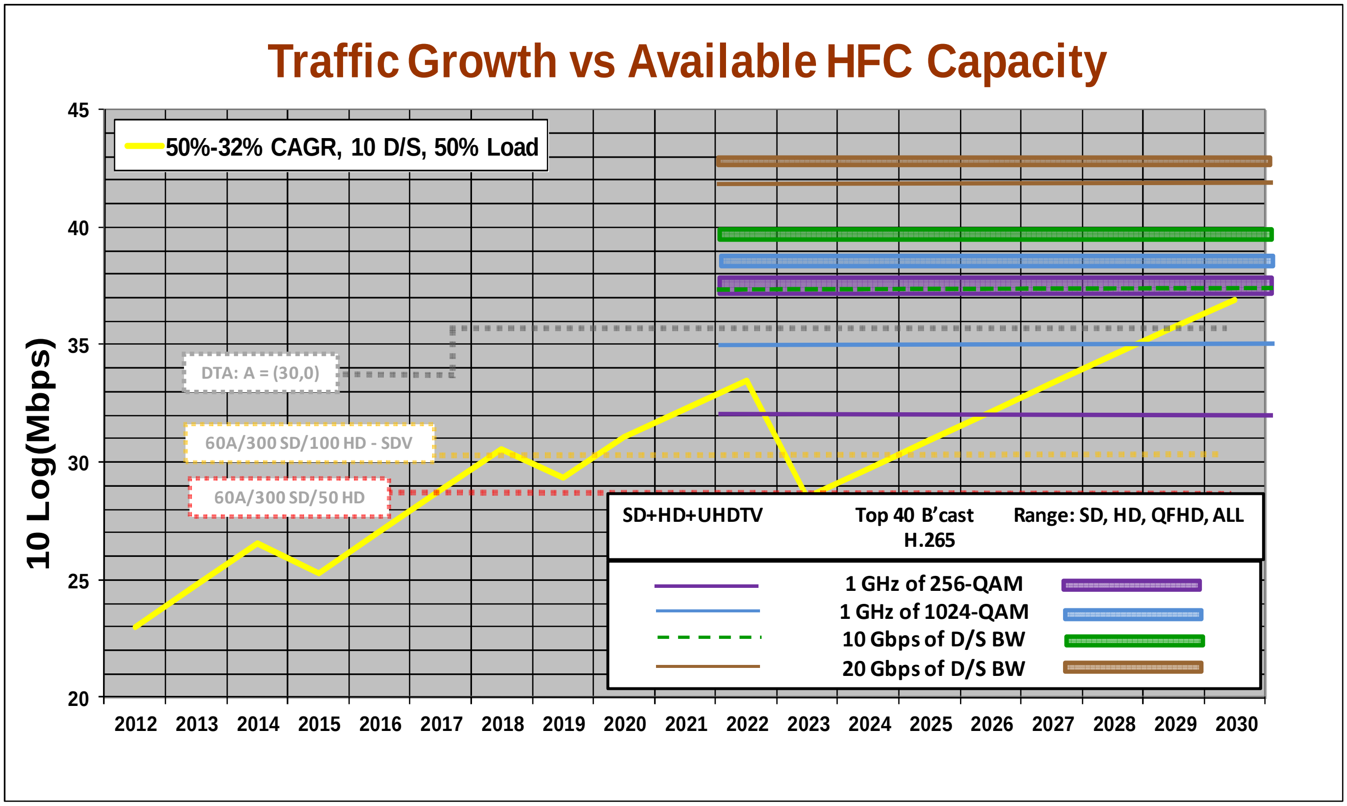


Figure 11 – Next Gen Video, Traffic Growth + CAGR Settling, and HFC Capacity

Conversely, this conclusion might more optimistically be stated by noting that HFC that manages a capacity of 10-20 Gbps *can* clearly support an early phase of UHDTV experimentation and deployment, and provide some cushion of years over which its significance as a scalable service can be evaluated. Does it become a niche scenario, where a very select number of channels become part of a programming lineup, much like 3D is today? For a mid-2020 time frame of experimentation, there are enough years of support in an early, modest phase of deployment where these kinds of questions can be asked and answers provided. These answers can then be used to guide a phase of network evolution, such as Fiber-to-the-Home, if scalability of the service is required. Or, it may lead to the conclusion that UDHTV is not an every-household type of consumer service, but associated with, for example, the penetration of home theatre-type owners. If so, it likely remains largely on the IP unicast service tier, and never become a broadcast scenario to worry about. Though, if this latter situation comes to pass, then this could exactly be the kind of thing that keeps CAGR chugging at 50%, while this model reflects the 18-year, 40% average case.

There are clearly many interrelated variables to consider and scenarios to quantify. Our assessment of the results leave inevitably to the conclusion that these projections are best viewed as living documents, and must be periodically re-assessed for the validity of the assumptions as trends and service mixes evolve over time. Advantageously, though, the projections indicate there are valuable windows of time near term, and again in the long term as efficiency improves. These windows offer the opportunity to observe and make methodical decisions to manage the evolution, without the pressure of an urgent congestion problem on the horizon.

THE EYES HAVE IT

While developing HEVC, compression science was not standing still elsewhere. Recently, a new technology – Perceptual Video Processing (PVP) -- was incorporated into broadcast encoders to improve the efficiency of both MPEG2 and AVC significantly. PVP technology [20] leverages the biology of *human vision* itself to enhance the encoding process. It can be thought of as a compression co-processor. Performance improvements typically range from 20% for moderately-easy-to-encode content to up to 50% for hard-to-encode content. Given the close familial resemblance of HEVC to its predecessors, it's quite possible that PVP could grant similar bonus improvements on top of HEVC's innate high compression efficiency as it has for MPEG-2 and AVC [6, 34, 35].

Signal Processing and Human Vision

Perceptual Video Processing (PVP) technology is an encapsulation of design principles that are thought to be at work in the visual system based on decades of research into the biology of human vision [2, 3, 4, 7, 9, 20, 24, 26]. Though biological in origin, these design principles are rooted in concepts that are familiar to signal processing engineers, namely, the ideas of noise reduction, signal estimation, and error signals. What is unique is that PVP is based on a model [21] of early visual signal processing, which has the following key components:

- Vision is tuned to the scale-invariant statistics of natural images [8]
- First stages of visual processing act as optimal filters designed to minimize the impact of noise
- A second stage of processing makes an estimate of the error associated with the first stage and uses that error signal to self-adapt to changing lighting conditions
- The output stage of processing is a coded form of the error signal, which can be thought of as a visual map of statistical uncertainty associated with the estimation process.

A key insight is that statistical uncertainty equals perceptual significance. The output error signal – the "uncertainty" signal -- highlights two kinds of information:

1. Image features that are uncertain because local correlations in the image are as likely to be attributable to noise as to actual variations in the signal. These are the features that are likely to be ambiguous from a signal estimation point of view and thus may require more attention.

2. Image features that contain local correlations that deviate from statistical expectations associated with natural scenes. In some sense, these are the "unexpected" correlations that might be worthy of closer inspection.

The notion that the output of early vision correlates with local statistical uncertainly provides a potential clue about higher-level perception and visual behavior. Eye-tracking and saccades, for example, might be considered behaviors intended to spend more time inspecting areas of high uncertainty to minimize overall uncertainty. Similarly, areas of high activity in retinal output might correlate with areas of high perceptual significance because they are the most suspicious in terms of statistical expectations – this is a clue that it may be worthy of special attention.

This model of the early visual system might also provide a context for understanding why edges are perceptually significant. According to the model's key components, edges are not perceptual important because they are edges, rather because they are localized correlations that deviate from the global expectation of scale invariance and thus require longer inspection to reduce uncertainty. It is not in fact the edge that has maximum uncertainty, rather it is the area around the edge, which itself might provide insights into the fundamental nature of perceptual masking and Mach bands – the illusion of heightened contrast near edges.

The Engineering View of Retinal Processing

The signal processing described occurs through the biological processing of the retina. The retina is made up of specialized cell layers, and each has a specific task. These can be classified as follows [20]:

Photoreceptors – The rods and cones we learned about in primary school health class. Photoreceptors are the first line of processing, are very densely aligned, and convert light (photons) into neuroelectrical signals.

Horizontal Cells – This second stage of processing cells collect the output of the photoreceptors and share them with adjacent horizontal cells as kind of a spatial low-pass filter operation on the discrete photoreceptor inputs.

Bipolar Cells – In the third stage of processing, bipolar cells collect both photoreceptor and horizontal cell inputs, and essentially acts to subtract the photoreceptor cell inputs, performing a differentiator type of mathematical operation.

Amacrine Cells – Bipolar cell inputs are received by amacrine cells, which come in different types. One important type acts as an electrical rectifier and gives a measure of the mean activity in the bipolar layer. A second type provides feedback to the first two layers to adjust their response properties according to this mean activity observed.

Ganglion Cells – The final stage of retinal processing, these cells take input from both bipolar cells and amacrine cells, and process and package them for delivery over the optical nerve to the brain.

Figure 12 illustrates the visual processing stages as a signal processing operation, described using tools analogous to functions common in signal estimation applications [20].

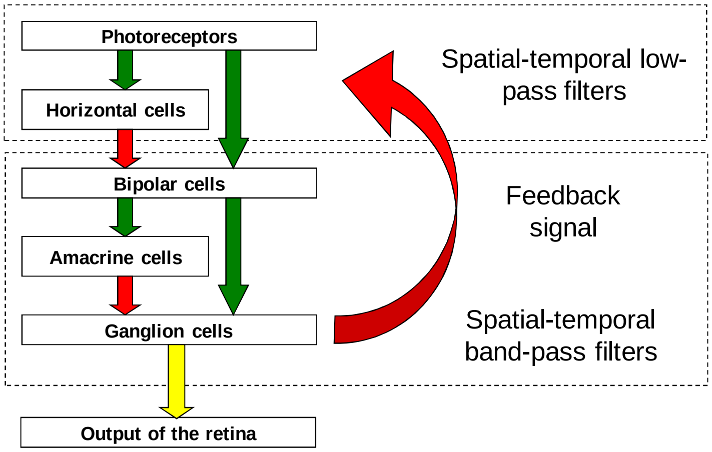


Figure 12 – Visual Cells as Signal Processing Functions

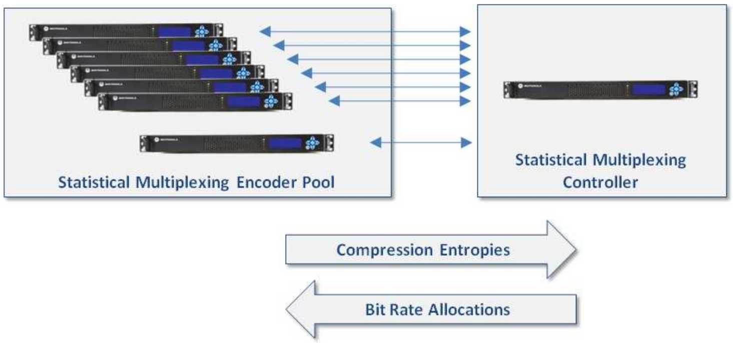
PVP Technology

Considerations for the biology of vision has proven to be very effective in improving compression efficiency in professional broadcast encoders. The key design principles have been extended to encompass space, time, and color and collected into a set of tools and software and hardware implementations collectively referred to as the Integrated Perceptual Engineering Guide (IPeG™). PVP is a particular commercial implementation of IPeG designed to operate in real time to reduce compression entropy and improve predictability in coding.

Internally, PVP identifies features in video that are likely to have high perceptual significance and modifies those features to reduce the number of bits required while preserving video quality. In its first commercial incarnation [20, 38], PVP performs two noteworthy complimentary operations: 3-Dimensional Noise Reduction (3DNR) and Adaptive Detail Preservation (ADP). The 3DNR operation is a combination spatial/temporal nonlinear adaptive filter that is very effective at reducing random noise in areas the eye may not notice. The ADP element preserves visually important detail and attenuates quantization noise, impulse noise, stochastic high-contrast features, and other hard-to-compress detail difficult for the eye to track.

An example of PVP used to improve compression efficiency for statistical multiplexing is illustrated in Figures 13 and Figure 14. The central concept in statistical multiplexing (aka "statmux") is that more and better channels can be delivered over a limited bandwidth by allocating bits intelligently between the various channels that comprise a statistical multiplexing pool. Channels that are easy to encode at a given point in time are given fewer bits than channels that are hard to encode. This traditional "statmux" operation is illustrated in Figure 13.

Using PVP, this operation is modified with this additional intelligent processing as shown in Figure 14. The statistical multiplexer still does its bit rate allocations, as always, but it now does so based on an enhanced set of inputs from the IPeG processor. PVP improves statistical multiplexing by selectively reducing the greediness of hard-to-encode channels in real time. High compression entropy means more bits would be needed to achieve a target video quality. Low compression entropy would require fewer bits to achieve the same video quality. PVP preferentially reduces the entropy of hard-to-encode features thereby making tough content kinder and more generous neighbors in the pool.


Statistical Multiplexing Encoder Pool
Statistical Multiplexing Controller
Compression Entropies
Bit Rate Allocations

.

Figure 13 – Traditional Statistical Multiplexing

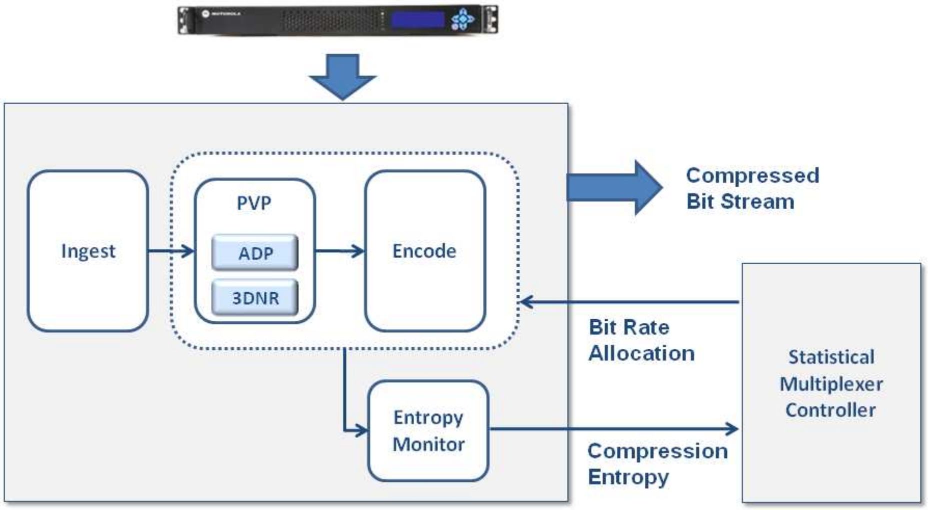


Figure 14 – PVP: Perception-Guided Adaptive Modification of Compression Entropy

An example of the graded impact of PVP on compression entropy is shown in Figure 15. Note that the relative impact of PVP is largely independent of the operational bit rate, which could prove to be a useful feature in statistical multiplexing pools that contain premium channels with higher targeted operating bit rates than other channels in the same pool. The data shown in Figure 15 are typical of moderate-to encode and difficult-to-encode broadcast content. The actual reduction in compression entropy may be optimized for particular use cases by adjusting the strength of PVP from weakest to strongest.

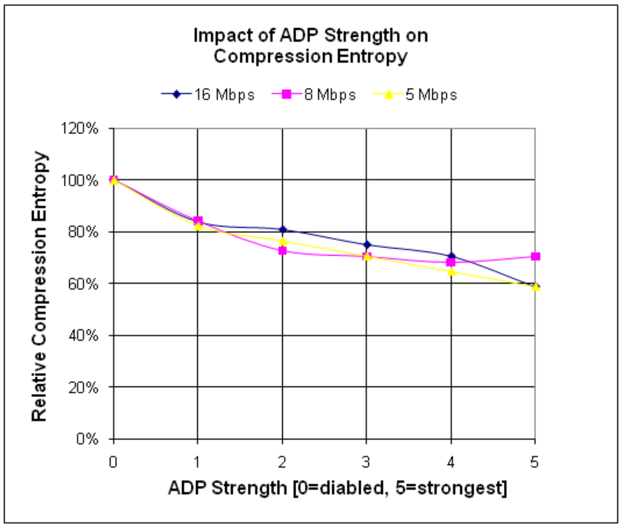


Figure 15 – PVP Reduces Compression and Can be Tuned to Requirements

Complementing HEVC

One of the key advances of HEVC is "just-the-right-size "processing in which each Coding, Prediction, and Transform Unit is sized precisely to capture the self-similarity within the picture detail they encode. It is without question a highly efficient way to squeeze bits -- but it's *not* the way the eye sees.

There are two key questions to examine to predict the impact of PVP on HEVC efficiency:

1) Would PVP enhance predictability and thus promote regions of self-similarity that can be captured efficiently by HEVC Units?

2) Are HEVC's "just-the-right-size" Coding, Prediction, and Transform Units also "just-the-right-size" for the natural scale of vision? If they are, then we would expect PVP to have less of an impact for HEVC than it does for AVC and MPEG-2.

The first question is straightforward, and the answer is *yes*. PVP nudges video towards statistics that would be expected of clean natural scenes when those nudges would not be very noticeable. In other words, the PVP promotes predictability and regional self-similarity. It does this by reducing unpredictable random noise and slightly modifying stochastic high-contrast features that are "unexpected" as described previously. On this basis, we would expect PVP to improve HEVC's innate compression efficiency to approximately the same extent that PVP improves AVC and MPEG-2 efficiency.

The second question is a bit more involved. Getting a handle on the natural scale of vision entails comparing the size of retinal images to the resolving power of the eye.

The visual angles subtended by various kinds of displays are listed in Table 5. The size of the visual field depends on the physical size of the display and its distance from the viewer. For QFHD (4k) and Ultra HDTV, we use the dimensions of recently announced displays [29] and predict that comfortable viewing distances will be only moderately larger than they are for traditional HDTV.

Table 5 -- Expectable Visual Angles for Various Display Types

Display Type	Format	Resolution		Dimensions (inches)			Viewing Distance (inches)	Visual Angle (degrees)
		Horizontal	Vertical	Diagonal	Width	Height		
Smartphone	QHD	960	544	5	4	2	12	19
Tablet	1080p	1920	1080	11	10	6	16	35
HDTV	1080p	1920	1080	55	48	27	76	35
Super HDTV	4K	4096	2160	70	62	33	88	39
Ultra HDTV	8k	7680	4320	85	74	42	90	45

The fovea of the retina sees the central 2 degrees of the visual field with high acuity [17]. It is the part of the retina with the greatest resolving power. We watch television by continually moving our eyes around to bring our fovea in line with particular features one after the other. Our brains integrate this sequence of focal observations into a unified seamless experience.

In Figure 16, we examine the size of the foveal image relative to the size of the visual field subtended by various display types. Our fovea spans only about 1/10th the width of a smartphone display, which means we must still move our eyes about even for the smallest display type. For 1080p and finer resolutions, our fovea sees at any moment in time only a disc of pixels having a diameter about 5% of the width of the whole display. It is worth noting that area of the disc comprises less than 1% of the total pixels in the display. We only see that small 1% of the display in detail at any instant. Research into bit rate reduction of video in other circles has been around trying to figure out how to take advantage of the fact that so little of a screen is actually processed at any given instant [5].

In Table 6, we quantify these relationships across a range of display types. An important insight comes about when we analyze the number of physical pixels seen by the fovea as a function of display size and resolution. A disc about 100 pixels in diameter contributes to foveal vision for smartphones, 1080p tablets, and HDTV. If brightness and contrast were put aside, the equal density of pixels would suggest that we would notice about the same level of visual detail – and same level of compression artifacts – on smartphones and tablets as we would see on HDTV when viewed from normal distances. Visual details and artifacts would likely be less noticeable for 4K and UHDTV because they would be 2-3x less magnified in the foveal image according to the pixel diameters shown in Table 6.

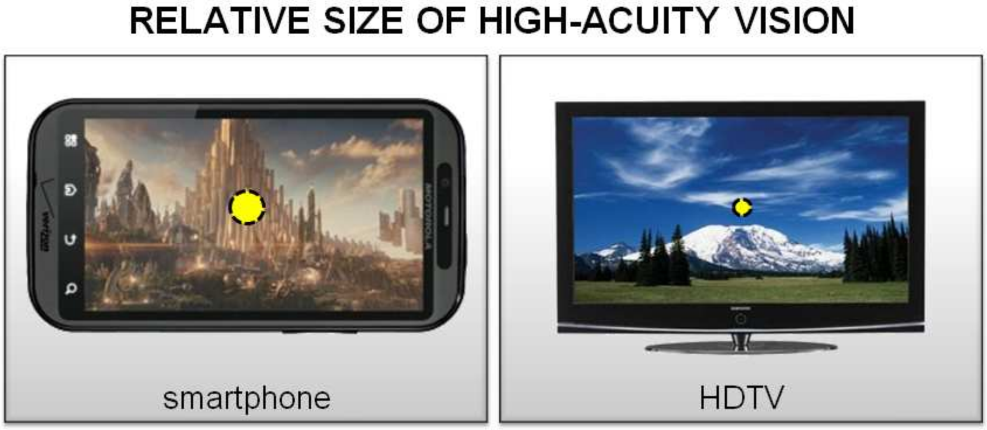


Figure 16 – Size of Projected Foveal Image (yellow) vs. Display Type

Table 6 -- Size of Foveal Field of View Relative to Size of Coding Units

Display Type	Format	Size of 2-degree Foveal Field of View						
		Percent of Screen Width	Pixels (dia.)	Macroblocks or Coding, Prediction, and Transform Units				
				4x4	8x8	16x16	32x32	64x64
Smartphone	QHD	11%	101	25	13	6	3	2
Tablet	1080p	6%	111	28	14	7	3	2
HDTV	1080p	6%	110	27	14	7	3	2
Super HDTV	4K	5%	211	53	26	13	7	3
Ultra HDTV	8k	4%	343	86	43	21	11	5

The scale of MPEG-2 and AVC macroblocks and sub-partitions relative to the size of the foveal image for smartphones, tablets, and HDTV is illustrated in Figure 17a. The homologous HEVC Coding, Prediction, and Transforms Units are depicted in Figure 17b. We noted previously that the fovea covers only a tiny fraction of a display screen at any moment. Smaller yet are macroblocks, sub-partitions, and HEVC Units. Even the Largest Coding Unit (LCU) presently allowed in HEVC (64x64) is significantly smaller than the fovea's field of view.

The homologous HEVC Units for 4k and UHDTV are illustrated in Figure 17c. The difference between HDTV and beyond-HD is a matter of visual scale. The foveal image of a LCU becomes 2-3 times smaller in 4k and UHDTV, respectively, compared to HDTV. Other smaller HEVC Units become visually diminutive, and the smallest 4x4 HEVC Units become tiny.

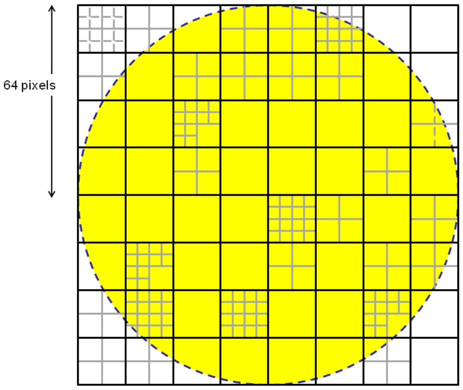


Figure 17a – MPEG-2 and AVC Macroblocks (dark) and Sub-partitions (light) Relative to Foveal Image (yellow) for smartphones, tablets, and HDTV

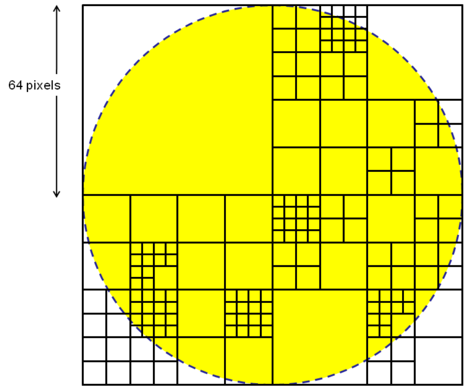


Figure 17b – HEVC Coding, Prediction, and Transform Units Relative to the Foveal Image for smartphones, tablets, and HDTV.

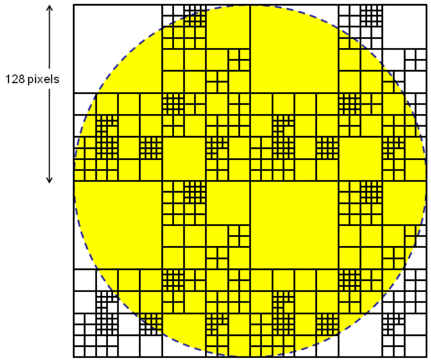


Figure 17c – HEVC Coding, Prediction, and Transform Units Relative to the Foveal Image for 4K and Ultra HD (note the relative size of the Units are smaller than in Figure 17b)

HEVC and AVC use rectilinear segmentation. The specific architecture is different, but the motivating philosophy is the same. More important, the visual scale of the rectangular segments is not dramatically different. HEVC provides a few larger block-size options that are better able to isolate regions of self-similarity without over segmentation, but those block sizes are still smaller that the fovea's field of view.

We can conclude from the above analysis that HEVC Units are, in fact, not always visually "just-the-right-size." Like AVC macroblocks and sub-partitions, HEVC Coding Units will have discrete boundaries within the foveal field of view even when encoding video that is visually smooth across the fovea. Compression artifacts tend to gather around discrete boundaries because those are the places that prediction is weakest. When those boundaries lay within the retina's high-acuity foveal field of view, they will be noticed. HEVC would need larger Largest Coding Units (LCU) to prevent over segmentation of the foveal image and meet the "just-the-right-size" visual ideal. For of smartphones, 1080p tablets, and HDTV the LCU would need to be at least 128x128. For 4K and Ultra HD, LCU would need to be at least 256 x256.

<u>PVP and HEVC Together</u>

Given the overall similarity of HEVC and AVC in terms of coding philosophy and visual scale, we project that PVP will improve HEVC coding efficiency to much the same extent that is improves AVC and MPEG-2 coding efficiency. HEVC's intrinsic compression efficiency is reported in [6, 34, 35]. Relative bit rates expected are listed in Table 7 and plotted in Figure 18. The impact of PVP is very content specific. Nonetheless we have found that PVP provides an overall average bit rate savings of ~20% in national-scale commercial deployments. We use that value in Table 3 to calculate the benefit of PVP to HEVC.

Table 7 -- Expected Bit Rate for Various Coding Modes and Display Types

Coding Method	Expected Bit Rate (Relative to AVC alone)		
	Smartphones	1080p Tablets & HDTV	4K & UHDTV
AVC	100%	100%	100%
AVC + PVP	80%	80%	
HEVC	66%	56%	50%
HEVC + PVP	53%	45%	40%

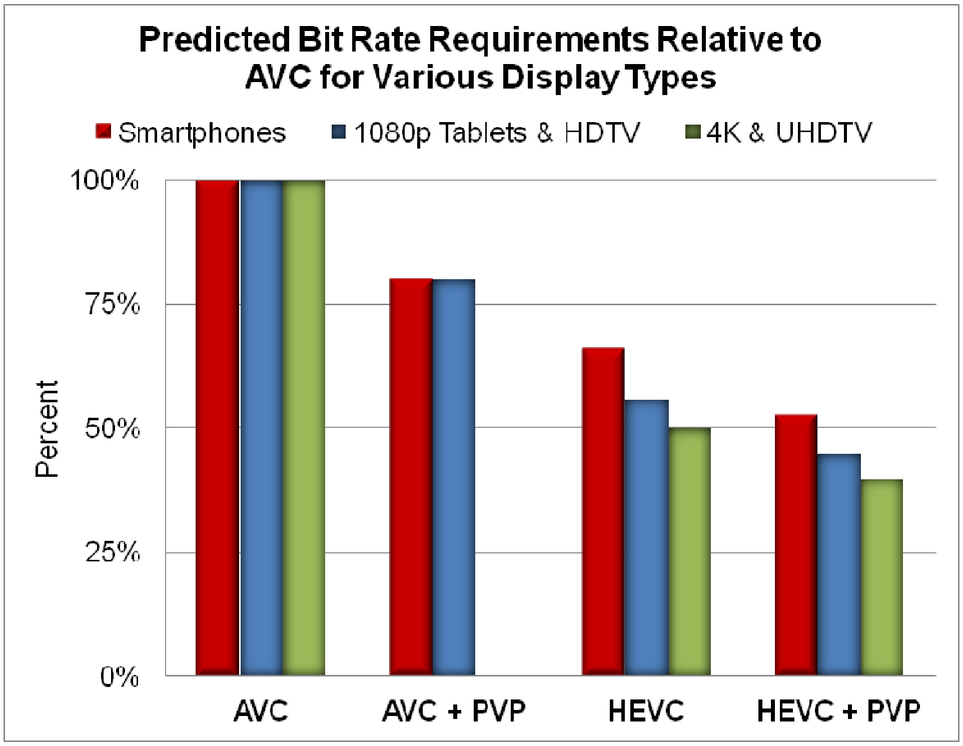


Figure 18 – Projected PVP Efficiencies Bit Rate for AVC and HEVC vs. Display Type

We can take this new knowledge and apply it to the prior figures that quantify traffic growth impacts. Figure 19 does so for the last case evaluated previously (Figure 11, 18-yr average CAGR of 40%). Of course, we do not anticipate tremendous new lifespan effects of PVP with a projected 20% of added efficiency. The expected value, at least early in PVPs evolution, is improved QoE of AVC and eventually HEVC video.

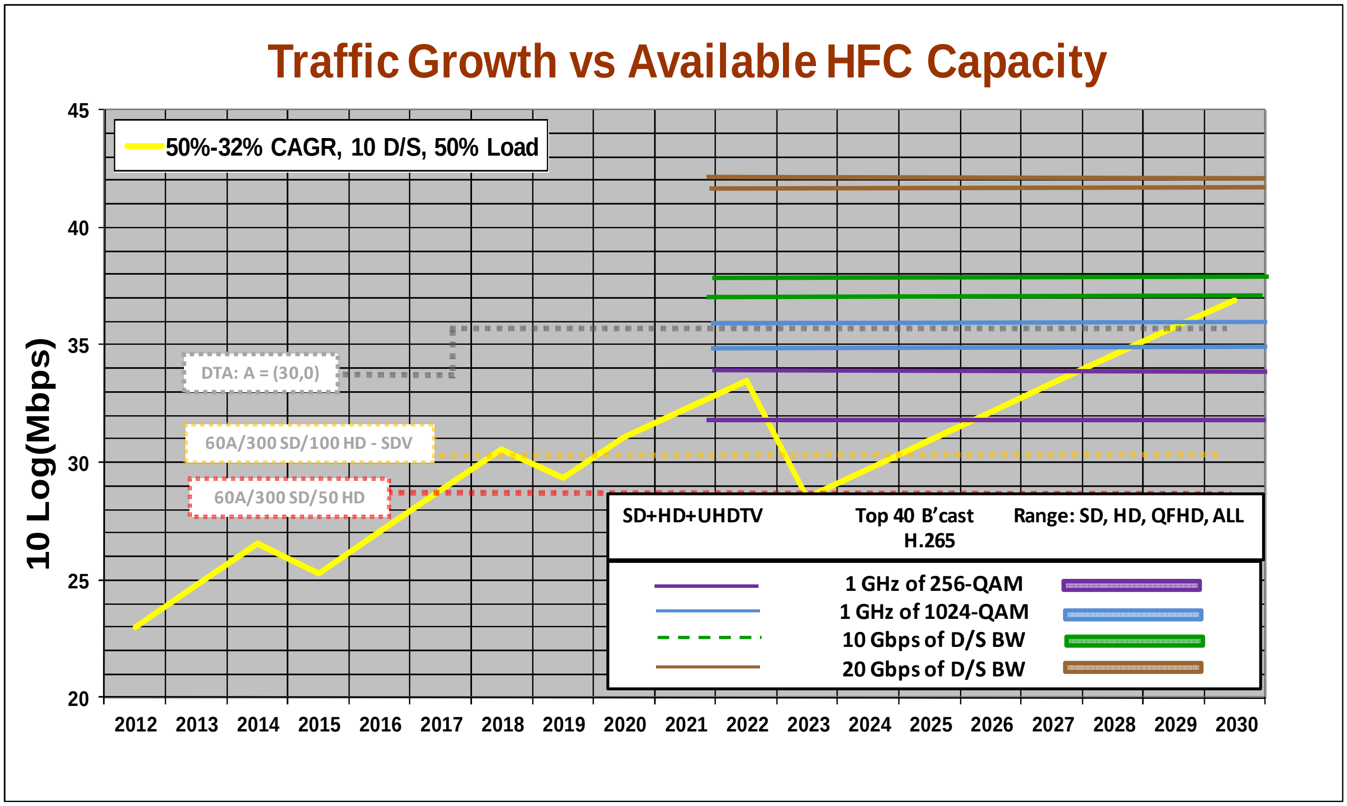


Figure 19 – HEVC + PVP, Traffic Growth and HFC Capacity (Settled CAGR Case)

Figure 19 indicates that the 20% of added efficiency at least has made the least-capable architecture evaluated (1 GHz of 256-QAM, purple) theoretically capable of weathering UHDTV services, or any substitute, similarly bandwidth-hogging applications that might beat it to market, into the middle of the next decade without the threat of breaching the threshold of capacity within a ten-year time frame under the assumptions used here. For that architecture, it also amounts to two extra years of lifespan, with the added burden on the non-PVP case that the final N+0 segmentation must also occur at least two years earlier.

For the higher capacity cases (1024-QAM, 10 Gbps, 20 Gbps), the impacts are less dramatic. Given that the existing network is, in fact, based on 256-QAM and outdoor plant equipment is 1 GHz capable only today, that impact carries more weight regarding preparation for a next generation of video bandwidth utilization.

Now consider Figure 20. Figure 20 is a redo of Figure 7, with the anticipated 20% benefits of PVP rolled up on the case of MPEG-4 AVC used in the Figure 7 analysis. In this case, we can observe a pretty significant impact of the extra 20%, largely because modest increases translate into large dividends when there is so little latent network capacity to begin with. These are shown in the upward pointing black arrows, which show the before/after of PVP being added for each scenario previous calculated. For example, in the worst case scenario in Figure 7 (and shown also in Figure 20) – QFHD in both the broadcast and the SDV tier as next generation HD, the network capacity was essentially completely consumed. Three available slots remained for IP traffic.

Because 20% of that tremendous amount of bandwidth is also a good chunk of bandwidth itself, adding it back to the pool for IP growth is pay substantial dividends as shown in Figure 20. With the savings, QFHD could actually be supported with some data growth runway. And, with a 1 GHz network, the network supports this level of enhanced HD with IP growth through 2020 under the migration assumptions used here of two segmentations. It is very unlikely that enhanced HD resolutions will be this pervasive in the market is such a short period of time. The introduction as VOD may be more practical in the timeframe of Figure 20. However, it is comforting to apply a bandwidth hungry, yet practical, "killer" application example to analyze in the projection analysis, and come out with a conclusion that the system does not only not break, but in fact enabling of such an application to a degree before any new steps or technologies are applied that could increase network capacity.

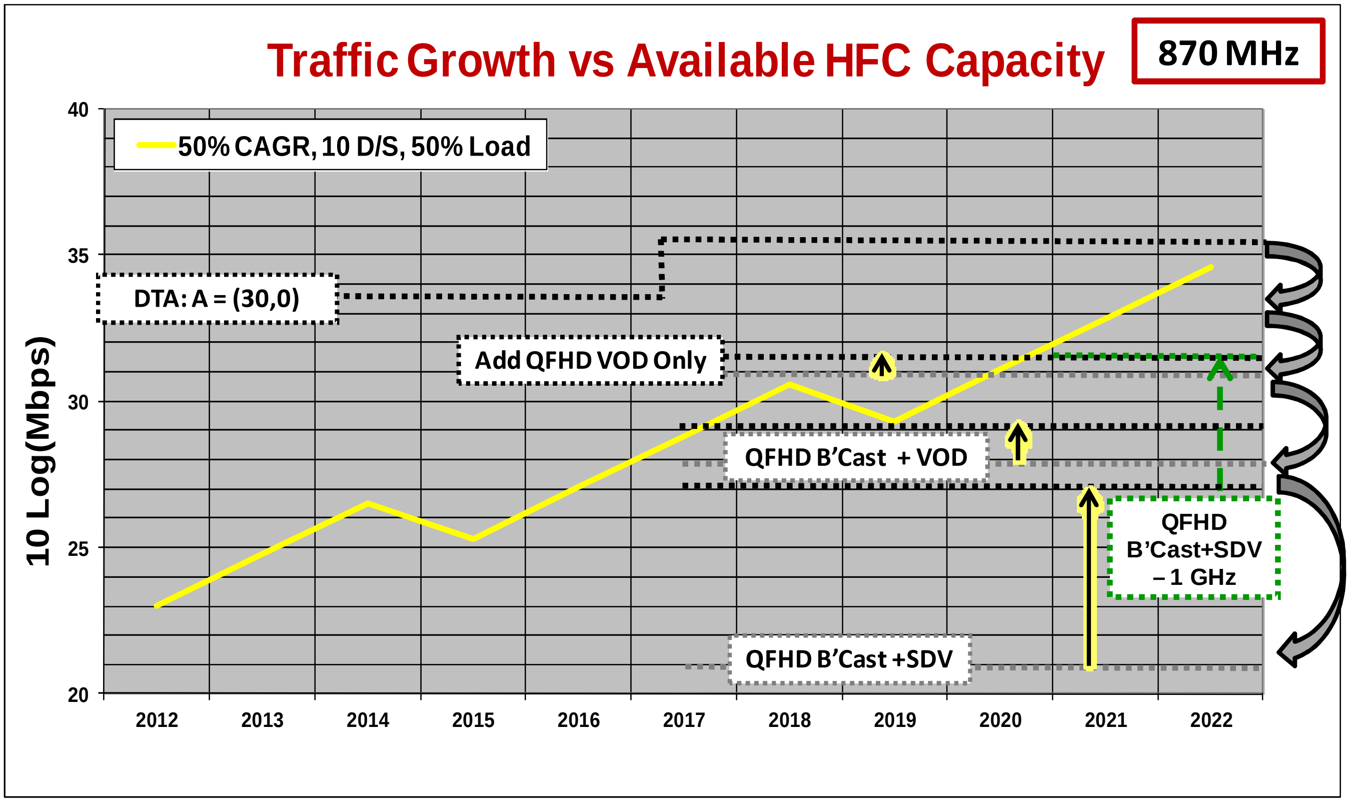


Figure 20 – Added Capacity with 20% PVP Efficiency, QFHD Format Cases (aggressive CAGR Case)

SUMMARY

In this paper, we evaluated network projections for the long-term, including many permutations of scenarios that included current and future services. We included technology and architecture options that are likely to come into play during the time windows observed, and applied these to quantify their effect. These include the shift to IP delivery, “beyond HD” video services, standards-based and innovative new encoding techniques, emerging use cases and delivery, and architecture, spectrum, and RF delivery enhancements. The result is a blueprint for an approach to preparing network service and migration plans – a blueprint that is, however, a “living document” given the accelerating pace of change in technology and services.

It is clear that there are many interrelated variables. However, any solution approach must include a comprehensive understanding that quantifiably describes the effects of network, technology, and service changes, such as shown in this paper. This is critical to properly engage in effective scenario planning, bound the problem, and prepare solution paths suited to an operator’s circumstances and expectations.

REFERENCES

[1] Armstrong, M and D Flynn, M Hammond, S Jolly R Salmon, *High Frame Rate Television*, BBC Research Whitepaper WHP 169, September 2008.

[2] Attneave, F., *Information Aspects of Visual Perception*, Psychol. Rev. 61 183-93, 1954.

[3] H.B. Barlow, The Coding of Sensory Messages: Current Problems in Animal Behaviour, Ed. W. H. Thorpe and O.L. Zangwill, Cambridge: Cambridge University Press, 331-360, 1961.

[4] Barlow, H.B., *Redundancy Reduction Revisited*, Network: Comput. Neural Syst. 12:241-253, 001.

[5] Deering, Michael F, “*The Limits of Human Vision*,” Sun Microsystems, 2nd International Immersive Projection Technology Workshop, 1998.

[6] De Simone, F et al., *Towards high efficiency video coding: Subjective evaluation of potential coding Technologies*, J. Vis. Commun. (2011), doi:10.1016/j.jvcir.2011.01.008

[7] Dowling, J.E., The Retina: An Approachable Part of the Brain, Harvard Univ. Press. 1987.

[8] Field, D.J., *Relationship Between the Statistics of Natural Images and the Response Properties of Cortical Cells,* JOSA A, 4 (12): 2379-2394, 987.

[9] Hare, W. A. and W.G. Owen, *Spatial Organization of the Biplolar Cell’s Receptive Field in the Retina of the Tiger Salamander,* J. Physiol. *421*:223-245, 990.

[10] Ho, Yo-Sung and Jung-Ah Choi, *Advanced Video Coding Techniques for Smart Phones*, 2012 International Conference on Embedded Systems and Intelligent Technology (ICESIT 2012), Jan. 27–29, 2012.

[11] Howald, Dr. Robert L, *Boundaries of Consumption for the Infinite Content World,* SCTE Cable-Tec Expo, New Orleans, LA, October 20-22, 2010.

[12] Howald, Dr. Robert L, *Fueling the Coaxial Last Mile*, SCTE Conference on Emerging Technologies, Washington DC, April 2, 2009.

[13] Howald, Dr. Robert L, *Looking to the Future: Service Growth, HFC Capacity, and Network Migration*, 2011 Cable-Tec Expo Capacity Management Seminar, sponsored by the Society for Cable Telecommunications Engineers (SCTE), Atlanta, Ga, November 14, 2011.

[14] Howald, Dr. Robert L, and Phil Miguelez, *Upstream 3.0: Cable's Response to Web 2.0*, The Cable Show Spring Technical Forum, June 14-16, 2011, Chicago, Il.

[15] Howald, Dr. Robert L**,** Michael Aviles, and Amarildo Vieira, *New Megabits, Same Megahertz: Plant Evolution Dividends*, 2009 Cable Show, Washington, DC, March 30-April 1.

[16] Howald, Dr. Robert L, Dr. Sebnem Zorlu-Ozer, Dr. Nagesh Nandiraju, *Delivering Pixel Perfect*, The Cable Show Spring Technical Forum, May 11-13, Los Angeles, CA.

[17] Helga Kolb, et al, *Webvision: The Organization of the Retina and Visual System. Part XIII: Facts and Figures Concerning the Human Retina*, WorldPress, http://webvision.med.utah.edu

[18] Marpe Detlev , et al., *Video Compression Using Nested Quadtree Structures, Leaf Merging, and Improved Techniques for Motion Representation and Entropy Coding*, IEEE Trans. Circuits Syst. Video Techn., Vol. 20, Nr. 12 (2010) , p. 1676-1687.

[19] McCann, Ken, and Jeff Gledhill, Adriana Mattei, Stuart Savage, *Beyond HDTV: Implications for Digital Delivery,* An Independent Report by ZetaCast Ltd, July 2009.

[20], *A Biological Framework for Perceptual Video Processing and Compression*, SMPTE Motion Imaging Journal, Nov/Dec 2010.

[21] McCarthy, Dr. Sean T., and W.G. Owen, "Apparatus and Methods for Image and Signal Processing,". US Pat. 6014468 (2000). US Pat. 6360021 (2002), US Pat. 7046852 (2006), 1998.

[22] Sullivan, Gary J. and Jens-Rainer Ohm, *Recent Developments in Standardization of High Efficiency Video Coding (HEVC),* SPIE Applications of Digital Image Processing XXXIII, Andrew G. Tescher (editor), Proceedings of SPIE Volume 7798, Paper number 7798-30, August, 2010.

[23] Ulm, John and Gerry White, *Architecture & Migration Strategies for Multi-screen IP Video Delivery***,** 2012 SCTE Canadian Summit, March 27-28, Toronto, CA.

[24] Vu, T.Q., S.T., McCarthy, and W.G Owen, *Linear Transduction of Natural Stimuli by Light-Adapted and Dark-adapted Rods of the Salamander, J. Physiol. 505(1):* 193-204, 1997.

[25] Wang, Zhou and Alan C. Bovik, *Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures*, IEEE Signal Processing Magazine, January 2009.

[26] Watanabe, S., *Information-Theoretic Aspects of Inductive and Deductive Inference*, IBM J. Res. Dev. 4. 208-231, 1960.

[27] Wiegand, T, G. Sullivan, G. Bjontegaard, and A. Luthra, *Overview of the H.264/AVC Video Coding Standard*, IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 560-576, July 2003.

[28] Yoshika Hara, *NHK Bets on Super Hi-Vision as Future TV*, EE Times, Sept 17, 2007.

[29] *CES 2012: 4K TV Sets Make Their Debut, Minus the Hoopla*, Los Angeles Times, January 11, 2012.

[30] ITU-T and ISO/IEC, ITU-T Rec. H.264 | ISO/IEC 14496-10 Advanced Video Coding (AVC), May 2003 (with subsequent editions and extensions).

[31] ISO/IEC JCT1/SC29/WG11 (MPEG), "Description of High Efficiency Video Coding (HEVC)," doc. no. N11822, Daegu, KR, January 2011.

[32] ISO/IEC JCT1/SC29/WG11 (MPEG), "Vision, Applications and Requirements for High Efficiency Video Coding (HEVC)", doc. no. N11872, Daegu, KR, January 2011.

[33] ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/16, "Joint Call For Proposals on Video Compression Technology", WG11 document N11113 and Q6/16 document VCEG-AM91, Kyoto, January 2010.

[34] JCTVC-A204, "Report of Subjective Test Results of Responses to the Joint Call for Proposals (CfP) on Video Coding Technology for High Efficiency Video Coding (HEVC)," Dresden, DE, April, 2010.

[35] JCTVC-G339, "Comparison of Compression Performance of HEVC Working Draft 4 with AVC High Profile," Geneva, Nov. 2011.

[36] JCTVC-F900, "Common test conditions and software reference configurations," Torino, IT, 14-22 July, 2011.

[37] JCTVC-H1003, "High efficiency video coding (HEVC) text specification draft 6," Geneva, CH, November, 2011.

[38] Motorola Mobility SE6601 Encoder, http://www.motorola.com/Video-Solutions/US-EN/Products-and-Services/Video-Infrastructure/Encoders/SE-6300-6500-Series-US-EN.

[39] "Video Quality Experts Group Report on the Validation of Video Quality Models for High Definition Video Content" VQEG HDTV Final Report, vers. 2, June 2011.

[40] www.carbonbale.com

[41] www.100fps.com

BUILDING A WEB SERVICES-BASED CONTROL PLANE FOR NEXT-GENERATION VIDEO EXPERIENCES

Yoav Schreiber, Sunil Mudholkar, Nadav Neufeld
Cisco

Abstract

Next-generation video services require solutions that are aware of real-time data and granular business rules encompassing identity, location, policy, etc., and can make decisions based on that data for a multitude of applications. In conventional video systems, however, this collection of data and business rules resides on disparate elements linked by closed, proprietary connections. This reliance on closed, "siloed" video systems impedes an operator's ability to develop new services and features, or to effectively scale cloud-based video services.

This paper presents a scalable, open-standards approach to orchestrating video services to allow for video control plane extensibility in multi-vendor ecosystems. It describes the architectural foundation for a loosely-coupled, modular video control plane with service provider-grade high availability and scalability. This approach enables video end-points to discover cloud functionality and external systems to expose services and communicate with video endpoints. Drawing on Internet communication methods, the proposed architecture enables a more flexible and scalable video services platform.

INTRODUCTION

An array of market forces is driving demand for new kinds of video services and, ultimately, profound changes to the service provider video systems delivering them. The Cisco Visual Networking Index (VNI) projects that more than 90 percent of consumer IP traffic and two thirds of the world's mobile traffic will be video by 2015.[1] The same study projects 10 billion mobile Internet-connected devices connected by the following year. And, the Cisco Global Cloud Index projects cloud IP traffic to reach 133 exabytes per month by 2015.[2]

These trends point toward a massive shift in consumer viewing habits from traditional, closed TV video systems to an open, cloud-based model. Consumers want the ability to access multiple types of content on multiple types of devices, regardless of the users' location or of the network over which they are connecting (e.g., the managed service provider footprint, an unmanaged Wi-Fi network, a cellular network, etc.). Consumers also seek new kinds of video experiences that integrate conventional video content with cloud services and interactive applications, and extend intuitively across multiple screens.

Service providers are well aware of these industry changes, and many are already moving to expand their video services to new screens and devices beyond the traditional set-top box (STB). However, the traditional service provider video architecture – designed to deliver legacy broadcast and on-demand video content, over a closed network, to a managed STB endpoint – is simply not equipped to support cloud and multi-screen delivery. These next-generation video services demand a level of service orchestration that conventional video platforms do not address. Consider: to deliver a personalized next-generation video experience to a subscriber, the video system must account for:

- Identity
- Device
- Content entitlement

- Location
- Bandwidth availability
- Past user activity
- Social network connectivity
- And much more…

Yet in today's video architectures, this information resides on several disparate, closed systems, including OSS/BSS, content management systems, session resource managers, client software, applications, middleware, etc. In addition, the isolated service "silos" on which conventional service provider control planes rely (i.e., treating managed and unmanaged clients, wired and wireless networks, etc., as entirely separate environments) further impede the service orchestration necessary to deliver a seamless, personalized multi-screen video experience. Conventional video system architectures are also ill-equipped to address the complexity inherent in serving multiple and changing consumer devices connecting to the service, or in optimizing the quality of experience (QoE) based on changing conditions.

Meeting the requirements of modern, cloud-based video delivery will require a new architectural model: a control plane for loosely coupled video systems that is designed specifically to meet the technical requirements of next-generation video services. This paper presents such architecture.

The proposed architecture is a scalable, standards-based approach to orchestrating video services to allow for video control plane extensibility in multi-vendor ecosystems. Drawing on proven Internet communication approaches used by some of the largest web companies in the world, it provides an architectural foundation for a loosely coupled, modular video system with service provider-grade high availability and scalability. Chiefly, this architectural approach:

- Enables video endpoints to discover cloud functionality in a loosely coupled system, and allows external systems to access exposed web services for communication with video endpoints
- Provides a platform to dynamically manage sessions, resources, and workflow in a loosely-coupled system incorporating both managed and unmanaged devices

ARCHITECTURAL ELEMENTS AND KEY CAPABILITIES

The proposed video control plane employs an Internet-based architecture, and as such, represents a significant departure from conventional video systems. Effectively, this approach applies the proven architectural model and design principles used by major web companies like Google and Facebook to contend with massive amounts of data and users, and applies them to video services. This gives service providers a more scalable video services platform, and affords them the same degree of flexibility and speed as web companies when designing, testing, and rolling out new features and applications.

At a high level, the proposed architecture encompasses the following building blocks (Figure 1):

- **Base platform:** The foundation of the video architecture is an open-source operating system, on top of which resides a distributed cache that acts as a shared data store accessible to all loosely coupled elements and workflows in the system.
- **Common messaging infrastructure:** A standards-based, highly-scalable, real-time messaging bus provides the communication framework over which distributed endpoints communicate.

- **Service infrastructure:** The architecture employs a standards-based, hardware-agnostic service infrastructure and workflow engine that enables complex service orchestration with the necessary performance for video services.
- **Session and resource management:** The architecture provides real-time session and resource management capabilities across multiple networks and devices, and is designed to support flexible policy enforcement and dynamic business rules.
- **Applications:** All video applications (i.e., service assurance, device authentication, emergency alert services, etc.), are built upon this platform.
- **Application programming interfaces (APIs) and web services:** the video control plane brokers communication among all elements of the system, including both legacy and cloud-based video applications, via APIs and web services.

Together, these building blocks create a next-generation video control plane that represents a fundamental shift from traditional video systems. Unlike legacy video systems, which rely on tightly coupled client/server communications, the proposed architecture employs a distributed model, similar to that used in web applications. The Internet has solved many of the problems of software resiliency, performance and scale and taking advantage of those is key to building this distributed control plane. In this distributed control plane, functionality and intelligence is allocated to various loosely coupled endpoints (i.e., clients, virtual machines, network elements, etc.), which then advertise, discover, and consume functionality from a shared cache of data.

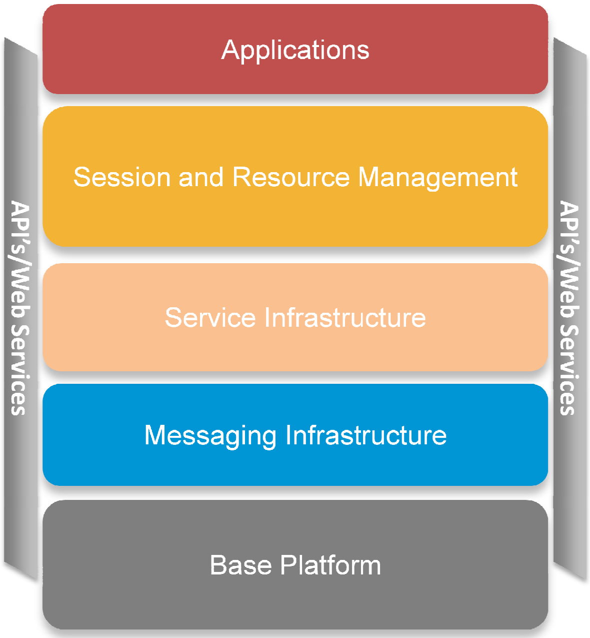


Figure 1. High-level architecture for next-generation video control plane

This data includes all of the essential information that applications need to deliver a video service via the cloud, including presence, state, entitlement, resource availability in the network, etc., all of it updated in real time. The data are stored in a high-speed shared cache, from which they are accessible to distributed applications in real time.

Clients, network elements, and applications can access data stored in the shared cache via a standards-based messaging bus. This common messaging infrastructure connects all elements to the cache via an encrypted, authenticated connection, and facilitates messaging back and forth among the various applications. Cisco has designed the architecture using a widely adopted communication protocol known for its scalability and performance in social presence and instant messaging applications.

Once this core framework is in place – open-source operating system, high-performance data store, and real-time messaging infrastructure – operators can build applications on top of it. These can include everything from straightforward core service functions like device authentication and service assurance, to advanced cloud service offerings such as cloud- or network-based DVR, social TV experiences, and synchronized companion device experiences.

Effectively, the intelligence in the proposed architecture is decentralized – residing in the applications. The proposed video control plane merely acts as a broker for these applications, providing all of the essential information they need to make real-time decisions and deliver cloud-based video services. Together, the shared data cache and common messaging bus functions almost like a web-based news feed: various elements throughout the system publish events or information, and every other element in the system can subscribe to any relevant information. This web services-based communication framework provides inherently more flexibility and scalability than a client/server model, and represents a significant departure from traditional closed video systems, and even some contemporary IP video systems. It should also be noted that the operator need not store every piece of data in the system in a single, centralized real-time cache. Less frequently used data can easily be stored elsewhere in the infrastructure, and remain accessible via the same messaging bus.

Employing this web-based infrastructure for brokering information between applications, the proposed architecture can:

- Orchestrate cloud-based services across multiple devices in real time
- Perform end-to-end session and client/device management for both managed STBs and cloud-connected endpoints
- Provide an interface between multivendor systems and applications
- Achieve service provider-grade scale and availability
- Allow for fully customizable user experience and applications

An Open System, Incorporating Standards and Web Design Approaches

An essential characteristic of the proposed next-generation video control plane is that it is based on open standards to allow for maximum flexibility. As a result, it integrates with legacy systems and with any standards-based third-party technology or application. It is also designed to allow operators to change technology vendors, equipment, systems, etc., as they choose. By avoiding proprietary standards that can age quickly, it provides a more future-ready architecture.

Along the same lines, the proposed architecture is designed to increase service velocity by incorporating web services and design approaches. It is modular and loosely-coupled, allowing for phased introduction of services and technologies. As described later in this paper, the architecture also facilitates service velocity and flexibility through its ability to dynamically manage workflows. The following sections describe this architecture in greater detail.

OPEN, STANDARDS-BASED SOFTWARE PLATFORM

In a legacy QAM-modulated video system, applications and middleware reside on STBs, and video content acquisition and provisioning systems populate back-offices. In this legacy environment, deploying and maintaining software, especially when proprietary, requires a substantial investment of time, skill, and financial resources. As operators transition to cloud distribution based on IP, however, they can take advantage of existing Internet-based standards to achieve greater scalability, and afford greater flexibility and service velocity.

The core technology envisioned in the proposed distributed video control plane has several facets. It is based on an open-source programming language such as Java, it uses web services-based SOA design principles, and it functions as a workflow engine.

Open-Source Technologies

As stated, the proposed architecture is fundamentally an Internet-based approach to video service delivery. As such, it should be based on an open-source or standards-based language such as Java. By using open standards, an operator has its choice of additional open-source libraries and frameworks that simplify the process of building, testing, piloting, and enhancing new services and features.

Java in particular is an excellent fit for a distributed video system. Known for its portability and openness, Java is an apt programming language for an extensible multi-screen video platform with automated workflow, session, content, and other video control plane functionalities. Because Java language code can be represented in the intermediate form Java bytecode, Java can run on a multitude of operating systems that otherwise would require platform-specific machine code. Other reasons for Java's popularity include its efficient memory management and relatively simple object model.

Service-Oriented Architecture

The proposed architecture is designed to operate on a service-oriented architecture (SOA) platform, and should be operable on any standards-based SOA platform. SOA principles allow for loose couplings between clients and servers, and facilitate the development of services independent of the client or underlying platform. As such, SOA design principles help facilitate the systems' ability to share information and functions among multiple distributed applications and system elements in a widespread and flexible manner.

Workflow Capabilities

At the top of the SOA stack is the workflow engine. Users can create workflows by using the workflow engine's graphical editor or by editing XML files directly. Workflows are not hard coded, are not compiled, and do not require any kind of system downtime for modification.

The system supports the real-time execution of multiple workflows, which can be easily extended to add new features rapidly in a controlled manner. These workflows can be characterized as:

- **Atomic:** Once a request begins executing a workflow, it will continue executing that workflow, without interruption.
- **Extensible:** Workflows are defined according to standards and can be modified using standard tools.
- **Flexible:** Workflows are built with a series of nodes that support the basic concepts of sequential operations (IF statement, multi-threading, etc.)

As with the rest of the proposed architecture, the control plane is designed to use a standards-based workflow engine. However, the workflow engine must be fine-tuned for speed to function effectively in a video system. After all, while a lag of a few seconds may be acceptable when initiating a video-on-demand (VOD) session, such a delay in more advanced real-time cloud applications (i.e., pausing or rewinding content in a cloud DVR service) would render the service unusable. Cisco worked to optimize open-source workflow engines to meet these demands, and contributed those gains back to the open-source community.

The communication with external or third-party application procedures also requires a new kind of workflow invocation service. This allows operators to deploy and test features to various overlapping subsets of subscribers, based on criteria such as set-top media access control (MAC) address, account number, service group, or requester IP address range. In the proposed control plane, options for selecting the workflow to execute also include service endpoint and designated market area (DMA) code.

Effectively, this workflow engine provides the tools to create rules, and supports tremendous customization. It also allows for greater flexibility, resiliency, and velocity when rolling out new functions or applications. Traditional systems require operators to shut down the system and come back up when implementing changes, and are unable to execute multiple workflows in parallel. The proposed control plane architecture supports multiple workflows, allowing operators to modify a workflow dynamically, without affecting connections already in use. This enables greater innovation and service velocity by giving operators web-like workflow capabilities, such as A/B testing, where a service provider can use multiple workflows that differ from one another in order to target a specific set of customers, endpoints, or video assets. For example: “Apply workflow (WF) 1 if customer lives in Massachusetts. Apply WF2 if customer is also a high-speed data customer.” The engine can even target individual endpoints, which is useful in beta-testing to familiar customers. Operators can test multiple similar versions of a function or application, and expand or roll them back with relative ease – providing a major boost in their ability to innovate, in less time, at a lower cost. This is typically not possible with a traditional video system.

In the same way, because the video control plane operates via virtualized software instances in a data center rather than tightly coupled hardware systems, the proposed architecture also gives operators greater ability to scale services dynamically. For example, the datacenter can dynamically spin up resources on the East Coast as prime time approaches, and shift those resources to the West Coast as the evening progresses.

VIDEO CONTROL PLANE

The next-generation video control plane is built upon the workflow system described above. It performs three key functions:

- Real-time session management
- Resource management
- Business policy management

Session Management

A next-generation video service must provide a framework for session management across multiple screens, in a variety of video applications. This can include “session-shifting” across devices and networks (i.e., beginning playback on the a TV via the STB, pausing, and then resuming playback later from a smartphone connecting over a cellular network), as well as more advanced applications. One example is a “companion screen” experience that integrates both a managed service (e.g., video content delivered to a managed STB over the service provider’s managed video network) and unmanaged services (e.g., IP data services that complement the STB video service and may be synchronized with it, but are delivered to an unmanaged device such as a tablet or smartphone).

The session management function of the proposed control plane architecture performs the majority of the core functions of a traditional video system back office, but includes support for both QAM and IP environments, and operates according to open, web-services practices.

The traditional approach to video services has been to perform all session processing within a proprietary environment, based on a client/server model with the STB endpoint tightly coupled with back-end servers controlling session and state. In other words, all of the logical software components (resource management, business policy, billing, entitlement, etc.) communicate with each other using closed protocols. The video control plane envisioned here functions differently: The client drives session and state, dictating the format and streaming bit rate required for a specific viewer using a specific device. As discussed, this model is based on the way software works on the web, where diverse applications and devices from multiple vendors share a common delivery language and services framework. By applying this model of session management to video services, operators can gain more flexibility and control.

Self-contained or siloed legacy video back office systems have also had trouble scaling to handle growing volumes of traffic. For a large service provider running a popular VOD service with millions of concurrent users may require dozens of separate session management systems just to handle the load. Additionally, if one part of the system was resource-constrained, the system needed to be expanded as a whole, rather than simply adding an additional node or virtual machine to support that function. Since the distributed control plane envisioned here uses SOA design principles, a shared data cache to store real-time state information, and a messaging infrastructure designed for Internet scale, a single session management system can theoretically scale to serve unlimited clients.

Resource Management

The objective of resource management is to manage video objects in the video distribution system and balance the load among video servers/streamers and networks. To achieve this, the proposed architecture employs resource management tools that intelligently and automatically consider such factors as allocation, video server selection, replication, and cache management to help ensure optimal load balancing among video servers, and to minimize the delay for video requests to be served.

In addition, the resource management function of the proposed architecture relies on the same shared cache as the session management function, and affords the same degree of scalability. The distributed control plane also uses a common resource manager for both legacy and cloud services.

The session and resource management functions of the proposed next-generation video control plane are implemented as logical individual components. Multiple component instances may be deployed throughout the operating environment as virtualized applications in varying degrees.

Flexible Policy Management and Dynamic Business Rules

A traditional TV video distribution system applies a variety of business rules to the delivery of content and services to consumer endpoints. These rules can encompass specific times content can be distributed, specific markets barred from receiving content (for example, blackout rules governing some sports broadcasts), rules barring a device from receiving content without the right content security, etc. For a cloud-based, multi-screen video service, creating explicit rules governing how content can be delivered to every possible IP endpoint in every possible location is simply not practical. The proposed distributed video control plane therefore includes a more flexible business rules engine capable of dictating rules for delivering content through the cloud (for example, allowing streaming of entitled content to any authenticated IP device that supports a particular digital rights management [DRM] system).

The business policy management function of the proposed architecture is powered by a next-generation business rules engine that brings deeper sophistication and intelligence to the process of delivering video services. The business policy management function gives operators a greater level of detail about how the content is going to be consumed. Effectively, it allows operators to store all business rules and policy logic in a centralized script or table, where it can be accessed by other elements in the system. This globally accessible data repository provides the system with vital information on sessions, devices, business rules, etc., and facilitates automated decision-making by the network in applying policy.

This repository is flexible, allowing the operator to define business rules within an XML-based searchable workflow. The workflow can incorporate not just native services associated with the control plane, but allow operators to make "off-board" calls to existing or third-party services. For example, a mobile operator could configure the workflow to make calls to an existing location service, instead of having to recreate that service for the next-generation video control plane. Furthermore, the operator can call entire off-board workflows (not just services), and take advantage of existing third-party business logic instead of having to recreate it. This is yet another example of the value of using an open workflow definition language.

Contrast that with current video systems, which give operators only the most rudimentary knowledge of where the content is being delivered. Essentially, operators know only whether content will be delivered on-net or off-net. With the business policy management tools in the proposed next-generation video control plane, operators can see beyond that, to know exactly what type of device the content is going to and the subscriber consuming it. This allows them to define more finely grained, sophisticated rules for delivering content, giving them an opportunity to generate additional outlet revenue and reduce capital expenditures.

APPLICATIONS AND USE CASES

Once the next-generation video control plane architecture is in place, operators can deploy all applications involved in the video service on top of this framework. This includes essential applications such as device authentication, service assurance, emergency alerts across multiple devices, etc. The proposed architecture is also well-suited to enabling the unique capabilities essential to delivering a cloud-based, multi-screen video service, including the ability to authenticate users and devices among multiple back-end systems (both legacy and IP), and the ability to manage and assure services across multiple devices and networks. The proposed video control plane architecture can also support more advanced applications that take full advantage of cloud capabilities, such as a synchronized companion device experience and a cloud DVR service.

Authentication Among Multiple Back-End Systems

A next-generation video system must communicate with varying types of back-end billing systems, from mainframe to web-based. For cable operators especially, authentication systems are based on a tightly coupled, usually proprietary authentication process between the STB and the back-end billing system, where the STB boots and queries the system for each subscriber's/STB's entitlements. These legacy authentication systems will likely remain in place for the foreseeable future, but they present a significant barrier for newer IP media delivery systems and endpoints, which authenticate in very different ways.

Some IP video delivery systems in use today have attempted to bridge this divide. Typically, however, this entails invasive changes to the core of the IP application to support communication with legacy authentication systems. This is to be expected: closed video systems communicate via closed, proprietary mechanisms. Exposing core software elements for authentication (or billing, or other services that require communication with a conventional video back-end system) is typically a significant development project, undertaken at a significant cost. In addition to the potentially onerous costs of this custom integration, this process also impedes an operator's ability to quickly design and deploy new service offerings that interconnect with legacy billing and authentication systems.

As discussed, the proposed distributed video control plane architecture is an open system. It provides a common infrastructure that can unify legacy authentication and billing systems with newer IP distribution services. In the proposed architecture, this is accomplished via protocol conversion mechanisms that broker this communication and provide the interface to various back-end systems. Rather than incorporating communication with legacy systems into the core of the video control plane, protocol conversion mechanisms deployed at the "edges" of the system communicate with legacy systems, while the core of the distributed services platform remains purely IP-based, and highly scalable. Effectively, this model preserves a kind of stateless, web-aware application core, even as the software communicates with older billing and authentication systems, and frees the video control plane from having to adapt to legacy systems. And, since these protocol conversion mechanisms are open and standards-based, operators need no proprietary intelligence to develop applications to communicate with the IP system.

Service Assurance and Management Across Multiple Device Types

An adaptive bit rate (ABR) streaming capability – the ability to optimize video streams for the specific connecting endpoint based on real-time network conditions – is an essential requirement of a next-generation video system. However, operators cannot rely on a standard, generic ABR functionality. More sophisticated ABR management is required for next-generation services, augmenting basic information about client and network with an additional layer of business rules and priorities, based on a more sophisticated awareness of the network and the subscriber.

Consider a premium cable customer streaming a high-definition program to a TV in her living room via a "smart" TV. Upstairs, her children begin streaming a movie on an iPad. The cable operator would not want an automated ABR function to downgrade the living room TV stream to standard-definition video halfway through the program. The video control plane envisioned here draws on network intelligence to inform ABR decisions, and uses the software control plane to effect quality-of-service (QoS) prioritization and bandwidth reservation in the network. In the scenario described above, the system can draw on operator-defined rules, as well as customizable rules defined by the subscriber, to assure that the large-screen TV in the living room takes priority over a mobile device, and that the primary subscriber takes priority over secondary users.

Synchronized Companion Device Experience

A session-aware control plane for both IP and QAM traffic can be used to deliver complementary viewing services consumed on two devices at the same time, such as viewing a live TV broadcast while using a synchronized application on a tablet. Synchronized companion services can include push applications for polling, alternative content, instant replay and other time-sensitive experiences. A viewer watching "American Idol," for example, could receive bios of contestants pushed to the companion screen when contestants come on stage, and a real-time voting application to vote for the winner during the show.

The proposed video control plane architecture is an ideal platform for deploying these types of synchronized multi-screen applications. The real-time messaging infrastructure allows the operator to synchronize the web applications, companion screen, and live TV stream to enable these and other real-time interactive multi-screen applications. The same messaging infrastructure can also support social applications, such as the ability to allow a subscriber watching a TV show to identify and chat with friends who are watching the same show. These and other synchronized multi-screen applications can benefit service providers by differentiating their services, enhancing subscriber loyalty, and driving customers to higher subscription tiers – and all are enabled by the proposed video control plane architecture.

Cloud DVR Service

Many operators are now seeking to move content storage back into the network, rather than on DVR appliances in the subscriber's home. These nDVR or cloud DVR services offer operational advantages over traditional DVR service offerings – most notably, the ability to reduce capital expenditures on costly DVR appliances. For a cloud DVR service to function effectively, however, operators need to manage those cloud-based resources and enforce business logic across unmanaged devices to deliver a seamless transition of experience for the subscriber. A cloud-based DVR service requires a system with extremely high performance that can scale sessions well beyond normal VOD utilization behavior, with hundreds of millions of assets.

The proposed video control plane architecture with its next-generation, high-performance workflow engine meets all of these requirements. In addition, because the architecture employs a web-services based control plane, it simplifies the process of extending the DVR service to other IP devices. These capabilities allow service providers to roll out differentiated, revenue-generating services like cloud DVR itself, but also allow for additional revenue-generating services. For example, operators can invoke a workflow that detects when a subscriber's DVR storage is nearly full, and pushes a message out to the user's companion device to ask if the subscriber would like to purchase additional cloud DVR storage space.

CONCLUSION

For many service providers, the shift from traditional QAM-based video architectures to an open, cloud-based distribution model is a five- to 10-year transitional exercise. But the video services landscape is rapidly evolving, and this transition already is well underway. Video service providers are faced with the need to adopt cloud distribution capabilities even as they serve existing customers over legacy infrastructure.

A key to navigating this transition is handling video session, resource, policy, and other functions from a common software control plane. The proposed next-generation video control plane architecture can manage this transition, offering a range of overlapping and mutually reinforceable benefits. These include:

- **Service velocity:** Transitioning to a video control plane based on open programming languages and web services offers a competitive advantage. By providing web services management capabilities for video services, the platform helps enable rapid testing and deployment of new features and applications to a large number of end devices and targeted sets of customers, while allowing operators to continue taking advantage of legacy systems in a graceful manner.
- **Flexibility:** A core value of SOA-based systems, flexibility is exemplified in the proposed control plane architecture and its real-time execution of multiple workflows. Operators have the ability to develop and test all manner of features and applications, in both targeted and large-scale deployments. The ability to dynamically manage multiple workflows in parallel also allows operators to roll out changes and upgrades with no downtime.
- **Scalability:** While legacy systems could expand to handle growing volumes of traffic only with difficulty and cost, the proposed architecture scales simply by adding another node or virtual machine to support any given function. Based on Internet design and communication principles, it provides a platform for video services at massive scale, capable of supporting theoretically unlimited clients.

Ultimately, the proposed video control plane fills a critical gap in the ongoing evolution of service provider networks into a cloud-based delivery framework. Legacy transport technology has many years of life remaining. This platform allows providers to continue supporting those services, while adapting to emergent market realities and using the most efficient cloud-oriented software architecture, techniques, and methods to deliver compelling new subscriber experiences.

REFERENCES

[1] Cisco. (2011). *Cisco Visual Networking Index: Forecast and Methodology, 2010-2015.* Cisco. Retrieved April 1, 2012, from Cisco.com: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

[2] Cisco. (2011). *Cisco Global Cloud Index: Forecast and Methodology, 2010-2015.* Cisco. Retrieved April 1, 2012, from Cisco.com: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

COMPLEXITY CONSIDERATIONS FOR CENTRALIZED PACKAGING VS. REMOTE PACKAGING

Brian Tarbox
Motorola Mobility

Robert Mack
Motorola Mobility

Abstract

Adaptive streaming protocols will be a critical component for operators offering IP video services. One of the key functions in Adaptive streaming is a "packaging" function that creates playlists/manifests, segments the video into chunks, and "wraps" the chunks to make them suitable for one of several protocols. There is an ongoing debate as to the merits of where to perform adaptive stream packaging within a service provider's content delivery network (CDN). Various analyses have considered centralized, distributed, and edge packaging architectures. These analyses primarily considered the CDN bandwidth and storage savings that could be attributed to distributed/edge packaging architectures versus the operational complexity that would likely result. In addition, these evaluations focused more on video on demand (VOD) rather than linear content distribution.

For many Service Providers the ability to centralize all transcoding and packaging operations is appealing, particularly if they own the CDN and are therefore less concerned with the per-bit content distribution costs. For other Service Providers, particularly those that want to augment their existing service with streaming capabilities, but are sensitive to these costs and the costs associated with standing up large centralized video processing centers, the ability to customize content at the edge may make more sense. So, for example, a Tier 2 or Tier 3 operator may want to augment his offerings out of an existing Regional Headend.

In addition, edge packaging may offer options that can reduce the complexity associated with providing certain desired system functions. For example, considering regional ad insertion and blackout, edge packagers can incorporate simple functions that emulate similar legacy system capabilities which minimize the impact to a service provider's network and operations.

Finally, this paper will also explore some of the unique functional capabilities that packagers can offer in support of centralized or regionalized architectures, including intelligent access network capacity management, playlist obfuscation for ad insertion, regionalized blackout, and support for both legacy and advanced advertising in adaptive environments. This will enable operators to fully understand the trade-offs of implementing various packaging architectures and make the right choices when rolling out IP video services.

INTRODUCTION

Previous papers examining where to perform packaging focused on network bandwidth and capital expenditures as the variables to measure. This paper examines other factors that must be explored in order to enable sound decision making in systems architecture design. These factors focus on the processing complexity required to output segments and manifests under various expected conditions. These complexity considerations may make the support for certain desired capabilities problematic. In particular, this paper examines the ability of the various packager configurations to support regionalized and targeted ad-insertion as well as to support blackout processing.

BACKGROUND

Packager Locations Within the Distribution Network

Until recently most discussions about packager locations listed center and edge as the options. In 2012, however, the concept of a hybrid center/edge packager has gained traction. Each will be described briefly.

In the center packager, the packaging function configuration is embedded within the transcoder, or the packager is connected to the output of the transcoder which, in turn, is connected to the origin server. All video stream processing, including fragmentation, ad-insertion, manifest creation, etc. is performed prior to any client request for the content. All fragments and manifests that might be required or requested are deposited onto the origin server.

In the edge packager configuration, the transcoder outputs its fragments (e.g., MP4 or FMP4) and optionally a mezzanine manifest file onto the origin server. One or more packagers are located "south" of the CDN. The client connects directly to the packager or to an optional edge server which redirects to the packager. The packager then requests fragments and the optional mezzanine manifest file from the origin server, transwrapping on-the-fly into the appropriate format for the requesting clients (e.g., HLS, HDS, HSS).

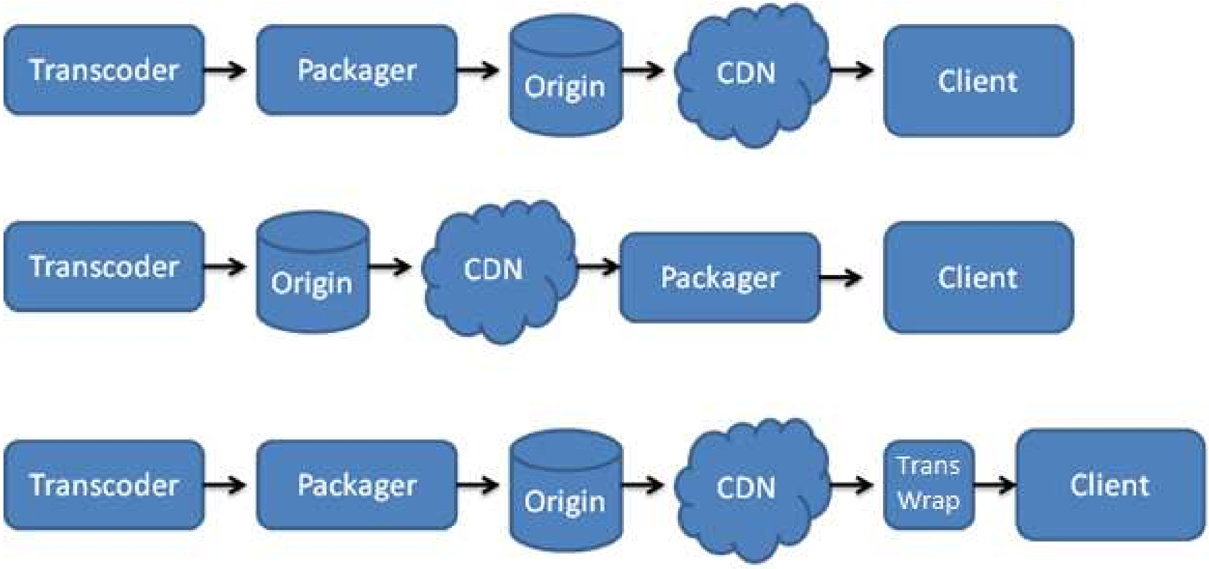


Figure 1: Three Styles of Packaging

The hybrid solution is similar to the center configuration except that it adds a transwrapper component located between the client and the CDN. This is a relatively new style of configuration and so many options are possible. The “north side” packager might transwrap to one of the four standard formats (HLS, HDS, IIS, DASH), leaving the transwrapper to re-wrap only if a given request was for a different format. The transwrapper component might also perform session or region specific operations such as ad-insertion or blackout.

Regionalization Perspective

Some service providers may want to augment their existing home cable/data service with streaming capabilities for in-home IP devices. They may not being trying to stand up an OTT type of service to off-net subscribers, rather they want to offer current subscribers the ability to use their portable devices or want to deploy IP STBs with streaming capabilities. Furthermore, they plan to perform multi-rate transcoding within their existing Central Headends and will deliver the multi-rate transcoded transport streams or mezzanie files to the Regional Headends via their IP distribution networks or potentially via CDN. Just as in their legacy systems, regional customization takes place at the Regional Headends. This is where an Edge Packager could be used to perform regional ad insertion or even blackout processing.

Tier 2 and 3 service providers may also wish to offer streaming services and are sensitive to CDN content distribution costs. Similarly, they may wish to receive multirate transcoded content directly from content providers or resellers and need to perform regional customization (e.g. regional ad insertion).

FUNCTIONAL CAPABILITIES SUPPORTED BY PACKAGERS

Any discussion of ad insertion must occur in the context that viewers do not, in general, want to watch ads. They will go to substantial lengths to avoid ads, with smartphone and tablet users running clients that have been specially designed to defeat ad presentation systems. If one builds a manifest that looks like the following, you can bet that someone will find a way to avoid watching fragment three.

```
<fragment time=1, length=10, uri=LOTR.mp4>
<fragment time=11, length=20, uri=LOTR.mp4>
<fragment time=21, length=80, uri=http:/www.ad-decision-system.com>
<fragment time=81, length=10, uri=LOTR.mp4>
```

Figure 2: Easy to Defeat Manifest

Center Packaging Ad Insertion

Central packaging implies that all of the work to create fragments and manifests is completed prior to any request for content. The transcoder outputs to the packager which outputs to the origin server and that “transaction” is complete. The transaction from the client to the origin server (through one or more CDNs) is a completely separate transaction. It may not be apparent, but this is true regardless if the content is VOD, linear, or network digital video recorder (nDVR). In both VOD and nDVR there is a gap between the recording/packaging of content and its eventual playback. Even in the linear case, however, the packager is essentially filling a bucket (the origin server) while the client is emptying that bucket. This becomes more clear if the linear case is expanded to consider StartOver TV. StartOver is basically linear with a limited ability to jump back to the start of a program. In adaptive bit rate (ABR) linear TV, there is what amounts to a 30-second jitter buffer to deal with packaging and manifest creation; StartOver just expands that buffer to 30 minutes.

This implies that all manifest information for regional or targeted advertisements must be built prior to content request. This can be accomplished in two ways. First, one could create multiple manifests, one for each ad region. The session manager then needs to point the client to the region-specific manifest. It does not appear that this approach is feasible for targeted advertizing. The second approach is to build a single manifest such that the ad decision manager (ADM) or some other component is invoked at playback time.

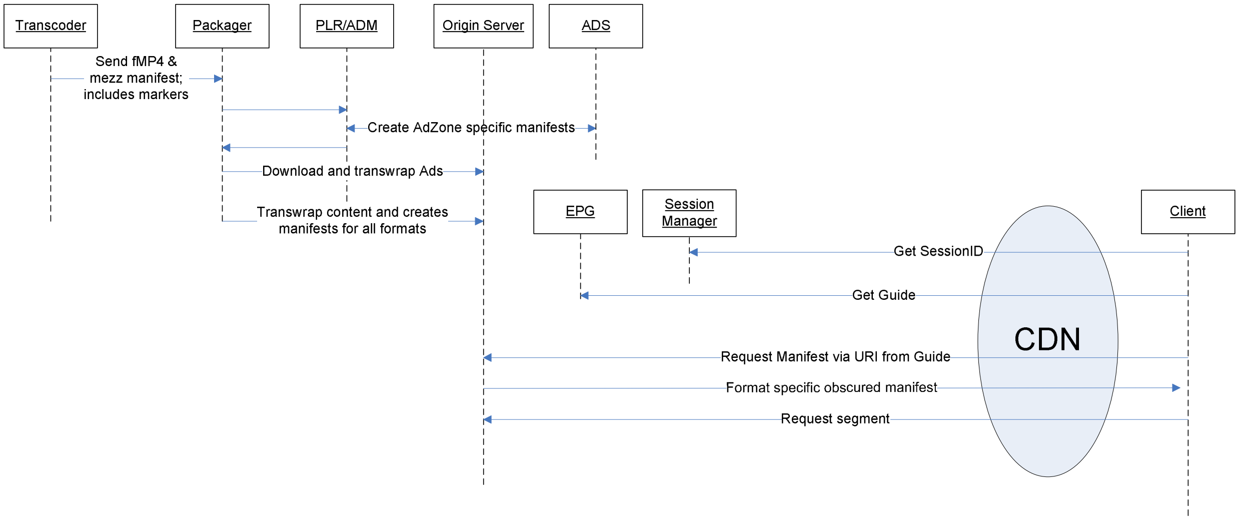

Transcoder
Packager
PLR/ADM
Origin Server
ADS
Send fMP4 & mezz manifest; includes markers
Create AdZone specific manifests
Download and transwrap Ads
Transwrap content and creates manifests for all formats
EPG
Session Manager
Client
Get SessionID
Get Guide
CDN
Request Manifest via URI from Guide
Format specific obscured manifest
Request segment

Figure 3: Center Packager for Ad Insertion

For the second approach to work, the manifest entry corresponding to an ad (or set of ad segments) must be a uniform resource identifier (URI) that can be processed by the origin server or by the ADM itself. One could imagine generating a URI pointing to the origin server that was encoded in such a way that the origin server could detect it, perform some processing on it, and issue an HTTP redirect to the ADM with some localization parameters supplied. Alternately, the URI could point to the ADM directly with some guarantee that the parameters required for localization would be directly supplied by the client. In this last case the URI cannot effectively be obscured. In addition, the client may be able to spoof the parameter(s) used to direct the system towards the targeted ads.

Edge Packaging Ad Insertion

Edge packaging defers the creation of manifests and fragments until the content is requested. This means that a session ID (targeted ads) and/or region ID (regionalized ads) are known at packaging time. Given this information, the packager can use its Play List Rebuilder (PLR) function to query the ADS during manifest creation. It can also cache those results. Since it does not have to pre-build the manifests, it can build regionalized manifests only as needed. This is a savings since it is unlikely that every program that is recorded will, in fact, be requested from all possible ad regions. Granted, manifest creation is not expensive, but managing an explosion of files that might never be used can add complexity to the overall solution.

Since manifests are only created as needed, it becomes possible to create targeted or per-user/per-session manifests. Depending on the ADM being used and the campaign that is in force, it may be desirable to create viewing experiences tailored for a single user at a particular time. While such a manifest cannot be cached since it is a onetime use artifact, at least it can be created. In the standard central packaging configuration, targeted manifests are simply not possible. It should also be observed that while the manifest for a targeted viewing cannot reasonably be cached, the content and ad fragments associated with that viewing may well be cachable.

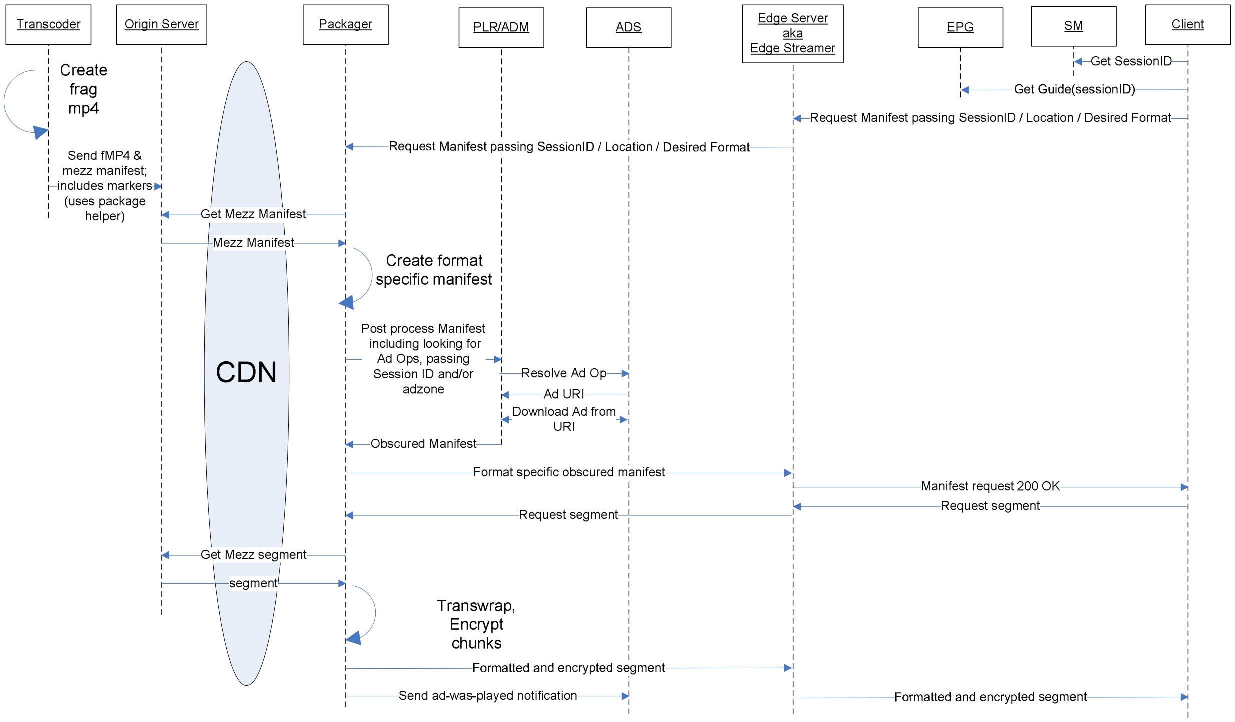

Transcoder
Origin Server
Packager
PLR/ADM
ADS
Edge Server aka Edge Streamer
EPG
SM
Client
Create frag mp4
Send fMP4 & mezz manifest; includes markers (uses package helper)
CDN
Get SessionID
Get Guide(sessionID)
Request Manifest passing SessionID / Location / Desired Format
Request Manifest passing SessionID / Location / Desired Format
Get Mezz Manifest
Mezz Manifest
Create format specific manifest
Post process Manifest including looking for Ad Ops, passing Session ID and/or adzone
Resolve Ad Op
Ad URI
Download Ad from URI
Obscured Manifest
Format specific obscured manifest
Manifest request 200 OK
Request segment
Request segment
Get Mezz segment
segment
Transwrap, Encrypt chunks
Formatted and encrypted segment
Send ad-was-played notification
Formatted and encrypted segment

Figure 4: Edge Packaging Ad Insertion

Viewership Management

Another significant problem with Central Packaging is related to fulfillment or viewership management. As seen in Figures 3 and 5, there is no obvious way to indicate when an ad is actually played by the client. Keep in mind that the origin server is acting as a simple web server and the distinction between content fragments and ad fragments has been obscured. This appears to leave us unable to inform the ADS when the fragments corresponding to a particular ad were ever actually viewed by a client. This would seem to imply the need to put some level of intelligence in the origin server, yet the proper level of that intelligence is elusive. If a pattern to the manifest entries is declared such that the fulfillment observer could determine which fragments were actually ads, one would have to assume that the clients could detect the pattern.

Alternately one could imagine creating a back channel from the origin server back to the packager providing fragment-requested information for all fragments. The packager presumably knows which fragments are ads and could be the component to send the fulfillment message to the ADS That begins to blur the lines of function design for the packager and also makes assumptions about packager knowledge.

To avoid this one could add a component to the control plane path such that whenever a fragment was fetched by a client this new component could detect when an ad fragment was requested and inform the Ad Management Service. This control plane component would have to deal with the fact that an ad fragment might be cached by the CDN and not actually fetched from the Origin Server.

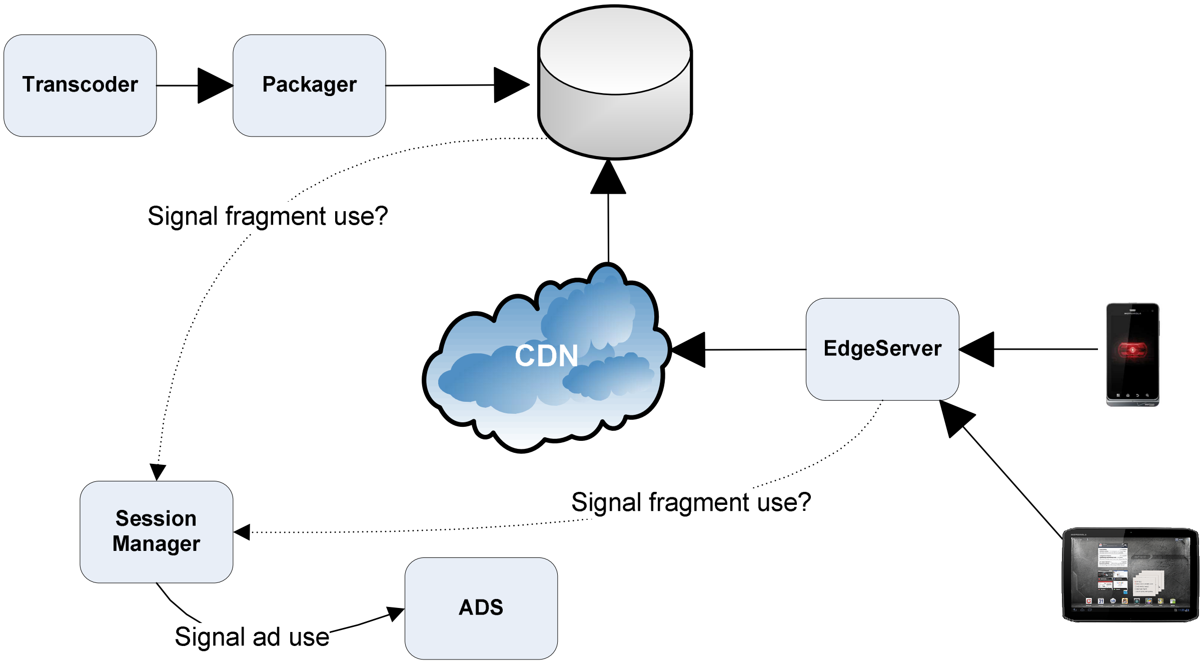


Figure 5: Center Packager Viewership Monitoring

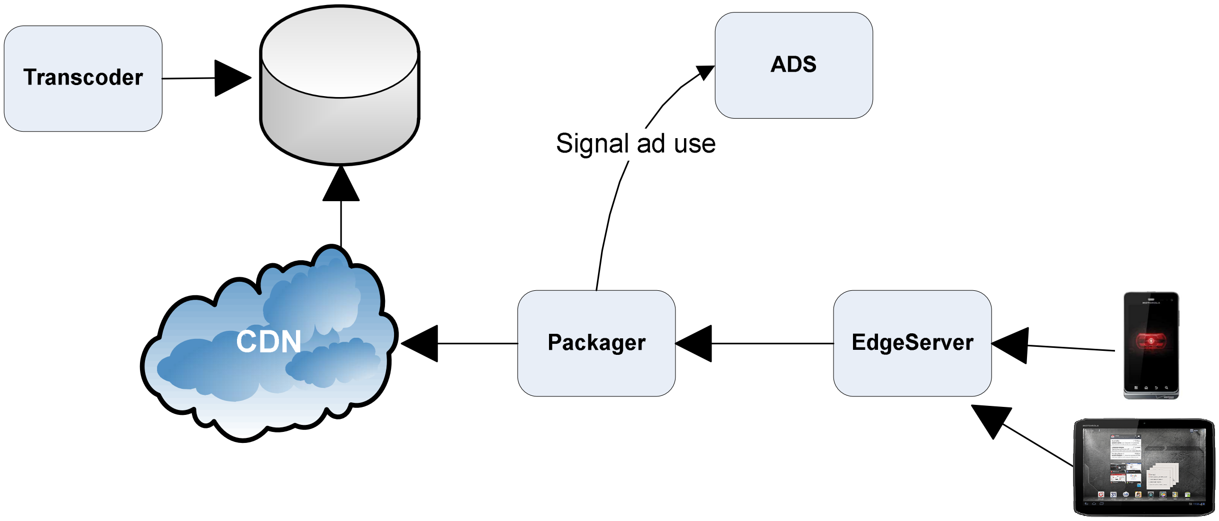


Figure 6: Edge Packaging Viewership Monitoring

Such a centralized control plane component would also seem to become a bottleneck and/or single point of failure as it would appear to need to be involved in every fragment request from every client.

Yet another approach would be to split the packager function into separate transwrapping and manifest generation components. One could then position the manifest generation component logically between the client and the origin server. This hybrid option will be discussed in more detail later

Each of these problem becomes simpler in the Edge Packaging deployment shown in Figure 6. As can be seen from the diagram the packager is involved in all fragment requests and can directly inform the ADS about fragment downloads. This is especially convenient since the Packager component is already well positioned to have knowledge of which fragments are and are not advertisements.

Ad Zone Dynamism

Late binding of manifest creation also allows for dynamism in the set of ad zones. Ad zones are only applied to manifest creation at session creation time. While this may not be a large difference for linear or nDVR applications, it can be for VOD content. VOD contents are ingested once and then may exist in the system for weeks or months. Some classic content might stay in the library for the life of the system. It is certainly imaginable that an operator might want to change the configuration of their ad zones during this time window. It’s not clear that there is a way to modify ad zones in a center packaging configuration.

Time Based Ad Selection

Edge packaging also provides more options for ad selection. For example, suppose the system wants to generate a targeted ad based on the playback time of the content. An example would be a show-teaser ad suggesting that the user watch the “next” program. Except in the case of linear content, “next” will mean something different at playback time than it did at record time. If a user records a show on Monday at 7:00 PM they may receive teaser ads for the 8:00 PM Monday show on the same channel. Playing that content back on Tuesday results in a teaser ad that is, at best, useless and, at worst, defeats the viewers quality of experience. Receiving promotions for a show that you missed and cannot, in fact, watch could frustrate viewers. Therefore, the ad decision that will be reflected in the manifest should be made at playback time, not at record time.

Hybrid Packaging

A hybrid approach to packaging should be explored to round out the possibilities. As alluded to earlier, this approach involves using a central packager to perform content and ad fragment chunking. All such fragments are loaded onto the origin server, along with one or more undifferentiated manifest files.

Between the origin server and the client is another component, the transwrapper. This component or software service may be co-located with the origin server or may be a deployed on separate hardware. The intention is that the URI supplied to the client for obtaining the manifest should resolve to the transwrapper component.

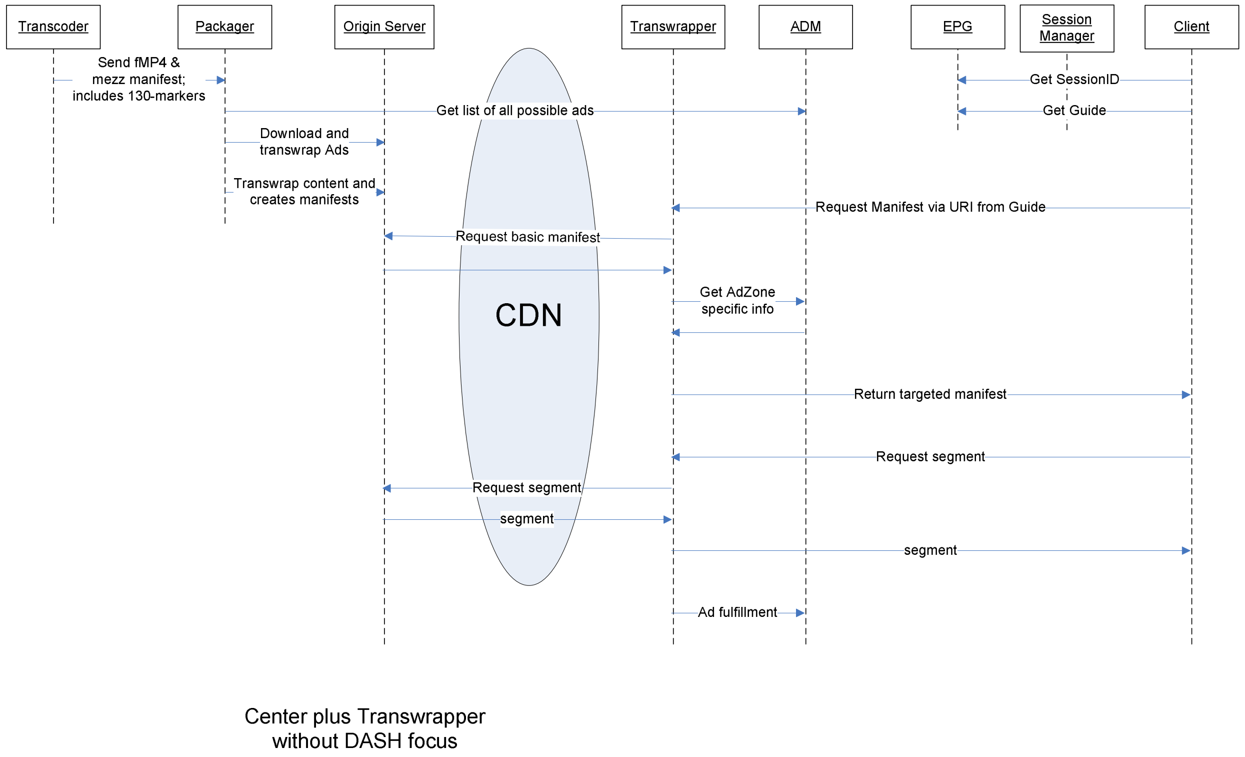


Figure 7: Hybrid Packaging

The transwrapper uses information in the client request to a) transwrap as needed to the client's requested format, and b) perform ad personalization. To do this, the client must supply a regionID for regionalized ads or a sessionID for personalized ads or equivalent data that can be resolved into an appropriate ad zone. Based on that derived ad zone, the transwrapper assembles a manifest tailored to clients within that ad zone.

At first glance, this seems like a reasonable compromise that achieves many goals. From another point of view, however, the system now has both center and edge packaging components. In other words, it's not clear that Hybrid packaging has any advantage over Edge Packaging.

CENTRAL & REGIONAL BLACKOUT APPROACHES

It is anticipated that blackout control will be a required function in multi-screen environments, wherein IP set-top boxes, in-home portable devices, and portable devices outside the home will have to be restricted from receiving content based upon their location (or the subscriber's home location) during a service substitution event. Today, blackout is enforced through the content provider's uplink control system. Normally, a retune command is inserted at the uplink and targeted to individual integrated receiver/decoders (IRDs) known to be operating within a specific region. When a specific IRD observes a retune command addressed to it, it mutes the video stream or replaces it with an alternate service for the duration of the blackout. Similar functionality can be provided in multi-screen systems by manipulating playlist/manifest files during a blackout event.

In a centralized architecture, where packaging and manifest creation is performed, a new manifest or sequence of manifest files needs to be created during the blackout event for clients within the affected region. The new sequence of manifest files will direct those affected clients to tune to alternative content for the duration of the blackout event. Through an element known as a Blackout Manager, unique regional manifest files are generated for all the blackout regions under its control. The Blackout Manager must have knowledge of the CDN topology and specifically the mapping of each edge cache to the specific geographic region it services. It is required to continuously monitor for blackout events by processing IRD retune messages for its regions. When a blackout is in effect for a given region, it requests manifest updates from the playlist rebuilding function, which are subsequently published to the CDN. The updated manifests reference new URLs that point to alternate content during a blackout for one or more affected regions.

Figure 8 illustrates the system. The blackout manager requests that the playlist rebuilder generate unique manifests for each of the three regions under its control, namely, Pittsburgh, Philadelphia, and State College. The manifest (M) is retrieved from the packager and the SportsNetwork.ServiceProvider.net/pitt, SportsNetwork.ServiceProvider.net/philly, and SportsNetwork.ServiceProvider.net/StateCollege manifests are created and published to the CDN. The content identified within the different regional manifests can be identical until such time as a blackout event is required to be enforced. At that time, the manifest file for that area is modified by replacing the blacked out content URLs with URLs for the content to be substituted.

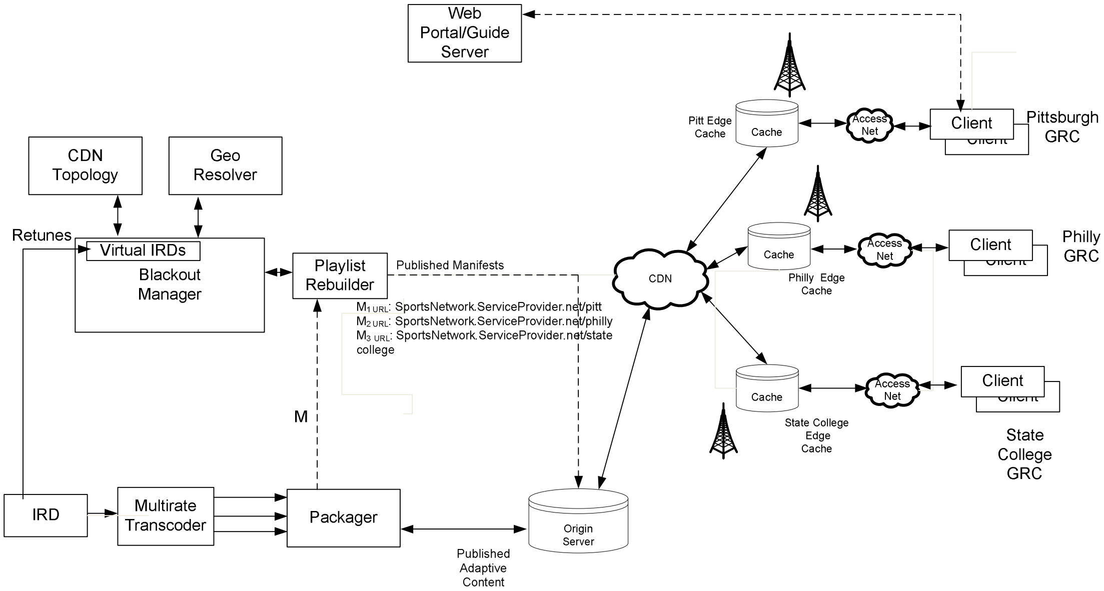


Figure 8: Centralized Blackout Management

A particular client within a given blackout region can retrieve the right manifest for that area through a number of techniques listed here:

1) Client GeoLocation: Client geolocates itself using embedded GPS technology or geo position services available on the network. The client reports its location to an upstream control plane element—for example, a session manager—which, in turn, identifies the appropriate URL for the manifest associated with that region. Alternatively, the client can construct an HTTP request that includes location metadata, which results in the return of a location-specific manifest.

2) Control Plane GeoLocation: The client is geolocated by control plane elements within the network. For example, a session manager that the client communicates with could use a geoLocation service to resolve the client's source IP to a location within the network. The session manager, in turn, identifies the appropriate URL for the manifest associated with that region to the client.

3) Edge Network GeoLocation: Edge network elements append location metadata into a client's HTTP request relying on the network's knowledge of where HTTP requests entered the network.

Each of these options has advantages and disadvantages as described below. For the first option, it is necessary to have clients that can perform geo location processing. Without this capability, they would be unable to request a manifest for their particular GRC and would likely receive the most restrictive manifest (blackout area), even if they were not located within a blacked out area. It's also not hard to imagine the development of downloadable applications that will allow clients to spoof their actual location within an HTTP request for content.

In the second option, control plane elements are in the critical path of determining the client's location at all times. This is particularly difficult if the client is mobile and is crossing different blackout zones. Each time the client enters a new zone it must be detected by the control plane elements so that a new manifest for that zone can be delivered.

The third option is the most ideal way. Here, a function exists within edge distribution network elements (e.g., CDN cache) that can append location-specific metadata into the client's HTTP request for the manifest file. For example, if Service_Provider is the subscriber's service provider, and if the subscriber is trying to acquire the SportsNetwork broadcast, the guide/navigation function would provide the SportsNetwork URL, SportsNetwork.ServiceProvider.net/index. The location- specific metadata would be inserted within the access network or at the boundary (edge cache) between the access network and CDN ingress point, so as to accurately identify the physical location of the client. This would result in a modified URL, SportsNetwork.ServiceProvider.net/pitt/index. This is a simple and reliable method that even works if mobile clients cross GRCs dynamically. As the client moves in and out of different access points, the network elements at the edges of the network add location-specific metadata that becomes part of the request, resulting in the return of the appropriate manifest. Of course this approach requires this functionality to be incorporated into access networks or CDNs in a standardized fashion.

An alternative that models today's blackout system solutions uses the packager to enforce blackouts within a given region. As illustrated in Figure 9, edge packagers are located within the different blackout regions.

Each live packager is configured to assume a virtual IRD identity. The IRD retune messages received from the satellite downlink are carried within metadata that is distributed to all the packagers in the regions. Each packager filters for retune messages destined for its VIRD identity. A live (edge) packager that observes a retune message addressed to it will update the manifest it is creating with URLs that point to alternate content during the duration of the blackout.

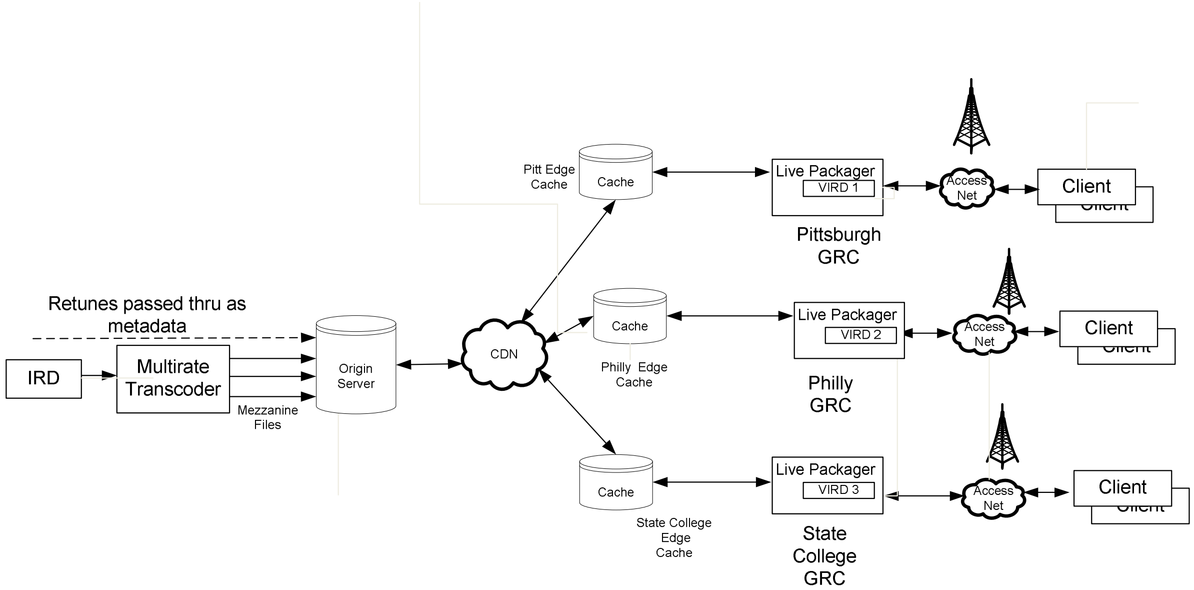


Figure 9: Edge Blackout Architecture

The following is a summary of the advantages and disadvantages provided by the centralized and edge/regional blackout solution options:

<u>Centralized</u>

Advantages

- All operations are managed centrally.
- All content processing equipment can be co-located.

Disadvantages

- A blackout management function is required that understands the GRCs it manages and CDN topology.
- A scalable playlist rebuilding function is required.
- Network changes may be required to append location information to HTTP requests or higher level managers may be required to point the client to the appropriate manifest URLs

<u>Edge</u>

Advantages

- There is no need for a centralized blackout management function.
- There is no need for a centralized, highly scalable playlist rebuilder function.

Disadvantages

- Requires deployment of edge packagers.
- During BO, content replacement has to be facilitated at the edge packagers.

CONCLUSIONS

The choice of where to locate the packaging function is a complex one with implications far beyond capital expenditure decisions such as how much hardware to purchase and where to locate it. Previous discussions of central versus edge packaging have focused on the costs of network infrastructure. There are however many other important factors that should be considered. The packager has been thought of as a simple component that provides chunking and creates manifests. As has been described in this paper the truth is that many of the high value features of the overall system are highly dependent on the packager configuration. Ad insertion and blackout control are two examples of such high value features.

As an operator, do you want to support generalized ads, regional ads, or targeted ads? How concerned are you about the race between ad providers and ad-defeating clients?

Are you focused exclusively on one of several possible subsets of the viewing experience (VOD, linear, nDVR, ABR, multi-screen)? Depending on the subset, the problem space changes. One may want to consider a phased approach or look at the entire problem when planning a deployment.

Of course, the choice may not be between center or edge packaging. It may well be between center and Hybrid packaging, and edge packaging. One could well argue that simply using edge packaging is the simpler solution.

CREATING CONTENT WITH EXTENDED COLOR GAMUT FOR FUTURE VIDEO FORMATS

J. Stauder, J. Kervec, P. Morvan, C. Porée, L. Blondé, P. Guillotel
Technicolor R&D France, jurgen.stauder[jonathan.kervec]@technicolor.com

Abstract

New technologies in capturing and displaying images with extended color gamut and new standards for wide gamut color encoding enable a new market of extended-color-gamut content (video, images, games, electronic documents). What is the challenge and what are the issues when feature film production goes for extended color gamut? This paper discusses two topics: digital capture of extended color gamut scenes and color correction of wide color gamut footage. In film production, proof viewing and initial color decisions migrate from the post-production facility to the production site. When capturing digitally scenes with extended color gamut, what can be expected to be seen on the proof monitor? This white paper discusses the issues of sensitivity metamerism, color resolution and color clipping. Once captured, color correction creates the aimed looks for digital cinema viewing, TV home viewing, and other possible means of consumption. This paper discusses the issue of color correction with the constraint of multiple means of color reproduction. A new method is presented that supports the colorist to handle multiple color gamuts using the concept of soft gamut alarm.

INTRODUCTION

When looking into history of motion pictures and technology of argentic film, people always tried to enhance image quality and user experience. In 1932, Technicolor invented the 3-color-dye system starting worldwide the transition from black and white to colored motion picture. More recent efforts aimed to enhance resolution and image size from 35mm to 70mm argentic film [1] or from classical 2D film projection to 3D projection [2]. In all these examples, people tried to enhance image quality while preserving as much as possible from existing infrastructure. The color print of 1932 could be projected using the state of the art film projectors of that time. The film reels were the same. When testing 70mm film stock, the constraint was to keep the Digital Intermediate workflow of 35mm technology. For 3D film projection, the inventors [2] used classical film projectors and same film stocks, they just added an optical system.

In television and video, current standardization efforts include the increase of fidelity of color reproduction and the extension of color gamut. Aiming the fidelity of color reproduction, the EBU specified recently the reference monitors to be used in production and post-production [3]. The IEC specified a metadata format called “Gamut ID” to transmit color gamut information for better color reproduction [4, 5]. In order to increase the color gamut (and the image resolution) from High Definition (HD) to Ultra High Definition Television (UHDTV), the ITU-R (WP6C) looks into extending the color gamut. More precisely, they specify a video signal encoding format [6, 7] that allows conveying colors that are more saturated than specified in current HDTV color encoding format ITU-R BT. 709 [8]. Similar efforts have been done in SMPTE and IEC [9,10,11] but these solutions are not widely used.

If the video industry intends to migrate from HDTV to UHDTV, *production*, *distribution* and *consumption* of video needs to be adapted. For *consumption* of extended color gamut, display makers announce for 2012 first OLED TV screens able to show 40% and more of all visible colors (current displays are limited to 33%). Video *distribution* is addressed by ITU-R.

This paper focuses on the *production* of video with extended color gamut and presents two aspects.

First, extended color gamut will have impact on acquisition using digital cameras. While sets usually are prepared in a way that illuminance of surfaces and colors keep within usual ranges, directors now start to use lights and colors with peaky spectrum, or higher saturation. Three issues of digital acquisition will be discussed: sensitivity metamerism, color resolution and color clipping.

The second topic concerns color correction aiming multiple color displays with different, extended, color gamut and viewing conditions. The concept of soft gamut alarm will be introduced and illustrated.

EXTENDED COLOR GAMUT IN DIGITAL ACQUISITION

New requirements in production using digital cameras include the capture of scenes showing colors with wider color gamut. Directors start to light scenes on production sets with colors that are out of the color gamut of usually used proof viewing devices (such as Rec. 709 monitors). For example in music life events, modern spot lights use programmable color filters able to generate light of high degree of saturation. In traditional production using digital cameras, such colors are avoided. In straight forward signal processing, illegal RGB values may be simply clipped somewhere in the imaging chain. This causes the color output on the reference screen to be widely different from the colors that can be seen in the scene. There is a need of controlled handling of out of gamut colors, in which the errors are minimized.

Color encoding

Before discussing camera specific issues, some basic terms are recalled. The skilled color scientist will skip this section. When a color is expressed by color space coordinates, this is called color representation. When color representation includes aspects such as binary encoding and reduced validity such as device or observer dependence, this is called color encoding.

One type of color encoding is scene-referred color encoding. The principle of color encoding has been structured by the ISO [12] for the field of digital photography and desktop publishing, but the definitions are valid for the video domain, too. Scene referred color encoding identifies color coordinates that are meant to be directly related to radiometric real world color values. The raw RGB output values of a digital camera are usually transformed to scene-referred RGB values, such as defined by ITU-R BT.709 [8]. However, we will see later that this relation is ambiguous due to sensitivity metamerism.

Another type of color encoding is output-referred color encoding. As opposed to scene-referred color encoding, output-referred color encoding is used to represent reproduced colors. Output-referred color encoding identifies color coordinates that are prepared for specific output devices with their defined characteristics and viewing conditions. For example, RGB values of a video can be said to be output-referred color encodings since they are intended for a reference display under reference viewing conditions. Well-known output-referred color encodings are for example sRGB display input values or CIE 1931 XYZ values.

Output-referred color encodings are obtained by color matching experiments. An output-referred color space and the related color matching experiment are characterized by:

- the characteristics of the output device driven by the output-referred color coordinates;
- the characteristics of the observer that perceives the colors reproduced by the output device.

Let us take as example the output-referred RGB coordinates being input to a display. The related trichromatic color matching experiment is classical [13] and involves the CIE 1931 standard (human) observer, corresponding to the average behavior of a small group of test persons. In the experiment, an observer compares the color reproduced by the display with the color of a monochromatic light of a specific wavelength. For each wavelength, he adjusts the RGB values such that both colors match. The result of a color matching experiment are three color matching functions (red, green and blue) indicating, for each wavelength, which RGB coordinates should be input to the display in order to match the monochromatic light.

The classical color matching function results in the output-referred RGB color space of the specific RGB display that was used at the time of the experiment. An RGB space can be defined for any other RGB display.

Better known is the output-referred CIE 1931 XYZ space based on an ideal display with XYZ input signals and mathematically derived XYZ primaries. XYZ coordinates encode a color according to these standardized primaries and according to the CIE 1931 standard observer.

Less known is that we could build an $R^C G^C B^C$ or $X^C Y^C Z^C$ output-referred color space that is based on a digital camera as observer. Let us recall that output-referred color spaces not only depend on the aimed display but also on the referred camera used as observer.

Linear output-referred color spaces can be transformed into each other using a linear coordinate transform as far as the same observer is considered. Hunt [13] shows this for RGB-XYZ transform and the SPMTE [14] for different RGB spaces of different displays. Trichromatic observers (such as the human eye or a digital RGB camera) are characterized by the spectral sensitivities of their photoreceptors. The set of three spectral sensitivities are directly linked to a set of three XYZ color matching functions. One set can be derived from the other but they are of different nature.

Color characteristics of digital cameras

The color performance of a camera is determined by a series of elements:

- Optical system (chromatic aberration, transmission);
- Color filters (shape and coverage of spectrum);
- Primaries separation (beam splitting or CCD RGB pattern);
- Color signal processing (noise, colorimetry transform).

From color science point of view, a classical color image camera is a trichromatic observer. Another well-known trichromatic observer is the human standard observer.

A digital camera is characterized by its spectral locus, defined by the coordinates of all responses to monochromatic light in $R^C G^C B^C$ or $X^C Y^C Z^C$ or even $x^C y^C$ spaces. The spectral locus is the characteristic of a camera that corresponds to the color gamut of a display. The camera spectral locus is less known than the spectral locus of the human observer, but is of the same nature since a classical color image camera is just another trichromatic observer. The spectral locus is represented in an output-referred color space and can be derived directly from the corresponding color matching experiment (see further below). For example, from CIE 1931 XYZ color matching functions, a pair of xy coordinates can be calculated for each wavelength. Plotted in the chromatic xy diagram, these points define all together the curve of the spectral locus. The spectral locus circumscribes all colors that are visible by the observer.

Sensitivity metamerism

Metamerism happens when different spectral power distributions result in the apparent matching of colors for a human eye, or matching of color coordinates for a camera acquisition.

A camera transforms a real-word color stimulus, defined by a spectrum, into three RGB tristimulus values. Similarly to human vision, cameras are subject to metamerism. This raises issues in two directions:

- A given camera may produce identical tristimulus values for two (or more) different spectral stimuli, called a metameric pair (or metameric set, respectively);
- A camera with sensitivity curves different from the human eye differs in their metameric pairs from a human observer.

The link between scene-referred camera RGB values and CIE 1931 XYZ coordinates cannot be trivial since two different spectral sensitivity curves sets are involved, that of the camera and that of the human eye, respectively. Camera and human eye may differ in their metameric pairs leading no non-invertible relations between RGB and XYZ coordinates such as illustrated in Figure 1. Distinct *rg* points can correspond to the same *xy* point and vice versa. *rg* and *xy* chromaticity coordinates are obtained from the RGB scene-referred camera output values and from the output-referred CIE 1931 XYZ values, respectively, by normalization [13].

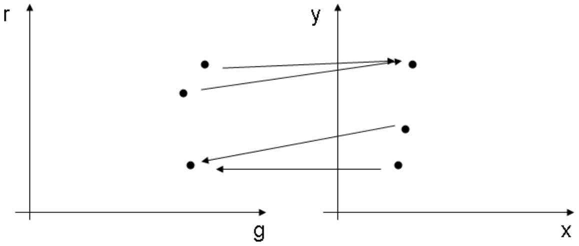


Figure 1: Non-invertible relation between rg and xy due to sensitivity metamerism

This problem is referred to as sensitivity or observer metamerism and can be avoided completely only if the camera satisfies the Luther condition [15] i.e. if its spectral sensitivities are linear combinations of the color matching functions of the CIE 1931 standard observer. Another solution is multispectral cameras [16].

Color clipping in proof viewing

A solution to the problem of sensitivity metamerism would require the estimation of scene-referred and human observer related color values, for example CIE 1931 XYZ values, from camera raw RGB output [15,17]. However, in proof viewing we have a different problem: How to reproduce captured colors on a given proof viewing monitor?

When proof viewing a camera raw RGB output signal on an RGB proof viewing monitor, the raw RGB values should be transformed into output-referred RGB values. We call this a proof viewing color transform. As shown in before, such a proof viewing color transform can exist only up to metamerism difference between the camera and the human eye.

For analysis, let us develop a straight forward proof viewing color transform. For presentation purpose we neglect any non-linearity. For a given camera and a given proof viewing monitor, a straight forward proof viewing color transform can be determined by the following steps:

- Determining the three scene colors that are within the color gamut of the proof viewing monitor;
- Measuring the camera output RGB values for these three colors;
- Determining the monitor input $R^mG^mB^m$ values for these three colors;
- Set a linear RGB transform RGB to $R^mG^mB^m$.

When applying this transform to the camera RGB output values, attention has to be paid to $R^mG^mB^m$ values that are outside of the valid coordinate range, for example [0;1] for normalized RGB values or [64;940] for 10 bit encoded RGB values in TV systems. The values should either be clipped, or soft clipped or compressed into the valid coordinate range.

Figure 2 shows an example for simple color clipping. We set a series of scene colors outside of the proof viewing monitor color gamut and captured them by a digital film stream camera. We applied the straight forward proof viewing color transform and RGB clipping. We displayed the processed RGB values on the proof viewing monitor and measured the CIE xy chromaticies on the monitor and in the scene.

As observed in Figure 2, color clipping modifies hue and saturation. While desaturation may be accepted by a director watching a proof viewing monitor, hue changes are not acceptable. A proof viewing color transform should address and solve this problem.

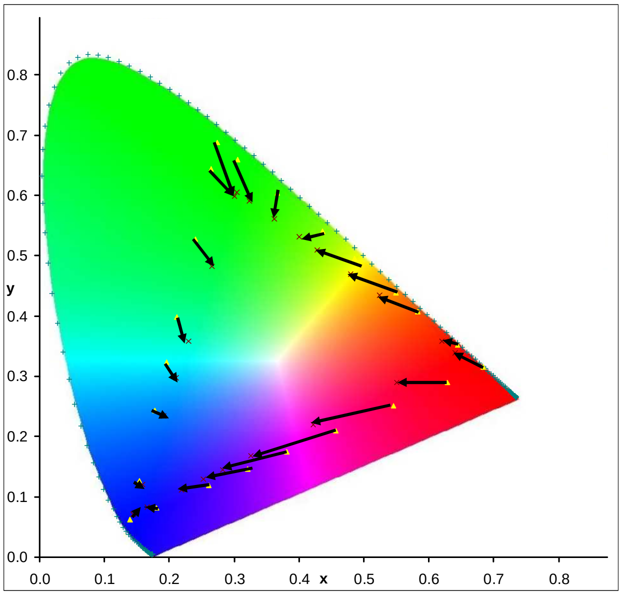


Figure 2: Color clipping (see arrows) of sample real scene colors when displayed on a Rec. 709 proof viewing monitor

Color resolution in digital acquisition

Another issue of digital acquisition when capturing scenes with extended color gamut is the color resolution:

- Difference of filter spectrum from spectral sensitivities of human eye;
- Restricted capacity to distinguish saturated colors;
- Impact on precision of captured hue.

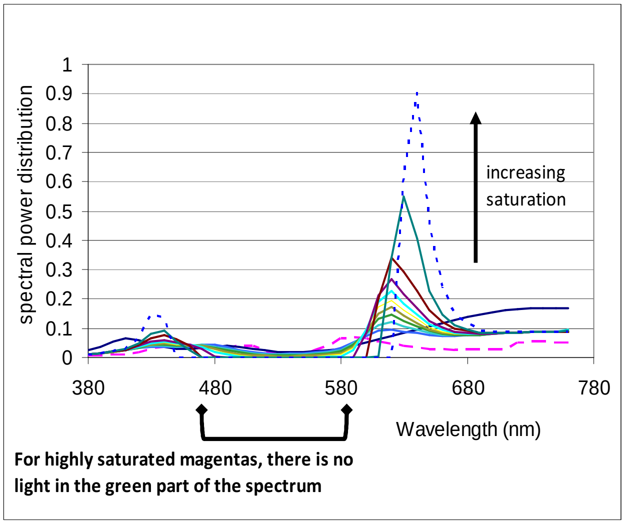
We want to show in the following that these issues result in additional errors on a proof viewing screen:

- Hue shift;
- De-saturation and color clipping.

We will use in the following an ideal proof viewing monitor without color gamut limitations. Color clipping errors such as discussed before are thus excluded.

Let’s take a series of test colors at constant magenta hue and with increasing saturation in perceptually uniform IPT color space [18].

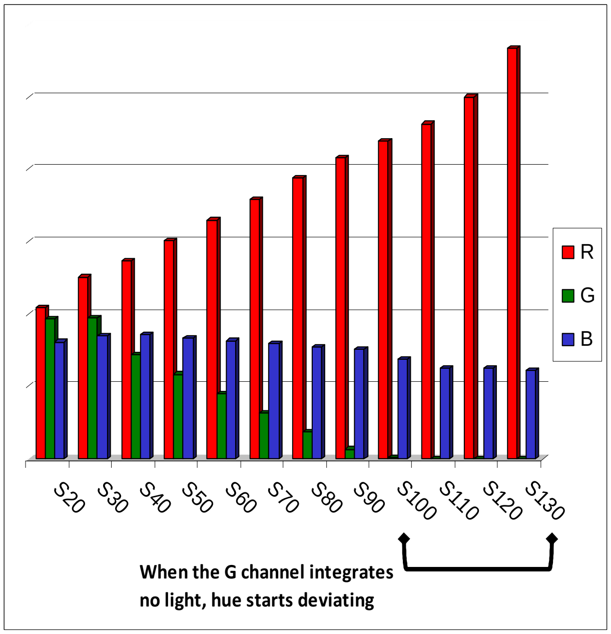
Figure 3 shows one of the possible sets of spectral power distributions that correspond to the chosen test colors. (Note that an infinite number of spectral power distributions may result in the hue and saturation of a given test color.) The spectral power distributions in Figure 3 are representative for spectra becoming sharper with increasing saturation. As observed in Figure 3, the luminous contribution of the spectrum for wavelengths between 480nm and 580nm decreases with increasing saturation. The four most saturated test colors have even zero contribution.



For highly saturated magentas, there is no light in the green part of the spectrum

Figure 3: A set of spectral power distributions corresponding to magenta test colors with increasing saturation from low (magenta dashed) to high (blue dotted)

In such a case, one channel of the camera (here the green G channel) will have no signal and then the camera no more exhibits trichromatic characteristics, but only two channels are active/excited. Figure 4 shows how R, G and B channels evolve with increasing saturation at constant hue according the stimuli from Figure 3. We see the system becoming di-chromatic for stimulus S100 and above, where only the R and B channels integrate light. For these stimuli, hue and saturation deviate as the acquisition system is no more coherent with the usual three channel system behaviour.



When the G channel integrates no light, hue starts deviating

Figure 4: RGB output with increasing R channel (red) decreasing B channel (blue) and decreasing and cropped G channel (green)

A solution to this problem involves the optimization of the spectral sensitivity curves and is beyond the scope of this paper. Such a solution should include an evaluation of color precision such as carried out by Pujol et al. [19] on the number of distinguishable colors inside the McAdam limits.

EXTENDED COLOR GAMUT IN COLOR CORRECTION

One of the artistic steps in production is color correction. Often a first phase is carried out to adjust roughly film footage or raw streams acquired by digital film stream cameras. Large mismatches in color balance and transfer function are compensated by linear matrices and non-linear one-dimensional transfer functions, respectively. Frequently, specific 3D Look-Up-Tables (LUT), also

called Cubes, are applied to produce a more pleasant version than the raw version. In a second phase, the director of photography and the colorist apply artistic color changes in order to obtain the desired look of the images.
In this artistic phase, the director of photography describes the intent of color correction while the colorist or a skilled operator has to translate the intent into an actual color transform applied to the footage. Such a color transform may include an increase of saturation, a change of color hue, a decrease of any RGB channel or an increase of contrast, for example. Color correction can be applied to an entire frame, to a set of frames, to a specific region in one single frame or even to all image regions in several frames corresponding to a specific color or semantic object (tracking).

Color reproduction during color correction

During this process, the director of photography and the color grading operator have to keep in mind what will be the impact of the applied color correction on the final reproduction medium. For example, if argentic film is first scanned and digitalized and then color corrected using a dedicated, digital proof-viewing projector, the operator verifies the applied color correction on the projection screen while the final reproduction is done by a film printer and then the film is projected.

Differences between the proof viewing display device (for example a digital proof-viewing projector) and the final reproduction device (for example a film printer followed by film projection) should be taken into account during color correction. Differences are due to different media, different equipment but also to different viewing conditions. Viewing conditions include ambient light, surround, background, reference white and adaptation state of the human eye. Differences between the proof viewing display device and the final color reproduction device can include objective, measurable differences of CIE 1976 hue angles, changes of CIE saturation, changes of contrast, differences in CIE 1976 luminance, differences in dynamic range, differences in color gamut as well as differences in color appearance such as changes in lightness, saturation and chroma. The latter three differences can not be photometrically measured.

A known solution to this problem is colorimetric color management (CMM) [14]. For CMM, the characteristics of the proof viewing device and the final reproduction device are measured, mathematically modelled and then compensated using a color transformation. CMM takes into account the color gamut of the devices. When an image contains colors outside of the color gamut of a display device or close to the border of the gamut, the applied color transform may contain color gamut compression, color clipping or other specific operations such that the transformed colors are inside of the device color gamut.

Issues of color correction

The difference of color gamuts of display devices is a problem for color correction. It may happen that the operator applies a color correction that generates the desired image on the proof-viewing device while the final reproduction device is not capable to reproduce some of the colors since the color gamut of the final reproduction device is different from the gamut of the proof-viewing device. It may happen that the operator wants to apply a specific color correction which would generate acceptable results on the final reproduction device but which cannot be visualized on a proof-viewing device with different color gamut.

A known solution is

- to detect out-of-gamut colors for the final reproduction device;

- to detect out-of-gamut colors on the proof view device;
- in the framework of CMM and
- to show a gamut alarm to the operator when an out-of-gamut color has been detected.

Figure 5 shows a typical example how gamut alarm is signaled to the operator. Each pixel that contains a detected out-of-gamut color is shown white.

Classical color correction systems offering gamut alarm functionality however do not address a series of problematic cases.

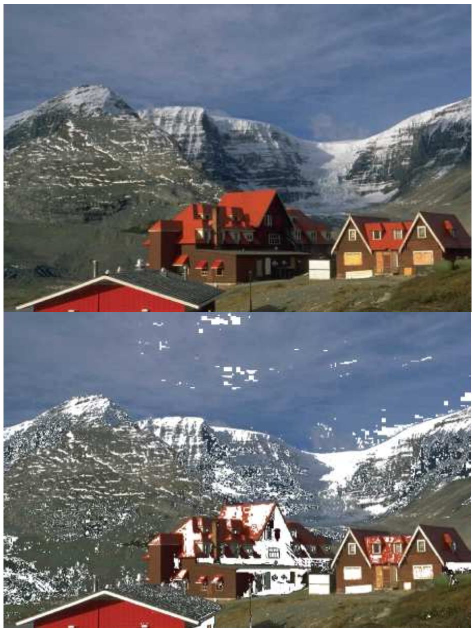


Figure 5: Original image on the screen of the colorist without gamut alarm (top) and with gamut alarm (bottom)

The first case is the difference in viewing conditions. The gamut alarm mechanisms are limited to colors that can not be rendered on a display in the framework of colorimetric color management. In this framework, colors are usually measured by CIE 1931 XYZ coordinates. These coordinates do not consider viewing conditions that influence the human observer while watching the display. In an appearance-based color management framework (appearance-based CMM), such influences are compensated. In such a case it may happen that a color that the operator desires on the proof viewing device can be reproduced on the final reproduction device in colorimetric terms but can not be reproduced when viewing conditions are compensated.

A second case is the consideration of an original reproduction device. When an operator works on footage that is aimed for a final reproduction device and proof viewed on a proof viewing device, it may be important to consider where the content comes from, i.e. for which device the content was originally prepared. This device is called here original reproduction device. It may happen that a color after color correction is well reproduced on the proof viewing and final reproduction devices but not on the original reproduction device. This case needs to be detected and indicated to the operator.

The third case is the uncertain nature of viewing conditions. In an appearance-based CMM framework, influences of viewing conditions are compensated. As soon as colors need to be modified since they are out of the gamut of reproducible colors taking into account viewing conditions, they should be indicated to the operator. This could be an advanced case of classical gamut alarm. Such colors could be marked on the proof viewing screen by specific false colors, for example red. Classical gamut alarm is binary: either on or off. This is well adapted for the case of out-of-gamut alarm considering well-defined color gamuts of display devices. A binary gamut alarm is not adapted to the gamut of reproducible colors considering viewing conditions since characteristics of viewing conditions are less well mastered and known than characteristics of display devices. A binary gamut alarm would be finally not useful for the daily work of the operator.

The fourth case is when the operator wants to modify out-of-gamut colors. There is a difficulty of interpretation of classical gamut alarm. If classical gamut alarm is shown on the proof viewing device, those regions of the image are marked with a false color that represents out-of-gamut colors. An example is shown in Figure 5. When the operator looks at the image with gamut alarm, he aims to identify the colors (their hue, their saturation, their luminance) that are out of gamut. Either he switches on and off the gamut alarm or he analyzes the image as it is.

There are situations where this is easy. In Figure 5, he will identify the blue tones in the sky that – once getting clearer – approach the gamut border and go slightly outside. The blue tones are easy to analyze since the blue sky region contains a variety of tones and transitions. By the position and shape of the out-of-gamut regions the operator can easily analyze the problem.

There are situations where the identification of out-of-gamut colors is difficult. In Figure 5, the red roofs and the brown walls are out out-gamut. Since transitions are lacking, the operator can not be aware which portion of red and brown tones is concerned. This problem is increased in animated and painted images where the color palette is often restricted. It is not visible whether the correction to be applied to these colors needs to be weak or strong. From the image in Figure 5, it is not clear to the operator what may happen to similar colors, those that may occur on the same objects but in following frames where light is slightly different.

This problem is solved today by trial and error as well as by switching on and off the gamut alarm. The operator applies corrections and verifies the gamut alarm. By “trying around” a couple of neighboured tones, he will understand the position of the concerned colors within the color gamut and apply an appropriate correction. This procedure takes time. Furthermore, the operator can not separate out colors being largely outside the gamut that need to be worked first. By watching the image in Figure 5, he can not establish a priority list for his work. This prevents from being quicker by neglecting colors which are only slightly out of gamut.

The fifth case is the growing variety of display technologies in the consumer world, when video productions are to be distributed to consumers with different display technologies, the color correction process using a single final reproduction device will fail to produce content that has controlled quality on displays with other characteristics than those of the targeted final reproduction device. In this case, there may be non-detected colors that are out of the color gamut of the actually used reproduction device.

METHOD OF SOFT GAMUT ALARM FOR COLOR CORRECTION

This section introduces the new concept of soft gamut alarm that assists the colorist in future tasks of color correction with extended color gamut.

Overview

The proposed method aims at proof viewing the visual content introducing the new concept of alarm.

The method has the four following advantages with respect to classical color correction:

- Differences between viewing conditions of different color reproduction devices are considered;
- The uncertain nature of knowledge about viewing conditions is taken into account and content can be created considering this uncertainty.
- The variety of final reproduction devices is considered and content can be created with regard to this variety;

- Reduction of degradations of content with respect to its original/raw version.

The proposed color correction method aims to correct original colors of original images targeting an original color reproduction device with respect to a set of final color reproduction devices. Each of these color reproduction devices is characterized by its color gamut of reproducible colors in device independent, absolute color space and its viewing conditions for color perception by human observers.

The method can be summarized by the following steps:

1. The original colors of the original images are displayed on a subset of the final color reproduction devices, these devices are called proof viewing color reproduction devices;
2. For each of the color reproduction devices, the distance of the original colors to the color gamut of the color reproduction device is determined;
3. For each of the color reproduction devices, the color appearance of the original colors and of the color gamut of the color reproduction device are determined, taking into account the viewing conditions of the color reproduction device;
4. For each of the reproduction devices, the visibility of the original colors is determined, each visibility being the distance of the color appearance of the original color to the color appearance of the color gamut;
5. On one of the proof viewing color reproduction devices, false colors are displayed instead of the original colors, where the false colors reflect the correspondent distance and visibility of the corresponding original color.

The original colors of the original images are color corrected by an operator. Original colors are replaced by modified original colors in a way that the corresponding distance is minimized and the corresponding visibility is maximized.

Figure 6 shows the color processing flow path according to the proposed system. From original colors, false colors are determined that depend on distances to color gamuts. Original and false colors are displayed.

The process can be assisted by automatic gamut mapping [20,21,22]. For all proof viewing color reproduction devices, gamut mapping is applied in such a way that the false colors can be switched off and a reproducible, mapped color is shown. Gamut mapping is preferably carried out in color coordinates representing the color appearance of the colors.

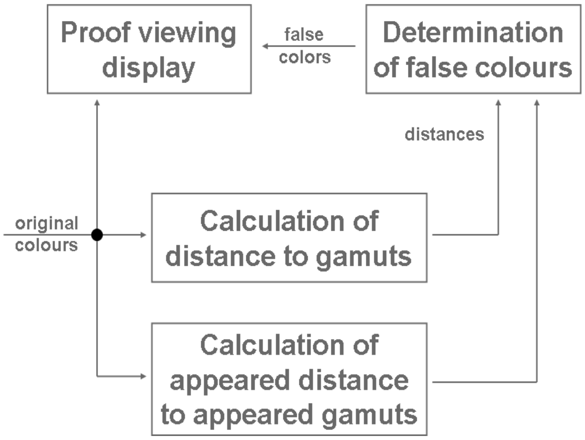

Proof viewing display
false colors
Determination of false colours
distances
original colours
Calculation of distance to gamuts
Calculation of appeared distance to appeared gamuts

Figure 6: Principle of the soft gamut alarm system

The distance to the color gamut is determined as follows. For each of the reproduction devices, the distance of the original colors to the color gamut of the reproduction device is determined using the Euclidean or a weighted Euclidean distance. The distance is forced to zero for original colors being inside the color gamut.

The visibility of an original color for a human observer is determined from the so-called appeared distance that is determined as follows. The original colors aimed for the original reproduction device are transformed into an original device independent color using the device profile of the original color reproduction device. The original device independent colors are transformed into original appeared colors according to the viewing conditions of the original reproduction device, where the appeared colors reflect the color appearance for a human observer. For each of the color reproduction devices, viewing conditions of the reproduction device, the color gamut is transformed into an appeared color gamut. The appeared distance is determined as distance of the original appeared color to the appeared color gamut. For original appeared colors being inside the appeared color gamut, the appeared distance is forced to zero. The visibility is a monotonic function of the appeared distance.

The concept of soft gamut alarm can include more than one false color to be calculated shown instead of one single. For example, two false colors can be calculated as follows. A first false color is calculated from the distance between the original color and the color gamut of a selected color reproduction device. A second false color is calculated from the appeared distance between the original appeared color and the appeared color gamut of the selected reproduction device.

In the following, the proposed method of soft gamut alarm is applied to the case of proof viewing for color correction during post-production of a digitalized film.

Reproduction devices

Three reproduction devices are considered:

- A proof viewing digital projector under dark conditions;
- A digital cinema projector under dark conditions;
- A broadcast reference monitor under dim lighting conditions.

All devices are fed with RGB color values. By device characterization, for each reproduction device, a forward and an inverse device model is established. The forward device model calculates device-independent XYZ color values from device-dependent RGB color values. The inverse device model realizes the inverse operation. The devices model provides also the color gamut of the device.

Consideration of color appearance

The appeared color values and appeared color gamuts are established in the perceptual color space JCh of CIECAM-02. In this color space, J is lightness, C is Chroma and h is hue angle perceptual estimate.

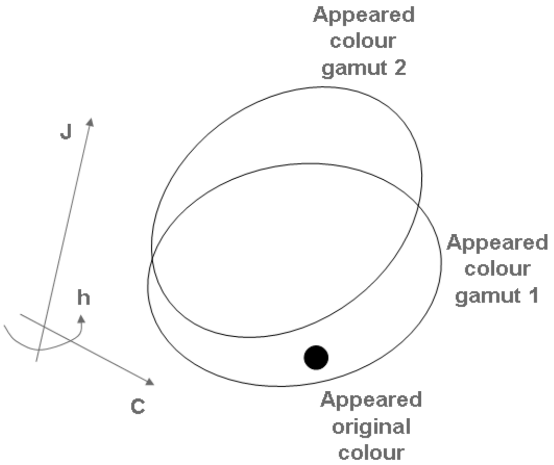


Figure 7: Example of an appeared color that cannot be reproduced on device no. 2

Figure 7 shows a sketch of an appeared original color and the appeared color gamut of two color reproduction devices no. 1 and no. 2 with different viewing conditions. On device no. 1, the appeared original color is close to the appeared gamut and has thus a bad visibility. On device no. 2, the appeared original color is outside of the gamut and is thus not reproducible.

The color appearance model (CAM) CIECAM02 is defined by the following viewing conditions parameters:

- The XwYwZw tristimulus values of the reference white; it can be set to the white point of the display obtained from the forward device model;
- La: this is the adapting luminance to which the observer is adapted; it is expressed as an absolute value in cd/m². It can be set to a value corresponding to 20% of the reference white luminance (mean video value).
- Yb: this is the background luminance which corresponds to the entire screen (or display) average white luminance. This value depends on the video content and may be specified as a percent of the reference white luminance. e.g. 20 for 20%.
- The surround type : there are four possible states:
- Average for day light vision (Yb>10cd/m²);
- Dim for dim viewing conditions (3-5 < Yb < 10 cd/m²);
- Dark for night viewing conditions (Yb<3-5 cd/m²);
- Intermediate this is a linear combination between each of the three other states.

For the use of CIECAM-02, all these parameters need to be known. For the three reproduction devices, the parameters are chosen as follows:

- Proof viewing digital projector
    - XwYwZw: display white measured in the center of the screen
    - Yb: 20% of Yw
    - Dark surround
- Digital cinema projector
    - XwYwZw: display white measured in the center of the screen
    - Yb: 20% of Yw
    - Dark surround
- Professional television monitor
    - XwYwZw: display white measured in the center of the screen
    - Yb: 20% of Yw
    - Dim surround

Generation of soft gamut alarm

The false colors showing the gamut alarm are calculated for the original colors of the images. For each image pixel, and for each of the two other color reproduction devices (the DC projector and the reference monitor), two false colors are calculated for the original color of the image pixel. For each pixel in total, four false colors are calculated. In the following is explained, how two of these false colors are calculated for one of the two reproduction device, selected by the operator.

A first false color is calculated from a function of the color components of the distance vector that is related to the distance between the original color and the color gamut of the selected color reproduction device. More precise, the distance describes the Euclidian distance between the original color and the closest point of the color gamut.

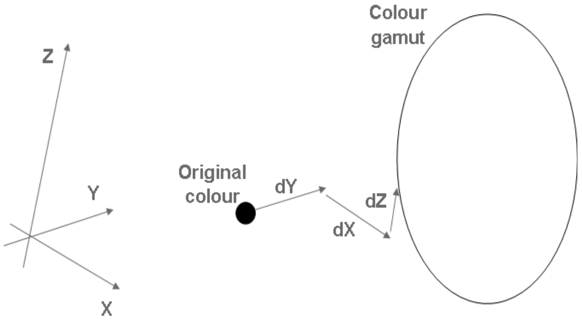


Figure 8: Calculation of a first false color in CIE XYZ space from the distance between the original color and the color gamut

For each color reproduction device, the distance of an original color to the color gamut of the color reproduction device is forced to zero for original colors being inside the color gamut. When the distance is zero, the related first false color is disabled and not calculated.

A second false color is calculated from a function of the color components of the distance vector that is related to the appeared distance between the appeared original color and the appeared color gamut of a color reproduction device. The components of the distance vector are calculated in the perceptual JCh color space of CIECAM-02 representing lightness, hue and saturation. By this choice, the second false color reflects the distance of the appeared original colors from the appeared color gamut of a reproduction device in aspects of lightness, hue and/or saturation, see Figure 9.

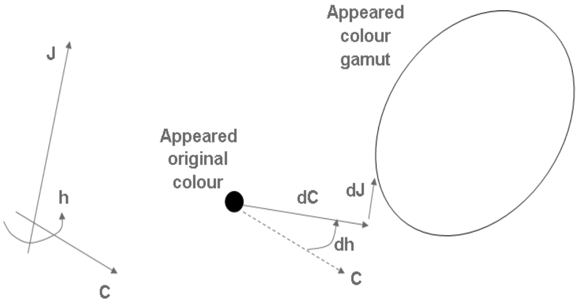


Figure 9: Calculation of second false color from the distance between the appeared original color and the appeared color gamut

The false colors are displayed according to the choice of the operator and will considerably help the management of wide color gamut.

CONCLUSIONS

This paper discusses issues in digital acquisition and color correction of images with extended color gamut such as camera sensitivity metamerism, proof viewing color clipping and gamut alarm in color correction. Production equipment builders should address the increasing demand of directors to capture and proof view scenes with extended color gamut. Optimized color filters and wide color gamut processing modes need to be developed for cameras. Post-production and color correction facilities should adapt color transforms and the related functions of gamut alarm to extended color gamut including evolving viewing conditions, new display technologies and color appearance.

This paper provides some inputs to ease the production of extended color gamut content. However the distribution of this content raises additional issues to be considered, such as the adaptation to the device characteristics or the viewing conditions. However, it is clear that future video formats will integrate extended color gamut so as to better approximate and serve the human visual system capabilities.

REEFERENCES

[1] R.R.A. Morton, M.A. Maurer, G. Fielding, C.L. DuMont, Using 35mm digital intermediate to provide 70mm quality in theaters, SMPTE 143rd Technical Conference and Exhibition, November 4-7, 2001.

[2] Technicolor 3D, www.technicolor.com

[3] EBU-Tech 3320, User requirements for Video Monitors in Television Production, Eurpean Broadcast Union (EBU), Version 2.0, October 2010.

[4] IEC, Multimedia systems and equipment - Color measurement and management - Part 12-1: Metadata for identification of color gamut (Gamut ID), 2011.

[5] A. Roberts, Coloring the future, tech-I, European Broadcast Union (EBU), March 2012.

[6] J.Stauder, C. Porée, P. Morvan, L. Blondé, A gamut boundary metadata format, 6th European Conference on Color in Graphics, Imaging, and Vision (CGIV), Amsterdam, May 2012.

[7] S. Y. Choi, H. Y. Lee, Y. T. Kim, J. Y. Hong, D. S. Park, C. Y. Kim, New Color Encoding Method and RGB Primaries for Ultrahigh-Definition Television (UHDTV), 18th Color Imaging Conference (CIC), San Antonio, USA, November 8-12, 2010.

[8] ITU-R BT.709-5, Parameter values for the HDTV* standards for production and international programme exchange.

[9] ITU-R BT.1361, Worldwide unified colorimetry and related characteristics of future television and imaging systems

[10] IEC, Multimedia systems and equipment – Color measurement and management - Part 2-4: Color management - Extended-gamut YCC color space for video applications – xvYCC, IEC 61966-2-4 Ed. 1.0, November 2006.

[11] Y. Xu, Y. Li, G. LI, Analysis and Comparison of extended color gamut in ITU-R BT.1361 and IEC 61966-2-4, Journal of Video Engineering, Vol. 33, No. 3, 2009.

[12] Photography and graphic technology - Extended color encodings for digital image storage, manipulation and interchange - Part 1: Architecture and requirements, ISO 22028-1.

[13] R.W.G. Hunt, The reproduction of color, Sixth Edition, Wiley, 2004.

[14] SMPTE, Derivation of Basic Television Color Equations, Recommended Practice RP177-1993.

[15] P. Urban, R. S. Berns, R.-R. Grigat, Color Correction by Considering the Distribution of Metamers within the Mismatch Gamut, Proc. 15th IS&T Color Imaging Conference, pages 222-227, 2007.

[16] Yuri Murakami, Keiko Iwase, Masahiro Yamaguchi, Nagaaki Ohyama, Evaluating Wide Gamut Color Capture of Multispectral Cameras, Proc. of 16th IS&T Color Imaging Conference, November 10-15, Portland, 2008.

[17] Jack Holm, Capture Color Analysis Gamuts, Proc. 14th Color Imaging Conference, pages 108-113, Scottsdale, Arizona, November 2006.

[18] N. Moroney, A radial sampling of the OSA uniform color scales, Proc. 11th IS&T Color Imaging Conference, pp. 175-180, 2003.

[19] J. Pujol, F. Martínez-Verdú, M. J. Luque, Cobija, P. Capilla, M. Vilaseca, Comparison between the number of discernible colors in a digital camera and the human eye, Proceedings of CGIV 2004, Second European Conference on Color in Graphics, Imaging, and Vision and Sixth International Symposium on Multispectral Color Science, April 5-8, 2004.

[20] J. Morovic and M. R. Luo, The Fundamentals of Gamut Mapping: A Survey, Journal of Imaging Science and Technology, 45/3:283-290, 2001.

[21] Montag E. D., Fairchild M. D, Psychophysical Evaluation of Gamut Mapping Techniques Using Simple Rendered Images and Artificial Gamut Boundaries, IEEE Trans. Image Processing, 6:977-989, 1997.

[22] P. Zolliker, M. Dätwyler, K. Simon, On the Continuity of Gamut Mapping Algorithms, Color Imaging X: Processing, Hardcopy, and Applications, edited by Eschbach, Reiner; Marcu, Gabriel G. Proceedings of the SPIE, Volume 5667, pp. 220-233, 2004.

Delinearizing Television – An Architectural Look at Bridging MSO Experiences with OTT Experiences

Bhavan Gandhi, Varma Chanderraju, & Jonathan Ruff
Motorola Mobility, Inc.

Abstract

Advent of over the top (OTT) content services by providers such as Netflix, Hulu and Vudu has dramatically altered the media consumption experience and with it the expectations of consumers. OTT services and in some cases cable provider services such as Xfinity TV supplant traditional linear and on-demand offerings. However, despite the availability of all these choices and services, the end-user's media consumption experience is disjoint and detracts from traditional lean-back TV watching. There is an opportunity to build solutions that provide a more cohesive, unified and intuitive user experience for the end-user. This paper describes architectural and system details of a system capable of delivering such an experience.

BACKGROUND

Consumer Experiences Trends

Traditional or linear TV watching, contrary to anecdotal views, is not dead. The ability to access over-the-top (OTT) content has not led to a mass exodus away from television[1]. Looking at absolute numbers, a recent Nielsen study found that people watch an average of 32 hours and 47 minutes per week of traditional TV compared to 27 minutes a week of watching video online[2]. Regardless of absolute numbers, the trend is that people are watching more video than ever before; this includes online, on portable devices, and traditional TV sets as well[3].

It is fair to say that the TV watching experience is evolving; it has been changing from a single-source, single-device experience to one of a multi-source, multi-device experience. In this new world, the incumbent service operator (i.e., MSO) is still the richest single source for linear, broadcast entertainment. Even in this ecosystem, the content delivery and experience infrastructure has been changing to accommodate and support varied devices and content formats.

The increasing number of content sources brings about the more radical experience changes. In some cases, the source is the content provider; other sources include the growing number of internet-based providers of entertainment content such as Netflix and Hulu, who we refer to as OTT operators. The addition of these content sources is fragmenting the user experience and taking it from a lean-back experience to forcing the consumer take an active role in discovering and consuming content from their various sources or subscriptions, while keeping mindful of the devices on which the content is playable. Experience fragmentation is caused by the user having to be cognizant of their subscriptions and the applications from which the content can be discovered and then played. These applications tend to be provider specific, so discovery of content within the provider’s offerings requires being in their application. In the confines of the living room, this experience starts diverging from being lean-back. Even outside of the living room, there is a meaningful need to have a central hub for content discovery and an easy way of consuming the content.

Towards Convergence

The objective of making entertainment content lean back in this changing marketplace is attained primarily by unifying the linear, on-demand, and OTT content discovery process. This should be without regard to the client device that is being used for consuming the content. And, once discovered, content playback should be easy too. Playback capabilities of the specific device should be transparent to the user.

There are a number of client-based applications on various devices that are attempting to achieve unification. Applications such as Fanhattan[4] are trying to unify the content discovery process across OTT content stores such as Netflix, Hulu, Amazon, iTunes, etc., as well as what is generally available linearly. One shortcoming is that linear TV is regional and subscriber dependent; therefore it does not reflect a subscriber's view into available content from their subscription. When it comes to fulfillment, OTT content can be consumed readily if the appropriate clients are supported on the device, however, fulfillment and consumption of broadcast content (through an MSO) is not supported even if the subscriber has a subscription.

Google TV[5] is a client device / application play in the living room to unify the content consumption experience. It is an application platform running Android OS that has the capability of interfacing to the Internet and to the set-top-box. Generally, OTT provider applications (e.g., Netflix) can run on the device to access the repository. There is also an attempt to converge the experience around television, movies, and shows through the Google TV application. This application strives to unify the content discovery from live content and the web; the application allows you to consume the content without regard to the content source (broadcast or internet). If the content is live, the Google TV box tunes the STB to the appropriate channel, and if it is available over the Internet it can be streamed to a client player that supports the appropriate security protocols. The disadvantage of this approach is that the convergence is done at the Google TV client; as such, this experience cannot be replicated across the increasing varying number of devices that are also being used to consume entertainment content.

Server-side (or cloud-based) unification of content discovery and federated content delivery and playback has the potential of bringing to bear the best of the Internet and marrying it with the best of broadcast television. Not only does it unify the content discovery, it also has the flexibility to support applications and experience across a variety of fixed and mobile devices. This is an evolution over unifying content discovery and centralizing content delivery that is espoused by Tranter[6]. We espouse centralizing discovery but federating the fulfillment.

ARCHITECTURAL CONVERGENCE

Architecturally modularizing and separating the content discovery (metadata), control, and data delivery provides flexibility in creating new services. This enables service providers to aggregate content information from multiple sources and to create tools and services that let end-users browse, search, discover, access and control content consumption across their ensemble of devices. This is key to creating a more unified user experience. The end-user is provided a unified content discovery mechanism through an application or guide interface. We have developed a cloud-hosted metadata service that aggregates, normalizes, and correlates content metadata from disparate sources and provides RESTful interfaces to applications. With the prescribed architecture, we allow for secure registration and sign-up to the user's set of subscribed services, whether they are linear broadcast or OTT, and secure access to the user's desired content. One of our primary goals is to achieve a unified discovery experience for the user while simultaneously ensuring that the content distributors or service operators can exercise and enforce their content rights over their media assets.

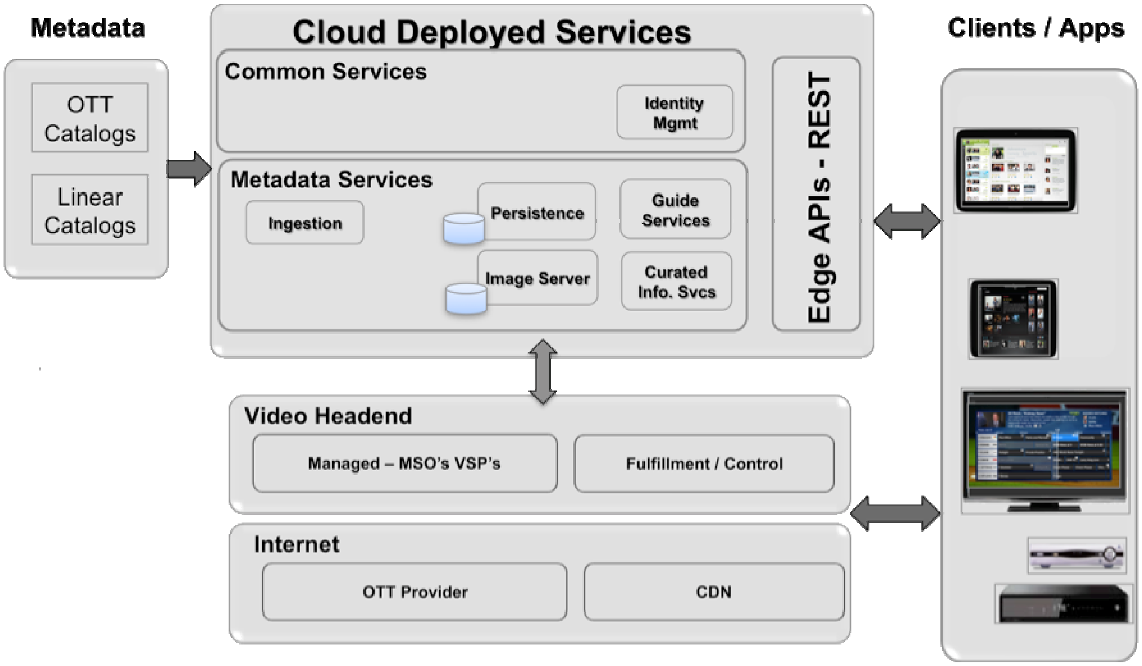


Figure 1. Modular Architecture supporting Converged Experiences

Figure 1 shows a high-level architectural system diagram that disaggregates content discovery from content fulfillment and control. The Cloud Deployed Services is responsible for hosting metadata across the varied sources and providing a unified view for discovery to client-facing applications.

For Internet based OTT content, the URL to play the content may be hosted as part of the Metadata Services or accessed using OTT provided APIs in the client application. OTT content is typically hosted and fulfilled through a content delivery network (CDN).

Taking a similar model with MSO hosted content channels and on-demand content, the client applications need a mechanism to access the content originating from the Video Headend. In our system, we use the Metadata Services as a proxy for passing minimal yet essential control information. Our approach is to publish the URL of the Fulfillment/Control services that the client application can access to discover the video channel (URL). This is modeled after the Internet OTT. By this approach, the data plane and control layers provide flexibility with how the content is delivered and what client devices are supported. The data plane layer is highly flexible and can exercise complete control over video playback on client devices. By minimizing the interdependencies between content discovery, delivery, and ultimately playback, we reduce the likelihood of linear and on-demand information sourced from a provider from getting stale.

In part, the function of an intelligent Fulfillment / Control module is to advertise its location and to enable appropriate content playback. Essentially it acts as a media broker between a user application(s) and supported clients. Once the client application / device accesses the location (URL) of this brokering service and stations that are supported, the client player can then tune to the selected content channels for the supported device type. The inherent assumption is that users are subscribers and the devices are registered with the service operator (and ultimately the Video Headend).

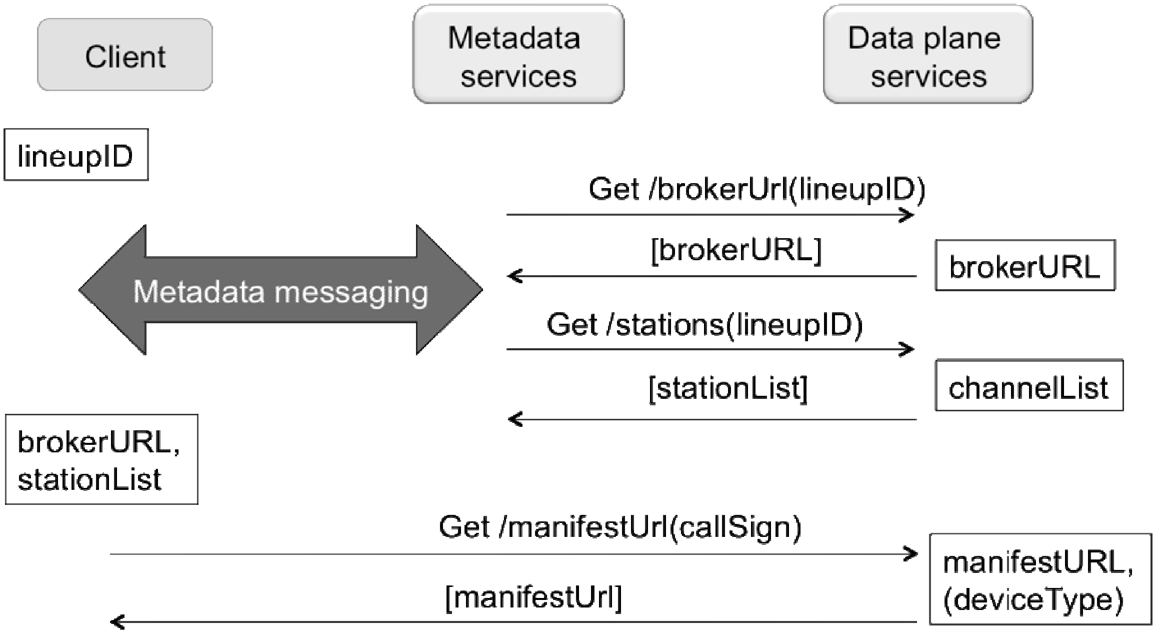


Figure 2. Interaction Diagram between Client, Metadata, and Data Services

Different models can be envisioned depending on whether all the services are hosted and operated by a single operator, or whether there are different operators for the different services.

An example set of interactions is shown in **Figure 2**. The client application first interacts with the Metadata Services to find the address of the appropriate Data plane that is servicing the delivery of the specific lineup (through lineupID). The Metadata services, through interaction with Data Plane services, can find out where the client application should point for getting access to the appropriate linear channels. Given this information (brokerURL, stationList), the application can interact directly with the data plane (Video Headend) to access the linear content (manifestURL) supported on the accessing device. This modularized system provides the advantages of unified content discovery as well as the ability to support an ensemble of devices. Also, since the service providers are ultimately responsible for fulfilling the content to the user / subscriber, they can control access. This allows flexibility in deploying and operating the system.

SOLUTION & DEPLOYMENT ECOSYSTEM

The entire system is built on a proven and robust technology stack that makes use of the latest web services technologies. The system has also been designed to support flexible deployment scenarios.

At a high level the de-linearizing (or converged) television ecosystem is comprised of several subsystems and logical modules that each encapsulate specific functionality. These include:

- Unified Metadata Services subsystem
- Portal & User Interface subsystem
- User Management subsystem
- Device Management subsystem
- Network DVR subsystem that includes scheduling, recording, and archiving
- Fulfillment subsystem
- Dataplane subsystem (Video Headend) that encapsulates transcoders, recorders & streamers

A key element of our solution is the ability to provide a set of unified metadata services to our clients. This allows the clients (internal and external) to rely on a single entity for all content metadata needs irrespective of where the source data is derived from and irrespective of the type of the source metadata (linear, non-linear, broadcast, VOD, OTT etc).

The unified metadata services subsystem is capable of assimilating highly unstructured, inconsistent and incomplete data sourced from dozens of individual metadata sources and turning the data into a consistent, complete and usable set of metadata services that can be consumed by a variety of clients. The unified metadata service exposes feeds and APIs for clients to access linear TV data including scheduling data, lineups, stations and non-linear data such as series, episodes, movies, news programs etc. The unified metadata service subsystem uses a host of techniques including customized data-collection agents (or ingestors) that are continuously tuned to be in sync with ever-changing data publishing formats used by the data-providers that we ingest data from. Examples of data sources include metadata providers such as TMS & Rovi, OTT providers such as Hulu & Netflix. Data-clustering and data-classification techniques are used to normalize and classify related data sourced from multiple metadata providers. The metadata capture and classification has evolved over time and has been enhanced using heuristic learning. In addition a powerful set of editorial tools provide the means and capability to further refine the data through manual oversight. Manual oversight is only required for a small fraction of the data. At the storage layer metadata services use a combination of SQL and NoSQL storage to ingest, classify, store and archive metadata.

Metadata services are deployed in the cloud and expose clean, simple and flexible REST APIs to consumers of the metadata services. JSON is the preferred format for the output of these REST APIs as most clients are browser-based rich-web applications.

Another key element of the solution is the usage of a powerful and flexible framework that is capable of driving the presentation elements on the client (device) side as well as providing an SDK (both on the client side and on the server side) that service operators (or 3rd parties) can use to extend, adapt and customize the look, feel and functionality of the solution. The framework helps abstract different form factors and input paradigms from user-experience developers and assists with ultimately creating uniform experiences across various device types (TVs, Tablets, Smartphones). The choice of client-side technologies (HTML5, JS, CSS) makes it possible to address devices running mobile operating systems such as Android and iOS to STBs running custom Linux images.

The server side components of the solution are deployed in Linux OS, virtualizable, developed on a Java EE platform, conform to the MVC architectural model, make heavy use of the Spring framework and use Hibernate to abstract the storage layer. The server side components are highly modular and communicate with each other mostly using REST APIs. This lends itself to flexible deployment options that can take into account different business and technical requirements that drive the deployment choices of service operators.

CONCLUSION

Being aware that video consumption is evolving and a continued user need for being easily entertained, we have architected and developed a system that allows content from multiple sources to be discovered and consumed on any device. Such a system requires the disaggregation and modularization of the discovery and fulfillment processes. We have used the latest in web technologies to create a flexible eco-system for deploying and operating the system. This allows content to be discovered and consumed from both traditional TV and Internet OTT sources in a unified way.

ACKNOWLEDGEMENTS

Our colleagues Anthony Braskich and Stephen Emeott contributed the specification of the interaction between the Metadata and Data Plane services, including **Figure 2**.

REFERENCES

1. Robertson, A. (Feb. 9, 2012). The Verge. In "*Nielsen: most Americans still pay for traditional TV, but a growing minority go broadband-only*." Retrieved March 22, 2012, from http://www.theverge.com/2012/2/9/2787037/tv-internet-streaming-video-viewing-survey-2012-nielsen

2. Schonfeld, E. (Jan. 8, 2012). Tech Crunch. In*"How People Watch TV Online And Off."* Retrieved March 22, 2012, from http://techcrunch.com/2012/01/08/how-people-watch-tv-online/

3. Indvik, L. (Oct. 20, 2011). Mashable Entertainment. In "*Americans Are Watching More Video Online – and Everywhere Else*." Retrieved March 22, 2012, from http://mashable.com/2011/10/20/nielsen-video-tv-study/

4. (n.d.). Fanhattan. Retrieved March 19, 2012, from http://www.fanhattan.com/

5. (n.d.). Google. Retrieved March 19, 2012, from http://www.google.com/tv

6. Tranter, Steve, 2011. *"Putting the Best of the Web into the Guide,"* SCTE Cable-Tec Expo Proceedings, Atlanta, GA.

ENGINEERING ECONOMICS – DOCSIS 3.0 CHANNEL BONDING FOR IMPROVED NETWORK ECONOMICS

Amit Garg, James Moon
Comcast Corporation

Abstract

DOCSIS 3.0 (D3) enables channel bonding, i.e. multiple downstream (DS) and multiple upstream (US) RF carriers that can be combined to provide a wideband service.

In this paper, we demonstrate that channel bonding not only provides multi-system operators (MSOs) with the opportunity to offer faster speeds to their customers, but also provides an opportunity to reduce capital required to meet the growing traffic demand. Combinatorial models were used to assess the opportunity for such load balancing gains. Later, empirical data was used to measure the load balancing gains achieved on a plant with D3 cable modem termination systems (CMTSs). Eventually, results from a trial with paying subscribers demonstrated the impact of providing D3 modems to select customers.

The paper demonstrates that instantaneous load balancing achieved through channel-bonding provides carriers with substantial improvement in engineering economics.

A NEED FOR SPEED

The Internet eco-system is flourishing; subscribers love the ease and convenience of broadband access, while content providers have embraced this new platform to provide an ever-increasing plethora of data intensive services – video email, video chat, video-conferencing, music, streaming video, cloud storage and cloud computing to name but a few.

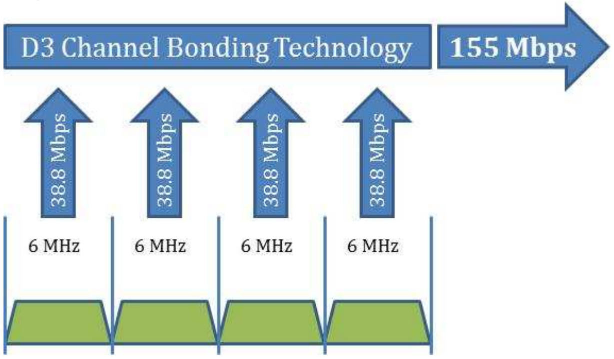
Over the last 15 years the Internet has grown from a novelty to a necessity. Be it communications, travel plans, information, education, news or entertainment, individuals are very likely to use the Internet. Over the same period, internet access has undergone a massive shift, from dial-up modems providing 14.4 kbps to always-on broadband access at DS speeds in excess of 100 Mbps as consumers have embraced faster and faster broadband speeds. MSOs have been leading the way in providing broadband access – by embracing DOCSIS as a way to provide broadband services to their customers. Until a few years ago, in the absence of cable's competitive broadband services, the only way to get a 1.5Mbps service was to pay upwards of $1,000 per month for a T1 from the local or competitive telephone company. Today, the most common broadband packages, with DS speeds in excess of 6Mbps, start at around $40 to $50 per month.

In the early days of DOCSIS, MSOs in North America provided broadband service by using a single 6 MHz channel for DS, and another 3.2 MHz carrier to provide US service. Improvements in modulation eventually enabled ~38 Mbps DS and ~ 10 Mbps US with each of these carriers. And, until recently, a single 38Mbps DS carrier was shared across a group of subscribers (service group) to provide customers with economical access to broadband speeds, while ensuring that customers received a desirable experience.

The development of the DOCSIS 3.0 standard changed that. D3 enabled multiple channels to be bonded into a single service group and was a direct result of subscribers' appetite for faster and faster speeds. To provide speeds in excess of 38 Mbps bonded RF channels become an absolute necessity as the service speeds exceed the offered line rate.

Today, MSOs in North America are typically bonding 4 to 8 DS channels and are beginning to bond 2 to 3 US channels. Channel bonding has enabled DS speed offers in excess of 100 Mbps, while demonstrating DS speeds of up to 1 Gbps. US speeds of 20 Mbps have been offered to customers; US speeds of up to 100 Mbps have been tested. This growing ecosystem has resulted both in an increase in number of subscribers using broadband, as well as increased demand per subscriber. In recent years, demand per subscriber has been growing at ~45 to 50% CAGR. For an MSO, this translates into the need to double the capacity of their high speed data (HSD) networks every 18-24 months.

Figure 1



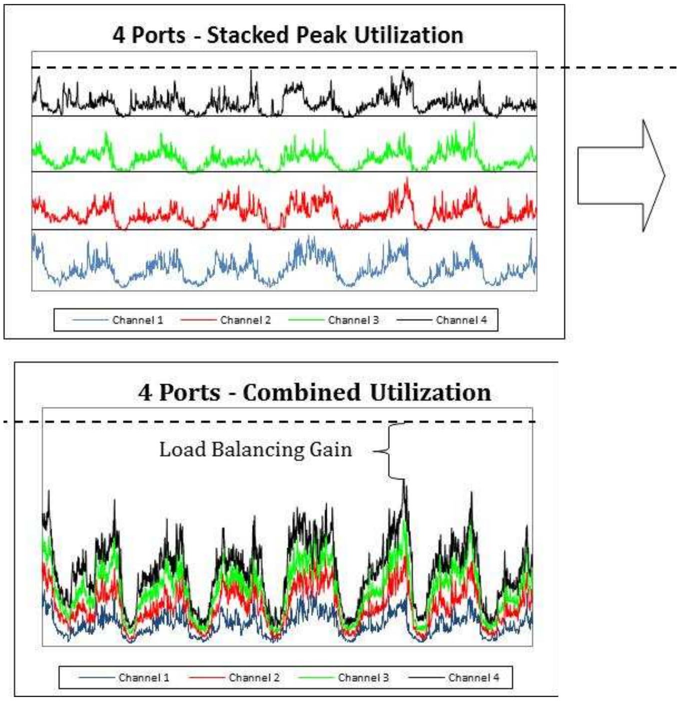
WHAT OTHER ADVANTAGES MIGHT ARISE OUT OF CHANGES IN ARCHITECTURE?

Once the D3 rollout began, there was an increased interest in understanding what other benefits might be derived from the bonded channels – something akin to increased operational efficiency of trunks as explained with Erlang math – fatter pipes are more efficient. Specifically, did channel bonding enable any statistical multiplexing or load balancing gains?

LOAD BALANCING GAINS

Load Balancing enables better use of network bandwidth by managing the network to the Peak of the bonded group and not by managing each port to its own peak. In our case, while observing the utilization of individual DS channels, it was noted that the peaks for multiple channels rarely occur at the same instance. For our purposes, the diagram below illustrates how we viewed the opportunity for statistical load balancing gains.

Figure 2



The top chart has peaks stacked, one upon the other; the bottom has traffic layered. The latter provides stat-mux gain over the former.

Early on, it was very clear to us that channel bonding could unlock some fairly significant network efficiencies as we increase the number of channels included in each SG. Our work helps determine the ranges of those gains and efforts that might be needed to capture those gains.

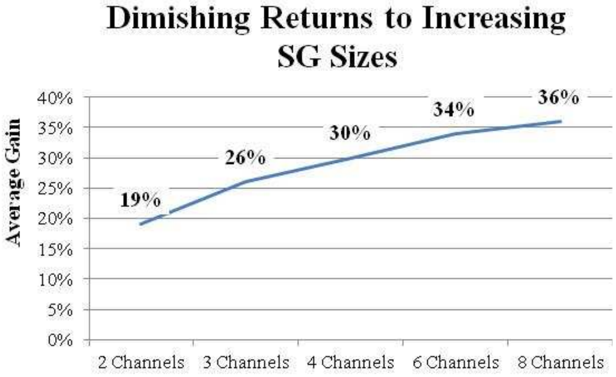
COMBINATORIAL ANALYSES TO APPROXIMATE STATISTICAL MULTIPLEXING GAINS

Prior to deployment of actual D3 networks, attempts were made to quantify the magnitude of hypothetical statistical multiplexing gains that could be possible. To that end, combinatorial models[1] were used to combine pools of existing DS channels into hypothetical service groups of 2, 3, 4, 6 and 8 channel combinations.

1. 5-minute channel utilization records were used.
2. All ports were combined into 2,3,4,8 channel bonded-groups.
3. Peak utilization for the period for each DS channel in the hypothetical SG used in the calculation was noted.
4. A SG peak for the combination was calculated by layering each of the 5-minute values for the channels that made up the hypothetical SG and finding the SG peak value. [Value A]
5. Gains were calculated by dividing the calculated SG peak by the sum of the individual peaks of the channels comprising the SG [Value B] and subtracting 1.
6. Distributions of these potential gains i.e. [(Value B-Value A)/Value B] are summarized in Figures 3 and 4.

Later, larger samples that included more channel/SG combinations from multiple markets were developed to evaluate the gains. Evidence indicates diminishing returns. 2 channels provide 19% gain, 3 channels provide 26% gain, or an incremental 7% points over 2 channel. 4 channels provide 30% gain, or an incremental 4% points over 3 channels, etc.:

Figure 3

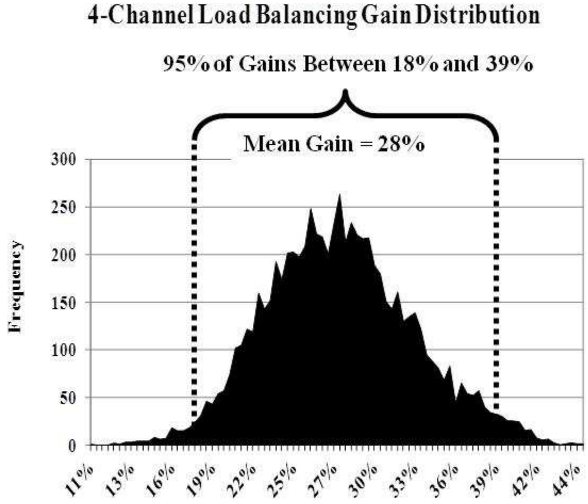


THERE IS A DISTRIBUTION WITH RESPECT TO LOAD BALANCING GAINS

While we were able to calculate the average gains from 2, 4, 6, and 8 port combinations, a quick glance at the distribution chart in Fig. 4 illustrates the fact that the gains are not uniform, but are normally distributed.

Our initial work focused on a single CMTS with 22 ports. We grouped the 22 ports into 7,315 4-Port SG combinations ($_{22}C_4$) and calculated the gain for each.

Figure 4



In this set, we found that the "worst" combination provided a gain of 11%, while the "best" was nearly 45%.

[1] Models implemented in Matlab. Neha Gadkari performed simulations in support of this analysis.

For our purposes, it appeared that network efficiency gains of approximately 25 to 30% could be realized for the (then) typical DS SG deployment of 3 or 4 channels.

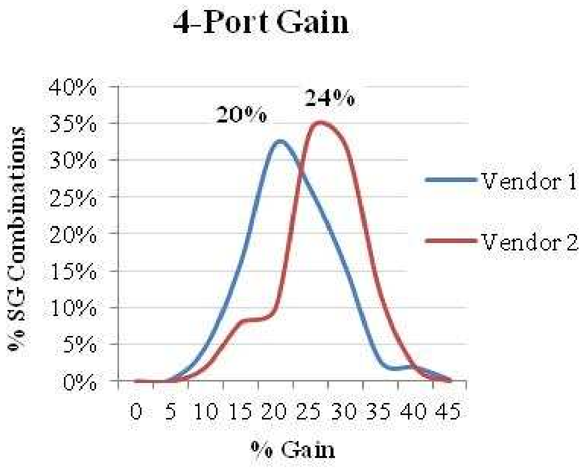
PRODUCTION D3 SERVICE GROUPS EVALUATED FOR REMAINING STAT MUX GAINS

As empirical data became available on production D3 service groups, additional combinatorial analyses were performed to determine if there was any load balancing opportunity remaining as the CMTS vendors had implemented load-sharing algorithms to balance traffic on SGs where majority of cable modems were still not D3. Our data set for this portion of the analysis consisted of over 300 4-channel service groups.

The data showed that most of the 25-30% load balancing gain was still on the table. For one vendor, the average opportunity was~ 20%, while it was closer to ~24% for the other vendor. This led us to conclude that significant gains could be achieved only through instantaneous statistical load balancing.

While vendors raced to develop various load-sharing algorithms to help balance demand across multiple RF channels, it was clear that without the deployment of significant numbers of D3-enabled devices that these significant statistical multiplexing gains would prove elusive as D3-devices enable instantaneous load balancing.

Figure 5



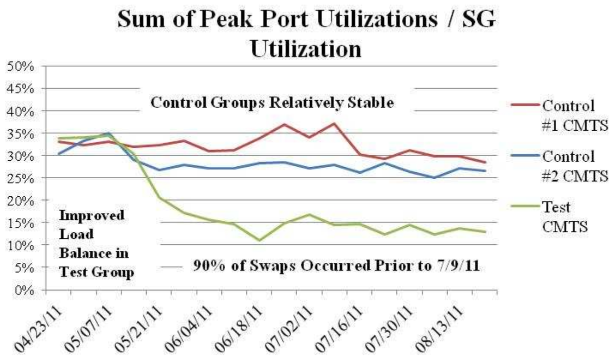
TESTING THE HYPOTHESIS

To further our understanding of what additional stat mux gains could be attained, we conducted an experiment where we provided D3-enabled gear to a large number of subscribers on a CMTS in one Comcast market. Two additional CMTSs in the same market were used as controls.

Over a period of approximately 2 months, select customers were provided new D3 CPE or modem and self-install kits. In addition, all new additions in the market (test as well as control CMTSs) were provided D3-CPE to prevent new users from inadvertently influencing results.

Measurement of the available gains on the SGs of the test CMTS as well as those on the controls indicated that there were about 30 to 35% gains available at the beginning of the study. As targeted modems on the Test CMTS were swapped by our customers and as new customers were supplied D3-enabled gear, we found that the deployment of the D3 gear was generating the desired effect – that load balancing gains were being generated (see Fig. 6). That is, the sum of the peaks of the individual channels that made up the SG and the actual SG peak were converging.

Figure 6

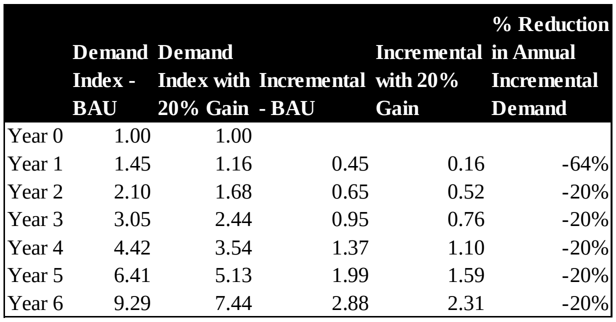


IMPLICATIONS FOR NETWORK EFFICIENCIES

Given a sufficient penetration of D3 gear that most, if not all of the hypothetical gains can be realized. A one-time gain of 20-30% in network capacity offers meaningful returns and can be exploited by MSOs to improve the bottom line as load balancing gains provide savings for years to come.

A simple model illustrates the annual network impacts of a 20%, 1-time gain (see Fig. 7.) In year 1, a 20% impact is recorded. Capital expenditures in that year plummet 64% vs. the business as usual (BAU) view.

Figure 7



	Demand Index - BAU	Demand Index with 20% Gain	Incremental - BAU	Incremental with 20% Gain	% Reduction in Annual Incremental Demand
Year 0	1.00	1.00			
Year 1	1.45	1.16	0.45	0.16	-64%
Year 2	2.10	1.68	0.65	0.52	-20%
Year 3	3.05	2.44	0.95	0.76	-20%
Year 4	4.42	3.54	1.37	1.10	-20%
Year 5	6.41	5.13	1.99	1.59	-20%
Year 6	9.29	7.44	2.88	2.31	-20%

However, the gain is the gift that keeps on giving; with annual expenditures continuing to track 20% below BAU figures.

CONCLUSIONS

While channel bonding has enabled MSOs to offer DS speeds in excess of 100 Mbps and US speeds in excess of 20 Mbps, it offers significant engineering economics. With 4 or 8 bonded channels, MSOs can expect 25-30% gain in network efficiencies through instantaneous load balancing. As more cable modems are upgraded to D3 over the next few years MSOs will benefit from these engineering efficiencies in their capital outlay for years to come.

EPoC APPLICATION & MAC PERFORMANCE

Edward Boyd
Broadcom Corporation
Kevin A. Noll
Time Warner Cable

Abstract

Ethernet Passive Optical Network (EPON) systems have been successfully deployed worldwide for high-speed access networks. EPON uses the 802.3 Ethernet MAC over optical fiber to provide high-speed IP connectivity to the home or business. In November 2011, the IEEE 802.3 formed a study group [3] to study the feasibility of creating a coax cable physical layer (PHY) for the EPON MAC. With the Ethernet-Protocol-over-Coax (EPoC) PHY, cable operators can deploy high speed IP connectivity using the EPON MAC over optical fiber or coaxial cable. Key criteria for selecting and evaluating a PHY layer will be the application in which it is used and the MAC performance over the system.

The MAC layer performance over a Coax PHY layer will be different than an optical fiber PHY layer. Emerging interactive services and higher speed data links will require shorter delays than today's services over low speed links. In this paper, the bandwidth, buffering requirements, and delay over an EPOC network will be predicted for different deployment scenarios and physical layer technologies for the EPOC PHY. The impact of increasing the round trip delay will be considered in a comparison between EPON and EPOC with expected services requirements.

Introduction

EPoC provides a solution for Cable TV operators to provide fiber performance over a coax network or Hybrid Fiber Coax (HFC) network. By re-using the EPON OLT, EPoC promises common head-end or hub site equipment for both fiber and coax customers. There are many architectural choices for EPoC to connect the OLT to the Coax Network Unit (CNU) in the customer's home.

The IEEE 802.3 working group will define a new physical layer to operate on the coax cable. During this process, decisions will be made to achieve reliable performance, high efficiency, and low delay. This paper will explore a set of service requirements for VoIP and Metro Ethernet Forum (MEF) services operating on a potential EPoC implementation. The coax physical layer will require additional functionality that will add delay and increase the round trip time. The efficiency, buffer requirements, frame delay, and frame delay variation (jitter) will be considered for a range of round trip times to understand the impact to the operator. In a point-to-multipoint network like EPON or EPoC, the shared upstream contains the highest frame delay and frame delay variation. The upstream MAC layer differences with DOCSIS and bandwidth requesting mechanisms will be considered. This paper focuses on upstream traffic performance since it is the most challenging.

EPOC Architecture

There are several possible architectures that EPoC could follow. All are rooted at an

EPON OLT and have Coax Network Units (CNU) at the leaves. The variations exist in the outside plant configuration and the implementation of the electrical interface.

Direct Coaxial Connection

One possibility removes optical fiber from the link and attaches the coaxial cable directly to the OLT system. This approach, pictured in Figure 1, mirrors what is implemented with DOCSIS CMTSes today. In DOCSIS, the electrical interface is a coaxial cable secured to the CMTS (or Edge QAM) chassis with an F-connector. It is easy to imagine that an EPOC implementation would have the same electrical interface and F-connector mechanical attachment.

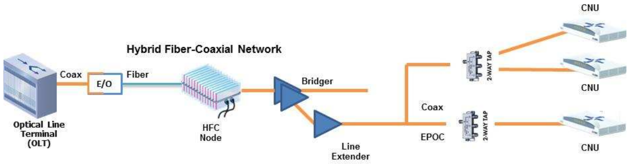


Figure 1: Direct Coaxial Connection

The practical application of this approach suffers from the fact that the bulk of coaxial plant is separated some distance from the hub site and connected via fiber optic cables. This means that the OLT would need to connect to a fiber optic link anyway. The development time and expense to develop a solution of this type is likely to be unproductive.

An alternate approach might carry the RF modulated EPoC signal over analog optics to an HFC node to be converted back to an electrical signal. This approach, however, does not provide the EPoC signal some easily realizable gains in the outside plant characteristics.

Baseband Signaling to Remote CMC

A more preferred architecture is one that uses baseband Ethernet or EPON signaling across the fiber plant. In this scenario, the hub site equipment might be (for example) an Ethernet switch containing WDM baseband optics connected to an OLT that is installed on the strand near an existing HFC node, or even in the HFC node. The OLT in this case could have a direct electrical connect to the coaxial cable and directly implement the EPOC PHY.

This architecture moves in a direction to reduce the use of expensive linear optics in the transmission path to support this type of application. However, the cost and operational complexity of installing an OLT in the outside plant is best avoided in most situations. In addition, for operators that already have EPON OLTs deployed in their hub sites, this approach is not a very effective use of capital.

A similar approach, and the one that is the focus of this analysis, uses the existing OLT and fiber plant to connect to an optical-electrical media converter that is installed in the coaxial plant. A typical configuration is shown in Figure 2.

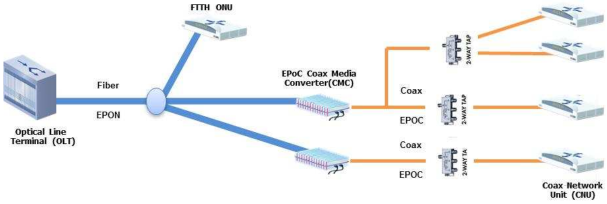


Figure 1: Baseband to Remote CMC

The EPON OLT provides the interface between the PON and external networks (the Internet, for example). It also is responsible for the well-known management functions in an EPON – admission control, station maintenance, scheduling upstream transmission, and other tasks. The role of the OLT in an EPOC network is no different than in an EPON and the CNUs appear to the OLT as if they are ONUs.

From the CMC to the CNU

As mentioned above, this chosen architecture, shown in Figure 2, requires an optical-electrical conversion. The implementation under study refers to this device as the Coaxial Media Converter (CMC). The CMC could be installed in or near an HFC node or somewhere closer to the subscriber. The CMC could be an Ethernet Switch or an Ethernet Repeater. The Ethernet Repeater could be a simpler and lower power device connecting the EPON optical PHY with the EPoC coax PHY. The Ethernet Switch would contain a bridge between an EPON MAC/PHY and an EPoC MAC/PHY.

Operators' Plant Characteristics

A coaxial cable plant, like any other transmission medium, has a set of characteristics that constrain the performance of the communication channel. The typical (but not exhaustive) list of physical-layer metrics includes signal-to-noise ratio or carrier-to-noise ratio, carrier-to-distortion ratios (Composite Triple Beat, Composite Second Order, etc.), carrier-to-interference ratio, group-delay, and micro-reflections. Each of these parameters varies based on operating frequency and bandwidth, so these two parameters must be specified as well.

Fully characterizing a coaxial cable-based network is a nearly intractable problem. Further complicating this is the variation in construction and operating practices from operator to operator and sometimes within a single operator's footprint. This study is focused on the MAC layer performance; therefore this study assumes that physical layer conditions are not a variable and circumstances allow the system to achieve the desired MAC signaling rates.

In addition to the physical-layer, the EPoC system will be expected to adapt to each operator's plant topological design and construction. The primary factors that characterize topology and affect capacity include the distance (which helps define loss characteristics and system timing constraints) from the CMC to the nearest and farthest subscriber, number of active subscribers, and offered subscription tiers (speeds).

Number of CNUs

The number of CNUs to be supported on the EPoC network needs to closely align with the number of active users on an HFC node today. This will help the operator avoid the cost of plant modifications required to deploy the EPoC system.

Today's HFC node-branch typically serves as few as 50 subscribers and as many as 500 subscribers (there are certainly cases where the node serves more or less than this). Based on this it is safe to require that the EPoC system support a similar range.

For the purpose of comparing EPOC performance to EPON performance, we should choose 32 CNUs. For the purpose of analyzing EPOC under conditions similar to today's average density this study will analyze network populations up to 512 CNUs and activity on up to 256 CNUs.

Distances

The propagation delay, that time required to transmit a frame across the coaxial cable plant, can have significant impact on the scheduler in the EPoC implementation. Therefore the distances spanned by fiber and coaxial cable in the network are an important parameter in the MAC performance.

Given the topology chosen for analysis – baseband EPON to a CMC located in or near an HFC node – we must consider two contributors to the distance from OLT to CNU. The first is the fiber from the OLT to the CMC. This distance can range from 0 meters (when the node is located in the hub site) to a typical maximum of 30km.

The second contributor to distance is the coaxial link from the CMC to the CNU. In an N+0 configuration the coaxial distance can range to around 150 meters. In an N+5 configuration with 1000-foot spacing the coaxial plant contributes about 1.7km to the total distance.

Subscription Tiers

Another factor in the system's ability to deliver traffic in a timely fashion is the speed tiers offered to subscribers. The typical Internet access service is a best effort service and ranges widely in offered data rates. A sampling of current offerings across the industry shows offered tiers 3x1Mbps (downstream x upstream bandwidth), 50x5Mbps, 60x6Mbps and as high as 100x10Mbps.

Operator Service Requirements

Operators offer many different services over their networks. Services include Internet access, Voice (VoIP), Video, Cellular Backhaul, Enterprise-class Ethernet circuits and more. Each service has its own set of network service level objectives.

Conveniently, there are two sets of specifications that can be referenced to cover the majority of these services and use cases. These specifications are the PacketCable specifications published by CableLabs and the MEF23 Implementation Agreement published by the Metro Ethernet Forum.

Packet Cable VOIP

MSOs provide packet cable VoIP service to residential and business subscribers. These are often a single line per home but multiple lines are possible, especially for business services customers.

Performance requirements for an access network supporting voice services are widely understood. Requirements specifications include packet loss, latency, and jitter. The major source of jitter in the EPON/EPOC network is scheduling the upstream transmission.

There are several sources of delay in the EPON/EPOC network. These include DSP processing and encryption, packetization, upstream transmission, and forwarding at the OLT.

Impairment	Value
Packetization Delay	20ms
Forwarding and Transmission Delay	< 10ms
Jitter	< 10ms

Table 1: VoIP Requirements

Table 1 summarizes these impairments and gives some typical tolerances in use by various service providers. In this analysis, we will assume that packet loss is trivial.

MEF 23H

The Metro Ethernet Forum defines a set of performance metrics that specify High, Medium and Low parameters that set the expectations for Ethernet services that traverse Metro, Regional, Continental, and Global distances (Performance Tiers). The general description of each performance tier (PT) is given in Table 2. In the context of this study, only the Metro PT is interesting and the EPON/EPoC network segment will generally only be a small portion of any one Ethernet service. The expected contribution of the EPON/EPOC link to the performance budget is expected to be small.

Performance Tier	Distance
PT1 (Metro)	< 250 km
PT2 (Regional)	< 1200 km
PT3 (Continental)	< 7000 km
PT4 (Global)	< 27500 km

Table 2: MEF Performance Tiers

Each PT definition includes a maximum frame delay (FD), mean frame delay (MFD) and a maximum inter-frame delay variation (IFDV).

The MEF 23 high quality service definition (H) is intended to carry delay sensitive traffic such as VoIP and financial trading transactions. These performance metrics for point-to-point delivery are summarized in **Error! Reference source not found.**.

Metric	Value
FD	≤10ms
MFD	≤7ms
IFDV	≤3ms

Table 3: MEF 23H Parameters [2]

MEF 23M

The MEF 23 medium quality service definition (M) is intended to carry traffic like Fax and network control traffic which are delay-sensitive but non-interactive. These performance metrics for point-to-point delivery are summarized in **Error! Reference source not found.**.

Metric	Value
FD	≤20ms
MFD	≤13ms
IFDV	≤8ms

Table 4: MEF 23M Parameters [2]

MEF 23L

The MEF 23 low quality service definition (L) is intended to carry Internet data service for business or residential where delay and jitter are not of any significant concern. These performance metrics for point-to-point delivery are summarized in Table 3.

Metric	Value
FD	≤37ms
MFD	≤28ms
IFDV	Unspecified

Table 3: MEF 23L Parameters [2]

EPoC System for Analysis

EPoC Sources of Delay

EPON Delays

In 1Gbps EPON, a round trip time of 250µs includes the propagation delay and physical layer delay for 20Km of fiber. The fiber propagation delay is about 100µs in each direction and 50µs covers the physical layer and synchronization delays in the OLT and ONU. For the analysis in this paper, the EPON round trip time of 250µs will be used as a baseline for comparison. EPoC bandwidth overhead (same FEC, 64/66) will be used on all RTT values so the difference is limited to the round trip delay.

EPoC Architecture

The MSO network has cable distances longer than the traditional TELCO network. While 20km may cover the entire network in EPON, EPoC will likely need to cover 30 km spans. The extended distance could add another 100µs of propagation delay.

EPoC PHY Functions

The EPoC PHY will require additional functionality to provide reliable performance when faced with burst or narrow band interference. A forward error correction (FEC) and interleaver will be selected to handle 25µs or more of burst error. The interleaver and FEC could add 400µs to 800µs delay to the round trip time.

Long symbol times of 20µs or 100µs will help combat multipath reflections. To gain better granularity, a block of symbols will be transmitted in selected carriers. Depending on the symbol and block size, an additional 400µs could easily be added.

Sharing Upstream & Downstream Frequency

Some operators like the option of using the same frequencies in the upstream and downstream in a Time Division Duplex (TDD) mode. While EPON is a full duplex protocol, half duplex operation to support TDD might be achieved by alternating between upstream and downstream transmissions in a fixed time block. To get reasonable efficiency on the upstream and downstream, a large block of transmission from each direction is needed. The larger block would be more efficient but it would add a significant amount of delay to the upstream and downstream. For example, an EPoC system that gave 1 millisecond of slot time to the upstream and 1 millisecond of slot time to the downstream would add 2 milliseconds of delay to the round trip time. The split between upstream and downstream maybe 50/50 or it might give a larger percentage to the downstream. In either case, the round trip time delay is the sum of the upstream block size and downstream block size. Small upstream block sizes would provide an additional restriction on the per-CNU upstream burst size. This paper will only consider the effect of the round trip time. An EPoC system using TDD would likely add 2 to 4 milliseconds of round trip time.

Switched or Repeated

The EPoC CMC provides a link between the optical fiber to an EPON OLT and the coax cable link to a CNU. The EPoC CMC could be defined as a switch or as a repeater.

An EPoC CMC Switch would contain an EPON ONU MAC layer connected to an EPoC OLT MAC layer through an 802.1D Ethernet Bridge. In this case, the access plant has two networks. The CMC will schedule and aggregate data from the CNUs and the OLT will schedule and aggregate data from the CMCs. The two layers of scheduling and aggregation allow for a more efficient use of the fiber. To determine the service delays, the fiber network frame delay and the coax network frame delay would be added together.

In an EPoC CMC Repeater, the EPON PHY and the EPoC PHY will be connected together in a fixed delay repeater. A single layer of scheduling and aggregation from the OLT handles upstream traffic. This system allows for a much simpler device but doesn't provide the second level of aggregation so it will not get full utilization of the fiber network when multiple CMCs share an OLT port. In networks with large Coax plants, the fiber to the OLT would likely be point-to-point so there is no needed for aggregation on the fiber. When there are very few CNUs connected to each CMC coax segment, data from the CNUs could be aggregated to the fiber as if the CNUs are on the same coax plant. For example, four CMCs with 10 CNUs each could share an OLT port as a single 40 CNU network. The EPoC CMC Repeater does not require QoS buffers, classification, SLAs, or scheduling in the CMC.

For round trip delay analysis, only the CMC Repeater is considered in this paper. The CMC Switch performance can be determined by assuming 300µs less round trip delay on the CMC Repeater RTT time and adding a second system with the EPON delay of 250µs. For example, the FD results for a CMC switch could be determined from the CMC repeater results by the following equation.

FD-Switch(RTT) = FD-Repeater(RTT-250us) + FD-repeater(250us)

The IFDV would follow the same equation since the delay frame variation from the coax scheduling would be added to the fiber network. The total delay budget for the access plant must be shared between the coax aggregation and fiber aggregation to guarantee compliance. In all cases, the CMC Switch will add delay to the access plant because of the two stages.

Delay Summary

The EPoC system could have delays from 1ms to 6ms based on decisions made in the standard and architecture deployed by the operator. In the performance analysis, a selected set of round trip times will be used to analyze the performance impacts. In most cases, the delay would be different for upstream and downstream. To simplify the analysis, the round trip time will be divided evenly between upstream and downstream.

EPoC MAC Layer Performance

EPoC MAC Layer Differences

Packet Fragmentation

Like other Ethernet MAC solutions, EPoC does not support layer 2 fragmentation of packets in multiple flows [1]. Fragmentation in ATM and other networking technologies allow for improved Quality of Service on low speed links along with a large unit of granularity. EPoC will need to support variable packet sizes and burst sizes with a finer granularity. On higher speed links like EPoC, the value of fragmentation and reassembly is questionable for the additional complexity. Since QoS is measured by frame delay variation and maximum frame delay, QoS on cells (fragments of packets) is misleading for packet analysis. The scheduling of cells can increase the worst-case delay and frame delay variation since a packet could span multiple upstream bursts. Even though fragmentation is not supported in EPON and EPOC, this paper will consider the impact of fragmentation on the performance when appropriate.

Stateless REPORT Frame

EPON and EPOC use a REPORT frame to pass queue information from the subscriber side CNU to the operator side OLT. The REPORT frame is not a request for bandwidth. It identifies the depth of the queues at the time of generation [1]. REPORT frame values will only change when data moves in and out of the queue. It is the responsibility of the OLT to track what has been granted in the past. This method is commonly referred to as stateless bandwidth reporting since the CNU doesn't hold state on the status of a bandwidth request. The CNU reports the queue size at the present time without regard to previous report frames.

DOCSIS systems use a stateful bandwidth request. The CM will generate a request for an upstream slot and it will not request for the same packets unless there is a timeout. The CMTS must grant the request or acknowledge it so the CM can update state on the request. A second request will not include the request in progress from an earlier request. The CM and CMTS must track the state of the request for the stateful system.

Stateful bandwidth requests were required for DOCSIS to support multicast bandwidth request slots. The multicast slots would only be used by a cable modem that hadn't already requested a bandwidth request. Stateful bandwidth requests are required for this function. EPON does not support multicast slots since the user count is lower and upstream bandwidth is higher. Performing a worst-case delay analysis is greatly simplified without multicast bandwidth request slots.

The stateless queue reporting of EPoC provides a simplification for a higher bandwidth upstream. It allows the CNU to avoid timers and long timeouts from a lost upstream request frame, downstream bandwidth acknowledge frame, or grant frame. In a stateless system, the polling interval determines the delay penalty for a lost upstream REPORT or gate frame. A timeout is considered in the delay penalty.

The REPORT frame provides a solution for reporting to frame boundaries. Since Ethernet doesn't support fragmentation, grants that aren't at frame boundaries will significantly decrease the efficiency. The REPORT frame contains one or multiple queue sets to define a queue's frame boundary at different thresholds. The queue set allows for the OLT to know a frame boundary at maximum size. A REPORT value for every frame in the queue would make a very large REPORT frame. The number of queue sets and maximum number of bytes can be configured with the SLA. For the analysis in this paper, a 4 queue set REPORT frame will be used. All 4 queue sets will have an equal limit. For example, queue set 1 will REPORT up to 4K bytes and queue set 2 will REPORT up to an additional 4K bytes. With a 4 queue set REPORT frame, the OLT can give 4 grants from a single REPORT frame before receiving the next REPORT frame. A smaller queue set will allow for smaller bursts and shorter delays for the upstream. Larger upstream queue sets will result in more efficient upstream bursts but longer delays.

Contention Slots

The EPoN MAC and EPoC system won't support contention or multicast slots. The lone exception to this rule is the discovery slot where multiple CNUs may respond. After discovery, all grants to an ONU or CNU will be unicast. Only one CNU or ONU will transmit in the slot. While the contention slots are useful in a large user network with many CMs, contention slots will prevent a smaller user network to reach high upstream data performance.

Since contention slots are not used in the EPoC based system, the worst-case delay is easier to determine and guarantee. It is also easier to show stable performance at close to or reaching 100% capacity.

The loss of contention bandwidth request slots also impacts the requirements for SLAs on the subscriber side. In DOCSIS, a cable modem will have an SLA to prevent it from over requesting bandwidth from the CMTS. The stateless REPORT frame of EPoC will only be sent by a CNU when requested by the OLT. The OLT has complete control over the

CNU for bandwidth granting and reporting so there is no need for an SLA on the CNU.

Piggybacking

REPORT frames can be sent in a single frame burst or as a frame in a longer burst with many frames. Since the REPORT frame contains the status of the upstream at the time of generation, it is normally sent as the last frame of the burst to exclude the frames in the burst. The OLT uses the force report indicator in the GATE frame to request a report frame in the burst. While a CNU could decide to send a REPORT frame in any burst, the normal practice is to send a REPORT frame only when requested by the OLT. The GATE frame with the force REPORT bit set is commonly referred to as piggybacking while the burst with only a REPORT frame in it is commonly referred to as a polling grant.

GATEs and MAPs

A MAP in DOCSIS provides a time slot description of the upstream with information for all stations. In EPoC, the GATE frame provides a unicast message to the CNU with a start time and length. In some cases, the MAP frame contains many grants over a significant portion of time. In the case of EPoC, the GATE frame will only contain a single grant to a single CNU. The GATE frame allows for up to four grants to the same CNU. In practice and in this analysis, a GATE will only contain a single grant. A MAP block delay or generation time does not exist for this reason.

Multiple LLIDs and Service Flows

A Cable Operator who provides multiple billed services to a single subscriber uses service flows to allow for different service level agreements. In EPON, the logical link identifier (LLID) provides a virtual point-to-point MAC connection between the OLT and CNU. A CNU with multiple LLIDs acts with multiple EPON MACs. With a MAC for each service, the OLT can monitor, enable, or grant the service independently of the other services on the CNU. By using multiple LLIDs, a cable operator can have multiple service flow like DOCSIS.

Activity based Polling

The large number of LLIDs or service flows on an OLT port will require a significant amount of bandwidth to query for status. Since service flows are often inactive for large residential systems, activity based polling can save bandwidth. Any service flow can have an active and inactive polling rate. The active polling rate would be much higher than the inactive rate. A simple example is a VoIP call where the active rate is used when a call is active and the inactive rate is used when no call is active. Activity can be determined by looking at the presence, rate, or type of frames on a link. The OLT system can determine the rules for activity and inactivity.

<u>EPoC PHY Parameters for Analysis</u>

The IEEE 802.3 will define overheads for the physical layer. Commonly suggested options for FEC and encoding burst overhead will be selected to get an estimate of the overhead. The Ethernet Frames will use the 64/66 encoding of 10G EPON and an 85% efficient LDPC FEC code. With these constant overheads, a fixed 20% overhead would be needed. A 1Gbps Ethernet MAC rate would require 1.2Gbps of Ethernet Line rate.

For bursts, a shortened FEC code word is allowed for end of bursts. A common burst overhead for EPON is 32 time quanta (time quanta are 16ns long) for sync time, 64 time quanta (TQ) for laser ON, and 64 TQ for laser OFF. At 1Gbps, the total burst overhead will be 1536 bits or 192 Bytes. EPoC will use the same burst overhead as EPON so the analysis can focus on performance differences due to round trip time. A larger EPoC burst overhead would reduce the performance and it should be considered in future analysis.

Packet Cable VoIP

Packet Cable VoIP service can be mapped to EPoC in a variety of ways. The most obvious is an unsolicited grant similar to DOCSIS. Another solution is a solicited granting based on polling.

For the analysis below, the G.711 codec will be assumed. Based on this code, a 218-byte packet will be generated every 20ms for each subscriber with an active voice call. A maximum FD and IFDV of 10 milliseconds will be required.

Unsolicited Grant Synchronization (UGS) Performance

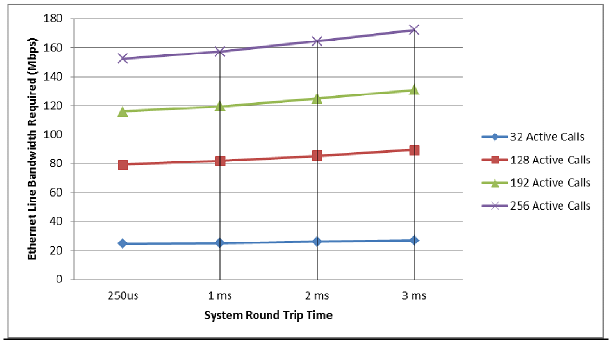
In the UGS solution, the EPoC system will establish two LLIDs. One LLID will carry signaling while the other LLID will carry the encoded voice. The encoded voice LLID will use unsolicited granting. Unsolicited granting is based on a timer at the OLT. A fixed size grant is given in a fixed time period. A REPORT frame with a non-zero queue set is not required for the grant generation. The signaling LLID will use solicited granting. Using activity based polling, the LLID will be polled at 17ms when the LLID is active and 100ms when inactive. The unsolicited granting could be enabled or disabled in the OLT based on the state of the voice call from observing the signaling channel.

The UGS slot will be sized large enough to carry a single 218 Byte Ethernet frame. The granting period of the UGS must guarantee a maximum delay of less than 10ms. The UGS slot is not aligned with the arrival time of the packet so the worst case scenario is an upstream frame just after the slot passed. The worst case delay of packet upstream will be the upstream transport delay plus the period of the UGS slot. The downstream delay does not factor into the UGS performance since the GATE is autonomously generated by the OLT. It is assumed that the worst case slot jitter from discovery slots is less than 500µs.

The period of UGS slots to a CNU must decrease with increased upstream delay. The period can be determined by subtracting the fixed delays from the worst case delay of 10ms. The equation below can be used to find the UGS period. As the UGS period decreases, the amount of upstream bandwidth consumed increases.

UGS-Period = MaxDelay – RTT/2 – SlotJitter

In the example scenario, each CNU will have a single VOIP session. The system is assumed to have 512 CNUs. The amount of Ethernet Line bandwidth required is shown for a different numbers of active voice calls and for different round trip times.



Graph 1: Required UGS Bandwidth

Graph 1 shows the bandwidth required for different round trip times and numbers of active voice calls. The System Round Trip Time of 250µs represents the performance of the all fiber 20Km EPON solution. The 1ms, 2ms, and 3ms show the performance of an

EPoC system with the corresponding total round trip time.

For UGS, the increase in bandwidth required due to longer round trip times is not significant for a small number of active calls. The increased RTT is more significant with 256 active callers.

The UGS efficiency is hurt by the single packet bursts. Additionally, the 20ms arrival time and sub-10ms delay will cause over half of the upstream slots to be empty.

Fragmentation or No Fragmentation

The UGS analysis assumes that EPoC does not allow fragmentation. Would fragmentation improve the capacity or decrease the delay? If packets were fragmented, they would need to wait for an additional UGS slot to be transported upstream. If the packet boundary and slots were miss-aligned, it would take up to 2 UGS slots for a frame to go upstream. In this case, the UGS slot would need to occur twice as often. The payload in the UGS slot could be divided in half. Since the overhead would double for the shorter interval, fragmentation would significantly increase the bandwidth required to transport the UGS flows.

Solicited Granting Performance

The UGS solution provides an adequate solution for transporting VoIP over EPON and EPoC. The UGS has the complexity of detecting the start and end of phone calls. UGS also requires a known packet interval and packet size. Additional phone lines at a CNU require more UGS flows or the complexity of detecting multiple phone calls in a single service flow. UGS is also not easily compatible with compressed voice or video conferencing. Solicited granting would greatly simplify the control and allow for other service options. Solicited is preferred if the performance is similar to UGS.

Solicited granting requires a REPORT frame to transmit upstream, a GATE frame downstream, and a data burst to be received upstream. The transport delay is therefore the downstream delay plus two times the upstream delay. Since data comes in asynchronous to the scheduler, the worst case delay should include the delay from simultaneous upstream slot requests from all active VoIP flows. For this analysis, we assume that VoIP is the highest priority. The polling period is the key parameter for the solicited solution. The following equation can be used to calculate the worst case delay.

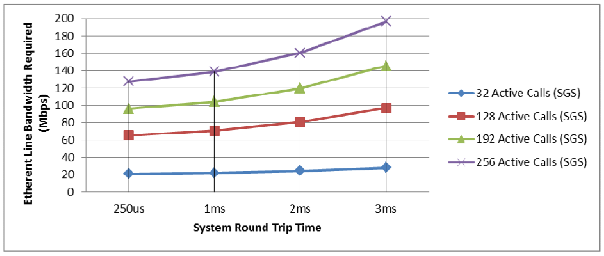
Tmax_delay = Tpolling + 2xTup + Tdown + Tall_service

The following equation solves for the polling interval.

Tpolling = Tmax_delay – 2xTup – Tdown – Tall_service

In the case of VoIP, piggybacking will not be used. While piggybacking would decrease the latency for arriving packets, it would not decrease the worst-case latency. In the case of the VoIP example, the packet spacing is larger than the maximum delay so a piggybacking would be useless to detect the next frame.

Graph 2 shows the bandwidth capacity required by the solicited VoIP solution.

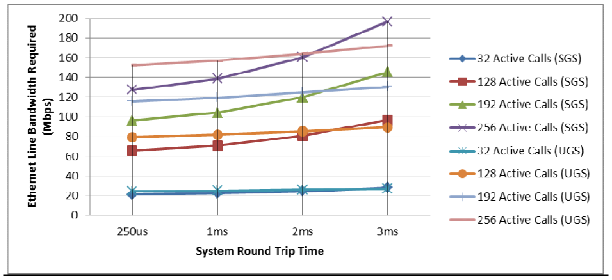


Graph 2: Required Solicited Bandwidth

The UGS bandwidth increase due to increased round trip time was much less than the solicited solution because of extra round trip in the delay equation. At a small number of active calls, the UGS shows little or no difference with a lower or higher round trip time.

The solicited solution is more efficient for the shorter round trip times. The solicited solution benefits from only granting data slots when a frame is present. As the RTT increases, the increasing polling rate to meet the maximum delay consumes more bandwidth than the wasted slot in the UGS solution.

When comparing the 250µs EPON data point, there is less than 10% increase in bandwidth to achieve the same delay performance if the round trip time is in the 1.5ms range. A 3ms round trip adds a 50% bandwidth penalty to achieve the same delays. It is clear that RTT delays beyond 3ms are unusable in the solicited.



Graph 3: UGS & Solicited Bandwidth

If the UGS and Solicited graphs are overlaid, it shows a cross over point between UGS and solicited around 2ms of round-trip time. A solicited solution is equal performance for a small number of users and it is better performance if the round trip time is less than 2ms. Since the solicited solution is more flexible for video or compressed content and simplifies controls, a lower round trip time that allows for efficient use of soliciting granting is preferred. For DOCSIS systems with many users and long delays, UGS must be used. For EPON systems with fewer users and shorter delays, solicited granting is clearly preferred.

Performance for MEF 23H

Requirement Overview

The MEF 23H service agreement is an example of a higher tier business or residential SLA. For the MEF 23H service, an IFDV of 3 milliseconds and a maximum FD of 8 milliseconds will be used as requirements. For the analysis, a 10Mbps-streaming load will be applied in the upstream direction. The 10Mbps load has a random packet size from 64 bytes to 1518 bytes.

Configuration to reach goals

With an unconstrained packet size, only solicited operation can be used since a UGS would require knowing the packet boundaries. In a system without fragmentation, the unknown packet boundary would be very inefficient. In a system with fragmentation, a packet spanning 2 grant slots would double the delay. In either case, UGS is not the preferred method.

Since the period of polling must be short to meet the IFDV requirement, there is no need to use piggybacking. While piggybacking may lower the average delay in some scenarios, it will not decrease the worst case IFDV or FD. Piggybacking would decrease the efficiency because of the additional REPORT frame in the burst.
The IFDV is the critical constraint in this system. The IFDV in the upstream will be sum of the polling interval and the scheduler delay. A packet arriving just before the

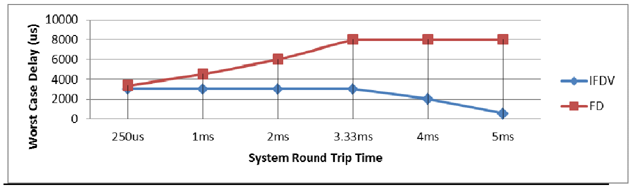
polling slot will have zero delay while a packet arriving just after the polling slot will wait the entire polling interval. The scheduler delay can be zero when only one CNU requests an upstream slot for shortest delay. The longest scheduler delay occurs when all CNUs need a slot at the same time. The maximum scheduler delay is number of CNUs times the maximum slot size.

The best efficiency can be found when the IFDV is equally split between polling and scheduler contention delay. For a 3ms IFDV, the scheduler delay of 1.5ms and a polling delay of 1.5ms are allowed.

Performance Analysis

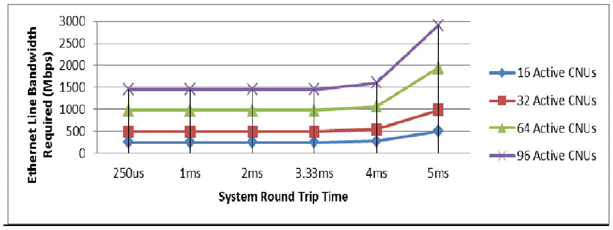
The maximum delay is defined by the same equation as the VOIP solicited grant example. It should be noted that this equation is the same as the IFDV plus the twice the upstream delay and downstream delay. The delay graph shows the relationship between the round trip time and the FD and IFDV. The bandwidth graph shows the best efficiency is found when the IFDV value is largest. A large IFDV allows for a lower polling rate and larger upstream data bursts which results in higher throughput.

At the EPON round trip time of 250µs, the maximum delay is far below the 8ms maximum. As constant delay is added for the round trip time increases, the IFDV and the efficiency remains the same. When the additional RTT delay causes the maximum delay to be exceeded, the polling period and burst size must be decreased. These decreases cause the bandwidth required to increase dramatically.



Graph 4: MEF 23H Delay

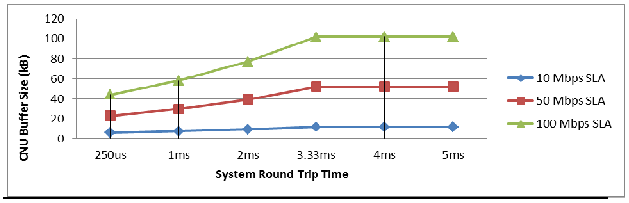
While the maximum delay increases for the RTT of 250µs to 3.33ms, delay is under the 8ms maximum and the efficiency is constant. If EPoC RTT delay is under 3.33ms, the MEF23H service can be supported without any additional bandwidth. Above 3.33ms, the penalty increases dramatically until the absolute limit of 5ms where the minimum polling period of 250µs is reached. At the 5ms limit, the bandwidth required to meet MEF 23H is more than double EPON at 250us.



Graph 5: MEF 23H Bandwidth

CNU Buffering Requirements

The additional delay will impact the buffering requirements for a CNU in the upstream direction. For a MEF 23H service, it is assumed that it is a guaranteed bandwidth without best effort data. The MEF 23M and MEF 23L will consider best effort data. With only guaranteed bandwidth to consider, the buffering required can be found by the multiplying the guaranteed rate by the frame delay (FD). Since the buffer is normally store-and-forward, 2000 bytes (the largest 802.3 packet size) is added.



Graph 6: MEF 23H Buffering

The results for MEF 23H buffer size required versus RTT has the same shape as the delay graph and the opposite shape of bandwidth graph. The increase in RTT increases the buffer size until the maximum delay is reached. After the maximum delay, additional RTT increases don't change the buffer size but bandwidth for higher polling rate climbs. Graph 6 shows that while the increase in the EPON fiber RTT delay from 250µs to 3.33ms does not hurt the efficiency, it more than doubles the upstream buffering requirements in the CNU.

Performance for MEF 23M

Requirement Overview

For MEF 23M, a worst case IFDV of 8 milliseconds and FD of 20 milliseconds will be used. For the analysis, a 10 Mbps and 50 Mbps streaming load will be applied in the upstream direction. The load has a random packet size from 64 bytes to 1518 bytes.

Configuration to reach goals

The MEF 23M traffic patterns would not normally be a traffic pattern compatible with a UGS flow. Variable sized bursts of unknown packet sizes are best handled by solicited granting.

The longer FD and IFDV limit allow the use of piggybacking for better efficiency than the polling only solution used in MEF 23H. The polling timer will be reset by the generation of a polling burst or a piggybacked REPORT frame. For a bursting station with a polling period greater than or equal to the scheduler contention delay, no polling bursts will be requested.

There are 2 equations to determine the maximum delay. The first equation is based on a station that has been active but not bursting. Polling will detect the packet in this case. This scenario will be referred to as "burst detection".

Tmax_delay = Tpolling + 2xTup + Tdown + Tall_service

The second equation is based on a CNU that is bursting and not reporting a zero length queue. In this case, the piggybacking will detect the packet arrival. This scenario assumes that the flight delay of 2xTup + Tdown is less than the time to service all stations. The scenario will be referred as "burst continuation".

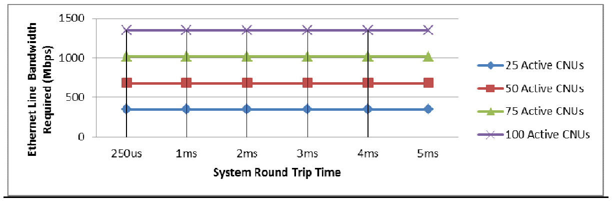
Tmax_delay = Tup + 2xTall_service

The worst case delay can be determined by taking the longer delay from the burst detection or burst continuation scenarios. For optimum performance, the polling interval should never be less than Tall_service and for best performance, they should be set equal. In this case, the burst continuation equation is not the worst case so the burst detection equation will be used for analysis. The Tpolling interval will be half the result of the maximum delay minus 2xTup + Tdown.

For a system mixed with higher priority services like MEF 23H, the Tall_service should include their burst interruptions. Tall_service should be the sum of all higher and same priority upstream slots. Since the MEF 23H IFDV is 3ms, the MEF 23M will assume no more than a 3ms disruption.

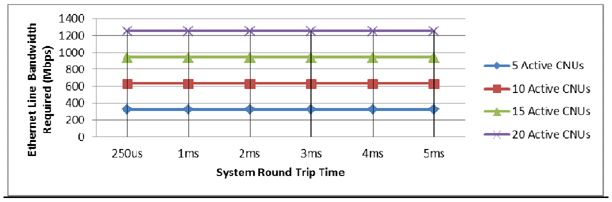
Performance Analysis

Graph 7 shows that as the round trip time increases no increase in the bandwidth required. Since the FD of 20ms is larger than the IFDV of 8ms plus 5ms RTT, there isn't a need to increase the granting rate. The bandwidth increase wouldn't occur in the MEF 23M until a RTT of around 12ms.



Graph 7: MEF 23M Bandwidth (10Mbps)

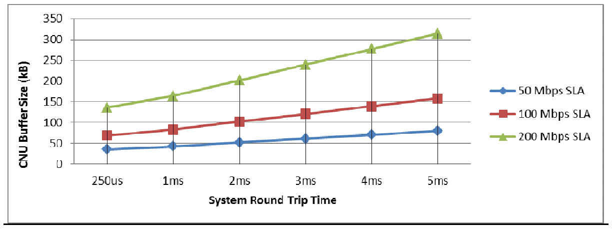
Graph 8 shows fewer active CNUs and therefore fewer bursts at a 50Mbps rate each. The charts show that the penalty for extended RTT is larger when there are more users and lower data rates. For a system with many users and higher data rates, the RTT doesn't have significant impact up to 5ms.



Graph 8: MEF 23M Bandwidth (50Mbps)

Buffering Requirements

While the efficiency of the system for MEF23M is constant from with the increased RTT, the buffering requirements on the CNU are not. The buffer required on a CNU to support the MEF 23M with data rates up to 200 Mbps would need to be more than double the EPON ONU. Graph 10 shows the buffering requirements to support average rates of 50, 100, and 200 Mbps.



Graph 10: MEF 23M Buffering

MEF 23L

Requirement Overview

The MEF 23L service agreement is an example of a best effort SLA. The MEF 23L specification contains a maximum FD requirement of 37 milliseconds. For the analysis, different data rate streaming load will be applied in the upstream direction. The load has a random packet size from 64 bytes to 1518 bytes.

Configuration to reach goals

For the same reasons as MEF 23M, a solicited granting with piggybacking will be used. A 37ms delay limit is very long for an EPON or EPOC system that is not oversubscribed. The RTT will be a small percentage of 37ms delay limit so it will not have a significant impact on the efficiency like the MEF 23M. The RTT will have a significant impact on the buffering requirements for a CNU to reach high bandwidth. In general, the MEF 23L needs to achieve high efficiency at a high data rate without requiring a large amount of upstream buffering.

The polling rate could be set for MEF 23L to 10ms and meet the FD requirement of 37ms. For the MEF 23L, different polling rates will be considered to balance efficiency with buffering requirements.

The burst detection condition will be considered for the same reason as MEF 23M.

The Tall_service delay is more difficult to determine at this priority level since many higher priority services could be active. The disruption from MEF 23H and MEF 23M services will be limited by the MEF23M IFDV of 8ms. Of course, this analysis assumes a round robin scheduler with guaranteed slots for lower priorities and shaping that streams the higher priority. Without these restrictions to the high priority, the delays to MEF 23L could be unbounded. Tpolling will be equal to Tall_service.

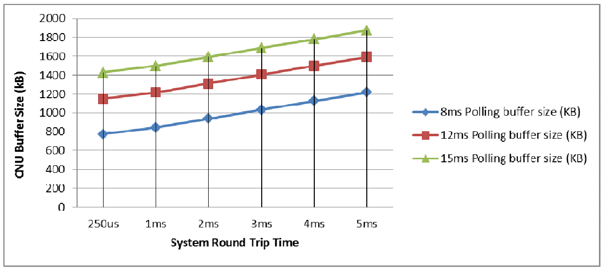
Tmax_delay = Tpolling + 2xTup + Tdown + Tall_service

The polling interval for the MEF 23L service can be determined subtracting the loop time and dividing by 2.
Tpolling = (Tmax_delay - 2xTup – Tdown)/2

To handle the disruption from high priority services, the MEF 23L polling rate shouldn't be set less than the 8ms IFDV of MEF 23M. For a 5ms delay, the maximum polling rate is just under 15ms. The analysis will look at polling rates from 8ms to 15ms.

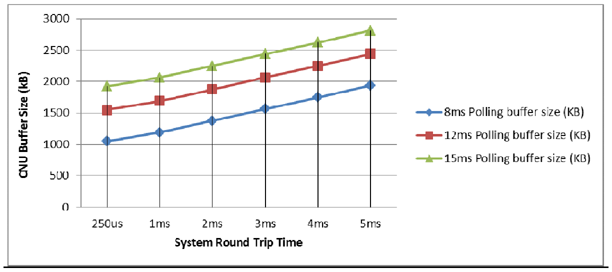
Performance Analysis

The MEF 23L buffer size is calculated for the different polling rates versus RTT. In Graph 11, the buffer size is considered for a sustained rate of 500 Mbps with 50% of the bandwidth taken by MEF 23M services. The buffer requirements are the maximum delay times the sustained input bandwidth.



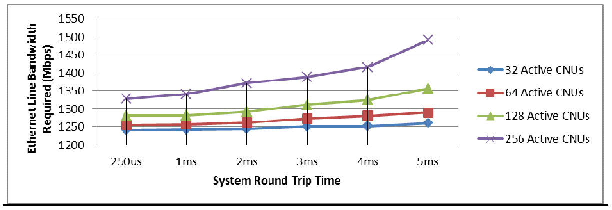
Graph 11: MEF 23L Buffer Size (500 Mbps)

Graph 12 shows the buffer size requirements for an empty system where a single MEF 23L CNU is bursting at 1 Gbps (100%) with no contention delay. Comparing Graph 11 and 12, it is clear that the worst case buffer requirement for MEF 23L is a single user with an SLA to reach maximum bandwidth. The buffer requirement decreases with more users sharing the upstream as the maximum data rate decreases.



Graph 12: MEF 23L Buffer Size (100% load)

Graph 12 shows that EPoC will require a significant amount of additional buffering (~1 MB) over EPON as the RTT time is increased. From Graph 12, the buffering requirements for EPON and the 5ms RTT are equivalent if the EPON system uses 15ms polling and the EPoC system uses 8ms polling. Since the buffering is a directly related to the delay, the EPON and EPoC would have the same delay as well. For a system with few CNUs, the penalty to compensate for RTT delay with polling will be small but a larger system will require significantly more bandwidth. Graph 13 shows the impact of increasing the polling rate to match the delay and buffer requirements of EPON. In the example for Graph 13, a fixed buffer size of 1.5 MB is used. The 1.5MB buffer represents the 12ms polling, 250µs RTT, and 25ms delay on Graph 12. Graph 13 assumes that the system will carry 1 Gbps of Ethernet traffic split evenly across the stations.



Graph 13: MEF 23L Bandwidth (100% load)

Graph 13 shows that the penalty for increased RTT multiples by the number of users. The system with 256 active users will have around a 15% penalty on bandwidth to match to match the EPON fiber based performance.

Conclusions

An EPoC PHY can be used by a cable operator in multiple network configurations. The EPoC PHY could be placed with an OLT in headend and operate over a traditional HFC network or the EPoC PHY could be placed in a CMC at a remote node and act as a switch or a repeater. The choice of architecture is dependent upon the individual operator's needs and plant design.

EPoC can provide a significant performance improvement over existing cable systems because of small service groups, a fast Ethernet MAC, and a single wide logical pipe. EPoC can provide VoIP, MEF 23H, MEF 23M, and MEF 23L services if the round trip time is low enough. RTT increases will impact the CNU cost dramatically if it requires an EPoC specific chip with more buffering than the standard EPON ONU. Increased polling rates can compensate for larger round trip times to certain limits and still meet MEF 23 requirements but bandwidth efficiency will be reduced. Going over 2ms, forces EPoC from a solicited VoIP into the less flexible UGS VoIP. At RTT's of 3.33ms and 5ms, some MEF 23 services become impossible. Solutions with a shorter round trip time will be more efficient and perform closer to the fiber solutions without additional hardware or bandwidth costs.

The IEEE 802.3 standard should seriously consider the round trip time impacts in selecting the solution. A solution that increases the bandwidth efficiency at the PHY layer by adding significantly delay could hurt the overall system efficiency.

References

[1] Glen Kramer, Ethernet Passive Optical Networks, McGraw-Hill, 2005
[2] MEF Technical Specification MEF 23.1, "Carrier Ethernet Class of Service - Phase 2"
[3] EPON Protocol over a Coax (EPoC) PHY Study Group. http://www.ieee802.org/3/epoc/index.html

EVOLVING THE HOME ROUTER TO AN APPLICATIONS DELIVERY GATEWAY

Joe Trujillo and Chris Kohler
Motorola Mobility, Inc

Abstract

The home router has become a power house of performance, enabling a dizzying number of devices in the home to communicate with each other and the internet at ever growing bandwidth and capacity. With all this impressive brawn, it is easy to overlook the router's potential for brains.

The home router is an always-on device that is completely intimate to the physical and logical connectivity between devices on the home network and their connections to the internet. That intimacy makes the home router uniquely positioned to host a variety of applications.

In this paper, the authors discuss some of the applications that can supply a brain to accompany the brawn for next generation routers. Some example applications discussed relate to Machine-to-Machine (M2M) communication for home control and security, Personal Content Management, and Advanced Home Network Management. While this list is not exhaustive, it gives a fair idea about the possibilities and opportunities for the Service Provider to move up the value chain, while continuing to delight the customer.

INTRODUCTION

Until now, the nearly complete focus of the home router's evolution has been on improvements in the performance of IP connectivity, while the router's own participation in using that connectivity has been suppressed, maybe even discouraged. One could say that the focus has been on brawn - faster speeds - over brains. The time has come to turn some of that focus towards developing gateway intelligence by way of hosted applications for which the home router is uniquely positioned and qualified.

WHAT KIND OF APPLICATIONS AND WHAT MAKES THE ROUTER QUALIFIED?

A home router is not suitable for every kind of application. It has no keyboard, no joystick, no screen nor speakers of its own. Hosting games, word processors or corporate payroll applications makes no sense at all. The best applications for it to host are those that leverage and extend its innate properties. Simply put, those key properties are 1) It is always on; 2) It is connected to the internet; 3) It is intimately connected to every IP device in the home. Taking the concept one step further, an integrated home router with built in broadband access, such as DOCSIS® 3.0, xPON or bonded DSL, would expand the reach of the hosted applications into the WAN (see Figure 1).

The always-on nature and therefore its ability to continuously access both the internet and devices on the home LAN make the integrated home router the perfect place to host applications that need to provide one or more of the following properties [brain functionalities]:

- Anytime or always-on availability [always thinking]
- On demand or near real time access/control to devices on the LAN [gross motor skills]
- On demand or near real time access to devices on the *once-removed** network [fine motor skills]

* The "once-removed" network is the collection of devices in the home/office that are not necessarily directly IP connected, but can be controlled and/or monitored by other devices that are in turn IP connected. Examples of some technologies that can act in the once-removed network are Bluetooth® (1), ZigBee® (2) and Z-wave® (3)

- A high degree of local abstraction to hide, when necessary, the complexity of the local network or the once-removed network. This allows for more uniform and less complicated communication protocols between the Cloud or other internet devices on the LAN or the once-removed network [can process the environment to abstract and simplify clutter]
- A high degree of local autonomy and in-depth local knowledge to discover WAN, LAN and once-removed topology [is self aware and can communicate it's condition]
- A high degree of local autonomy to help in scaling or offloading from the Cloud or management system [thinks for itself, but is a member of a community]

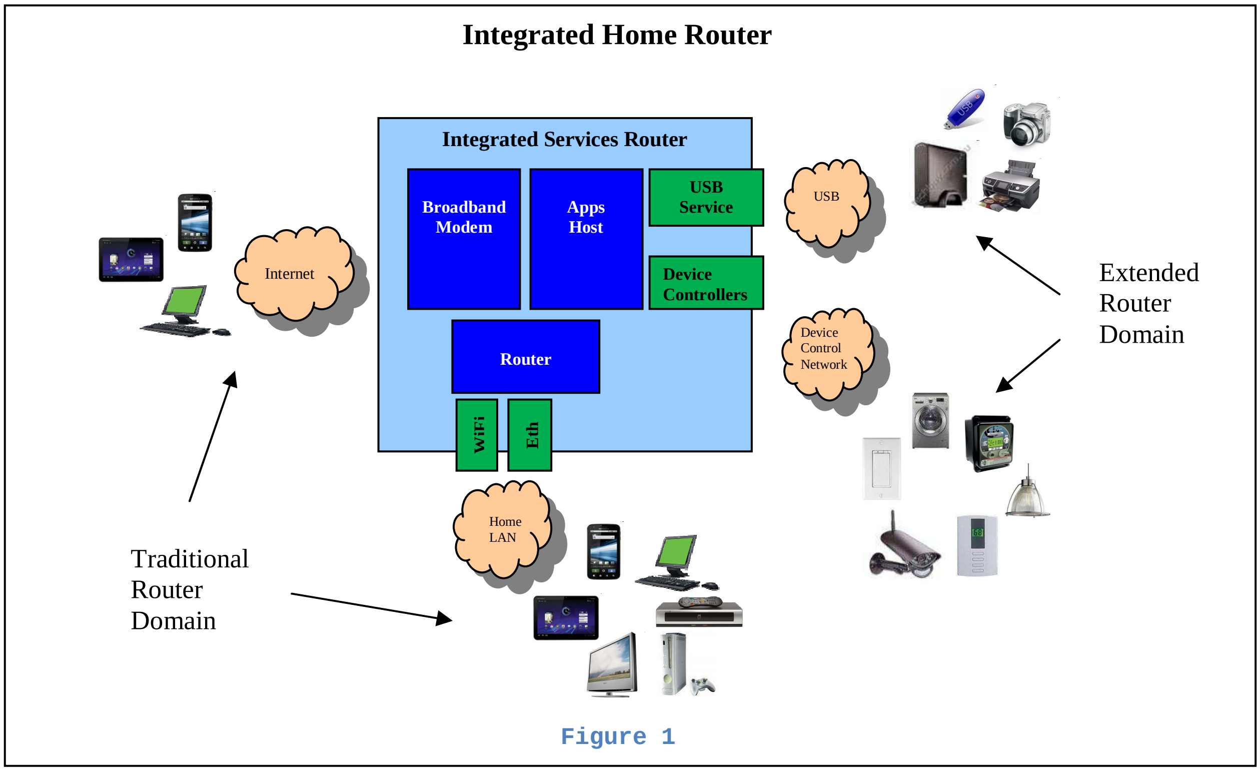


Figure 1

M2M CONTROL POINT APPLICATIONS

Some of the most interesting examples of applications which are ideal for integration into the home router are Machine-to-Machine (M2M) control points. Of course the concept of one device “remote-controlling” another device is not new to the internet. In the most basic sense, M2M is one smart device talking to another smart device via a communication network (4) . In industrial applications, such as on a complex factory floor, M2M has had natural and wide adoption, albeit for a closed environment and a non-consumer market. For the home/consumer market new possibilities are just beginning to open up.

There are several emerging genres of M2M applications for the home. Each of these genres is best serviced from a Cloud portal vantage that can homogenize the presentation to the end users, simplify presence and discovery of devices from across the internet, and be the integration and launching point for service extensions or other services supplied by the service provider. That said, hosting the control point portion of the application in a home router with its integrated WAN or broadband access and direct connectivity to the LAN and once-removed network provides the best solution for the service provider to deliver, control and manage the entire experience.

Home Automation

Examples of features in this genre include the ability to remotely turn on your sprinklers, turn off your air conditioner, turn on or off lights or even unlatch the dog door from a smart phone, computer or hosted scheduler. These are convenience features once only available to high end homes via highly custom installations.

Home Security

Features in this genre would include remote enable/disable of the alarm system, monitor/control of individual sensors (window, door, and motion), and control of camera pointing, scanning and live/recorded viewing of video feeds.

Home Energy Management

Features in this genre would include remote monitoring of total home power usage and/or usage on a per-device basis. Historical analysis of telemetry can be used to detect and correct consumption patterns. Available interfaces to the utility company's portal can be utilized to create useful correlations and validations of power consumption, including actual costs incurred due to specific power consuming devices such as air conditioning, clothes driers and entertainment clusters. Triggers could be used to inform the homeowner of a "violation" in progress, such as the drier being turned on during peak usage or peak billing hours. That alert could come, for instance, as an SMS to a cell phone or an alert ring and pop-up on a custom smart phone app.

Senior Care Monitoring

Features in this genre would include monitoring door sensors, motion sensors and pressure sensors to allow passive monitoring of the elderly or infirmed. Cameras could be added for more complete, but more intrusive monitoring. Triggers such as lack of motion for a prolonged period (have they fallen?) or opening of an off limits door (the front door leads to traffic or dangerous stairs) could alert a care giver and prompt a phone call, visit or emergency action.

Advanced Medical Monitoring

Features in this genre would include gathering telemetry from scales or other medical equipment such as heart rate monitors and glucose meters. For advanced medical monitoring, security, senior care and home automation could be combined. For critical care, perhaps FDA certified/approved devices for M2M applications have a market place.

It is important to note that these home oriented M2M features are not just about one-way remote control services into the home. Their best utilization is when a diversity of machines takes advantage of their local capabilities to build something more useful.

Here's an example of a fully automated M2M scenario that one could envision being easily "programmed" by an end user from a properly equipped smart phone. Using the phone's GPS, the phone can detect when it has moved one mile away from home. Using this event as a trigger, the phone can interact with the M2M network (via the Cloud to the home M2M control point) and cause the home doors to lock, the home alarm system to enable, verify and close the garage door, send an SMS or email from the phone to the elderly care service provider that the person has left the house and even pop open the doggy door to prevent an embarrassing accident.

Some major operators have already entered the home M2M market place and are deploying solutions. These kinds of engagements are expected to grow and help drive technologies and monetized deployments at an accelerated rate. Industry initiatives, such as the TIA's TR-50 (5) and ETSI M2M (6) promise to further standardize the M2M ecosystem and bring a plethora of interoperable service opportunities to the telecommunications industry.

PERSONAL CONTENT APPLICATIONS

Another natural set of applications for an always-on home router have to do with file storage and media access. Network attached storage (NAS) systems for the home are not new, but their presence in the marketplace appears to be growing. Digital photo, music and movie collections grow rapidly, but are almost always spread out over many devices (phone, tablets, cameras, computers). The desire to ease the ability to collect files from these devices to a central location is becoming more urgent.

Collecting the media (copying) to a central location provides a back up to the phone or camera against disaster and provides a place to store when the internal storage of the device becomes full. When a consumer consolidates media, they typically choose to use a home computer's hard drive. This approach is fine for back up and overflow storage, however, it can have some serious limitations.

Setting up an environment where other devices on the home LAN can access that computer's hard drive is complicated and not guaranteed to interoperate across varying devices' operating systems. Remote access from the internet to the computer's hard drive is not possible without special software on the computer. Maybe most limiting is that a computer can be turned off or in the case of a lap top, not even be at home. An always-on integrated home router with attached storage capabilities provides a platform to overcome these limitations.

There are several NAS devices in the market today that can be plugged into the Ethernet port of an existing home router. With enough patience to configure the NAS and the IT properties of the router, many solid features become available to the end user. These features typically include: SAMBA (LAN) access to files available on the NAS; DLNA-Server streaming of media files stored on the NAS to the growing list of compatible devices on the home LAN, including game consoles, MAC and Microsoft OS computers, Wi-Fi™ enabled TVs and Blu-ray Disc™ players; and remote access to files on the NAS from the internet. Remote access capabilities can be extended to social media and file sharing features, with mailing lists and automated posts to social media outlets.

All these features can be supported with a NAS application integrated in a home router. Several additional benefits over a standalone NAS are available if the Router/NAS combination also contains an integrated broadband modem.

Automated Configuration

Since the NAS, router and broadband access are integrated into a single box the configuration is automated. The user doesn't need to know how to configure the router to grant the NAS access, configure DHCP to get it on the network, or assign ports and port forwarding rules to allow internet access.

Advanced Management

It was noted above how an integrated device can automate the configuration tasks. In a service provider deployment there are additional advantages in the ability to manage and monitor the modem, router and integrated NAS as a single entry. A standard retail standalone NAS has no remote management capacities, such as TR-69 or SNMP. A full integration eliminates this problem, enabling the operator to have a much better position to manage a deployment. The combination of automated configuration and advanced management can be a great aid in customer satisfaction and customer loyalty.

Hardware Cost

The cost savings to the operator or as passed on to the end user of a consolidated box could be significant. The cost of buying separate modem, home router and NAS devices can stack up as compared to buying an all-in-one integrated router/NAS device.

Converged Commercial Media Routers

The advantages above will become even more pronounced as the traditional video set-top box continues its evolution towards the IP video gateway. The need to distribute live, on-demand or recoded video to devices on the home LAN will magnify the need for an integrated home router. SOCs which enable IP video distribution capabilities inside an integrated home router will start to appear in the market place in 2012.

ADVANCED HOME NETWORK DISCOVERY AND MANAGEMENT APPLICIATIONS

As stated at the beginning of this paper, until now the focus of the home router's evolution has been on improvements in the performance of IP connectivity. This performance increase is the great enabler of our time. The importance and continued evolution of throughput performance can't be overlooked and must continue for the foreseeable future. However, with all these improvements comes a drawback that must be overcome – high complexity.

Year over year the worry has been stated that lack of bandwidth would cripple quality of service (QOS). There have been numerous strategies to head this crisis off with advanced QOS methodologies, only to find that timely, cost effective technology advances in performance bail us out. It seems that a lack of bandwidth may not be the killer of QOS. The pipes keep getting bigger, symbol rates denser, spectrum more available and diversity transmission techniques ever more standard. However, it may be the complexity and digital clutter associated with this level of improved performance which could be the killer of QOS.

Bonded DOCSIS® 3.0, bonded DSL, 3G, 4G, Gigabit Ethernet with more ports, MoCA®, HomePNA®, multiple SSIDs per multiple Wi-Fi radios, HomePlug®, L2 tunnels, VLANs, VPNs, dual homed WAN - the list seems endless. The technological complexities and home-by-home variations of devices and interfaces have exploded. A typical home is starting to look like an enterprise deployment. But unlike an enterprise, every household cannot afford its own IT department. Compound this with the fact that traditional TR-69, SNMP and other call center techniques are insufficient to scale to the situation without some paradigm shifts. For the most part, current management systems are set up to query the discrete values of pre-known parameters internal to the router's configuration. These techniques are almost blind to the fluid nature of the devices on the home LAN.

A solution to solve this scalability and variability problem is to put much more intelligence and autonomy in the home router. This locally hosted application can analyze the network, detect issues and alert the user or customer care agent of a problem and where to fix it. Better yet, take this local intelligence to the next step for analyzing trends and alert and/or correct an impending problem before it becomes service affecting.

Network Discovery

Keeping track of what devices are on the home LAN can be a challenge. IP enabled computers, tablets, phones, games, set-tops, TVs, Blu-rays, printers, file/media servers and many other devices are popular in the home. How does an operator, customer service agent or even the home user know what devices are connected right now and what the expected properties these devices have so they can help setup or debug the home network? Current TR-69 or SNMP techniques can query some standardized MIBs to get some modem, DCHP, Wi-Fi information and perhaps a few more general router stats and try to interpolate a bigger picture. This can take many queries and still leave the agent without critical information.

Of course the integrated home router is the perfect location for hosting a Network Discovery application. It intrinsically has access to many pieces of information such as DHCP lease table and switch/Wi-Fi learning tables. It can ARP scan for devices that may have statically joined a subnet. Further probes and traffic monitors can discover UPnP devices and their capabilities and probe local IP devices for HTTP Web page capability. This gathered data can be used to create a small database representing the discovered nodes on the home LAN, how they are connected and most important, useful information on each device.

This database is easily exposed for use on the router's local UI to draw a network map that can be drilled down with mouse clicks or as a file which is available to a management system for it to draw the map for a customer service agent. The management system application that uses the topology database can then further augment diagnostics and corrective action by using traditional SNMP or TR-69 management objects.

Network Histogram

The Network Discovery application embedded in the router can automatically refresh the topology database at regular intervals. Changes in targeted parameters from a baseline can be recorded at regular intervals. With this method the database then becomes a histogram that can be useful in capturing variations and instabilities in the network. For instance, it could see that a fixed position Wi-Fi device intermittently drops on and off the network. Imagine the frustration saved by the customer care agent who can actually react with more than just sympathy to a customer saying "Well, it was happening this morning before I called!"

Trend Analysis and Alert Triggering

This application realm dives deeper and takes a running statistical look at the core access technology interfaces. Using various interfaces' instantaneous measurements and counters available on the integrated router, the application can collect, record, average, filter and analyze trends that can be used to take preventive action before an outage can occur.

Let's take a DOCSIS® 3.0 bonded downstream connection as an example. In a typical DOCSIS® 3.0 modem the downstream could consist of data distributed on 8 individual channels (QAM modulated data on 8 frequencies) that are captured and re-sequenced in the modem to create a 300Mbps connection. Each channel is subject to its own analog variations in signal quality due to minor Tx power fluctuations and interferences. In nominal operating conditions, digital receiver techniques are transparent to this "noise". However, if one or more channels degrade such that transmission errors become significant, then performance and connectivity will quickly degrade and perhaps result in an outage and a truck roll. Having a remote management system poll many measurements 24/7 across 8 channels is neither realistic nor scalable across a large population of devices. Furthermore, any single sample measurement has almost no meaning as far as "good" or "bad".

A statistical application local to the integrated router could monitor a history of vital signs like raw Frame Error Count (FEC), corrected errors and downstream power. For example, on a per channel basis, a rolling database window could record averaged samples over a statistically significant period of time and show if the frame error count, translated to a frame error rate, is trending up indicating a problem on any channel. It could be useful to graph the table to show this trend visually. Better yet, the application can track the trend itself and on a threshold, send an alert (SNMP trap, TR-69 inform) informing the management system or customer care proactively. This technique could be extended to Wi-Fi, Ethernet, MoCA®, HomePNA® and other interface types in the system.

SUMMARY

We've stated that the integrated router is the best choice for hosting applications needing the properties described in the opening paragraphs. The always-on nature guarantees access *when* it's needed. Its connectivity to the internet guarantees access from *where* it's needed. And its intimacy to all devices on the extended home network guarantee access to *what* is needed - simply, conveniently and at high quality. The example applications outlined reinforce this point of view.

M2M applications demand all the brain qualities the router can provide. They need to always be on and ready, connected through internet and provide on demand access to devices on the extended home network. These applications need a high degree of local abstraction and autonomy to hide complexities from the user experience and scale to the Cloud.

Personal Content applications are more valuable when the content can be accessed and exchanged from anywhere and anytime. The local autonomy and intimacy of the application with the router make configurations automatic and remote management seamless.

Advanced Home Network Discovery and Management applications take great advantage of the intimacy between the router and broadband modem systems, performing continual measurements and diagnostics not available or scalable from traditional management systems alone. This helps ensure the technological complexity of the networking environment doesn't subtract from the reliability and usability of the connection.

This is also a good time to circle back and thank our friend, performance. Thirst for greater performance has driven the silicon industry to higher densities making more processing power available to router applications in the form of faster CPUs and multiple cores. In older generations of silicon the desire may have been there for hosted applications on the router, but the processing platform was not. It's the brawn of the modern integrated router that has made the brain possible.

References

1. **Bluetooth Special Interest Group (SIG).** Specification: Adopted Documents. *Bluetooth Special Interest Group (SIG).* [Online] Bluetooth SIG. www.bluetooth.org/Technical/Specifications/adopted.htm.
2. **ZigBee Alliance.** ZigBee Standards Overview. *ZigBee Alliance.* [Online] ZigBee Alliance, 2012. [Cited:] http://www.zigbee.org/Specifications.aspx.
3. **Z-Wave Alliance.** Z-Wave Products. *Z-Wave.* [Online] Z-Wave Alliance, 2011 . http://www.z-wave.com/modules/Products/.
4. *The Promise of M2M: How Pervasive Connected Machines are Fueling The Next Wireless Revolution.* **Syed Gilani.** 2009, Embedded Systems Magazine - White Paper
5. **Telecommunications Industry Association (TIA) .** TR-50 - SMART DEVICE COMMUNICATIONS. [Online] http://www.tiaonline.org/all-standards/committees/tr-50.
6. Machine to Machine Communications. *ETSI - World Class Standards.* [Online] 2011. http://portal.etsi.org/m2m.

HFC NETWORK CAPACITY EXPANSION OPTIONS

Jorge D. Salinger
VP, Access Architecture
Comcast Cable

<u>NOTE:</u> The concepts and proposals presented in this paper are for discussion purposes only and do not reflect actual plans from Comcast. Similarly, all examples presented are only provided for illustrative purposes.

Abstract

MSOs are deploying more narrowcast capacity than ever before, and there is no evidence of a change in this trend.

- *DOCSIS® 3.0 is widely deployed, with 4 and 8 downstream channel bonding groups becoming the norm. A continual annual growth of 40-60%, observed industry-wide for over 10 years, would require many more channels over time*
- *8-channel service groups for video on-demand (VOD) are commonplace. Growth rate increasing due to both higher usage and higher bitrate (high definition)*
- *10, 20 or even more channels for switched digital video (SDV) are frequently used for longer tail content*
- *Growth in business service applications requires additional increased capacity*
- *The advent of IP video services and network-based digital video recorder (DVR), which are anticipated to be very popular amongst current and potential subscribers, will compound the need for additional narrowcast capacity.*

The effect of the above trends, combined with the need to simultaneously support a full set of legacy broadcast services, including digital, analog and/or both, would likely require additional hybrid fiber-coax (HFC) network capacity.

While it seems conceivable that a transition from legacy and broadcast services to an all-narrowcast/IP services infrastructure could be established, the industry as a whole is looking for options that would provide additional capacity to support simultaneous uses, and increased capacity beyond such transition. These options include:

A. *Traditional service group segmentation*

B. *Move quadrature amplitude modulation (QAM) generation downstream into the network*

C. *Implement higher modulation physical layer (PHY) and/or more efficient media access control (MAC) protocols*

D. *Increase HFC downstream capacity beyond currently deployed, and/or move split to higher spectrum for increased upstream capacity*

E. *Develop technology that would operate in unused portions of the spectrum, and even unleash spectrum above current top range (e.g., above 1 GHz)*

Each of the above options has benefits and drawbacks. Each approach offers different network engineering and operational simplifications and complexities. And, the relative improvements in offered capacity versus cost and customer impact can be significantly different.

This paper will provide, from an operator's perspective:

1. *A technical overview of each of the options outlined above, describing how each would be deployed and evolved over time,*

2. *The key benefits/drawbacks for each of the options, including engineering and operational pros and cons for each option,*

3. *Possible implementation approaches for various applications, including residential and commercial services.*

TYPICAL HFC NETWORKS TODAY

Most MSO's hybrid fiber-coax (HFC) networks have been designed to either 750 or 860 MHz of spectrum capacity. If not fully utilized, it is expected that use of their capacity will be increased to the point of exhaustion as the use of DOCSIS® increases for the higher high-speed data (HSD) service tiers, additional high-definition (HD) programs for both broadcast (BC) and especially narrowcast (NC) services such as video on demand (VOD) and switched digital video (SDV) are deployed, or new services such as internet protocol (IP) video and network-based digital video recorder (n-DVR) are added.

Proportionally few HFC networks have been deployed to operate up to 1 GHz, although all equipment available today can support the use of spectrum up to 1 GHz and even 3 GHz for some components.

In recent years the growth in, and demand for, HD programming has resulted in the need for allocation of large numbers of EIA channels for HD services, both for BC and NC, which has filled every available portion of the spectrum. This is especially true for BC, where large numbers of programs are offered in HD format, while simultaneously the need for distributing the standard definition (SD) version has persisted. This has resulted in the need for use of 3x to 5x the number of EIA channels than previously required. For example, a typical digital multiplex including 10 to 15 programs would require an additional 3 to 5 EIA channels for the HD equivalent streams, even assuming the newer, more sophisticated multiplexing schemes available in the market. Of course not every program is available, or still sought by subscribers, in HD format. But very large numbers of them are, including 100 to 150 BC programs.

The above is also applicable to a great extent in systems utilizing SDV technology for distribution of its content. The difference is that the SD version of the program is not distributed unless a subscriber is requesting it, which reduces the marginal increase in capacity. Assuming that all programs are distributed in only one format, which is certainly a valid expectation for programs of low viewership, then the increase in capacity for a conversion from SD to HD would just be the increase in capacity required for the transmission of the HD program without requiring the simultaneous use of bandwidth for both formats.

Additionally, considerable spectrum is needed to deploy high-capacity narrowcast legacy video services, especially n-DVR, and a full-array of HD video-on-demand services. For the former, initial observations suggest that network requirements for n-DVR may be as high as 4x to 5x that of VOD, and that peak utilization overlaps, at least partially, with that of peak use for other narrowcast services.

Finally, the growth in HSD services shows no sign of letting up. Network operators have observed an increase use of HSD service capacity for well over a decade now, which amounts to a year-over-year compounded growth of 40% to 60%. The applications have changed throughout this time, but the demand has continued to increase at the same relentless rate.

In fact, such increase in demand for HSD capacity shows no evidence of decreasing. Should that trend continue, MSOs would be in a position to increase access network capacity through either one or more of the existing capacity tools and/or through one or more of the new capacity tools outlined in this paper.

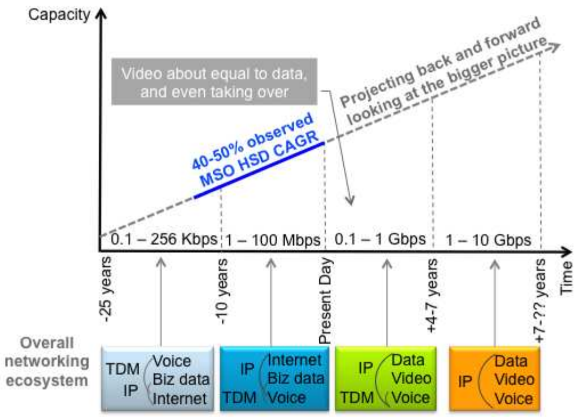

Capacity
Video about equal to data, and even taking over
40-50% observed MSO HSD CAGR
Projecting back and forward looking at the bigger picture
0.1 – 256 Kbps
1 – 100 Mbps
0.1 – 1 Gbps
1 – 10 Gbps
-25 years
-10 years
Present Day
+4-7 years
+7-?? years
Time
Overall networking ecosystem
TDM IP Voice Biz data Internet
IP TDM Internet Biz data Voice
IP TDM Data Video Voice
IP Data Video Voice

Figure 1: Example of HFC capacity utilization over time

How does this compare to other operator's data services and a longer period? As shown in Figure 1, projecting the MSO's HSD service growth back in time to when Internet services started as shown in the diagram, 25 years ago services should have been about 100 bps. This coincides with the history of telephone modems from 110 and 300 baud modems from the mid-80s, to 56 Kbps/V.42, into ISDN services.

This demonstrates that the growth seen in MSO's HSD services is typical over a much longer period of time, rather than an exception observed by MSOs in recent years.

GROWTH PROJECTIONS

From all of the above, it then follows that, should the usage growth pattern continue at the past experienced pace, networks will be required to provide HSD services in the range approximating 1 Gbps within the next few years. This growth, coupled with the surge in HD video formats, and more personalized narrowcast services, will result in a significant growth in NC capacity, as shown in Figure 2 below.

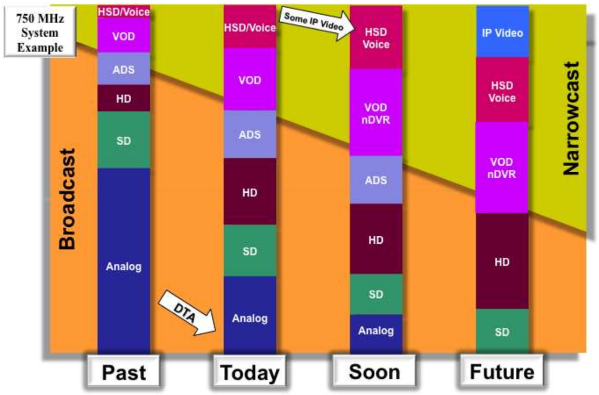


Figure 2: Example of narrowcast service growth over time

To support this growth, MSOs have deployed, or are considering deployment of, bandwidth reclamation tools such as SDV for digital broadcast, digital terminal adapters (DTAs) for analog services, or a combination of both. These tools have been extremely valuable to MSOs, which have seen their operational complexity and cost to be well justified.

In the case of SDV, early predictions several years back from industry analysts projected that the efficiency of SDV would reach 40% (e.g., programs requiring 10 EIA channels could be carried in 6). This has proven to be understated, since it was based on the use of SDV for reduction in bandwidth required for existing services. As SDV's role in the network grew, the efficiencies have been even greater, especially as SDV has been used to introduce niche services that have low viewership and would have otherwise been difficult to deploy.

The benefit of DTAs has been just as, or perhaps even more, striking. MSOs deploying DTA devices are able to eliminate the need to distribute the analog channels in the network. Even if DTAs are distributed to top analog tier customers, such as only to subscribers of the traditional expanded basic subscribers, such deployment would reduce a channel line up from perhaps 50 EIA channels dedicated to 50 analog programs to perhaps as little as 4 EIA channels dedicated to transport the 50 programs in their equivalent digital transport. Using the same comparison method as the above SDV case, this is a >90% efficiency. If extended to the entire analog tier the efficiency gains are very significant.

Despite the availability of these tools, they are not universally applicable. With respect to SDV, in general it is not likely that all broadcast programs will be switched since experience shows that many broadcast programs are constantly viewed by someone in the service group during peak hours, which will leave a large portion of the spectrum still used for broadcast. Similarly, not all analog channels can be removed in the short term due to operational and/or cost constraints.

Additionally, while many MSOs will use one or both tools, in general these tools won't be used by every MSO for all applications. Finally, there are also significant potential gains to be achieved from the use of advanced video CODECs (AVCs) and variable bit-rate (VBR). In the case of AVCs, coding efficiencies of approximately 50%, depending on implementation and content type, can be obtain with H.264[1] and/or MPEG-4 Part 10[2]. And the use of VBR promises to result in a capacity efficiency gain of as much as 70% versus CBR[3]. The combined gains from using both approaches could be very significant.

However, these are difficult tools to take advantage on the network since proportionally relatively few legacy set-tops still support AVCs and VBR, especially the latter. These tools will likely enjoy significant support in newer, IP-video based services equipment moving forward.

But, this approach will require additional capacity on the network. This is especially true when considering that the deployment of these advanced video services will result in an additional simulcast of video programs, at least initially, which is expected since its deployment will not at least initially replace the currently deployed services. Furthermore, ubiquitous support for such devices would require considerable spectrum if the legacy services are maintained for an extended period, as it is expected since legacy devices are and will continue to be deployed. Moreover, this increase in simultaneous use of advanced, IP video services while maintaining legacy services will be especially impacting over time as its penetration increases.

All of the above, coupled with the success experienced by MSOs in recent with business services, will likely require the deployment of IP capacity beyond what can be supported

[1] ITU-T Recommendation H.264: 2005, Advanced Video Coding for generic audio-visual services

[2] ISO/IEC 14496-10: 2005, Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding

[3] Capacity, Admission Control, and Variability of VBR Flows, CableLabs Winter Conference, February, 2009

today, requiring the development of tools for increased efficiency in the use of spectrum and/or unlashing of additional spectrum in the HFC network. The following sections of this paper will enumerate ways in which this can be achieved.

OPTIONS BEING CONSIDERED

Let us review the categories of options being considered throughout the industry, and evaluate how each one fulfills the above desirable targets. In the process, let us review the key implementation aspects of each option, leaving for another opportunity the details of the options and on how these could be deployed.

The categories of options are:

1. Traditional service group segmentation

2. Move QAM generation downstream into the network

3. Implement PHY and MAC improvements

4. Increase HFC downstream capacity beyond currently deployed, and/or move split to higher spectrum for increased upstream capacity

5. Develop technology that would operate in unused portions of the spectrum, and even unleash spectrum above current top range (e.g., above 1 GHz)

1. Traditional service group segmentation

This option is readily available and has been in use for many years. It consists of decombining service groups (SGs) when possible, or dividing nodes into smaller groups when decombining SGs is no longer viable.

Traditionally SGs have consisted on a number of nodes combined together in the cable headend, and nodes include a number of homes passed and corresponding subscribers. Therefore, service group segmentation normally is achieved initially by separating nodes into smaller SGs, and when SGs consist of a single node these are segmented further by separating a number of the homes in a node into a new, separate node.

For example, assume that a SG consists of 2,000 homes passed (HHP), which results from combining 4 nodes, each with 500 homes passed. The SG decombining could be initially achieved by dividing the SG into 2 SGs, each consisting of 1,000 HHP. The segmentation could continue by separating each of the 4 nodes into a separate SG, consisting of 500 HHP/SG. Beyond that, SG segmentation would include “breaking up” each of the nodes into a smaller group by adding 1 additional node, creating nodes (and SGs) consisting of 250 HHP.

The following a key options for SG segmentation:

I. SG decombining is generally achieved by adding equipment in the cable headend. This re-uses the spectral HFC network capacity in smaller SGs.

II. Node segmentation requires the same additional equipment in the headend, but also requires that additional nodes, and/or fiber be installed in the plant.

And, the following are key factors to consider regarding SG segmentation:

A. SG segmentation usually involves the same decomposition in the upstream (US) and downstream (DS).

B. The relative cost of SG segmentation is higher for node segmentation than for SG decombining. This is because the work requires for the former requires the installation of additional nodes and/or fiber in the network (node splits), which in some cases is substantially more expensive. Conversely, in general SG decomposition is significantly less expensive than node segmentation.

C. However, when additional peak capacity is needed, such as in high-speed data (HSD) services, the SG segmentation is not a viable solution since it does not inherently add peak capacity.

2. Move QAM generation downstream into the network

This option would require including the PHY, part, or all, the MAC, or all of the CCAP functionality into a line-gear device which would be installed in the HFC network. Depending on the functionality being 'remoted' into the HFC network and the desired interoperability, this option would require the creation of specification. Connectivity back to the headend would be achieved via a baseband laser, such as point-to-point Ethernet, as opposed to an analog modulated laser as used now in HFC network.

The advantage of this approach is the migration to a baseband laser, and the operational simplifications that this entails. This approach would also result in additional capacity given the inherent segmentation that would be implemented. And, given the reduction in noise sources (e.g., removal of the analog laser, shortening of the links especially upstream, and reduction in the number of components), it should be possible to achieve higher order modulation rates than are possible to achieve with the PHY located in the headend.

From an operational perspective, however, the proliferation of intelligent devices that would need to be maintained, upgraded, and supported, might result in an increased complexity.

3. Implement PHY and MAC improvements

Clearly, cable systems today are capable of supporting higher order modulations, resulting in greater bit transmission capacity in the same spectrum. For example, it is considered possible to support 1,024 QAM downstream modulation in current cable systems. In fact, it should be possible to support even higher downstream modulations such as 2,048 and perhaps even 4,096 QAM. In addition, it should be possible to support 256 QAM in the upstream, and perhaps even higher order modulation rates. These improvements would come at a cost of higher signal-to-noise requirements, which are believed possible to achieve in today's cable systems.

Additionally, given advances in CPU performance in DOCSIS components, both in the CPE and the CMTS, it should be possible to replace the currently used Reed-Solomon forward error correction (FEC) for Low-density Parity Check (LDPC) FEC. This change is expected to provide an improvement in bitrate equivalent to 2 bits/Hz.

Additionally, it appears that it would be beneficial to migrate to multicarrier modulation techniques, such as orthogonal frequency division multiplexing (OFDM) for the downstream, and orthogonal frequency division multiple access (ODFMA) for the upstream, as opposed to the currently used single-carrier approach.

OFDM and OFDMA offer superior performance and benefits over the older, more traditional single-carrier QAM modulation methods because it is a better fit with today's high-speed data requirements. The use of OFDM, and OFDMA, has become widespread and their implementation well understood in recent years, which was not the case when DOCSIS was initially conceived 15 years ago OFDM when it was extremely difficult to implement with the electronic hardware of the time. These techniques remained a research curiosity until semiconductor and computer technology made it a practical method in recent years, and extensively used for cellular and Wi-Fi transmission. OFDM, and OFDMA, is perhaps the most spectrally efficient method discovered and implemented so far.

4. Increase HFC downstream capacity beyond currently deployed, and/or move split to higher spectrum for increased upstream capacity

From a headend equipment perspective, this option is generally readily available to MSOs. However, CPE equipment would have to be implemented to support the new enhanced upstream and/or downstream spectrum.

From a network perspective, this option involves the change of the diplexers throughout the network such that the frequency division crossover is moved from the 42-50 MHz up to a higher portion of the spectrum, plus the simultaneous expansion of the network capacity to 1 GHz via a retrofit of the active components with minimal changes to the plant spacing and passive components. However, from an operational perspective, this option requires perhaps the most operational change to existing services, such as the removal of analog channels in that portion of the spectrum. That may not be possible for many MSOs that are either required to maintain support for analog TVs directly (e.g., without DTAs), or are unable to remove the analog channel for contractual reasons, or some combination of the above two reasons.

Even if removing the analog channels is possible, this option seems to require the installation of CPE filters in most or perhaps all home CPE devices (e.g., TVs, VCRs, etc.) to both protect that portion of the spectrum from emissions from such home devices and to protect the devices themselves from the levels of transmission of the new CPE that would use that portion of the spectrum for transmission.

And, even if removing the analog channels and deploying the necessary filters were possible, this solution alone provides limited additional US capacity in the network, as follows:

- A move of the split to 65 MHz provides an additional capacity of just 15 MHz, which less than doubles the current capacity. By all accounts, this is a change not worth embarking on.

- A move of the split to 85 MHz almost triples the US capacity, and the simultaneous expansion of the DS network capacity to 1 GHz would add a net 15-30 new DS QAMs (this calculation considers the combined effect of expanding the capacity of the network to 1 GHz from 860 MHz or 750 MHz respectively, and the loss of DS spectrum with the move of the split into the current DS region).

- The shift of the split up to the 200 MHz is also being considered, but while this change would provide much more US capacity, it would reduce the next number of DS capacity significantly and would require the change of large numbers of non-DSG STBs (most of the STBs deployed to date) because the existing and extensively deployed OOB carriers would become inoperable since the region of the spectrum these utilize would be used for the US. Additionally, this change has other plant implications, such as the US equipment currently deployed would not support such extensive US, and thus a new HFC return strategy/equipment would be required.

5. Develop technology that would operate in unused portions of the spectrum, and even unleash spectrum above current top range (e.g., above 1 GHz)

Unlike option #4, this approach involves equipment not currently available. Instead, implementation of this option will require the development of network components and corresponding equipment that would make use of the existing forward spectrum but would use an unused portion of the spectrum, above 750, 860 MHz, or even 1 GHz. This new technique, which we will call High Spectrum Overlay, would require new equipment that could be built in the form of a new network gateway that could be installed in the headend, or in the vicinity of the node, or even deep within the HFC network. This new equipment would provide the 'translation' from the optical transmission generated at the headend into electrical signals, and RF transmission from the location of the converter to the coaxial portion of the HFC network.

This approach would increase US and DS capacity considerably, likely providing multiple Gbps of net additional US and DS capacity. In the process it leaves legacy services and existing CPE untouched.

However, this approach will require considerable equipment development before it would become available for deployment. Such equipment would use spectrum above that being used today for both additional US and DS capacity.

This option could be implemented in three fundamental ways: where the network gateway is located in the headend, or where the network gateway is deployed in the vicinity of the node, or where the network gateway is deployed throughout the HFC network.

In the first case, the RF signals would have to traverse the entire HFC network, including the forward and return analog modulated lasers and receivers, thereby being limited to the spectrum manageable by the analog modulated lasers and receivers.

In the second, the RF signals would traverse the various amplifiers within the coaxial part of the network, but would not require of transmission via the analog modulated lasers and receivers.

And, in the third, the network gateway would be installed in the vicinity of each active component where advanced services are to be provided. This option is known as a Passive High-Spectrum Overlay system. Therefore, this option would require the deployment of additional fiber beyond what's already installed in the network, namely between the existing node and each of the active components in the HFC network. In that way WDM would be used to carry baseband signals up to the node, from which traditional PON technology would be used to interconnect each of the new network gateways back to the HE.

Any modern HFC network should support a Passive High-Spectrum Overlay. Figure 3 depicts an initial deployment of Passive High-Spectrum gateways, for which EPON equipment is deployed in the headend, a separate optical wavelength is used in the trunk fiber to carry the EPON signals up to the node (shown in dashed blue lines), additional fiber is deployed in the distribution portion of the network (shown in solid blue lines), and new Network Gateways that provide optical-to-electrical signal conversion are installed to provide the overlay within an HFC segment between amplifiers.

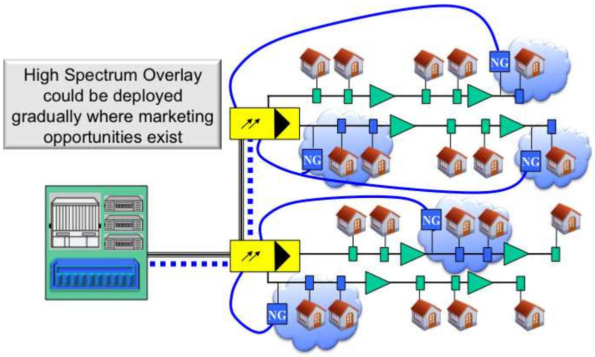

High Spectrum Overlay could be deployed gradually where marketing opportunities exist

Figure 3: Initial High Spectrum Overlay

This approach should not be construed as resulting in a Node + 0 HFC cascade reduction. This is because the cascade of HFC actives is not modified. Instead the RF output of the gateways deployed in the HFC network and operating above 1 GHz are combined with the RF signals existing in the coaxial network which operate below 1 GHz, much in the same way as narrowcasting a set of signals on a per service group basis where the other signals are broadcasted to the set of service groups.

The following categories of work would need to be performed in the plant in order to achieve the above:

- WDM could be used from the headed to the location of the node to reuse the existing long-haul fiber.

- To provide the remaining optical link from the node to the location of each active, additional fiber would be over-lashed to the distribution coaxial hardline cable, which is generally a short to medium length span.

- Finally, in order to pass RF signals above 1 GHz on the distribution network, it is likely that a proportion of the tap faceplates would need to be replaced, although it is expected that the tap housing will likely support these new faceplates, and that only faceplates serving subscribers and upstream from it would need to be replaced.

Assuming a high-bandwidth optical network from the headend to the network gateway, such as 10 Gbps EPON, and a high-order modulation and encoding scheme, it is expected that a transmission achieving 8-10 b/Hz might be possible, therefore resulting in a combined US/DS payload transport capacity of approximately 3-5 Gbps.

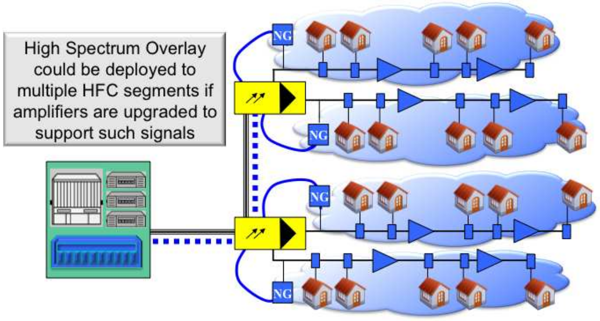

High Spectrum Overlay could be deployed to multiple HFC segments if amplifiers are upgraded to support such signals

Figure 4: Multi-segment High Spectrum Overlay

Figure 4 depicts the case of a deploying Network Gateways at node locations. This option would require less fiber, but would necessitate a rebuilt with amplifiers that would pass the new RF signals.

POSSIBLE IMPLEMENTATIONS

In evaluating the possible approaches outlined above, and taking into account the technologies available to date, it makes sense to consider the following implementations:

enhancement to DOCSIS for residential applications, and development of a new transport alternative to EPON over HFC for commercial applications. Naturally, despite the primary target services, either of the two technologies could be used for either or both services.

DOCSIS Enhancements

Given the success and widespread use of DOCSIS-based services to date, and the advent of the technical advances outlined above, it seems plausible to consider the following enhancements to DOCSIS to enable additional capacity and a more efficient use of HFC spectrum:

- Use of higher-order (1/2/4K QAM) and modulation techniques (OFDM/OFDMA) to improve throughput and simultaneously reduce spectrum utilization by as much as 50%,
- Replace the current Reed-Solomon FEC technique with a more modern Low Density Parity Check (LDPC) FEC, which would improve overall efficiency by as much as 25%,
- Enable use of additional spectrum for the US, beyond the current 5-42 MHz, up to 100 MHz or even higher spectrum, to increase US transmissions by a 3x to 5x factor, and
- As capacity is enhanced, consider simplifications of the DOCSIS protocol that may reduce implementation complexity, accelerate the availability of newer implementations, and reduce costs.

Implementation of the above new functionality will have to be done taking into account backward and forward compatibility to maximize the benefit for current and new equipment.

New HFC transport for EPON

Similarly to the enhancements now available for DOCSIS, it seems possible to implement a new transport for EPON over coax. Envisioned in the past as a component of Comcast’s Next Generation Access Architecture[4], a new transport for EPON Protocol over Coax (EPoC) is now under development at IEEE. This new transport will make it possible to provide EPON services to end-devices attached via cable operator’s coax network rather than only via fiber cable.

The work currently underway, known as an IEEE 802.3 Study Group, is intended to demonstrate the feasibility of implementing a coax transport for EPON using technologies and approaches similar to those that would be applicable to DOCSIS. Once completed the work of the Study Group, a Task Force would be formed to define the new PHY for EPON over coax.

This work would lead to the availability of a coaxial-attached alterative to EPON, which would enable MSOs to deploy EPON services to customers already served by its HFC network. This should result in a more economical and operationally simpler way to provide Metro Ethernet (MetroE) services to business customers without having to deploy fiber to each potential customer premise.

OVERALL ACCESS ARCHITECTURE

The new edge platform devices currently under development by vendors, as specified by the CCAP architecture, will support either of the approaches described above. The

[4] What is CMAP? Jorge Salinger and John Leddy, CED Magazine, February 2010

CCAP architecture already supports the modularity necessary to upgrade line cards progressively as new technologies become available.

For the case of the enhancements to DOCSIS outlined above, it seems reasonable to expect that the current downstream line cards could be updated via field programmable gate array (FPGA) programming changes, such as Hardware Descriptor Language (HDL) or Register Transfer Level (RTL) programmable changes. In the case of the upstream, it is expected that new line cards could be developed that would take advantage of the new technologies.

Furthermore, the CCAP architecture provides support for EPON, such that even EPoC is supported in the overall access architecture.

SILICON DEVELOPMENT

One important consideration in evaluating the benefits of each approach is the need and availability of silicon components, or on the flip side the need for its development.

This is critical for the following fundamental reasons:

a. When silicon exists the availability of the system solution is quicker, whereas when it needs to be developed the timeline is significantly longer, and

b. If silicon devices, or at least some of their components, are used for multiple purposes, especially for multiple industries, then their production increase rapidly and costs decrease considerably.

Some of the new technology enhancements will likely require silicon development, but others would not, for which technology design decisions would be important.

CONCLUSIONS

Additional HFC network capacity will be required for narrowcast services for both residential and commercial service applications in years to come. The expected growth appears to be quite large.

New technologies are now becoming available that would make it possible to achieve higher throughout and more efficient use of spectrum. This includes higher-order and more modern modulation techniques, more sophisticated forward error correction, and the use of more spectrum than currently utilized.

This paper presented an analysis of these technology options and their corresponding pros and cons, and outlined how these technologies could be used to enhance the current transport options available to MSOs, such as DOCSIS, and to create new infrastructure options, such as EPoC.

ACKNOWLEDGEMENTS

Thanks to my colleagues Sam Chernak, Doug Jones, Saif Rahman, John Bevilacqua, Chris Bastian, and Dave Urban from Comcast for their contributions to the discussions leading to the proposals and conclusions presented in this paper.

Also, many thanks to my colleagues from CableLabs®, the many vendors and advisors of Comcast, and especially fellow MSOs, for the many discussions leading to the materials presented.

HTML5 Framework and Gateway Caching Scheme for Cloud Based UIs

Mike McMahon, VP of Web Experience and Application Strategy
Charter Communications

Abstract

Recent advances in our industry such as TV Everywhere and second screen, companion apps merge video delivery and consumption with web technologies. Similarly, much progress has been made in introducing service-oriented architectures, exposing common web services and enabling a high degree of consistency and re-usability in backend systems.

Video is now clearly being consumed on a wide range of devices and these devices can vary wildly in terms of screen size, capabilities, development platforms, etc. Tablets, game consoles, smart TVs, mobile phones and a number of other devices are all viable video terminals. Processing and delivering video into a variety of flavors, bit rates and such is non-trivial but is generally well understood and now fairly commonplace. In order for the Cable industry to fully embrace an already highly fragmented client platform landscape and position itself to exploit new devices as they become available it is necessary to achieve a similar level of abstraction and re-use in the way user interfaces are built, delivered and maintained. This paper presents an HTML5 based UI framework, built on open standards but optimized and configured specifically for the needs of TV centric applications.

THE HTML5 OPPORTUNITY

Although HTML5 remains a maturing technology, the web development community has actively embraced it and most modern web browsers already support it. In our industry, there has been some speculation surrounding the video tag and the current lack of DRM. The premise here is not “HTML5 video,” rather the use of cutting edge web techniques associated with establishing a user interface, built from a singular and re-usable code base which is common and consistent across a wide range of devices. For the Cable Industry, moving the user interface code into the Cloud in this way not only represents an opportunity to address a variety of devices, but additionally empowers us to embrace retail devices as well as add features and extend functionality at web-like velocity, removing the burden of code downloads and complex provisioning scenarios.

Open Source Frameworks

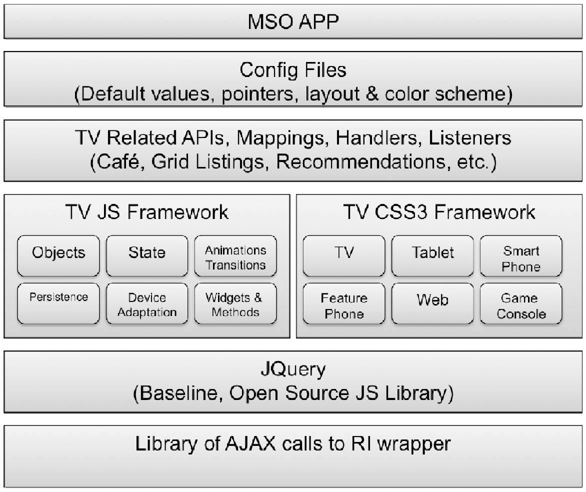
There are countless examples of extremely powerful applications, written entirely as web applications that are as rich in functionality, animation effects and behavior as desktop applications. In practice, these are written as a combination of HTML5 along with an aggressive use of JavaScript and CSS3. It is important to recognize that it is this collection of technologies, rather than HTML5 itself that enable these types of user interfaces. A variety of JavaScript and CSS3 frameworks such as jQuery or Sencha exist for HTML5 development. These typically abstract away platform idiosyncrasies, establish object and state models, provide a variety of animation libraries and generally simplify the development of a single web application to run across a variety of devices.

THE GAP

Without doubt, individual providers will seek to differentiate their brand through unique designs, features and interactivity. While each individual service provider could certainly select a given framework and develop its own, unique cross platform user interface there would be little commonality across the industry and much duplication of effort. We will all have linear listings and VOD search. We will all have cover art and DVR scheduling. Grids and a baseline set of animations are inevitable. We will share the need to support the same range of devices. Furthermore, each provider would be burdened with updating to versions of the framework and addressing new devices, screen resolutions, etc.

AN INDUSTRY FRAMEWORK

Envisioned here is a Cable Industry UI Framework. The ambition is to select among the various open source HTML5 frameworks the core aspects most beneficial to the generic needs of TV centric user interfaces. There would likely be several components involved, the particular assembly of which would constitute an MVC type construct with particular focus on the device and object abstractions required to represent "TV." This baseline component assembly would constitute the core foundation but would require an additional layer of CSS3 and JavaScript abstraction for the Cable specific UI components and underlying object model. The high level stack is represented below:


MSO APP
Config Files
(Default values, pointers, layout & color scheme)
TV Related APIs, Mappings, Handlers, Listeners
(Café, Grid Listings, Recommendations, etc.)
TV JS Framework
Objects
State
Animations Transitions
Persistence
Device Adaptation
Widgets & Methods
TV CSS3 Framework
TV
Tablet
Smart Phone
Feature Phone
Web
Game Console
JQuery
(Baseline, Open Source JS Library)
Library of AJAX calls to RI wrapper

The overriding purpose of this stack is to leverage the generalized foundations of an underlying open source framework such as jQuery and build on top of it the necessary specifics relating the Cable industry. These specifics would include such things as objects for TV listings, recommendations, actors, movies as well as standardized methods and callback routines for fetching recommendations, content searches, etc. Likewise, a variety of UI components representing things like an actual TV listings grid and animation effects such as a cover art carousel would be optimized. CSS3 style sheets for each device family or particular model would cater to the specifics needed in each rendered component. Each layer of this stack is intended to be extensible.

MSO Customization and Extensibility

Each MSO would benefit for the shared plumbing in the underlying framework. Configuration files, unique to each MSO would map to their web service endpoints, define the specific assembly of the various UI components into their presentation and provide a skinning capability via CSS3 overrides.

As new objects, event handlers or animations were envisioned and required; an MSO or third party would develop them within the overall framework. Ideally, these would be contributed back into the community such that other MSOs would benefit. There would likely, for example, be several variations of a TV listings component to choose from as well as useful extensions by way of animation effects.

Inclusion in App Stores

There is often confusion between "Apps" and "HTML5." The two are indeed different things as "Apps" are compiled, installable binaries and "HTML5" represents web pages. App stores and HTML5 are, however, perfectly compatible. There is, of course, very good reason to place applications in app stores. Most users of iOS and Android devices in particular are now familiar with app stores and this is the dominant avenue by which they are likely to search for and discover an MSO application. Applications available in these stores are compiled natively for the specific platform. In order to achieve the benefits of app store inclusion as well as the ability to re-purpose HTML5 across platforms a "native wrapper application" is written which essentially compiles a rudimentary shell for the specific platform and uses the device's underlying web browser to render all actual user interface screens. This is, for example, the way in which Netflix develops its applications.

ADDRESSING THE BIG SCREEN

With regards to the common retail devices of today such as iOS and Android powered smartphones and tablets, laptops and PCs this framework would provide a robust mechanism to deliver a common and consistent user interface as well as minimize the associated code. The UI is, effectively, a giant web site delivered from the Cloud. Changes made to a single file would propagate to all devices and users would not need to download any updates, they would simply benefit from the new experience during their next session. This is all well and good, but to what extent could the framework be used to deliver the same experience to a STB connected to a 60-inch plasma?

Relevance to the RDK

The RDK recently introduced by Comcast includes a Webkit implementation. Webkit is an open source HTML5 compliant web browser, used in both Apple's Safari and Google's Chrome browsers. This provides an alternative to Java as the presentation environment on the CableLabs <tru2way> reference implementation.

This stack can be used in a master-slave in-home architecture whereby a gateway device running the full RDK stack serves as the service termination point within the home, commanding control of tuners, handling conditional access, etc. Additional devices such as laptops, tablets and smartphones can connect to the gateway and both consume tuners and receive the user interface, which is delivered as HTML5 via a web server running inside the gateway. The gateway can be "headed" meaning a television display is actually connected to it or "headless" meaning it serves as the termination point but exclusively provides the UI and services to other devices within the home. Additional outlets need only be very dumb, thin IP STBs, which run Webkit.

To be sure, this description of the RDK does not do it full justice as it is, truly, a very compelling development and much more significant than the brief description above. The point here, however, is that there is an HTML5 presentation layer available and that it can render to the big screen.

Thus, a modern HTML5 compliant web browser is available through the RDK. The RDK, however, does not provide any specific UI or further framework, just that open book upon which things could be written. As is the case with other HTML5 devices, each MSO would need to develop and maintain its own specific UI.

The proposal is to extend the RDK to include the same UI framework discussed in this paper.

GATEWAY CACHING SCHEME

The UI framework would necessarily be hosted in the Cloud. This would allow it to be used independent of RDK gateway architectures as well as ensure that the UI could be rendered outside of the home over any network. By including the framework within the RDK, however, there are additional benefits relating to caching and performance to explore.

Insofar as an RDK based gateway acts as a web server (both to itself and other devices within the home) it is, in effect, a proxy to the actual remotely hosted Cloud. Like all proxies, it acts as the source of truth from the perspective of the client. This presents potential challenges by way of ensuring the gateway is, in fact, up to date but also represents a significant opportunity to be used as a caching node within a distributed architecture. Open source caches such as Varnish could be additionally included within the RDK and would provide a tremendous performance benefit. Specifically, the gateway could be configured to proactively cache guide and VOD listings, cover art, network images, user profiles and targeted ads. This could be done relatively easily via a lazy cache whereby the gateway deferred to the Cloud and simply stored content and data as it passed it through to the client, making it locally available for subsequent requests. It could also take on a more elaborate form whereby server side algorithms proactively pushed information to the gateway, likely during dark hours and with certain targeting parameters designed towards personalization of the UI.

ADDITIONAL CONSIDERATIONS

While I believe the industry would benefit significantly from a common, shared core UI framework it still assumes HTML5 and relatively modern web browsers. What about older PCs or even fairly modern devices with limited rendering capabilities like Smart TVs or game consoles?

HTML5 does not render everywhere. Older browsers like those in many PCs or early incarnations of Smart TVs are capable of rendering simpler versions of HTML. It is necessary that the framework can degrade gracefully by recognizing these devices and rendering a simpler, less animated form of the UI. This would require a somewhat more complex abstraction than would otherwise be necessary but is perfectly feasible. Other devices, like a current Xbox, will require platform specific applications. These devices will benefit from at least a general commonality of the UI in terms of data and objects as served from backend web services, but would still require platform specific, native applications to be written. HTML5 will not currently provide a UI on every device; although it will address a wide range of current devices and it is likely Webkit will continue to proliferate to things like Smart TVs and game consoles.

The Vulgarity

More challenging to the notion of a common UI framework is the fact that there is currently very little consistency in the backend of MSOs. While most of us are embracing web services and these web services are notionally representing very similar things they are far from standardized. The grid for example, is assumed to be in most operators' UI in some form or another. The data used to populate the grid would be fetched through a web service along the lines of:

```
http://operator.com/apis/getGrid();
```

The host and specific method call, of course, would be perfectly configurable and each MSO would have their own unique endpoint. This is not a problem. The syntax and structure of the response, however, is a challenge. There are differences in semantic naming conventions as well as overall object models. The semantics of one MSO labeling HBO a “network” and another “provider” or “programmer” are somewhat easier to deal with. Structural differences in the objects or varied sets of interfaces are far more troublesome. One MSO might include actors and detailed descriptions in the getGrid() response. Another might have a secondary web service for getAssetDetails() and a third that does not include actors at all.

These backend variations are not insurmountable but they do require some additional thought. Standardization of core web services is, of course, the ideal solution. It is also possible to establish a JavaScript mapping layer within the framework, although that will likely lead to poor performance. A possible middle ground scenario would involve each MSO establishing a server side transformation layer to its existing web services.

CONCLUSION

As MSOs, we face a similar challenge in providing consistent user interfaces to a growing set of devices. Cloud based UIs allow us to more uniformly deliver and present a user interface as well as extend new features and services in a coherent and efficient manner unlike with traditional STBs. This is something we can immediately explore online and through smartphones and tablet devices and will increasingly become viable on leased CPE through initiatives like the RDK. A common, shared industry UI framework would allow us to further exploit the opportunity and reduce the individual burden of redundant web development. Such a framework could be developed based on best in breed, open source efforts from the web community but configured specifically to suit the needs of TV centric user interfaces.

Intelligent Caching In An ABR Multi-Format CDN World

Patrick Wright-Riley, Brian Tarbox
Motorola Mobility, Inc.

Abstract

In their infancy, content libraries contained a few thousand pieces of content and most vendors put a copy of everything everywhere. As the contents grew to tens of thousands of titles, Central Libraries were added and Least Recently Used (LRU), then intelligent caching, was employed. As content libraries have grown by orders of magnitude and now adding Adaptive Bit Rate / multi-format copies to the mix, some suggest intelligent caching is no longer possible. Motorola asserts that intelligent caching is both possible and even more critical today than ever. Intelligent caching still plays a valuable role in the ABR Multi-Format world.

INTRODUCTION

In the last ten years the industry has experienced at least three distinct generations of thinking on the approach to placement and duplication of content. We define the first generation as a time when content libraries were small enough that each Video On Demand (VOD) system maintained its own copy of each piece of content. These libraries were stored on spinning media and were served either directly via disc arrays or DRAM. These libraries tended to contain a few thousand titles of standard definition content. Caching in these systems was something that happened in the disk driver or the VOD server's memory backplane.

Generation two can be characterized by the slow introduction of high definition content and libraries of tens of thousands titles. This increase drove the capital expenditure equation high enough to discourage the placement of all content at every site. Thus, the Central Library approach containing the "Gold Copy" along with smaller edge libraries that maintained copies of the commonly viewed content being watched by subscribers within their domain was introduced. Many VOD systems were constructed with the 80/20 rule where it was assumed that 80% of the subscribers viewed the same 20% of content. Given this assumption, distributed edge libraries used a simple Least Recently Used (LRU) caching algorithm to determine which 20% of the content from the edge library was essential to maintain. As it turned out, this content distribution model did not produce the content storage and reduction in network congestion operators expected. This dilemma led to the development of an alternative approach called intelligent caching. Intelligent caching (IC) incorporated additional information about content viewing behaviors beyond what LRU could provide. From there IC became the norm for caching at the edge. However not so far down the road, the explosion in SD and more HD content storage requirements combined with a growing number of smart devices and tablets, Adaptive Bit Rate/Multi-Rate was destined to become part of the picture.

In the third generation, content libraries jumped again to hundreds of thousands of titles, with HD now dominating the content scene. Today this content is now chunked and replicated into several bit rates and wrapped in several formats. Thus, hundreds of thousands of titles can easily become millions or billions of file chunks linked by manifest files.

Conventional wisdom suggested that reaching these levels of processing and file management rendered intelligent caching obsolete. It's suggested LRU caching within the content delivery network (CDN) is both the best that can be done and is enough. Given that intelligent caching increased the efficiency of edge content retention such that 98% of the content was properly retained, it seems reasonable to explore if those benefits can be retained in an ABR, multi-bit rate environment.

PROBLEM DEFINITION

What is Caching and What Drives It?

Caching is a predictive activity. When caching, the system uses data about past viewing behaviors to make assumptions about future viewing behavior associated with a given channel. The assumption typically results in the allocation of a scarce resource before it is actually needed. The "hit rate" of the cache is the percent of time that the assumption turns out to be correct. The impact of the hit rate is based on the comparative cost of permanently reserving the resource against the cost of allocating the resource on-the-fly. To achieve this second, more efficient method, caching is most powerful, especially given that in terms of network usage and related congestion, when the cost of real-time allocation is high.

In order to quantify the value of caching we have to look at the differential cost of resource allocation. Disks are substantially slower than memory and networks are substantially slower than disks. Some VOD vendors made a business out of this differential by attempting to build systems where the entire active portion of the content library lived in memory, or DRAM. This was a successful strategy until the growth of the library outpaced the growth in memory chip size. The battle then moved from DRAM vs. Disk to Disk vs. Library where the relative cost of late allocation was even higher.

Basic Caching

Caching algorithms are characterized by the predictive algorithms employed within the CDN. Caches are assumed to be filled with content at all times. Thus, the critical decision is actually which content item to remove from the edge storage. The most basic type of algorithm supporting this capability is the Least Recently Used or LRU. This algorithm maintains a usage timestamp for entries in the cache and when content removal is required will eject the item with the oldest usage time and replace it with new content. Such systems update the usage timestamp whenever there is a "hit" on the item. These algorithms can be compared to the psychological principal of "Win-Stay, Lose-Shift" where a successful outcome will cause a subject to make the same choice again and an unsuccessful outcome will cause the subject to make a different decision.

This leads to the need to understand the content viewing behavior attempting to be predicted. In the case of content viewership there are two approaches: attempting to predict the future behavior of a given viewer or the future behavior of a group of viewers. By tracking content watched by a particular viewer, inferences can be drawn regarding the potential viewing of that content by the set of all other viewers. However, when using an LRU algorithm to do so, the system simplifies the analysis to a single parameter—time last used—and may not be fully representative of the likelihood of future viewing by a group. Thus, although LRU has some value, it is greatly limited when compared to more intelligent, multi-parameter caching algorithms.

Comparison of LRU with Garbage Collection

Java is the world's most popular computer language and its performance is largely dictated by the behavior of its memory reclamation or garbage collection system. The Java computer language's Garbage Collection (GC) system is one of the world's most studied caching systems. Valuable insights may be gleaned by comparing GC with various other caching algorithms.

One of the primary drawbacks of a simple LRU approach is that it understates or ignores the effect of what GC calls infant mortality of reference. Many objects have a usage model of initial creation followed by limited use, ending with no further activity. In a computer program a variable might be declared, used in a single computation and then discarded. Similarly, in a television viewing experience a user might tune to a channel, watch for a few seconds and then move on. In this scenario the content would have a very high LRU score. In essence, the naïve algorithm employed by LRU would preserve the item in cache in spite of its low actual usage. This confirms the low predictive strength of LRU. This is important when we consider that most CDN "intelligent caching" systems are based on LRU approaches.

To achieve greater predictive power, an algorithm must incorporate a more sophisticated object lifecycle model; an object being a piece of content or a chunk of content carrying specific bite rate and format characteristics. Such a lifecycle model is typically generational. The GC partitions the cache into three generations: 1) Eden, 2) Tenured, and 3) Permanent. When objects are first created they live in Eden. The system periodically scans the memory list looking for items to eject. Items in Eden that are not ejected after two passes, meaning they still have active references to them, are promoted to Tenured. Items living in Tenure that survive more passes are promoted to Permanent and thus remain much longer in cache. There are actually two types of GC passes: full and partial. Partial collections are run quite frequently, have relatively little impact on system throughput and do not examine the Permanent cache. Full collections on the other hand are comparatively rare, can often affect system throughput and do look at the Permanent cache. So, content that exists in the Permanent cache are only occasionally examined for ejection.

Segmented LRU cache uses a similar (though limited) system. There are two LRU lists. Items initially live on the first list and after a second "hit" get promoted to the second list. While this is certainly better than a simple LRU, there is still a world of difference between noticing a second hit and true intelligent caching.

Advantage of Intelligent Caching

Intelligent Caching is a term we reserve for systems incorporating a more sophisticated object usage model. Such a model must acknowledge the realities of content viewing such as channel surfing, free content preview, time of day and day of week viewership patterns and other patterns of apparent viewership that may or may not represent true viewing of content.

The bottom line is that content hits, initial or passive, are not predictive or representative of actual viewership until the aggregate viewing time has exceeded a certain quantum of time. Once the aggregated viewing time of the content has passed a threshold (which may be dynamic and involve multiple analytic parameters) then statistical inferences may be made about the future likelihood of additional views. This is the basis for intelligent caching algorithms and where their value lies above LRU algorithms.

Factors That MAY Diminish Predictability

If in a multi-bit rate, multi-format world where content is delivered over a CDN, one could argue that caching, in its entirety, is unnecessary. There are factors that are commonly cited as evidence that caching is no longer possible or valuable. These may or may not eliminate the usefulness of all caching algorithms, but they certainly provide a challenge to the usefulness of some caching techniques. It is helpful to remember that the caching algorithm attempts to extrapolate from the past exposure or access of content the future possibility of that content being viewed again, possibly by another viewer. The problem in understanding and valuing cache is that "the same content" may now exist in multiple copies, in different formats and bit rates, with chunks spread across multiple edge streaming servers. (See section Content Affinity for more details about chunk distribution).

Multiple Formats

As we enter into the second half of 2012 the video format battle is raging on. Apple's HLS format appears to be dominant, but the Microsoft and Adobe formats are still contenders. Although it is unclear what position these latter companies are taking with respect to future support, they cannot be discounted. At the same time the DASH specification is evolving and may, over time, acquire significant share. While many hope to support fewer than four formats, that time is not yet here (and may never arrive).

There are at least two ways to address the question of how multiple formats effect cache predictability.

One way is to ask the question, "Does the fact that a piece of content was viewed in a particular format "enough" provide any evidence that it will be viewed again in the future…in that same format and/or in other formats?" Since the multi-format ABR world is so new it's hard to anticipate future usage patterns. To the extent that we can extrapolate from existing usage patterns it seems safe to assert that content reaching a threshold of use is in general more likely to receive future plays than content that has not reached the threshold. It is also reasonable, though untested, that reaching the threshold on a particular piece of content in one format is at least weak evidence for the future popularity of that content under a different format. To state the opposite one would have to assert that popularity in one format provided zero evidence of possible future popularity under another format which is unreasonable.

A second approach to this problem is to think about multistage packaging and common formats. As has been discussed in other papers, there is an ongoing debate of the merits of packaging in various locations. Some lobby the benefits of Central Packaging. Others point out the potential benefits of customization from Edge Packaging. An interesting hybrid approach is to perform an initial round of chunking and manifest creation in the center, followed by a real-time component that transwraps content and performs unique, targeted manifest generation. From a caching point of view this approach defers the combinatorics of multiple formats until well downstream of the CDN. A cache element located "upstream" of this real-time transwrapper might see just a single format, thus diminishing its value.

Multiple Bit Rates

The key to adaptive bit rate streaming is the availability of multiple representations of each piece of content. This can be seen from at least two points of view. On one hand the same content might well be viewed at a different bit rate on a phone, a tablet and a big-screen LCD based simply on the capability of the various devices. In this slice of the world each stream might have a different bit rate, but does not necessarily change its bit rate during the presentation. In the other slice, each client responds to the ever-changing load on the network by asking for smaller content when the network is slow and larger content when the network is fast. This is the grand assumption behind most ABR streaming.

The problem is that it invokes the dilemma of the commons: when there is a shared and limited resource, the greater good is often different from the individual good. When the network is congested, every viewer will fully support the idea that everyone else should limit their bandwidth such that "I" can continue streaming the highest quality experience. And everyone else feels the same way. This can be controlled if the client software is controlled by the infrastructure providers in that their client software can enforce the self-limiting behavior. On the other hand, does anyone doubt that clients will be made available that attempt to game the system to consume more than "their fair share" when the network is congested? We assert that it remains to be seen just how many distinct bit rates are actually active for a given content. So, while at first blush ABR might multiply the number of different copies of each piece of content by a factor of six to ten per format, the actual number may be significantly lower, perhaps three or four per format.

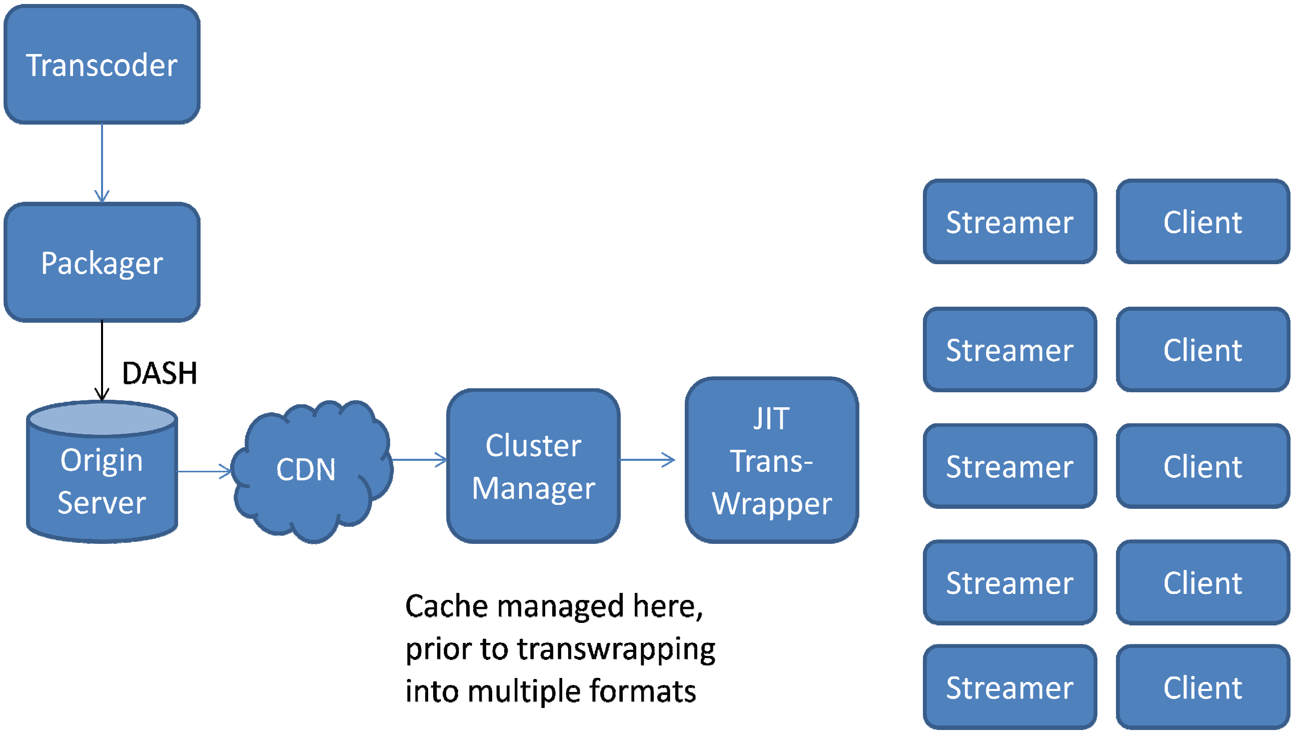


Figure 1: Possible Caching Prior to Transwrapping

NDVR – Unique Copy

Unique copy basically eliminates the ability to do caching at all. For those unfamiliar with the concept, a legal ruling has declared that if some number of viewers record the same content, the NDVR system must store a unique and distinct copy of that content for each of those viewers. In the systems, operators are explicitly forbidden to store a single copy and manage viewer access to that copy. So the only opportunity for re-use of segments or manifests would be if an individual user watched a recorded show multiple times—probably not sufficient to take advantage of caching.

While this might be seen as ending any discussion of caching, keep in mind that Unique Copy presents problems for many aspects of the system. It is anticipated that some vendors may push the envelope of mixed common copy / unique copy systems, especially outside North America. In this scenario, caching may have a larger role to play.

Personalization

Personalization is the process of converting a general video stream into one tailored for a particular viewer or group of viewers. Two main categories exist here; ad insertion and blackout (both are discussed in more detail in the paper "*Complexity Considerations for Centralized Packaging vs. Remote Packaging*" being presented at this conference.) In each case a stream that logically could be used to satisfy many stream requests is turned into one that is usable for a subset of those requests. To the degree that this personalization happens upstream of the caching system it will naturally render the caching system useless.

Factors That May Enhance Predictability

While many types of systems suffer from added scale, caching algorithms actually tend to work better in larger environment, if simply because there is more data to use for decision making and there is more content to provide a better opportunity to employ caching to enhance performance. There will undoubtedly be many different sized deployments of video systems, now and in the future. CDN-enabled, multi-format, multi-bit rate systems will be overwhelmingly biased towards the larger of these deployments; the cost of the complexity associated with such CDN systems precludes them from the smaller tier two and tier three deployments.

This then leads to the next important question which is, "Where does the caching engine live in the CDN architecture?" If it lives on the edge server, then it is limited to the total number of streams supported by that server. Many edge servers are relatively small devices supporting only a few thousand streams. The chances of getting meaningful hit rates in such a small environment are correspondingly low. On the other hand, if the caching engine lives near the edge, but in the CDN it might well be able to see dozens or hundreds of the edge servers. This scale changes everything. The chances of getting several play requests for a given content out of several hundred thousand streams is quite reasonable.

The Role of Content Affinity in CDN Caching

Most diagrams of ABR streaming show the client talking directly to an edge Packager or the CDN; the role of any edge server is not discussed. Motorola believes this is a mistake and causes large opportunities for caching via the use of Content Affinity to possibly be overlooked. If the diagrams do show an edge streamer they tend to show only a single one. In almost all cases any reasonably sized deployment will involve dozens or hundreds of edge streamers since each such device typically only supports a few thousand streams at a time.

Technical papers that have included a multiplicity of edge streamers have tended to view them as interchangeable, even on the per stream basis. It has been asserted that each chunk request from a client might be serviced from a different edge streamer, assuming that every edge streamer has the same chunks. This is then described as a resilient stateless design that can trivially survive the loss of one or more edge streamers. Some of that is true, but at a cost. The cost is that by making server selection stateless we remove the possibility of using knowledge from previous states to improve our caching.

Content Affinity is the process whereby all streams for the same content are directed to the same edge streamer. This can result in enormous savings in both disk space and network bandwidth utilization. If all streams for Spiderman, as an example, go to the same streamer, there is a far greater opportunity for fragment re-use than if the streams for Spiderman are distributed randomly to several dozen streamers.

If we accept the gains that can be realized from Content Affinity then we must look to see which deployment models give the best chance of using Affinity to our advantage. Figure 2 shows one such configuration.

The client makes its initial request to a Cluster Manager (CM) which is a control plane application that maintains the knowledge of which edge streamer has which content. The CM selects a streamer and issues an HTTP redirect message to that device. The client re-issues the request to the streamer which either services it directly if possible or defers to the Edge Packager to create the manifest, if necessary.

Note that Content Affinity is a separate concept from caching and the CM contains no storage of manifests or content chunks. The CM simply directs streaming requests in such a way as to increase the likelihood that the target Edge Streamer will already contain the required chunks for a stream.

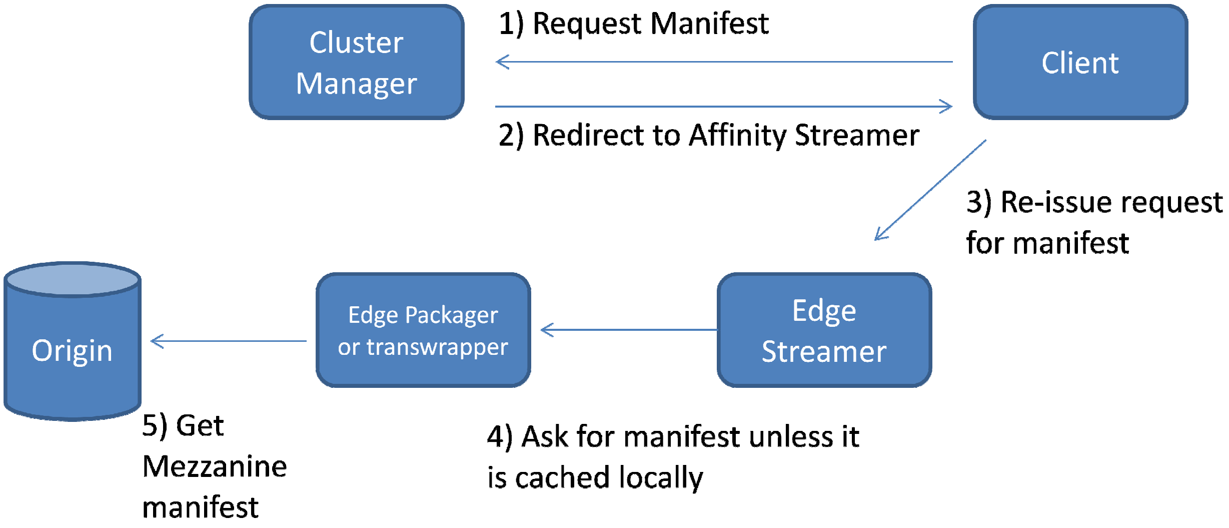

Cluster Manager
1) Request Manifest
Client
2) Redirect to Affinity Streamer
3) Re-issue request for manifest
Origin
Edge Packager or transwrapper
Edge Streamer
5) Get Mezzanine manifest
4) Ask for manifest unless it is cached locally

Figure 2: Content Affinity Deployment

Comparing Intelligent Caching with LRU Caching in a CDN

LRU-based caching in a CDN uses no intelligence about the content, its placement, or its usage. The algorithm simply notices which chunks were the least recently used and discards them when it determines that it needs to create space for new chunks. There is a single ordered list of the chunks logically maintained at the edge of the CDN. This single list covers all chunks sent to all edge streamers. It also makes no use of the fact that chunks may actually be related, i.e., being part of larger piece of content. This can be a benefit as well as a drawback.

If a viewer is channel surfing and briefly visits 20 different channels for 5 seconds each, then the system will likely generate the highest time-last-used values for those chunks and so they will remain stored in cache over other content that should be kept instead. A more intelligent system would never have promoted those chunks as they are clearly of transitory usage. On the positive side, since the system views each chunk individually it would not use those minor play times to promote later chunks from the same pieces of content.

An intelligent caching system would tend to treat such channel surfing as below the threshold for promotion within the cache and would thus not eject other, more popular content.

To put it another way, consider the case where three clients sampled a piece of content but ultimately were watching a different content and a fourth client sampled and then watched the content the others sampled. Putting aside the bit rate and format questions for a moment, we should objectively conclude that the program being watched by three viewers was more popular than the other content and should bias any limited resources such as caching towards the more popular program. The CDN/LRU-based cache cannot do that as it uses a strictly time-last-used algorithm rather than a hit counting-based algorithm. The Content Affinity-based system, on the other hand, allows for the direction of the common content to a common edge streamer and the one-off content to a different edge streamer. This automatically increases the locality of usage of each piece of content to a given pump and thus increases the hit rate of the particular pump's cache.

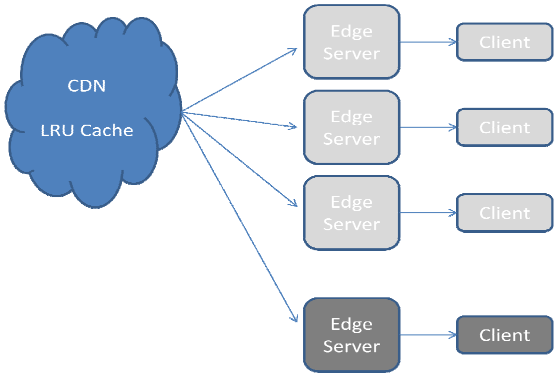


Figure 3: LRU-based Caching of Popular/UnPopular Content

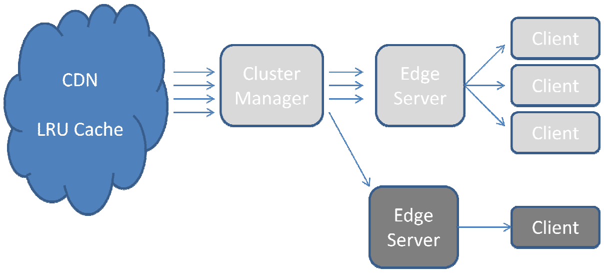


Figure 4: Affinity-based Caching of Popular/UnPopular Content

CONCLUSION

Historically, Intelligent Caching has been shown to provide significant reductions in the need for potentially expensive content storage. This benefit should not be discounted lightly. We have described several of the challenges facing intelligent caching in a multi-format ABR streaming environment. Some of these challenges such as the legal requirement for unique copy NDVR may prove insurmountable. We have, however, shown several opportunities that may allow the use of intelligent caching in other domains to have significant benefits over LRU caching. In particular, we have shown that the affects of Content Affinity can be profoundly and positively affected by efficient, intelligent caching algorithms.

Just-In-Time Packaging vs. CDN Storage

Yuval Fisher
RGB Networks

Abstract

Operators delivering video-on-demand (VoD) to multiple devices using HTTP streaming must select between two options: store assets in multiple formats to be delivered via a content delivery network (CDN), or utilize an on-the-fly, or just-in-time (JIT), packaging to convert VoD assets into the required client format when it's requested. This paper discusses the benefits of JIT packaging and then proposes a model to evaluate the costs associated with each approach, discussing the parameters associated with various use cases. We also discuss the implications of the cost model for more general edge processing, such as just in time transcoding.

INTRODUCTION

HTTP streaming of video based on protocols defined by Apple, Microsoft and Adobe (see [HLS], [MSS], and [HDS]) has led to the development of a new component in the video delivery chain – the packager (sometimes also called a segmenter or fragmentor). This component creates the segmented video files that are delivered over HTTP to clients that then stitch the segments together to form a contiguous video stream. The packager may be integrated into the encoder/transcoder that creates the digital encoding of the video, but often it is a separate component. Separating the components has various advantages, including the ability to capture the output of the encoder/transcoder as a mezzanine format that can be reused for packaging in both live and off-line scenarios.

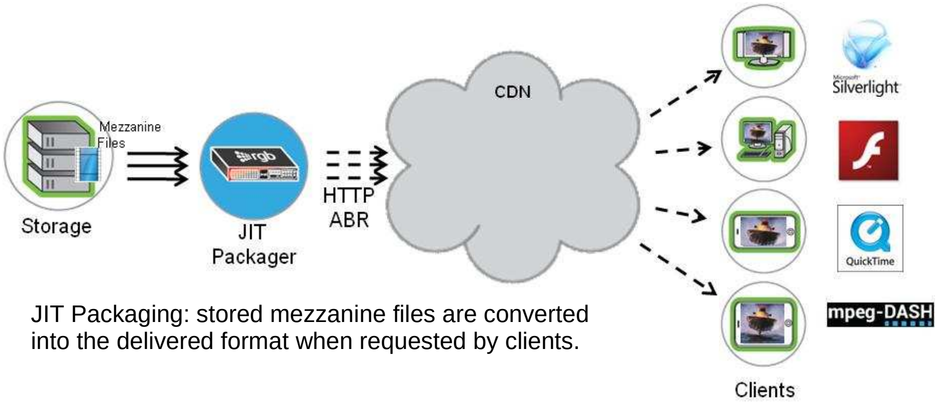
The emerging MPEG DASH (see [DASH]) standard attempts to standardize and unify these protocols under one open specification umbrella; but in the near term, DASH adds more formats that service providers may need to address, since HLS, MSS, and HDS will not disappear immediately, if ever. In fact, DASH has several profiles that have very different underlying delivery formats, so that it may be necessary for packagers to serve not just HLS, MSS and HDS, but an MPEG-2 TS DASH profile and a base media file format DASH profile as well.

In this paper, we focus on one specific use case: just-in-time packaging (JITP), which is applicable for VoD and network digital video recorder (nDVR) applications, including catch-up and restart TV. In all of these applications, each client makes a separate request to view video content (typically from its beginning), so that unlike broadcast video, viewing sessions are independent.

When delivering HTTP streams, two options are possible: either the assets are stored in an HTTP-ready format, so that clients can make HTTP requests for video segments directly from a plain HTTP server. Or, assets can be stored in a canonical (or mezzanine) format which is then converted to HTTP segments as the client makes requests for them – just-in-time. The first option is more disk storage intensive, while the second is more computationally intensive.

Just-in-Time Packaging

In a typical JITP use case, VoD assets are created from live content that is first transcoded into MBR outputs and captured by a "catcher" component that converts the live streams into files stored in a chosen mezzanine format. Alternatively, file assets, rather than live streams, are transcoded into a mezzanine format which uses H.264/AAC for the video/audio codecs and a pre-selected container format. MPEG-2 TS container



JIT Packaging: stored mezzanine files are converted into the delivered format when requested by clients.

format is a natural choice for the mezzanine files, since it can contain much of the signaling present in the original signals in an industry-standard way.

Clients that request a stream from the JIT packager first receive a client-manifest describing the available profiles (bitrates, resolutions, etc). The JIT packager will create the manifest when it is requested the first time; subsequent requests are served from a cached copy. Clients subsequently request specific chunks from the packager which extracts the requested chunks from the mezzanine files and delivers them to the clients. Thus, each client request is served from the JIT packager – the more subscribers that exist, the more JITP capacity is needed.

Selecting a Mezzanine Format

What characteristics should the mezzanine format have? It should:

- be computationally simple to package just-in-time;
- retain metadata in the input streams;
- be a commonly used format with an ecosystem of creation and diagnostic tools.

There are two commonly used mezzanine formats: ISO MPEG file format and MPEG-2 TS files. The first has the advantage that multi-bitrate output can be stored in just one file, as opposed to as many files as profiles, as happens in the MPEG-2 TS case. This makes management of files easier. However, MPEG-2 TS files can provide standardized ways to store many types of commonly used metadata, e.g. SCTE-35 cues for ad insertion points or various forms of closed captions and subtitling, and these are not standardized in the MPEG file case. Moreover, MPEG-2 TS would normally be the format captured in the NDVR use case, and the ecosystem of support tools (e.g. catchers, stream validation tools, stream indexing) is larger in the MPEG-2 TS case. Thus, MPEG-2 TS files make a better mezzanine format than ISO MPEG files in most cases.

WHY USE JITP?

There are a number of reasons why JITP may be a better alternative to pre-positioning assets in all final delivery formats.

Storage Cost Favings

When multiple HTTP streaming formats are used, every asset must be stored in multiple formats, with associated storage costs. This is especially true for network DVR where legal requirements in some regions mandate that separate copies are stored for each customer.

Format future-proofing

The HTTP streaming protocols in use today are still evolving; using JITP of mezzanine-format assets eliminates the need to re-package VoD libraries when these formats change. Changes in formats can be addressed via software updates of the JIT packager, which can then also manage a heterogeneous ecosystem of different format versions (e.g. various flavors of HLS). This is a huge boon to operators who must otherwise decide on a specific version of a format and thus potentially miss features in new format versions or not serve subscribers who haven't updated their video players.

Single Workflow

Using JITP for VoD with a caching CDN can automatically lead to an efficient distribution of contents in the CDN – that is, the caching of short tail (or commonly viewed) assets in the CDN and the use of JITP for un-cached long tail (rarely viewed) assets. This ensures that new assets automatically migrate into the CDN without requiring a separate offline packaging step in the workflow, as well as a separate, offline determination of which assets are short tail and which are long tail.

Graduated Investment

New VoD service offering using storage rather than JITP would require all assets to be stored in all formats up front, leading to large initial capital expenditure. With JITP, operators can add VoD capacity as the number of subscribers grows with capital expenses that match subscriber growth and revenue.

Unicast Relationship

Because the JIT Packager has a unicast session with the client, it can be used to encrypt VoD sessions uniquely for each client. Moreover, other unicast services, such as targeted ad insertion, can be integrated into the packager. Note that when chunks are encrypted per user, they cannot be cached in the CDN.

COST MODEL

In this section we describe a cost model for comparing storage with JITP. The cost model depends on whether the VoD streams are CDN-cachable or not, as could be the case, for example, if they are encrypted per user. If they are cacheable, the storage in the core used to store the assets, as well as the storage in the tiers of the CDN, can be compared to an equivalent JITP capacity. When the assets are not cacheable, the JITP cost is higher, since both the short and long tail content must be packaged just-in-time.

Cacheable Assets: Storage vs. JITP

A simple cost model (see also [Fisher]) can be created based on a few assumptions. First, we assume that short tail content will be served from the CDN and will not require JITP.

The cost of storing the complete library in multiple formats depends on multiple factors listed in the table below:

Description		Values
L	Library size (hours)	10K-150K
B	MBR bitrate (Mbps)	10
S	Number of subscribers	100K-10M
P_c	Peak concurrency	5%
P_L	Percentage of long tail requests	10%
C_s	Cost of storage ($/TB)	US $2,000
T	Number of CDN storage tiers	2
F	Number of ABR formats	3

The total cost of storage C_{ts} is then:

$$C_{ts} = C_S \times T \times F \times 3600 \times L \times B \times 10^{-6} \times 1/8$$

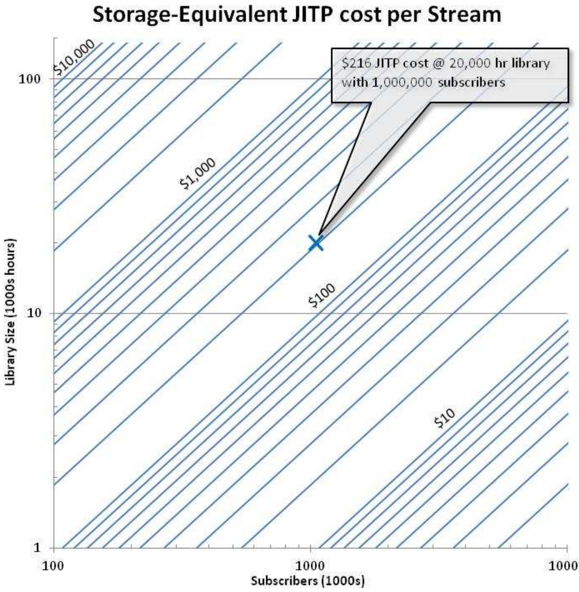
For example, a library of 20,000 hours stored in three formats at the core with two CDN points of presence (or CDN roots or different CDN tiers) would cost $1.08M.

The equivalent cost C_{jitp} of serving a JITP stream rather than using storage is the total storage cost divided by the number of long tail stream requests:

$$C_{jitp} = C_{ts} / (S \times P_c \times P_s)$$

So, in the example above, a million users would have an equivalent JITP cost per stream of $216.

We can look at the parameter space of library sizes and subscriber count to see where JITP provides value. Given that a high-end server can deliver hundreds of simultaneous JITP streams, the graph shows that the range of storage-equivalent JITP cost ranges from low (not even sustaining hardware costs) to very high (where significant savings can be achieved by delivering JITP streams rather than storage). Roughly speaking, the region where JITP leads to cost savings over storage is the upper left triangular half of the graph.



$216 JITP cost @ 20,000 hr library with 1,000,000 subscribers

It's worth noting that JITP may incur an additional cost in inbound network traffic, at least when it is centralized. Of course, if JITP is not centralized, then the library must be stored multiple times at the edge, mitigating JITP's value. A complete analysis of every variation is beyond the scope of this paper, but the model described above can be easily modified and used in each situation.

Non-cacheable Assets

When VoD assets are not cacheable, the cost model can still be used by considering 100% of the assets to be long tail. This eliminates the benefit (and cost savings) of caching the short tail in the CDN.

Just-In-Time Transcoding

The cost model does not discuss what type of processing is done in the network – only its cost compared to storage. Since the computational density of transcoding is about two orders of magnitude less than for packaging, the cost graph shows which regions in the subscriber library parameter space are suitable for transcoding as well; this is (roughly) the upper-left triangular portion of the graph that supports processing costs above $1000 per stream.

CONCLUSION

JITP may offer significant cost savings over storage, but its real value may be in other benefits: a simplified workflow, per-subscriber encryption based on unicast delivery, future-proofing against the evolution of formats, and investment and growth in capacity that is commensurate with subscriber growth.

REFERENCES

[Cablevision] 2nd Circuit Court ruling on network DVR
http://www.ca2.uscourts.gov/decisions/isysquery/339edb6b-4e83-47b5-8caa-4864e5504e8f/1/doc/07-1480-cv_opn.pdf

[Fisher] Comparing Just-in-Time Packaging with CDN Storage for VoD and nDVR Applications, Proceedings of the Canadien SCTE, March 2012.

[HLS] HTTP Live Streaming, R. Pantos,
http://tools.ietf.org/html/draft-pantos-http-live-streaming-06

[MSS] IIS Smooth Streaming Transport Protocol, http://www.iis.net/community/files/media/smoothspecs/%5BMS-SMTH%5D.pdf

[HDS] HTTP Dynamic Streaming on the Adobe Flash Platform, http://www.adobe.com/products/httpdynamicstreaming/pdfs/httpdynamicstreaming_wp_ue.pdf

[DASH] ISO MPEG 23009-1 Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats

Leveraging Time-Based Metadata to Enhance Content Discovery and Viewing Experiences

Ben Weinberger
Digitalsmiths

Abstract

Today's consumers have more choices than ever before for video entertainment and viewing devices. But with this explosion of choice has come complexity. Finding engaging entertainment has become a time-consuming and frustrating endeavor, resulting in decreased engagement and satisfaction.

The key to overcoming this discovery challenge lies in rich, time-based metadata. By creating time-based metadata at the production level and leveraging metadata-driven solutions to build best-in-class search and recommendation applications, stakeholders can create additional value at every stage of the video content lifecycle.

This paper discusses the superior metadata technologies and how they can be applied to solve today's toughest discovery challenges.

MAKING DATA RELEVANT

To enable state-of-the-art video search and recommendation tools, you need state-of-the-art data. And you need the ability to access, integrate and normalize data from disparate sources.

CREATING THE DATA SET

From dialog to set design, anything about a scene can be tagged. Efficiently and accurately creating this rich time-based data requires advanced algorithms for facial recognition, scene classification, speech recognition, natural language processing, closed-caption time alignment and ad break detection.

To understand the granularity of the resulting metadata, it is worth drilling down to the details. Each video (an asset) may have metadata associated with it. This asset-level metadata can be human authored or directly imported from 3rd party sources. Each asset in turn consists of a series of contiguous scenes. Each scene is named and is time-bound. This is a human authored process.

Metadata is tracked throughout an asset. Individual metadata elements, such as an actor, location, rights, score, etc., are associated with a scene (container) and specific frames of video (location within an asset). A metadata track may be subdivided into subtracks. For example, an objects track could be defined for tracking specific branded elements in an asset, such as cars. The process to create metadata tracks and subtracks is both automated and human authored. For example, once an actor is tagged by name, facial recognition software can tag the appearance of the actor throughout the asset.

A segment refers to a time-bound portion of the asset containing a metadata element. The segment can also have metadata associated with it (referred to as segment attributes). All metadata is automatically tracked to a frame-level timestamp, providing the ability to display the exact frame in which the metadata track or element occurs. The depth of metadata that can be associated within a single frame is shown in Figure 1, a frame from *Spiderman 3*.

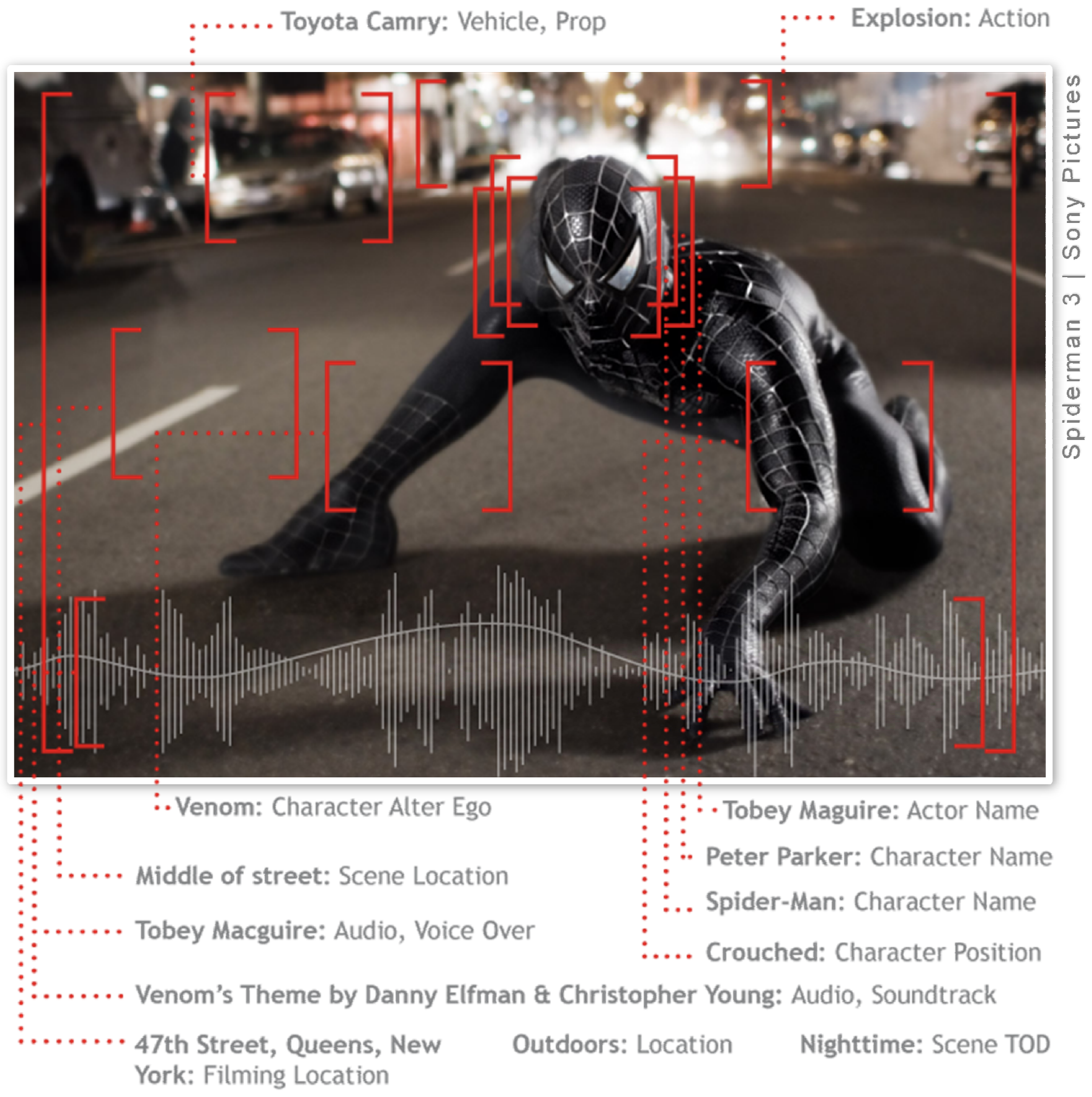

Toyota Camry: Vehicle, Prop
Explosion: Action
Spiderman 3 | Sony Pictures
Venom: Character Alter Ego
Tobey Maguire: Actor Name
Peter Parker: Character Name
Middle of street: Scene Location
Spider-Man: Character Name
Tobey Macguire: Audio, Voice Over
Crouched: Character Position
Venom's Theme by Danny Elfman & Christopher Young: Audio, Soundtrack
47th Street, Queens, New York: Filming Location
Outdoors: Location
Nighttime: Scene TOD

Figure 1: All metadata created around individual frame

Given the depth of information that can be created, it is preferable to tag the data at the time of production when much of the information is known by those closest to the creation of the asset. For example, in Figure 3, the filming location is tagged as 47th Street in Queens with a Toyota Camry in the shot. If the video were tagged by someone other than the production unit, this level of information might be lost.

Information depth and granularity provides for much stronger ability to search and find the specific video segment you are searching for. While tagging at the production level is ideal, post-production tagging also yields deep, rich information that goes beyond the typical descriptions of title, major actors, and plot.

The ideal metadata solution should also support real-time tagging for live content. With the 92nd PGA Championship as an example, time-based metadata was married to the scoring feed to video-enable the Leader Board and the Scorecard. This enhanced viewing experience allowed fans to click directly on the golfer or hole to replay specific shots, increasing viewer engagement and creating new sponsorship and advertising opportunities.

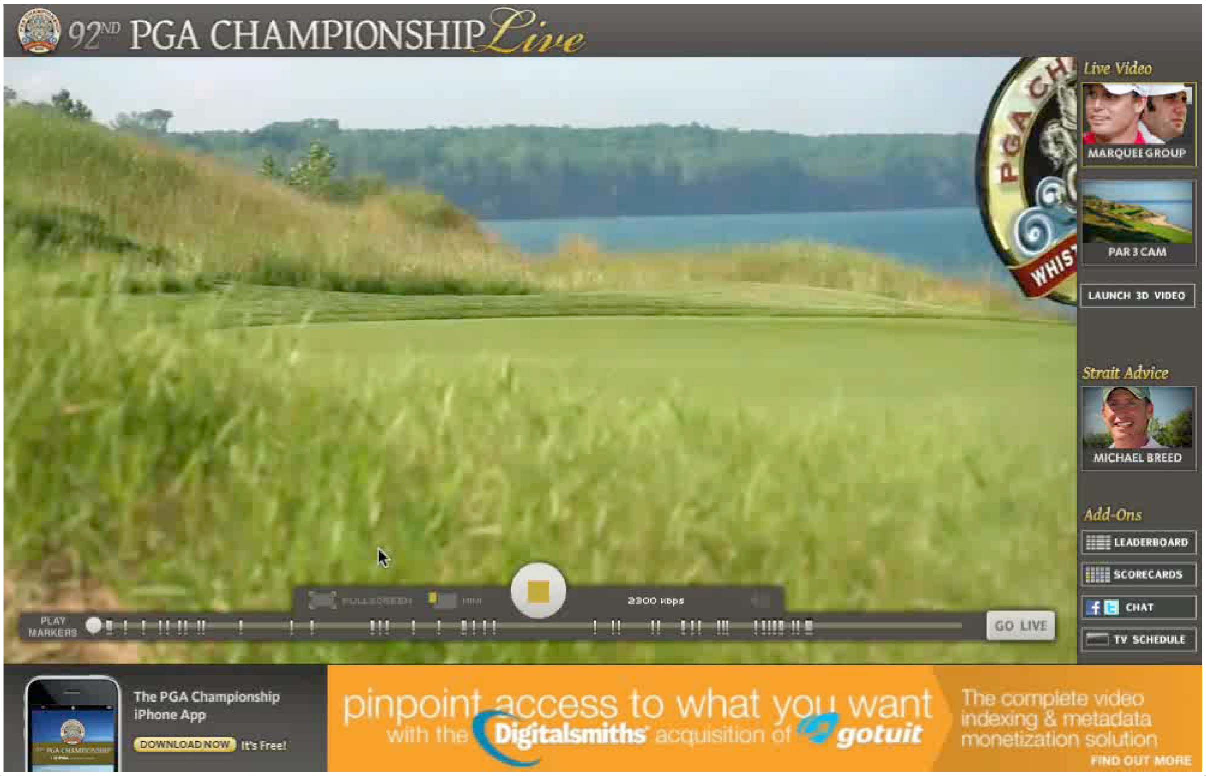


Figure 2: Metadata-driven live viewing experience

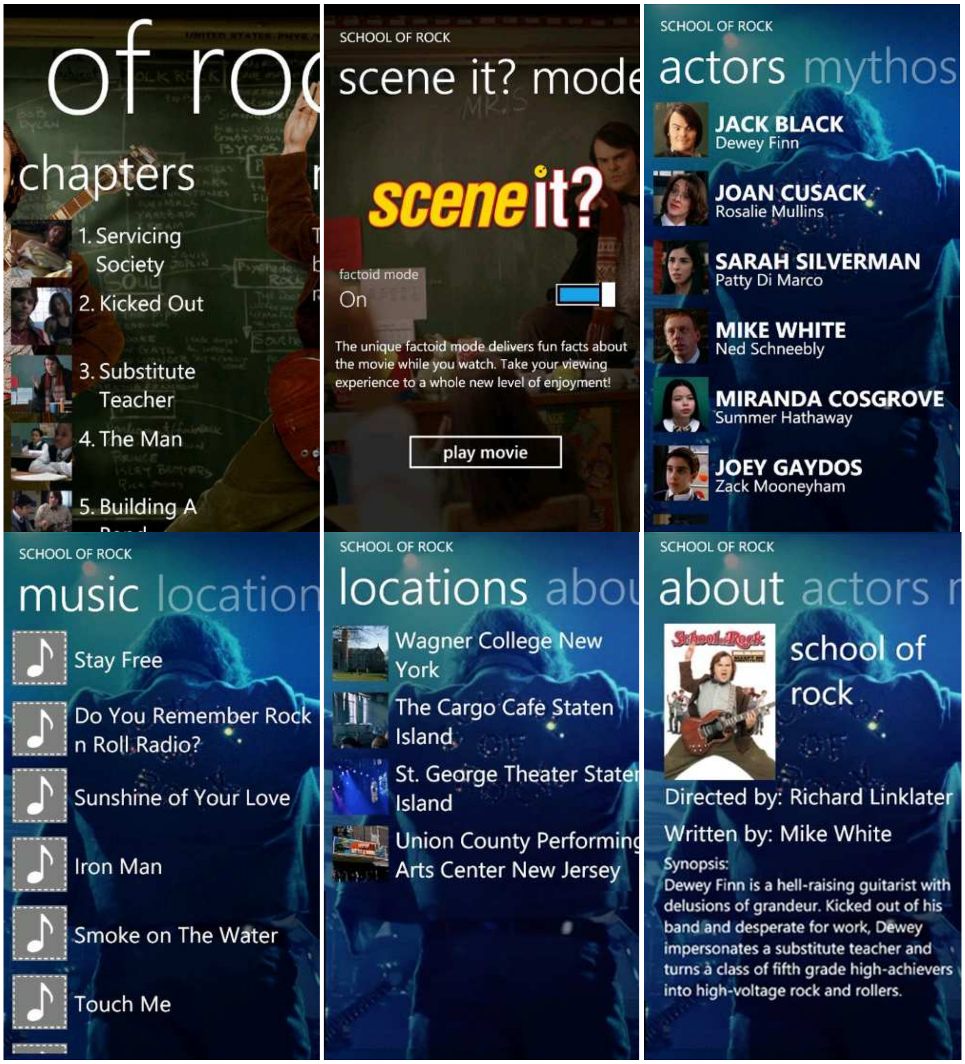
The full potential of creating rich metadata sets is achieved by creating large libraries of tagged assets. This deeper level of intelligence opens the door to unparalleled search and recommendation functionality and accuracy.

For example, while it may be common knowledge that Tom Cruise danced in *Risky Business*, a metadata-driven search for “Tom Cruise dancing” will also deliver *Tropic Thunder*, a movie in which Tom Cruise briefly danced but is not even listed in the credits.

ACCESSING THE DATA

Creating the dataset is the primary task; however, equally important is providing access to the dataset. As stated, the full power of a search and recommendation engine is found in the size of the libraries. But there will be multiple libraries available as they are created by production teams, post-production studios, and third parties (well after post-production). The search engine must be able to interface with multiple libraries to maximize the value of rich metadata and to provide the ability to recommend videos across production houses, studios, and movie libraries.

ENHANCING THE USER EXPERIENCE



The unique factoid mode delivers fun facts about the movie while you watch. Take your viewing experience to a whole new level of enjoyment!

Directed by: Richard Linklater

Written by: Mike White

Synopsis:
Dewey Finn is a hell-raising guitarist with delusions of grandeur. Kicked out of his band and desperate for work, Dewey impersonates a substitute teacher and turns a class of fifth grade high-achievers into high-voltage rock and rollers.

Figure 3: Second screens

With the growing popularity of connected devices, actionable, accurate metadata is needed to deliver the interactive applications that users have come to expect.

Being able to create a unique look and feel for the user that meets the specific need of the licensee is equally important to the success of a video search engine. For example, the needs of the PGA PC application differs widely from the phone application (Figure 3) that leverages time-based metadata from *School of Rock.*

The same dataset could drive a connected television application, a set-top box application or a tablet application.

SUMMARY

Efficiently creating rich time-based metadata around each frame of a TV show, movie or live event requires advanced algorithms for facial recognition, scene classification, speech recognition, natural language processing, closed-caption time alignment and ad break detection.

The ideal discovery platform then integrates and normalizes scene-level metadata with rich 3rd party data from disparate sources to create a deeper level of intelligence around video content, enabling unparalleled accuracy and personalization in search results and recommendations.

With these time-based metadata solutions, stakeholders can develop best-in-class enhanced discovery and viewing experiences that drive engagement and better monetization of video assets.

Managed IP Video Service: Making the Most of Adaptive Streaming

John Ulm & John Holobinko
Motorola Mobility

Abstract

The paper describes how an operator can leverage adaptive streaming protocols that are used today for unmanaged over-the-top (OTT) content for a complete managed IP video service. The paper describes how this solution is simpler and without some of the challenges imposed by implementing multicast delivery. Motorola's IP video modeling data shows compelling results regarding the relative benefits of adaptive versus multicast.

The conclusions and illustrations presented in this paper will help operators better understand how to: 1) initially deploy managed IP video services via DOCSIS, 2) plan their bandwidth and network resource requirements, 3) support existing video services in IP, and 4) optimize the network resources required as IP video viewership grows from small numbers to ultimately become the predominant means of video delivery in cable networks.

INTRODUCTION

Adaptive streaming is the primary technology for delivering over-the-top (i.e., unmanaged) IP video content to IP devices such as tablets, smartphones and gaming devices through the operator's Data Over Cable Service Interface Specification (DOCSIS) network. Adaptive streaming is the defacto delivery mechanism for OTT services. For managed services however, there is a popular assumption that multicast streaming video should be the principal delivery format to primary screens, not adaptive streaming. However, delivery of managed video in multicast format creates significant complexities for the operator, not the least of which are how to duplicate existing and planned services such as targeted advertising and network-based DVR, amongst others, and managing different segregated service group sizes compared to data services.

This paper presents a proposal to employ a comprehensive *managed* IP video services solution using adaptive streaming protocols with appropriate enhancements. An end-to-end multi-screen IP video architecture is presented, including the role of these adaptive bit rate (ABR) protocols.

The trade-offs of using adaptive streaming versus multicast for delivering managed video services are discussed. One of the other major concerns of operators is the bandwidth that will be required to deliver managed IP video services. Many factors come into play with the introduction of IP video, and our modeling results show that multicast gains may evaporate, so there is no penalty for using unicast-based adaptive protocols.

MANAGED IP VIDEO ARCHITECTURE

Multi-screen IP video delivery requires an end-to-end ecosystem that must encompass data, control and management planes. It must interact with legacy encoding, ad insertion, and content management systems while operating in parallel with traditional linear broadcasting. Operators will migrate towards multi-screen IP video to deliver content to a new generation of consumer devices such as tablets, smartphones and gaming devices; and to enable new cloud based services to attract and retain customers.

[Ulm_CS_2012] described an end-to-end conceptual architecture to support the evolution to IP video delivery. This architecture is segmented into Application, Services & Control and Media Infrastructure layers. Each of these layers is further decomposed into functional blocks.

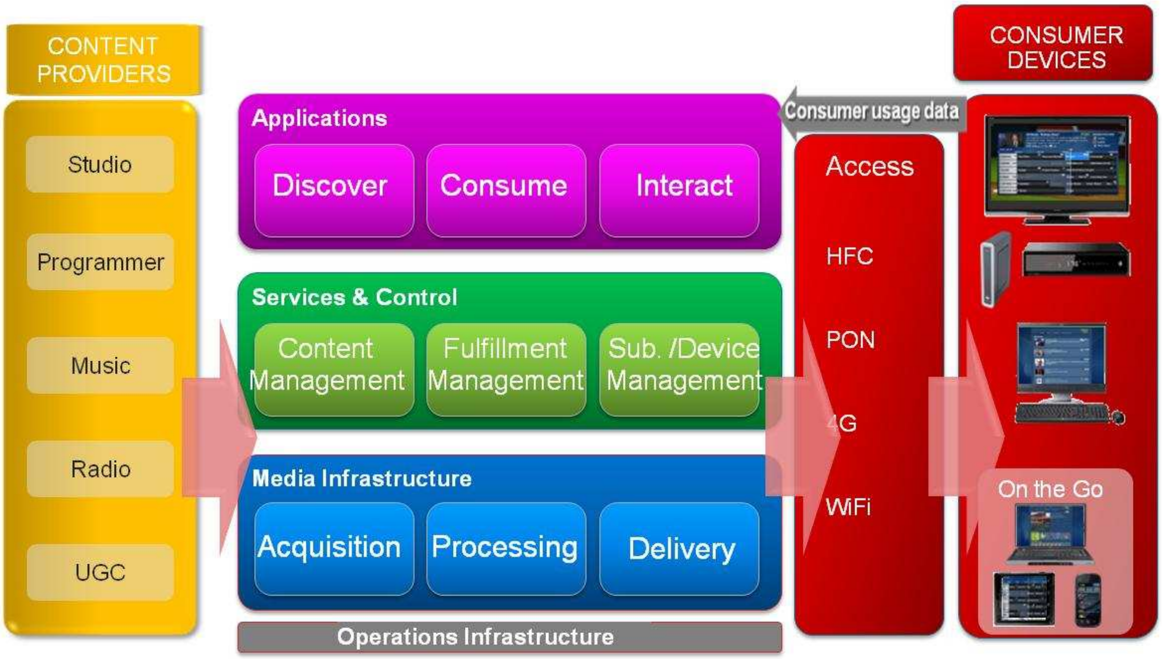


Figure 1: High Level Conceptual Architecture

Figure 1 shows a high-level abstraction of an end-to-end functional architecture for the delivery of IP video from content providers to content consumers. The video service provider must ingest content from multiple content providers, process it appropriately and then transport it over multiple types of access networks to the destination consumer devices.

The representation breaks the functions into three primary layers: Applications layer; Service & Control layer; and Media Infrastructure layer. A fourth functional block called Operations Infrastructure overlays the three primary layers.

Application Layer

The Applications layer provides interaction with the end user and is largely responsible for the user experience. It includes functions that discover content through multiple navigation options such as user interfaces (UI), channel guides, interactive search, recommendation engines and social networking links. It enables the user to consume content by providing applications for video streaming, video on demand (VOD) and network DVR (nDVR) consumption. These applications integrate with the Service & Control layer to authenticate the user, confirm access rights, establish content protection parameters and obtain resources for delivery as required.

The Application layer also provides companion applications which enable user interaction in conjunction with media programs. These may be as simple as allowing interactive chat sessions among viewers watching the same program or enable more complex integration with social media applications. It also enables enhanced monetization with new advanced advertising capabilities such as telescoping ads.

Services & Control Layer

The Services & Control layer is responsible for assigning resources within the network and for enforcing rules on content consumption that ensure compliance from a legal or contractual perspective. It includes functions that manage content work flow through all phases of its lifetime including ingest, transcoding, digital rights management (DRM) and advertising insertion policy. Other functions manage the fulfillment of user requests for content delivery by providing resource and session management, nDVR and VOD management and Emergency Alert System (EAS) and blackout support. Finally, it must manage subscribers and devices to ensure content delivery to authorized consumers in a format compatible with the consuming device.

The Services & Control Layer provides a unified approach for managing entitlements, rights, policies and services for the multitude of devices and DRM domains expected in the emerging adaptive streaming IP video service model. This solution must provide a mapping function between the billing system and the DRM system interfaces, recognizing that leveraging existing billing interfaces provides for a more seamless transition from legacy solutions. Billing should focus on account level transactions – allowing the network and associated DRMs to determine if content viewing is allowed on a specific account or a specific device. A tight integration with compelling DRM solutions is a necessity. By abstracting the complexity of a multi-DRM system, the Service & Control layer efficiently manages entitlements, rights, policies and services for a multitude of devices across a number of DRM domains. These unified provisioning functions will provide an essential building block for end-to-end multi-screen video solutions. For a detailed discussion on this topic, see [Falvo_2011].

Media Infrastructure Layer

The Media Infrastructure layer is responsible for managing video content flow and delivering the media. It includes content ingest, preparations, and delivery to the devices. Functions in this layer acquire content from satellite or terrestrial sources as either program streams or files and encode it for ingest into the system. It processes the content to prepare it for delivery. This includes functions such as transcoding, multiplexing, advertising insertion, EAS, black outs and encryption. Finally, this layer delivers the content to the target device through mechanisms such as Web servers, content delivery networks (CDNs), and streaming servers.

It is in the Media Infrastructure layer where the decision is made on video delivery protocols. For ABR distribution models, this layer includes packaging into appropriate file formats, manifest creation and publishing to a CDN origin server.

The remainder of this paper takes a detailed look at managed content delivery using adaptive bit rate (ABR) protocols.

ABR BENEFITS FOR MANAGED IP VIDEO SERVICE

Using ABR for IP video delivery can be considered a “pull” delivery model in which the end client requests the video data. With ABR, the video content is broken up and stored in a CDN as a series of small files at multiple different bit rates. The end client uses standard HTTP “get” requests to download each file segment into a local buffer from which the content is played out. The client monitors the rate at which downloads are occurring and the available locally buffered content to determine which bit rate to request. If the network is fast, a high quality high bit rate will be selected. If the network is slow, a lower quality, lower bit rate option will be requested. This is an inherently unicast service as there is no coordination between clients (even if they are watching the same content at the same time, two clients would download it independently). A tutorial on ABR for cable may be found in [Ulm_2010]. Below is an in-depth look at many key considerations and benefits in using ABR for a managed service.

CPE: Right Choice for Second/Third Screens

A key driver for migrating to IP video delivery is the ability to deliver services to a wide range of IP devices, in particular personal computers, tablets, smartphones and gaming devices. Operators want to offer these services to remote subscribers who are "off-net" as well as managed IP video services to devices inside their own network. The protocols are applicable to both linear television and on-demand delivery.

ABR protocols are the best choice for these smaller screen devices and off-net operations. They have very simple customer premises equipment (CPE) clients that adapt dynamically to changing internet resource availability. With extremely high churn on CPE devices, it is very important from an operational perspective to support the embedded client on new devices. ABR protocols are becoming the de facto standard for IP video delivery to these devices. With ABR, the operator will not become the long pole in the tent while trying to provide device drivers for the newest gadget of the week.

In-Home delivery of managed IP Video

Since ABR protocols use HTTP, they are extremely well suited for traversing home firewalls. This is in stark contrast to multicast delivery through consumer owned routers. This means that ABR is much better from an operations and support perspective.

The other issue with in-home delivery is that it may span a consumer's home wireless network with unpredictable latency and throughput. The ABR protocols are also well suited to adapt to this environment.

CDN Considerations

There are some CDN considerations that the operator must review when architecting an IP video delivery system. Traditional VOD systems today use a "push" model where streaming content 'pushes' through the system in real time. This approach supports multicast delivery, but requires session management and admission control to secure resources, guaranteed bandwidth from the server to the client, CBR-based video, and dedicated servers.

The server has the added constraints of maintaining correct timing for transmitting content. Any network-induced jitter must be removed by the edge device (edge QAM or set-top box). This approach uses a non-robust transport (e.g. UDP or RTP) which requires added complexity to detect and recover from errors. Because of all of this, a push CDN model cannot exploit general internet CDN technologies for access network delivery.

In an adaptive streaming world, clients "pull" content from the CDN as files or file segments using a reliable HTTP over TCP transport. The client pull approach is CDN friendly and allows operators to re-use HTTP-based Web caching technology that uses standard servers. The CDN caching reduces backbone capacity requirements for both linear and on-demand content. Multicast only reduces backbone traffic for linear content. All of this gives the operator significant cost benefits by leveraging internet technologies. Its state-less architecture also readily scales as needed.

To summarize, a pull CDN model provides the operator with a simpler, more cost-effective system that uses a single IP infrastructure. It leverages internet technologies for performance and resiliency. It supports ABR and enhanced quality of experience (QoE) from a common infrastructure. The operator is able to incorporate public and third party CDN services with its private CDN. Finally, this scales to a global delivery model.

Quality of Experience Considerations

In offering a managed IP video service, QoE is an important consideration for operators. One of the key factors is how the system reacts to congestion. With the high levels of compression in today's video streams, any lost packets can have severe impact on the user's experience. Implementing a multicast- based streaming service puts significant additional burdens on the operator's system. As mentioned earlier, multicast streaming is based on non-robust protocols, so in a heavily congested environment they might lose packets. The operator could choose to over provision the amount of bandwidth needed to prevent these conditions, in which case they are throwing away potential capacity gains from using multicast. The alternatives are to implement some combination of admission control and/or error recovery. An admission control algorithm will be further complicated if variable bit rate (VBR) video delivery is used to maximize bandwidth savings rather than constant bit rate (CBR). An error recovery system introduces new servers into the network and requires custom clients in the consumer devices. Overall, the design, deployment and operation of a multicast-based system are inherently complex.

ABR protocols were developed for Internet delivery with its constantly changing throughput. ABR seamlessly adapts to this varying environment. In a managed network with infrequent periods of congestion, ABR reduces its bit rates during these periods to compensate. The impact on QoE might be comparable to that of running legacy MPEG video through a statistical multiplexer (statmux), which is familiar to operators. ABR also is based on a reliable TCP protocol that has error recovery already built into it, so any packets lost during congestion are automatically retransmitted. Thus, it prevents blocking and other video artifacts that significantly impact QoE. In this case, no network resources need be reserved in advance for the service and ABR reduces or eliminates the potential for blocking. Using adaptive protocols for all IP video delivery helps the operator's overall system become much simpler. More on this topic can be found in [White_2012].

Another QoE consideration is the impact of channel change time. ABR protocols are well suited to fast change times as they can quickly load lower bit rate streams and then switch to higher bit rates as bandwidth is available. Using multicast delivery requires separate additional bandwidth and a proprietary protocol to quick start the video delivery in parallel with the multicast video.

Advanced Services

Another key reason for migrating to IP video services is the ability to offer new advanced services. In particular, this might include highly targeted advertising such as personalized advertisements and telescoping. The system must also support EAS and blackout identical to legacy video services. Using its playlist manipulation, ABR provides the service provider with tremendous capability to re-direct a client on-the-fly with minimal effort and equipment. Supporting these advanced services using multicast delivery becomes problematic.

Miscellaneous Considerations

IP video penetration will occur over a long period of time. This means that the operator's home gateway will continually change during that time as well. Today cable operators have DOCSIS D3.0 devices in the field with 3, 4 or 8 downstream channels. Over the next several years we will see this expand to include 16, 24 and perhaps 32 downstream channels. The operator needs to manage this DOCSIS modem transition. Using ABR and its unicast delivery allows every modem to be in a bonding group suited to its capabilities; multiple bonding groups can then overlap, allowing the cable modem termination system (CMTS) to fully utilize the bandwidth. Multicast delivery runs into multiple problems in a mixed bonding group environment as discussed in [Ulm_2009].

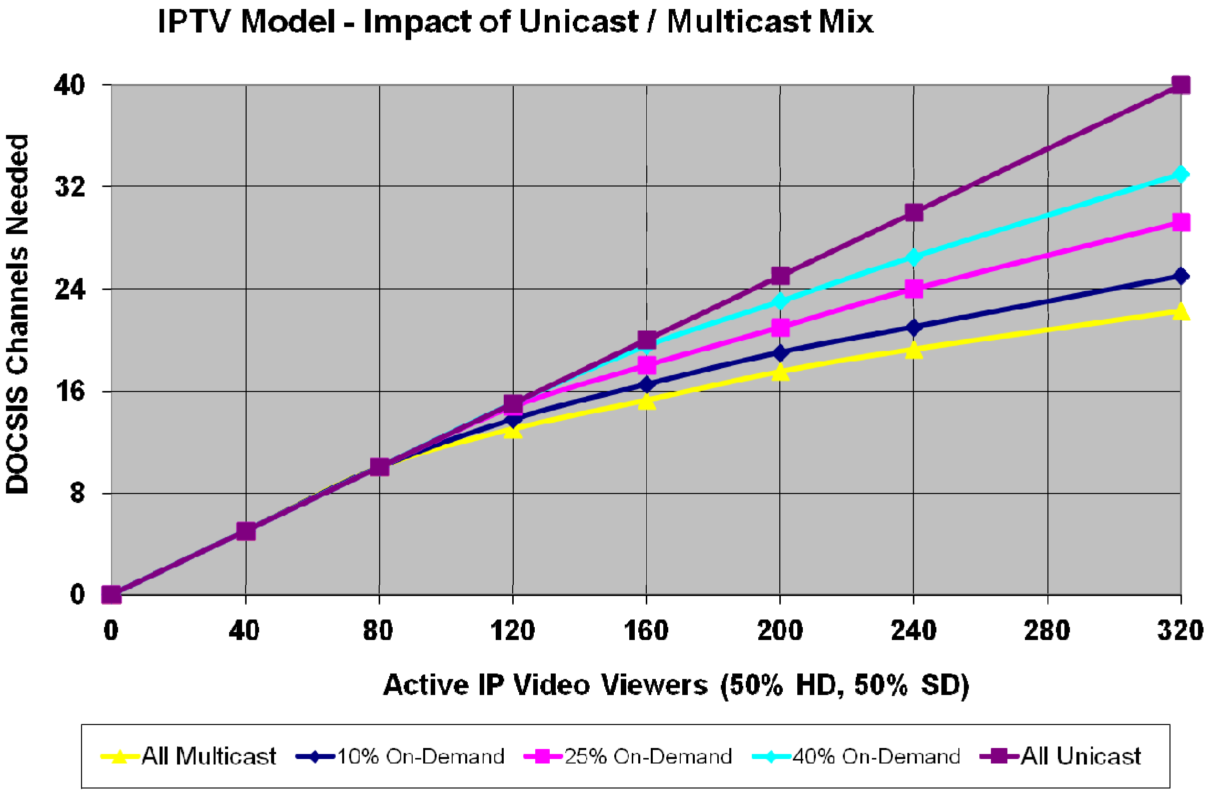


Figure 2: Impact of Unicast / Multicast Mix

ABR BANDWIDTH CONSIDERATIONS

A detailed analysis of bandwidth requirements for ABR compared to multicast was given in [Ulm_CS_2012]. The findings were that, under most conditions, multicast delivery will have little or no bandwidth capacity advantages over ABR unicast delivery. Figure 2 shows some results from that paper.

For early deployments of IP video, the penetration rate will be low. As indicated in this figure, there is no multicast benefit below 120 active viewers. With many operators considering phasing in IP video gradually, the operator also needs to factor in their plans for service group sizes. If the phasing takes 5-7 years, will the operator initiate node splits and cut service group sizes in half during that time? At the same time, increased VOD usage and the introduction of nDVR services might cause a shift from 10% to 25% or even 40% unicast usage. Figure 2 clearly shows what happens when the number of active viewers drops from 320 to 160 or 240 to 120 viewers.

Impact of Multi-Screen Delivery

This analysis was done for a two screen system: 50% of viewers watching high definition (HD) TV content and 50% of viewers watching standard definition (SD) TV content. With multi-screen delivery being a key impetus for IP video services, Motorola extended the IP video capacity modeling to see the effect of multi-screen viewing on capacity requirements.

Below are some sample outputs from the enhanced IP video capacity modeling. This looks at the bandwidth requirements for IP video for two different sized service groups as penetration grows.

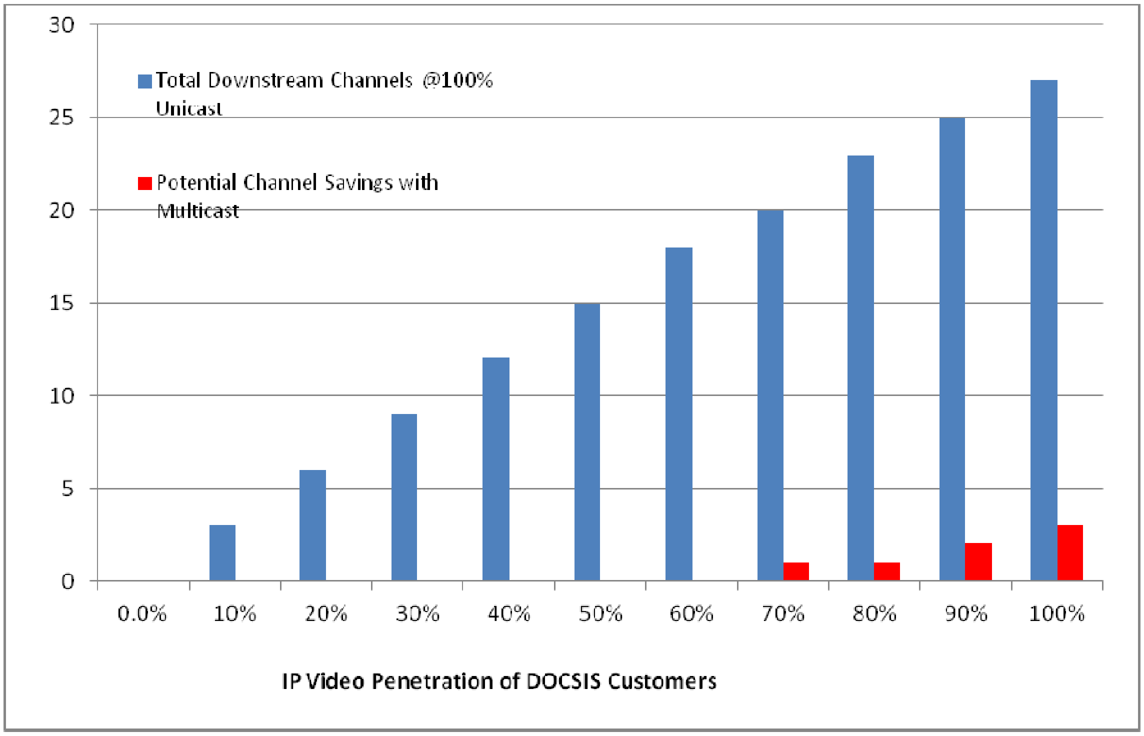


Figure 3: IP Video Bandwidth & Multicast Savings: 320 Active Viewers

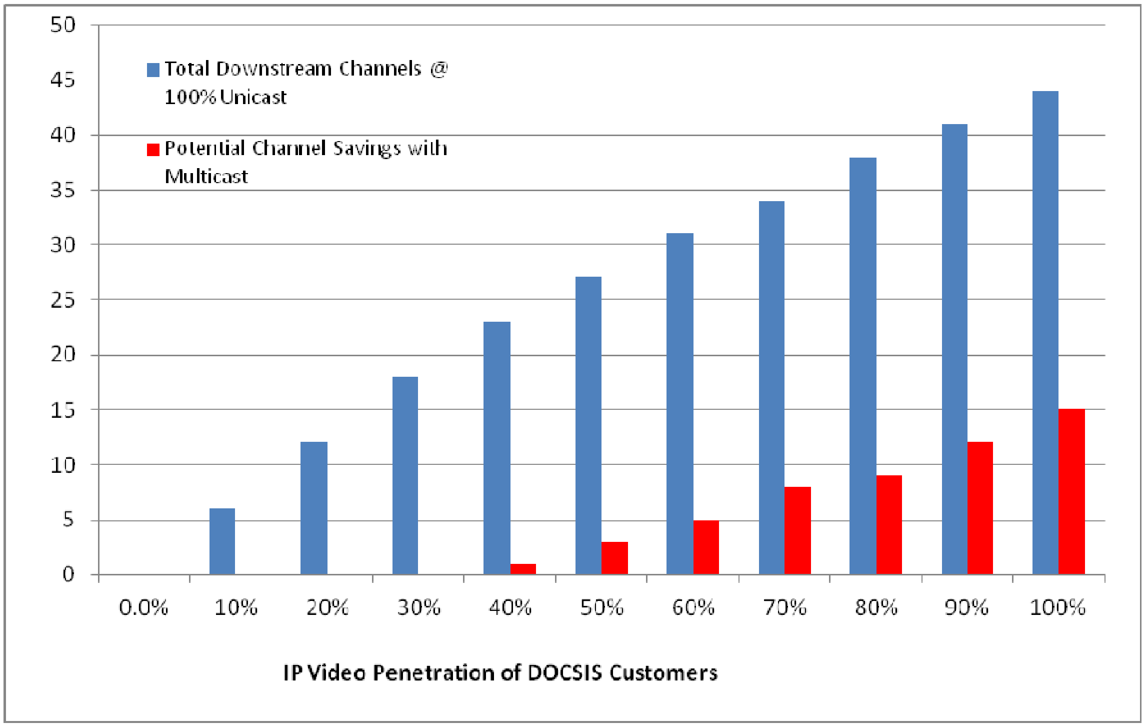


Figure 4: IP Video Bandwidth & Multicast Savings: 640 Active Viewers

In Figure 3, 100% IP video penetration corresponds to 320 active viewers which might represent a 500 homes passed (HP) service group, identical to the analysis above. Figure 4 doubles the service group size to 640 active viewers. In both these examples, viewership is spread across five different screen sizes: 30% HDTV, 30% SDTV; 20% tablets; and 10% each for two smaller screen sizes. It also assumes 25% on-demand usage which is reasonable if nDVR is deployed for the IP devices.

As indicated in Figure 3, the potential multicast gain is non-existent until the operator has reached 70% IP subscriber penetration. Even at 100% penetration, the multicast gain is only 3 channels or ~10% of capacity. This amount is almost negligible in a converged cable access platform (CCAP) environment capable of 64 channels per port.

In Figure 4, the serving group size is doubled. Perhaps the operator combined two fiber nodes to the same CCAP port to get additional multicast gains. Even with this extremely large service group of ~1000 HP, the multicast savings is still less than 20% at 70% IP penetration, yet it requires 34 DOCSIS channels of capacity for the large serving group. The small savings for multicast comes at a significant cost in spectrum used. It also comes in the late stages of the IP video deployment.

QoS in a Multicast Implementation

The purpose of implementing Multicast for delivering managed video content is to save bandwidth. By its very nature, a multicast system only makes sense if fewer channels of spectrum are required than a unicast implementation. Multicast designs are wholly dependent on the assumptions of multicast viewership during peak. At peak viewership, if more programs are being requested than the multicast service group was designed for, blocking occurs resulting in a denial of service. Therefore a prudent design calls for a safety factor in the number of QAMs reserved for the multicast service group. However this flies in the face of the rationale for implementing multicast, which is bandwidth savings.

In contrast, in a unicast implementation, if the bandwidth peak is achieved, the adaptive bit rates are lowered for the viewers in the service group. While video quality may lessen slightly in these cases, there is no denial of service. Therefore, unicast is a better choice for insuring a non-blocking service at peak usage times.

SPECTRUM MIGRATION STRATEGIES

Another very important aspect to IP video migration is finding sufficient spectrum. Some operators have already made more spectrum available by recovering analog TV channels using digital TV terminal adapters (DTA) while other operators have upgraded their hybrid fiber coaxial (HFC) to 1GHz or turned to Switched Digital Video (SDV). This available spectrum is being gobbled up today as more HD content is deployed, VOD requirements continue to increase and high speed data (HSD) services continue to grow at 50% annual rates. So there may still be a need for additional spectrum to ramp up IP video services with a corresponding economic impact.

Early Transition Plans – Hybrid Gateways

One way to significantly reduce spectrum requirements is to convert legacy MPEG-2 linear TV to IP video in a video gateway device that includes a transcoder. This approach requires no new spectrum for linear TV as this video gateway device appears as a set-top box (STB) to the system and uses legacy broadcast content.

The video gateway also has the advantage that it is the single point of entry for video services and allows IP STBs to be deployed elsewhere in the home behind it. These hybrid devices can also operate as IP devices and are pivotal in the transition to an all IP system. Longer term, the transcoding capability and adaptive protocols supported by the gateway may limit the quantity and type of IP devices supported in the home. Eventually the operator will want to support IP devices directly from the "cloud" using their network infrastructure.

A detailed discussion of the home gateway migration is given in [Ulm_CS_2012].

Complete Recovery of Legacy Bandwidth

The previous section on video gateway migration plans helps the operator as they begin the IP video transition. However, the end game is to eventually get to an all-IP system. Legacy MPEG digital TV services may continue to consume 50% to 80% of the available spectrum even after DTA and 1GHz upgrades. Regardless of which path the operator initially took to free up spectrum, eventually they will need to install switched digital video (SDV) to reclaim all of the legacy digital TV bandwidth.

Adding SDV to the mix also increases the need for narrowcast QAM channels. This plays well into a CCAP migration. As the mix between legacy and IP subscribers changes, an operator will need to re-assign SDV bandwidth to IP video bandwidth. This is well suited for CCAP. For a detailed discussion on IP video economics in a CCAP world see [Ulm_NCTA_2012].

Some SDV capacity reclamation modeling results are shown in Figures 5 and 6. Figure 5 shows the total spectrum required for legacy video services as the number of legacy viewers is reduced to zero. It assumes a video service with 180 HD programs (3 per QAM) and 200 SD programs (10 per QAM), so full broadcast requires 80 QAM channels. Figure 6 shows the corresponding SDV narrowcast requirements.

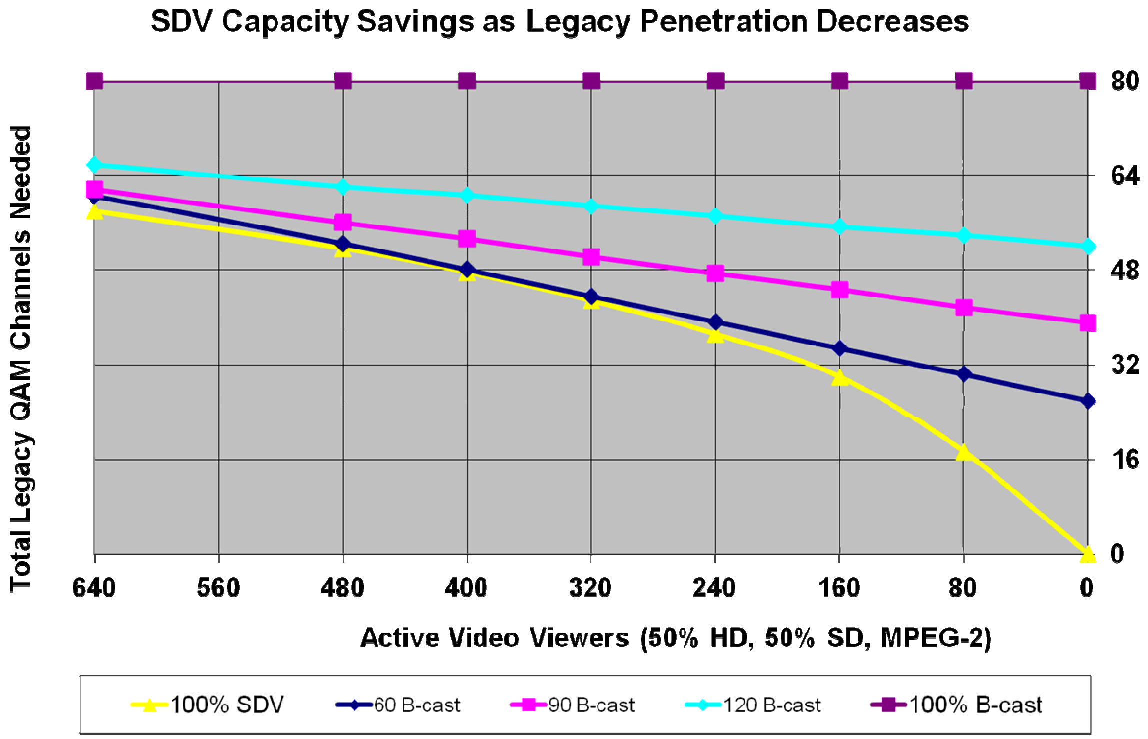


Figure 5: SDV – Total Capacity Savings with Decreasing Penetration

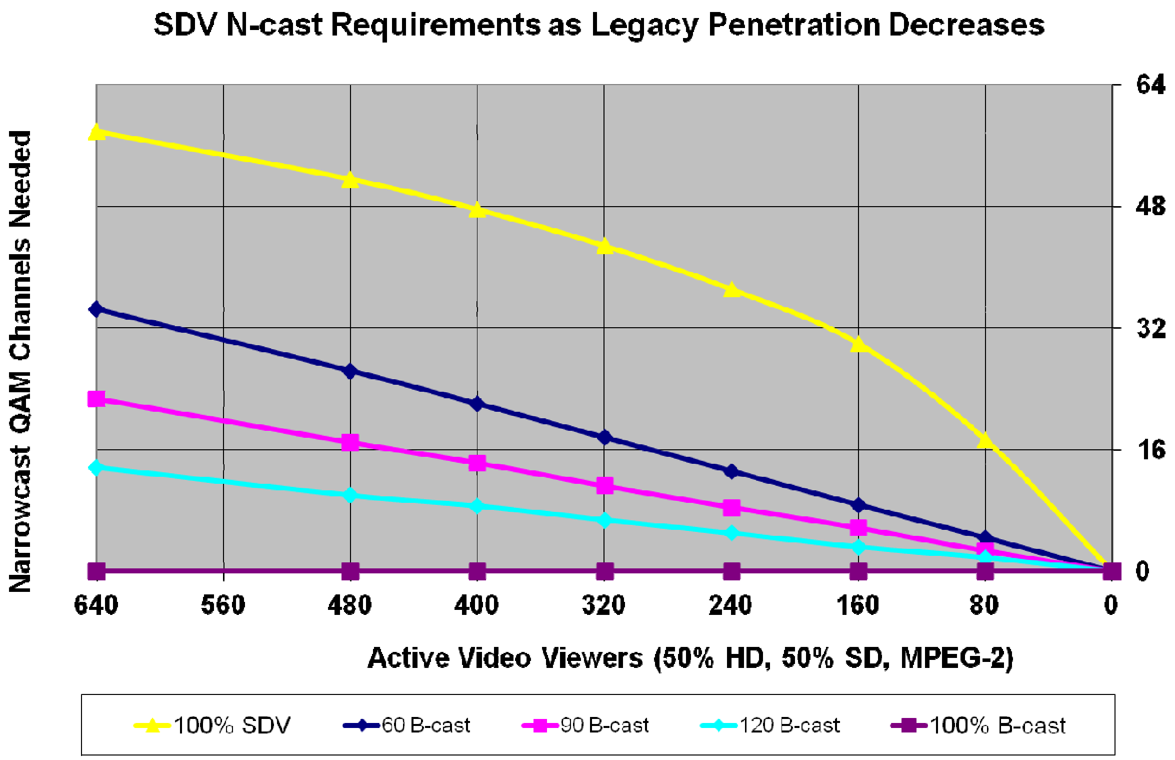


Figure 6: SDV – Total Narrowcast Requirements with Decreasing Penetration

Four scenarios are given varying the amount of switched content up to 100% switched. As shown, 100% switched provides the most bandwidth savings, but requires significantly more narrowcast. The operator has complete flexibility in trading off between spectrum saved and narrowcast QAM requirements. As can be seen in Figure 6, as the number of legacy viewers decreases, there is a corresponding decrease in narrowcast QAM requirements. This allows the operator to repurpose SDV QAM channels as they become freed for DOCSIS channels (HSD or IP video) or additional SDV savings.

It is informative to look at an example where the operator allocates twelve QAM channels for SDV and watch the impact as their legacy viewers are reduced. From Figure 6, the curve representing 120 broadcast programs and 60HD/80SD switched programs crosses 12 QAMs at 560 active viewers. Now looking at Figure 5, this scenario (i.e. 560 viewers, 120 B-cast) requires 64 channels of spectrum, freeing 16 channels (compared to 80 channels for 100% broadcast) for other usage such as IP video growth. As IP video penetration grows, legacy penetration shrinks. The next curve (90 broadcast with 90HD/110SD switched) on Figure 6 crosses 12 QAMs at 320 viewers. Mapping to Figure 5, this scenario (i.e. 320 viewers, 90 B-cast) only requires 50 channels of spectrum, so 30 channels are now available. The next scenario (60 broadcast with 120HD/140SD switched) crosses 12 QAMs around 200 viewers and requires ~36 channels for more savings.

As a result, the SDV spectrum savings are significantly more than multicast gains seen in the previous section. The SDV benefits are also available for small and large service groups. Every operator needs to consider SDV as a crucial part of its IP video migration.

CONCLUSION

Cable service providers will migrate from existing legacy video networks to a full end-to-end IP video system in a number of stages as new services are rolled out. They need to leverage the technology used for these intermediate stages into the final end-to-end system. Therefore, it is critical to have a layered architecture approach as presented in this paper that can isolate the changes between the various components.

Selecting the correct technology is particularly important for the delivery component of the Media Infrastructure layer as it is hardware centric, widely deployed and capital intensive. In particular, this paper focuses on the selection of adaptive protocols as the primary video delivery mechanism and discusses its benefits. ABR enables:

- A wealth of new and constantly changing IP devices
- Easily handles the home environment
- Provides excellent QoE to consumers
- Adapts to congestion without requiring complex admission control or re-try mechanisms
- Leverages internet CDN technology
- Readily supports advanced services including personalized advertising.

The updated IP video capacity modeling results shows the impact of migrating to a multi-screen environment. A 500HP service group may only get 10% multicast gain even once its switched to all IP video delivery.

Understanding the migration plan is a critical piece of the IP video architecture, especially with respect to managing available spectrum. Hybrid video gateways enable the introduction of IP video delivery with minimal impact on an operator’s infrastructure. As the system scales, these devices transition to full IP video delivery.

Finally, the operator needs to plan the reclamation of legacy spectrum as they migrate to an all-IP world. This migration will eventually require the use of SDV. The modeling results show that the benefits of SDV are actually greater than the savings from multicast delivery.

In conclusion, the operator needs ABR for its first IP video steps when delivering content to second and third screens; i.e., tablets, smartphones, PCs and gaming devices. Adaptive streaming is the final solution the operator needs once there is an all-IP world with any content, anywhere, anytime, anyplace. We have shown that ABR also handles the transition years and is the only delivery mechanism needed for a managed IP video service.

REFERENCES

[Ulm 2009]	J. Ulm, P. Maurer, “IP Video Guide – Avoiding Pot Holes on the Cable IPTV Highway”, SCTE Cable-Tec Expo, 2009.
[Ulm 2010]	J. Ulm, T. du Breuil, G. Hughes, S. McCarthy, “Adaptive Streaming – New Approaches For Cable IP Video Delivery”, The Cable Show NCTA/SCTE Technical Sessions, spring 2010.
[Falvo 2011]	B. Falvo, D. Clarke, C. Poli, *“Supporting Multi-CAS and DRM Entitlements”*, SCTE Cable-Tec Expo, Nov 2011.
[Ulm CS 2012]	J. Ulm, G. White, *“Architectures & Migration Strategies for Multi-Screen IP Video Delivery”*, SCTE Canadian Summit, March 2012.
[Ulm NCTA 2012]	J. Ulm, G. White, *“The Economics of IP Video in a CCAP World”*, NCTA Technical Sessions, May 2012.
[White 2012]	G. White, J. Ulm, *“Reclaimng Control of the Network from Adaptive Bit Rate Video Clients”*, NCTA Technical Sessions, May 2012.

Mission is Possible:
An Evolutionary Approach to Gigabit-Class DOCSIS

John T. Chapman, CTO Cable Access BU & Cisco Fellow,
Cisco, jchapman@cisco.com

Mike Emmendorfer, Sr. Director, Solution Architecture and Strategy,
Arris, Mike.Emmendorfer@arrisi.com

Robert Howald, Ph.D., Fellow of Technical Staff, Customer Architecture,
Motorola Mobility, rob.howald@motorola.com

Shaul Shulman, System Architect,
Intel, shaul.shulman@intel.com

Abstract

This paper is a joint paper presented by four leading suppliers to the cable industry, with the intent to move the industry forward in the area of next generation cable access network migration. To our knowledge, it is a first for four such suppliers to collaborate in this manner on a topic of such critical industry importance.

Cable operators are facing a rising threat associated with the limitations of today's 5 to 42 MHz return path. Constraints on capacity and peak service rate call for finding additional return spectrum to manage this emerging challenge.

We will explain how and why an approach based on the principle of an expanded diplex architecture, and using a "high-split" of up to 300 MHz, is the best path for operators to manage this growth. This includes considering the simultaneous expansion of the downstream capacity.

We will describe obstacles associated with legacy CPE in both Motorola and Cisco video architectures and propose solutions to these issues.

To use the reallocated HFC spectrum most effectively, we will consider an evolutionary strategy for DOCSIS and show how it capably meets the requirements ahead.

We will contemplate the application of new generations of communications technology, including a comparison of single-carrier approaches implemented today to multi-carrier techniques such as OFDM, including channelization options. We will consider higher order QAM formats as well as modern FEC tools such as LDPC.

We will discuss how these evolution alternatives can be harnessed to best extract network capacity. We will consider how evolution of the access architecture enables this new capacity, and how the end-to-end network components develop to support this growth.

In summary, we will present a strategy that preserves network investment, enables a versatile evolutionary path, and positions operators to create an enduring lifespan to meet the demands of current and future services.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The evolution of DOCSIS is bounded only by technology and imagination - both of which themselves are unbounded.

This white paper takes an in depth look into the technologies that are available to DOCSIS and then makes concrete recommendations on how DOCSIS should be taken to a new level of performance.

DOCSIS to Date

The original DOCSIS 1.0 I01 (Interim version 1) specification was released on March 26, 1997. DOCSIS technology has evolved very well since its inception over 15 years ago. Here are some of the interesting milestones from those first 15 years.

- 1997 Mar – DOCSIS 1.0 I01 released. Features basic data service.
- 1997 Dec – Cogeco has the first large scale DOCSIS 1.0 deployments
- 1999 Mar – First certified CM and qualified CMTS
- 1999 Apr – DOCSIS 1.1 released. Adds QoS.
- 1999 Dec – PacketCable 1.0 released. Adds voice over IP (VoIP)
- 2001 Dec – DOCSIS 2.0 released. Adds ATDMA and SCDMA.
- 2002 Feb – DSG released. Adds STB control channel to DOCSIS
- 2005 Aug – Modular CMTS (MHA) released. Shared EQAM between DOCSIS and video is added.
- 2006 Aug – DOCSIS 3.0 released. Adds bonding, IPv6, and multicast.

In the first phase of its life, DOCSIS focused on a moderately dense and complex MAC and PHY with a comprehensive set of features and services. DOCSIS now has a very rich and mature service layer.

If this was the first 15 years of DOCSIS, then what is the next 15 years of DOCSIS going to look like? How well will DOCSIS compete with other broadband technologies?

The Future Potential of DOCSIS

The next phase of DOCSIS will take it to gigabit speeds. DOCSIS needs to scale from a few RF channels within a CATV spectrum to being able to inherit the entire spectrum. And DOCSIS may not even stop there.

In the upstream, in an effort to get to gigabit speeds and beyond, DOCSIS needs to scale beyond its current 5 – 42 MHz (65 MHz In Europe) to multiple hundreds of megahertz. In the downstream, DOCSIS needs to extend beyond the current 1 GHz limit and set a new upper RF boundary for HFC Plant.

Table 1 shows where DOCSIS technology is today and where it is going.

Today, the deployed DOCSIS 3.0 cable modems have eight downstream channels (6 or 8 MHz) and four upstream channels (6.4 MHz). This provides an aggregate downstream data capacity of about 300 Mbps and an aggregated upstream data capacity of 100 Mbps.

Next year (2013), the market will see cable modems that have on the order of 24

Table 1 – The Future Potential of DOCSIS

	Parameter	Now	Phase 1	Phase 2	Phase 3
Downstream	Frequency Band	54 - 1002 MHz	108 - 1002 MHz	300 - 1002 MHz	500 - 1700 MHz
	Assumed Modulation	256-QAM	256-QAM	≥ 1024-QAM	≥ 1024-QAM
	Chan (or equiv)	8	24	116	200
	Data Capacity	300 Mbps	1 Gbps	5 Gbps	10 Gbps
Upstream	Frequency Band	5 - 42 MHz	5 - 85 MHz	5 - (230) MHz	5 - (400) MHz
	Assumed Modulation	64-QAM	64-QAM	≥ 256-QAM	≥ 1024-QAM
	Chan (or equiv)	4	12	33	55
	Data Capacity	100 Mbps	300 Mbps	1 Gbps	(2) Gbps

downstream channels and 8 upstream channels. DOCSIS 3.0 defines a mid-split upstream that takes the upstream spectrum up to 85 MHz and could contain at least 10 channels. That provides an aggregate data capacity of almost 1 Gbps in the downstream and 300 Mbps in the upstream.

The goal for the next generation of DOCSIS is to achieve 1 Gbps of data capacity in the upstream and to be able to scale to the full spectrum of the existing downstream. While the final spectrum plan has not been determined yet, an estimate would be a 5 Gbps down, 1 Gbps up system. That would maintain a 5:1 ratio between upstream and downstream bandwidth that is good for TCP.

As a stretch goal, there is additional spectrum above 1 GHz. If the downstream expanded into that spectrum, and the upstream spectrum was increased even further to keep the same 5:1 ratio, DOCSIS could become a 10 Gbps down and 2 Gbps up technology.   This would enable cable data capacity equivalent to next generation PON systems.

While the final choices for these numbers (indicated with "( )") still needs to be made, there seems to be at least three progressions of technology. Phase 1 upgrades the upstream to 85 MHz and takes advantage of technology available today. Phase 2 upgrades the upstream to 1 Gbps and the downstream to 1 GHz if it is not there already. Phase 3 extends the downstream to 1.7 GHz and gives a second boost to the upstream.

Now that we have established our goals, let’s look at how to achieve them.

2 CABLE SPECTRUM ANALYSIS

The spectrum allocation options should consider the impact to the overall end-to-end system architecture and cost. The solutions should also consider the timing of these changes as this may impact cost. The end-state architecture should be considered for this next touch to the HFC. We do not need to solve next decade's problems now, however we should consider them as part of the analysis.

The cable operator has several spectrum split options available and some are examined in this analysis. [33] [34] [35] Figure 1 below is an illustration of some of the spectrum split options; it also depicts a few other options, such as Top-split with Mid-split. In Figure 1, the Top-split (900-1050) options has a 150 MHz block of spectrum allocated for guard band between 750-900 MHz and 150 MHz block of spectrum between 900-1050 MHz for upstream.

2.1 Mid-split (85)

Overview

The Mid-split Architecture is defined as 5-85 MHz upstream with the downstream starting at approximately 105 MHz; this may also be referred to as the 85/105 split. The mid-split architecture essentially doubles the current upstream spectrum allocation

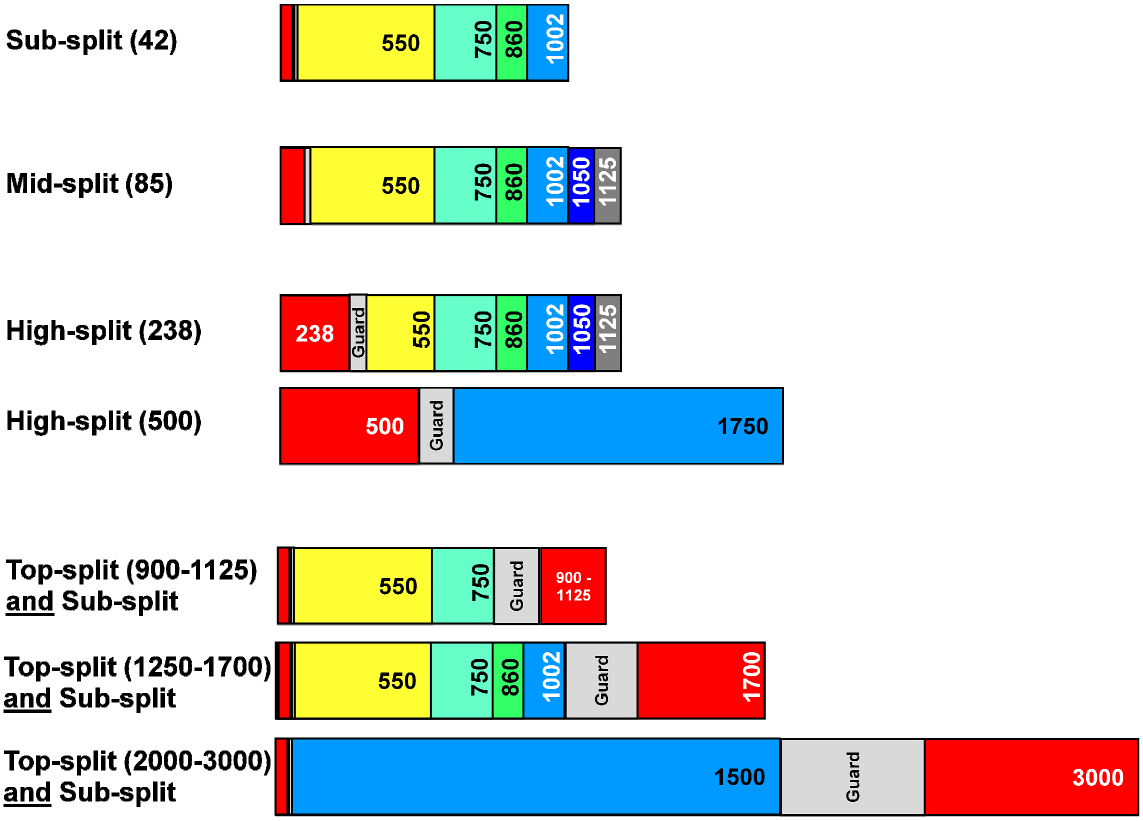


Figure 1 – Spectrum Allocation Options

however this may triple or even quadruple the IP based capacity.

The capacity increase in data throughput is a result of the high-order modulation and all of the new spectrum may be used for DOCSIS services, which is not the case with the sub-split spectrum that has generally accepted unusable spectrum and legacy devices consuming spectrum as well.

Pros

- Sufficient bandwidth to last nearly the entire decade
- DOCSIS QAM MAC layer capacity estimated at ~310 Mbps
- Avoids conflict with OOB STB Communications
- Lowest cost option
- High order modulation possible 256-QAM perhaps higher
- The use of 256-QAM translates to fewer CMTS ports and spectrum (using 64-QAM would require approximately 33% more CMTS ports and spectrum)
- DOCISIS systems already support this spectrum (5-85)
- MSOs that have already deployed DTAs (Digital Terminal Adapters) should strongly consider thing approach
- Some amplifiers support pluggable diplexer filter swap
- Some existing node transmitters and headend receives may be leveraged
- Does not touch the passives
- Upstream path level control is similar to the Sub-split (~1.4 times the loss change w/temp); Thermal Equalizers EQT-85 enables +/-0.5 dB/amp delta

Cons

- Impacts Video Service (in low channels)
- Reduces low VHF video spectrum
- Throughput of 310 Mbps is less than the newer PON technologies

Assessment

The selection of Mid-split seems like an excellent first step for the MSOs. This split option has little impact to the video services and does not impact the OOB STB commutations. This spectrum split may last nearly the entire decade, allowing time for the MSOs to assess future splits, if required, and the impacts to other split option at that time. The Mid-split appears to be an excellent first step. MSOs that have already deployed DTAs should strongly consider using this approach.

2.2 High-split (200, 238, or 500)

Overview

The High-split Architecture has generally been defined as 5-200 MHz with the downstream starting at approximately 250-258 MHz crossover for the downstream. However, we believe that a High-split (238) or even High-split (270) options should be considered, as this will have enough spectrum capacity to reach the desired 1 Gbps data rate, with reasonable PHY and MAC layer overhead removed. [33] [34] [35]

Also it is uncertain if the entire region of spectrum between 5-238 may be used as there could be legacy channels in service as well as frequency bands undesirable performance or usable for interference reasons. The use of High-split (500) has been mentioned as a possible long-term migration strategy if coaxial network want to offer the capacity of XG-PON1 systems.

In the case of 5-500 MHz our capacity targets assume a digital return HFC style optical connection and as will all architectures the paper model begins at a 500 HHP node to a 16 HHP node to determine capacity.

Pros

- High-split is far more predictable from an MSO deployment, operational, and service ability perspectives when compared with Top-split, as Top-split options have much tighter cable architecture requirements (refer to Cons of Top-split).
- Operates effectively at a typical 500 HHP node group using 256-QAM (see details in the sections later in this analysis)
- The use of 256-QAM translates to fewer CMTS ports and spectrum (using 64-QAM would require approximately 33% more CMTS ports and spectrum)
- High-split (238) using DOCSIS QAM reaches an estimated MAC layer capacity 1 Gbps
- However High-split (270) may be needed to allow for operational overhead
- High-split (500) at a 250 HHP through a 16 HHP optical node service group with digital return HFC optics is estimated to reach 2.2 Gbps DOCSIS QAM MAC layer capacity
- DOCSIS OFDM with LDPC may be able to use 2 orders higher modulation in same SNR environment
- Very low cost spectrum expansion option, especially considering similar capacity Top-split options (STB OOB cost was not considered in the analysis)
- The OOB STB problems will likely be reduced over time, and with the STB costs declining over time this will remove or reduce this issue to High-split adoption
- If DTAs are deployed or plan on being deployed High-split should be considered strongly, because DTA remove the Analog Video Service impact obstacle from High-split
- Lowest cost per Mbps of throughput
- Some existing HFC Equipment supports High-split like node transmitters and headend receivers
- DOCISIS systems already support some of this spectrum (5-85)
- Passives are untouched
- High-split provides sufficient upstream capacity and the ability to maximize the spectrum with very high order modulation
- High-split does not waste a lot of capacity on guard band
- Level control using Thermal Equalizers EQT-200 (~2.2 times Sub-split cable loss)
- Downstream could expand to 1050 MHz or even 1125 MHz perhaps using the existing passives

Cons

- Conflicts with OOB STB Communications if DOCSIS Set-top box Gateway (DSG) is not possible

- Takes away spectrum from Video Services (54-258 MHz or higher if the upstream stops at 238 MHz)
- Takes away spectrum from Video devices (TVs and STBs)
- Potentially revenue impacting because of spectrum loss supporting analog video service tier
- Downstream capacity upgrade from 750 MHz to 1 GHz to gain back capacity lost to upstream

Assessment

The use of high-split has several key challenges or cons listed above, and the major concerns include 1) the impact OOB Set-top Box communications for non-DOCSIS Set-top Gateways, 2) the analog video service tier and the simplicity of connecting to an subscribers TV to enable services, and 3) we takeaway valuable capacity from existing video devices like STBs and existing TVs.

However, if the deployment of High-split (238) is planned later in time, this may allow these older STBs to be phased out or redeployed to other markets. There may also be workarounds to enable high-split and keep the legacy OOB in place. The impact to the analog service tier is a major concern, this accounts for a large portion of how customers received video services.

If a customer is a digital video subscriber they likely have TVs, in fact likely more TVs, which are served with no STB at all, and receive a direct coax connection. This is a valuable service feature for the MSO. However, we do recognized that many MSOs are considering the deployment of DTAs to recover analog spectrum, if the MSOs do a full all digital service and have no analog, this will make a migration to High-split a stronger consideration.

Additionally, MSOs could expand to 1050 MHz or even 1125 MHz perhaps using the existing passives, this very important because the technical benefits of using the bandwidth around 1 GHz are superior for the forward path compared with placing the return approaching or above 1GHz, discussed in detail in this analysis.

If the main challenges with the use of High-split are overcome, this seems like the ideal location for the new upstream (technically). The economics are also compelling for High-split against the other split options considering just the network access layer.

If the STB Out of Band (OOB) and analog recovery need to be factored into to the High-split, the cost analysis will change, however these will continue to be phased out of the network. The costs to move analog services, which are non-STB subscribers, were not considered in the model. However many MSOs are already planning to use DTAs to reclaim the analog spectrum, this would make a migration to High-split more obtainable.

The High-split option may need to exceed 200 MHz and move to approximately 5-238 MHz to achieve a MAC Layer throughput around 1 Gbps. This would use the 22.4 MHz of spectrum in the existing Sub-split band and the new spectrum up to 238 MHz to allow thirty-three (33) 6.4 MHz wide DOCSIS 3.0 channels all using single carrier 256-QAM all in a channel bonding group.

2.3 Top-split (900-1125) Plus the use of Sub-split

Overview

A new spectrum split called Top-split (900-1125) defines two separate spectrum bands, which may use sub-split plus the new spectrum region of 900-1125 MHz for a combined upstream band. The total upstream capacity may be 262 MHz depending on the lower band frequency return selected and if the passives will allow 1125 MHz to be reached. The downstream would begin at either 54 MHz or 105 MHz and terminate at 750 MHz in the current specification.

All of these architectures will share a 150 MHz guard band between 750-900 MHz, this may vary in the end-state proposal however these defined spectrum splits will be used for our analysis. The placement of additional upstream atop the downstream has been considered for many years.

The Top-split (900-1125) approach may be similar to a Time Warner Cable trial called the Full Service Network in the mid 1990's, which is believed to have placed the upstream above the 750 MHz downstream. These are some of the pros and cons of Top-split (900-1125):

Pros

- Operates at a typical 500 HHP node group but with no more than QPSK (see details in the sections later in this analysis)
- Top-split with Sub-split DOCSIS QAM MAC layer capacity ~315 Mbps given a 500 HHP Node/Service Group
- Top-split with Mid-split DOCSIS QAM MAC layer capacity ~582 Mbps given a 500 HHP Node/Service Group (less than High-split)
- Top-split 900-1125 does operate at a 500 HHP node but may operate at not full spectrum and will only be able to utilize 24 channels at 6.4 widths.
- Top-split (900-1125) plus Sub-split using DOCSIS QAM has an estimated MAC layer capacity of ~932 Mbps given a 16 HHP Node/Service Group
- With Sub-split "no" video services, devices, and capacity is touched
- STB OOB Communications are not affected
- Estimated that most passives will not be untouched (only Top-split that avoids touching passives)
- Existing 750 MHz forward transmitters are leveraged

Con

- The absolute major disadvantage for Top-split is cable network architecture requirements to make the solutions possible and the demands to reach high data capacity push FTTLA.
- A major finding of this report found that the effects of noise funneling force smaller and smaller node service groups to increase data capacity regardless if this is a DOCSIS / HFC solution or Ethernet over Coax (EoC) solution
- FTTLA is really fiber to All Actives, this will increase the number of node (HFC or EoC) to approximately 30 times the level they are now to reach the capacity level that High-split can reach with just the existing 500 HHP node location

- High-split can work at a 500 HHP node and while Top-splits must reach 16 HHP (FTTLA) depending on spectrum/cable architecture more HHP or even less than 16 HHP to reach the equivalent data capacity, lots of dependencies.
- Top-split from deployment perspective can be a challenge different cable type and distances play a major role is the architectures performance even if FTTLA is deployed
- No products in the market place to determine performance or accurate cost impacts
- 16 HHP upstream Service Groups will be required to approach 1 Gbps speeds comparable to High-split (238)
- Spectrum Efficiency is a concern because of guard band (wasted spectrum) and lower order modulation (less bits per Hz) resulting in lower throughput when measured by summing the upstream and downstream of Top-split (900-1125) and High-split using similar spectral range.
- High-split has nearly 20% more capacity for revenue generation when compared to Top-split (900-1050) plus Mid-split at a 500 HHP node, this is because the guard band requirements waste bandwidth and low order modulation for Top-split
- Upstream is more of a challenge compared to using that same spectrum on the forward path
- Upstream is more of a challenge compared to using that same spectrum on the forward path (cable loss ~5x Sub-split, 2.3x High-split; ~+/-1 dB/amp level delta w/EQTs is unknown)
- Interference concerns with MoCA (simply unknown scale of impact but may affect downstream in same spectrum range)

Assessment

The major consequence of the Top-split approaches, which use frequencies that approach or exceed 1 GHz, will have significant network cost impacts when compared with High-split. The number of nodes will increase 30 times to yield same capacity of High-split.

However, the Top-split (900-1125) options are being considered because option keeps the video network "as is" when considering sub-split and has marginal impact if mid-split is used. The Top-split 900-11125 option has additional benefits in that the Set-top box out of band (OOB) challenge is avoided and this option does not touch the passives.

This Top-split is estimated to cost more than the High-split. However, not included in this analysis is an economic forecast of the cost for Top-split to reach 1 Gbps upstream capacity which is estimated to be a 16 HHP architecture, the analysis examined economics 500 HHP and 125 HHP node architecture.

The migration for FTTLA to achieve 1 Gbps, would be 16 HHP and require all amplifier locations, thirty (30) in our model, to be a node location and this will require unground and aerial fiber builds to all locations. The MSOs will just begin to evaluate this option against the others.

2.4 Top-split (1250-1550) with Sub-split Overview and Top-split (2000-3000)

Systems designed to leverage unused coaxial bandwidth above 1 GHz have been around for many years. New iterations of these approaches could be considered to activate currently unoccupied spectrum for adding upstream.

The primary advantages of the top split are operational considerations – leaving current service alone – and the potential of 1 Gbps capacity or peak service rates in unused spectrum. In theory, not interrupting legacy services makes an IP transition path non-intrusive to customers, although the plant implications likely challenge that assertion.

The Top-split (1250-1700) Architecture will be defined as part of the 1250 – 1750 MHz spectrum band. Top-split (2000-3000) In our analysis we limited the amount of spectrum allocated for data usage and transport to 450 MHz and defined the placement in the 1250–1700 MHz spectrum band.

The allocation of 450 MHz provides similar capacity when compared to the other split option. The main consideration for this Top-split option is that it avoids consuming existing downstream spectrum for upstream and avoids the OOB STB communication channel

2.4.1 Implementation Complexity

A key additional complexity to the top split is working the spectrum around or through existing plant actives, all of which are low-split diplex architectures. For top split, a new set of actives supporting a triplex, or a bypass approach, or an N+0/FTLA are necessary to make the architecture functional.

All of these are intrusive, and have heavy investment implications, with the latter at least consistent with business-as-usual HFC migration planning. The top split is best suited to N+0 due to the complexity of dealing with current plant actives as well as for link budget considerations. N+0 at least removes the need to developing new amplifiers for the cable plant.

By contrast, node platforms have been and continue to evolve towards more features, functions, and flexibility. Of course, N+0 can be leveraged as a high-performance architecture whether or not a top split is implemented – top split, however, practically requires it to succeed as an architecture.

The outside plant architecture is not the only architecture affected by the approach. With the emphasis on upstream loss and degraded SNR as a primary issue for top split, a top split also virtually demands a point-of-entry (POE) Home Gateway architecture.

The variability of in-home losses in today’s cable systems would seriously compound the problem if a top split CPE was required to drive through an unpredictable combination of splitters and amplifiers within a home.

The above issues apply to the case of Top-Split (900-1125) as well, but to a lesser degree with respect to RF attenuation and the inherent bandwidth capabilities of today’s passives.

2.4.2 Spectrum Inefficiency

The penalty of the triplex architecture in terms of RF bandwidth and capacity can be substantial. A triplex used to separate current downstream from new top split bandwidth removes 100-200 MHz of prime

CATV spectrum from use in order that a less capable band can be enabled.

This spectrum trade reduces the total aggregate capacity of the plant. Under the assumption used (MPEG-4 HD/IPV), approximately 90 channels of 1080i HD programming are lost to guard band loss in a top split implementation compared to a high split alternative.

A primary objective of an HFC migration plan is to optimize the available spectrum, extending the lifespan of the network in the face of traffic growth for as long as possible, perhaps even a “forever” end state for all practical purposes that is competitive with fiber. RF spectrum in the prime part of the forward band is the highest capacity spectrum in the cable architecture.

To architect a system that removes on the order of 100 MHz from use is a loss of significant capacity, as quantified above, and works against the objective of optimizing the long-term spectrum efficiency.

The above issues apply to the case of Top-Split (900-1125) as well, but to a somewhat lesser degree associated with the percentage of crossover bandwidth required – that number is slight lower when the top split band chosen is slightly lower.

Pros

- Top-split 1250-1700 with Sub-split DOCSIS QAM MAC layer capacity ~516 Mbps given a 125 HHP Node/Service Group
- Top-split 1250-1700 with Mid-split DOCSIS QAM MAC layer capacity ~720 Mbps given a 125 HHP Node/Service Group
- Top-split (1250-1700) plus Sub-split using DOCSIS QAM has an estimated MAC layer capacity of ~883 Mbps given a 16 HHP Node/Service Group
- Top-split (1250-1700) plus Sub-split using DOCSIS QAM has 716 Mbps MAC layer capacity of ~1.08 Gbps given a 16 HHP Node/Service Group
- With Sub-split “no” video services, devices, and capacity is touched
- STB OOB Communication is not affected
- Placing the upstream spectrum beginning at 1250 MHz and up allows for the expansion of capacity without impacting the downstream

Cons

- Much higher upstream loss = significantly more CPE power = lower modulation efficiency (less bps/Hz) for equivalent physical architecture
- Need to work around legacy plant devices incapable of processing signals in this band
- Altogether new CPE RF type
- New technology development and deployment risk
- Large lost capacity associated with triplexed frequency bands
- Bottlenecks downstream growth when used as an upstream-only architecture
- Let’s elaborate on some of the key disadvantages identified above for an upstream top split
- Will operate at a typical 500 HHP node group but only capable of three of the

- 16 HHP Node and Use Mid-split and Sub-split spectrum meet the 1 Gbps capacity
- Highest cost solution compared with High-split and Top-Split (900-1050)
- The Top-split (1250-1700) with Sub-split cost more than High-split (200) and requires FTTLA
- No products in the market place to determine performance or accurate cost impacts.
- Return Path Gain Level Control: (cable loss >6x Sub-split, 2.8x High-split; +/-2 dB/amp w/EQTs is unknown)
- Interference concerns with MoCA (simply unknown scale of impact but may affect downstream in same spectrum range)

Assessment

The Top-split (1250-1550) with Sub-split is far more costly of High-split for the same capacity. The placement of the return above 1 GHz requires the passives to be replaced or upgraded with a faceplate change. There are approximately 180-220 passives per 500 HHP node service group.

A 500 HHP will not support Top-split 1250-1550, so the initial architecture will have to be a 125 HHP. However the requirements for higher capacity will force smaller node service group, which will add to the cost of the solution. The use of lower order modulations will require more CMTS upstream ports and more spectrum, which will impact the costs of the solution as well.

Additionally, the conditioning of the RF components to support above 1 GHz may add to the costs of the solution. However determining the financial impacts of performing "Above 1 GHz plant conditioning" is unknown and was not considered in the financial assessment found later in this report.

The economic estimate used for Top-split was for 500 HHP and 125 HHP node architecture. The migration for FTTLA to achieve 1 Gbps, would be 16 HHP and require all amplifier locations, thirty (30) in our model, to be a node location and this will require unground and aerial fiber builds to all locations. This was not provided in the analysis.

Lastly, there is a significant penalty to downstream bandwidth in the form of triplex guard band – on the order of 100 MHz of RF spectrum is made unavailable for use. In the case of Top Split (900-1125), the band eliminated consists entirely of prime, very high quality forward path spectrum.

If we consider the service and network capacity requirements for the upstream and downstream for the next decade and beyond, the cable industry should have sufficient capacity under 1 GHz, which is the capacity of their existing network.

2.5 Summaries for Cable Spectrum Band Plan

Continuing to leverage the current downstream and upstream spectrum will force operators to reduce service group size by using node splits and/or segmentation. This is ideal for MSOs that want to avoid re-spacing the amplifier network.

Additionally, spectrum changes will undoubtedly require service outages, because all the electronics and even passives (if above 1 GHz is selected) would have to be touched. Spectral changes may have higher service down time compared with node segmentation or node splits.

MSOs may want to consider spectrum expansion where node splits are costly. Depending on spectrum selection, the MSO could maintain large service group in the optical domain. In others words, the optical node could service a larger area and number of customers, if the MSO selects low frequency returns such as Sub-split, Mid-split, or High-split and if additional downstream spectrum is selected this will increase the length of time a optical node can support a given service group.

The channel allocation of video and data services will define the spectrum needs and node migration timing. Additionally, the service offering, such as network based PVR, will impact the spectral usages; thus drives toward more spectrum or smaller services groups.

There really are lots of levers that will drive the MSOs to changing spectrum and/or service group reductions, predicting with all certainty of how long a given network will last is greatly influenced by services and legacy devices that may need to be supported.

The legacy STB out of band (OOB) communications which uses spectrum in the High-split area will be a problem for this split options; however a mid-split as the first step will provide sufficient capacity for nearly the entire decade according to our service and capacity predictions. The thinking is that another decade goes by and the legacy STBs may be few or out of the network all together.

If the STBs still remain in service, another consideration is that these legacy STB may be retrieved and relocated to markets that may not need the advanced upstream spectrum options. Yet, another consideration is a down conversion of the OOB communications channel at the last amp or homes that have legacy two-way non-DOCSIS set-tops.

2.6 Spectrum Options, Capacity, and Timing Implications

We have discussed the Pros and Cons of the various upstream spectrum options. As discussed in Section 2.1, it is well-understood that a limitation of the 85 MHz mid-split architecture is that it cannot achieve 1 Gbps of capacity, at least not easily or in the near term. We will discuss upstream capacity itself in detail in Section 9.6 "Upstream Capacity".

While 85 MHz cannot achieve 1 Gbps of capacity, it is also not reasonable to jump to high-split in the near term because a plan must be in place to deal with the OOB channel, as shall be further described in Section 3.3.5 "Legacy OOB" and Section 3.4 "The Legacy Mediation Adapter (LMA)". As such, MSOs appear to be in a bind for handling upstream growth. Or, are they?

Let's consider defining the 1 Gbps requirement for upstream data capacity. How would such a system fare in supporting long-term capacity requirements? We can easily quantify how this would help manage long-term traffic growth and compare it to examples like the 85 MHz Mid-Split.

This comparison is examined in Figure 2. It shows three threshold cases – 100 Mbps (A-TDMA only), 85 MHz Mid-Split (in this case, including use of S-CDMA), and the case of 1 Gbps of capacity, however we manage to achieve it (high-split or top-split).

Zeroing in on the red arrow identifying the gap between Mid-Split and 1 Gbps at 40% CAGR – very aggressive relative to 2011 observed growth rates – in each case with a node split assumed in the intervening years, we see that there exists about 2.5 years of additional growth. When we think of 1 Gbps, this intuitively seems odd. Why does migrating to Mid-Split buy a decade or more of traffic growth coverage, yet implementing a 1 Gbps system offers only a couple more years of survival on top of that decade?

This "linear" time scale on the y-axis is simply exemplifying how multiplicative compounding works. It is up to our own judgment and historical experiences to consider how valid it is to be guided by the compounding rules of CAGR originally identified by Nielsen, and if so what reasonable year-on-year (YOY) behavior assumption to assume.

However, the mathematical facts of CAGR-based analysis are quite straightforward: with CAGR behavior, it takes many YOY periods to grow from, for example, 5 Mbps services today, consuming or engineered for perhaps tens of Mbps of average return capacity, up nearly 400 Mbps or more. We will outline the data capacity possibilities for 85 MHz Mid-Split in Section 9.6, and then show a specific implementation in Section 7.1.2. However, once a 400 Mbps pipe has been filled, the subsequent annual steps sizes are now large. Because of this, consuming 1 Gbps is not many YOY periods of growth afterwards.

To demonstrate, we can calculate an example using 20 Mbps of average capacity satisfying demand today. At this aggregate demand, traffic can double four times and not eclipse 400 Mbps. It eclipses it in the 5th traffic doubling period. For ~40% CAGR (two years doubling), that's a total of ten years. For a CAGR of 25%, its about 15 years.

This is what Figure 2 is pointing out graphically. As such, relative to a solution that provides 1 Gbps, Mid-Split gets us through 80% of that lifespan under the assumption of an aggressive 40% CAGR and an intervening node split.

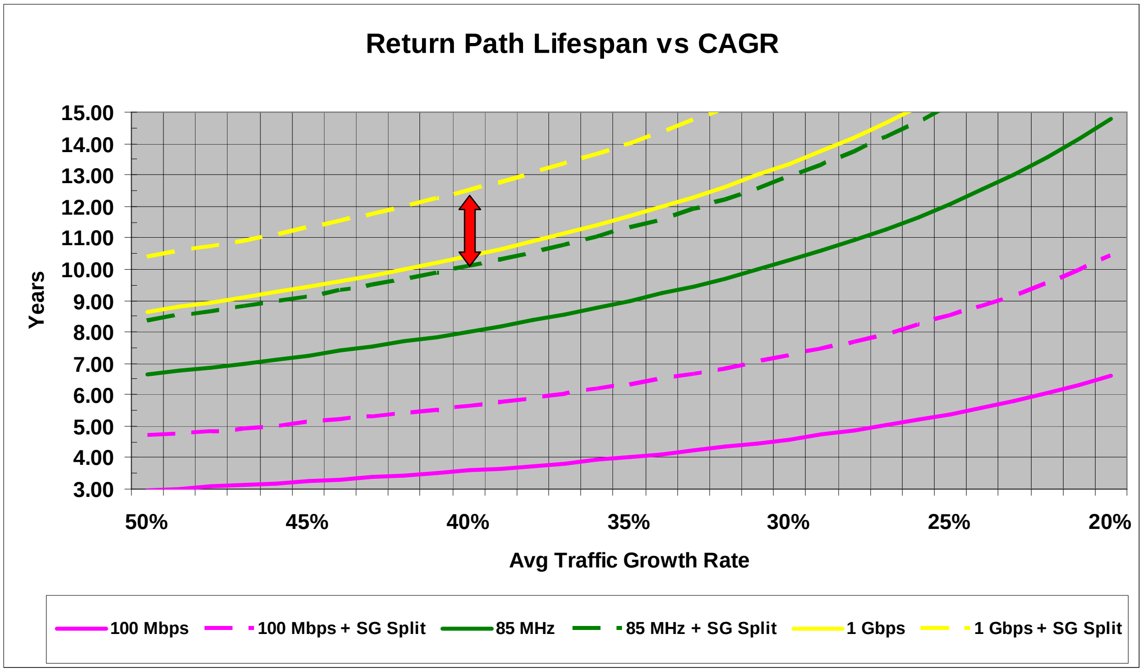


Figure 2 – Years of Growth: A-TDMA Only, 85 MHz Mid-Split, 200 MHz High Split

This Mid-Split vs. 1 Gbps lifespan analysis is an illustrative one in recognizing the long-term power of the 85 MHz Mid-Split. It provides nearly the same growth protection as a 1 Gbps solution would, if there even was one available. This means that the 1 Gbps requirement comes down to an operator's own considerations regarding the competitive environment, and whether a 1 Gbps market presence or service rate is important to their positioning for residential services.

3 SOLVING LEGACY ISSUES

3.1 Introduction

In order to significantly increase the upstream throughput in a DOCSIS system, more upstream spectrum is needed. That spectrum has to go somewhere. This white paper has examined multiple spectrum solutions and then different technology options within each spectrum solution.

Solutions are needed that allow an HFC plant to be migrated over to the next generation of DOCSIS without a full-scale replacement of subscriber equipment. Legacy and new equipment must co-exist in the same network.

The high level summary of the different spectrum solutions and their challenges is shown in Table 2.

This paper recommends mid-split and high-split as the best technical solutions. The attractiveness of top-split is that it interferes less with existing services. If the logistical problems of mid-split and high-split could be solved, then cable operators would be able to choose the best technical solution.

This section is going to specifically look at addressing the major logistical problems that the mid-split and high-split band plans face.

3.2 Summary of Operational Issues

Table 3 is a summary of the operational issue faced by each of the four upstream bandwidth solutions. This table is taken from [21].

There are several logistical challenges that are obstacles to the deployment of mid-split and high-split systems into an HFC plant that was designed for sub-split. The challenges include:

- Analog video
- FM band
- Aeronautical band interference
- Adjacent device interference
- Legacy OOB

Let's look at each one of these challenges in more detail.

Table 2 – Upstream Spectrum Comparison

Approach	Frequency	Comments
Sub-Split	5 - 42 MHz	Existing installed HFC plant. Add bandwidth with node splits.
Mid-Split	5 - 85 MHz	Technology available today with DOCSIS 3.0 CMTS and CM.
High-Split	5 - 200+ MHz	Best technical solution but challenging logistical solution
Top-Split	> 1 GHz	Tough technical solution but more attractive logistical solution

3.3 Analysis and Solutions

3.3.1 Analog Video

Problem Definition

There are many different channel plans in use around the world today. This white paper will choose the North American cable television plan as a specific example. This channel plan is defined in [20] and described in [18]. The upstream frequency cut-off is a maximum of 42 MHz. Some systems use a lower cutoff, depending upon the age of the system.

The downstream frequency range starts at 54 MHz. By convention, the analog

Table 3 – Summary of Operational Issues

Approach	Pros	Cons
Sub-Split	• All equipment already exists • No disturbance to spectrum • Simple	• Cost: Requires deeper fiber. • Cost: Requires more CMTS ports • Cannot hit peak rates over 100 Mbps of return path throughput
Mid-Split	• Supported by DOCSIS 3.0 equipment • Works with DS OOB	• All actives and some passives in HFC plant need to be upgraded • Cost about the same as high-split and only doubles the US throughput • Removes ch 2-6 of analog TV
High-Split	• Supports 1 Gbps throughput • Can co-exist with earlier versions of DOCSIS.	• All actives and some passives in HFC plant need to be upgraded • Does not work with DS OOB • New CM and CMTS components • Removes ch 2-36 analog TV • Removes FM band (issue in Europe)
Top-Split	• Leaves existing plant in place. • No impact to existing legacy customer CPE • Only customer taking new tiers would require new HGW CPE	• Requires triplexers • New active return path has to be built on top • High attenuation requires high RF power. Existing amplifier spacing may not be sufficient • Blocks expansion of downstream bandwidth directly above 1 GHz

channels are first in the spectrum followed later in frequency by the digital channels. The classic analog line-up is contained in channels 2 through 78 that occupy the spectrum from 54 MHz to 550 MHz. Within this spectrum are also channels 1 and 95 to 99.

The definition of the frequencies for a mid-split system has changed over the years. The mid-split for DOCSIS 3.0 is not exactly the same as legacy systems that used a return path upper frequency limit of 108 MHz ~ 116 MHz, with the downstream spectrum starting at 162 MHz~ 174 MHz (the actual frequencies varied among vendors).

The DOCSIS mid-split downstream frequency range starts at 108 MHz, which disrupts channels 1, 2-6 (54 MHz-88 MHz), and 95-97 (90 MHz-108 MHz) would be disrupted. A natural break point from a channel perspective would be to start the mid-split lineup at channel 14 a(120 MHz-126 MHz). If so, then channels 98-99 (108 MHz-120 MHz) would also be disrupted. Note that channels 7 through 13 (174 MHz-216 MHz) are located above channels 14 through 22 (120 MHz-174 MHz).

The upstream frequency range for high-split has not been chosen yet. If the high-split downstream frequency spectrum started at 300 MHz, then channels 1-36 and 95-99 would be lost.

Solutions

The first solution is to get rid of analog TV altogether on the cable spectrum. Any legacy TV that cannot receive direct digital QAM would have to be serviced with a digital transport adapter (DTA) or a conventional set-top box (STB). As radical as this idea may seem, several cable operators such as Comcast and CableVision are already free of analog channels on parts of their plants with plans to expand their no-analog foot print. The governments of many countries, including the USA, have already turned off most over the air analog broadcasts.

It costs money to retain analog channels. It is not that the money is spent on the analog channel equipment - which obviously is already paid for - it is that money needs to spent elsewhere to improve spectral efficiency. This may include plant upgrades, equipment upgrades or both.

Analog TV has only 5% of the efficiency of an MPEG-4 over IP video signal, yet analog TV typically occupies over 50% of the downstream spectrum. RF spectrum is always a scarce commodity, and this is a good example of where there can be a significant efficiency improvement.

The second solution would be to reduce the analog channels down to a smaller group of, say, 25 core channels. Then remap those analog channels into a higher channel space. For mid-split, only channels 2-6 need to be remapped. For high-split, it would be channels 2-36.

This may cause some channel confusion to the subscriber, but such a remapping trick has been done for high definition channels on STBs.

A semblance of continuity can be maintained by keeping the least significant digit the same. Remapping channel 2 to channel 62 is one example.

There are often contractual issues quoted, such as franchise agreements, market recognition, must-carry agreements, etc. These may have to be renegotiated. The driving force for doing so is a gigabit or more upstream speed. To the extent that these legal requirements are driven by the requirements of the community, then which is the bigger market - analog TV or an incredibly fast Internet access? The answer has to be a fast Internet service or there would not be a need to upgrade in the first place.

Finally, now that the government has shut down most over-the-air analog TV, the cable operators are the last service provider to have analog TV. The telco and satellite service providers are all digital.

There are two perspectives that can be taken on this. The first is that having analog TV makes the cable operators unique in being able to offer analog TV, and this differentiates them from all the other providers. The second is that the cable operators are the last to move to all digital, and that the other service providers may have more spectrum or resources as a result.

So, again, if the costs are equal, does analog TV with a lower Internet access speed beat out a competitor who has a significantly higher speed Internet service? What if the competitor is a fiber-to-the-home company with gigabit-per-second service?

The choice is somewhat obvious, but also very painful. It requires pain of some sort. But, the new upstream spectrum has to come from somewhere. Keeping analog TV spectrum indirectly costs money due to investment on alternative solutions.

3.3.2 FM Band

Problem Definition

The FM radio band is from 88 MHz to 108 MHz. There are two potential concerns.

The first concern is the loss of the ability for the cable operator to provide FM radio service over the cable system. This is not much of an issue in North America, but it is a concern in Europe and elsewhere.

The second concern is if interference generated by the HFC plant that might interfere with the FM band (signal leakage) or if the FM band might interfere with the with the HFC plant (ingress).

Solutions

As with analog TV, the easiest solution to the first requirement is to no longer carry the content. For Europe, this may require some regulatory work. The worst case would be to carry the FM band at a higher frequency on the HFC plant and down-convert it locally with the LMA. Refer to Section 3.4.

As far the HFC plant interfering with local FM reception, this should not be a problem. The capture effect of FM receivers [24] will most likely reject noise-like digital signals leaking from a cable network as a weaker signal. A strong FM signal might interfere with the upstream signal on the HFC plant. This can be mitigated with good plant shielding, ingress cancellation techniques, or OFDM noise/ingress mediation.

3.3.3 Aeronautical Interference

Problem Definition

The new CM will be transmitting at frequencies above 54 MHz at a higher power level than when the frequencies are transmitted as part of the downstream spectrum. The inherent leakage in the plant might be sufficient enough to cause interference with existing services.

For example, the frequencies from 108 MHz to 137 MHz are used for Aeronautical Mobile and Aeronautical Radio Navigation.

The radio frequency spectrum usage is shown in Figure 3. [23]

Specifically, the 108-118 MHz band has always been problematic because any CATV signal leakage here could interfere with aviation localizer (108-110 MHz) and VOR signals (110-118 MHz). Hence, sometimes channels 98 and 99 (also called A-2 and A-1) are not used to avoid this problem. The localizer is especially important, as it is responsible for providing the left/right guidance in an ILS approach; VORs are also important but more often used at longer ranges as navigation beacons.

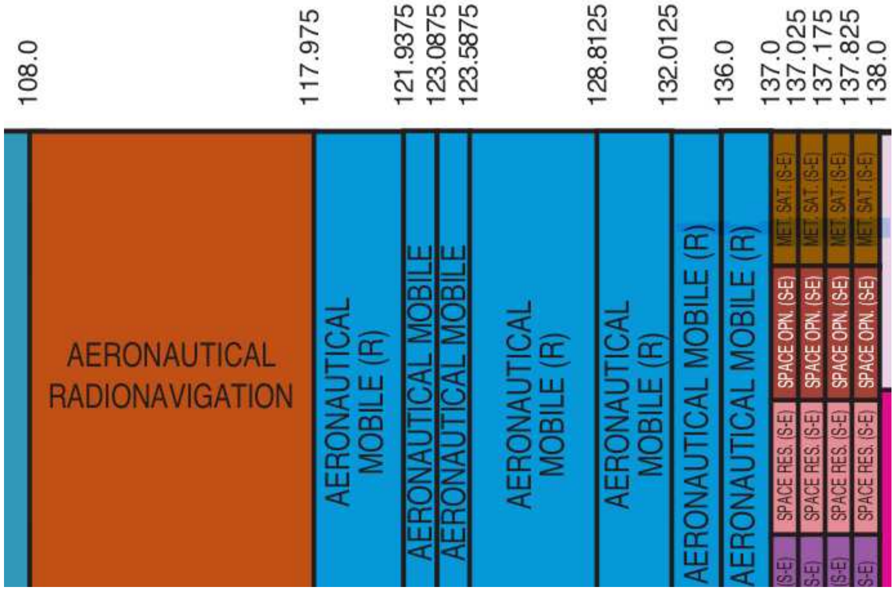


Figure 3 – Government Spectrum Allocation from 108 MHz to 138 MHz

There is also the 121.5 MHz aeronautical emergency frequency, and the 243.0 MHz distress (SAR) that may be of concern.

If the upstream spectrum expands above 300 MHz, another sensitive aviation band comes into play. The glideslope frequencies are in the 328-335 MHz band. The glideslope is the x-y counterpart to the localizer as it provides up/down guidance in an ILS approach.

Solutions

Research would have to be done to validate these concerns. If it is a problem, then the plant will have to be cleaned up to reduce this leakage. Some of this leakage may come from bad home wiring. That makes it even more important that the CM installation is done professionally.

In the absolute worst case, some or all of these frequencies would have to be avoided. The impact of that is that a larger upstream spectrum would have to be dedicated to DOCSIS. This would be a loss of up to 29 MHz or more in some networks.

Some of these interfering carriers are quite narrow. Current DOCSIS tools handles very narrow interferers better than modulated, but increasingly struggles as multiple interferers occupy a single carrier band. OFDM will be quite useful for notching these out.

This concern also existed 15 years ago prior to the deployment of DOCSIS. The plant did require cleaning up in many cases. It was done and the result was a more reliable HFC plant. So, it is doable, but must be planned and budgeted for.

3.3.4 Adjacent Device Interference (ADI)

Problem Definition

ADI refers to the situation where the operation of one device - such as a high-split cable modem - interferes with another device - such as a legacy TV or legacy set-top box. This is not an official abbreviation (yet). We are borrowing the concept from the term adjacent channel interference that describes a similar phenomenon, except ACI is in the frequency domain, and ADI is in domain of physical space.

For the sake of example, let's assume the high-split spectrum goes up to 230 MHz, and the downstream starts at 300 MHz.

Tuners in STBs and TVs in North American receive above 54 MHz with an expected maximum per-channel input power of +17 dBmV. Low-split and top-split can thus co-exist fine with legacy tuners. Mid-split and high-split systems carry RF energy in the upstream direction that is within the downstream operating range of the legacy STB and TVs.

If those devices are located near a CM that is blasting out energy above 54 MHz at levels approaching +57 dBmV (DOCSIS 3.0 max power for single 64-QAM), poor isolation and/or return loss in splitters and other devices could cause a significant amount of that upstream power to appear at the input connector of the legacy devices, which might saturate their RF input circuits, thus preventing the devices from receiving a signal at any frequency.

The typical North American legacy tuner has an output intermediate frequency (IF) centered at 44 MHz. If 44 MHz was applied to the input of a tuner with poor IF rejection, that signal might cause interference in the tuner, even through the tuner is tuned to another band. How much of a problem this is requires more research.

There is some evidence that shows that the sensitivity of the video signal to ADI decreases significantly as analog signals are replaced with digital. This is a somewhat intuitive conclusion, but validating data to this effect is important.

Solutions

So, what to do?

One solution is to put a filter in front of the legacy devices that filters out all content below the high-split cut-off frequency (85 MHz or 230 MHz in this example). But, is this filter needed in all cases? And where would the filter go? Let's look at this problem in more detail.

The general problem is best split up into two smaller scenarios:

- Impact within the same home as the new high-split DOCSIS CM.
- Impact to adjacent homes that do not have the new high-split DOCSIS CM

Same Home:

When a home is upgraded, the new DOCSIS CM will likely be installed as a home gateway (HGW). There are at least two scenarios. The first is a home with MPEG video STBs, and the second scenario is an all IP video home.

In the home that requires digital MPEG video, the HGW can receive the spectrum from the plant, filter the signal below 200 MHz, and pass the filtered spectrum into the home. The main filtering it is trying to achieve is from its own upstream transmitter. If the upstream transmitter is

+50 dBmV, the internal combiner has 20 dB of signal rejection, and the max signal level allowed is +15 dBmV, then the additional filtering has to provide 15 dB of attenuation. This filter could be located internal to the HGW or be an external inline filter in order to manage HGW costs.

For this to work, the HGW would have to be wired in-line with the home. That is not how CMs are installed today. CMs today are installed using a home run system. The drop cable from the street is split between the CM and the home. In this new configuration, the CM would have to have a return cable that then fed the home. This could add additional loss to the video path. However, it could be a workable solution.

In the home where there are only IP STBs, the downstream from the HFC plant does not have to be connected to the home. DOCSIS could be terminated at the HGW and the HGW would drive the coax in the house with MoCA. Video and data would be deployed with IP STBs that interfaced to the MoCA network.

The HGW becomes a demarcation point between DOCSIS and the cable plant on one side, and MoCA and the home network on the other side. Again, the CM would have to be in-line with the coax from the drop cable and the home. This does imply the need for a professional installation.

This is an interesting proposal in several ways. First, it solves the in home legacy tuner interference problem. Second, it isolates all the return path noise generated by the home network and prevents it from entering the HFC plant.

Adjacent Home

The other half of the problem is the impact to adjacent homes. While the installer has access to the home he is upgrading and has several options available to him, the home next door may not be part of the upgrade.

The energy from the new high-split CM would have to travel up the drop cable from the home, travel between the output ports on the tap plate, back down the drop cable to the next house, and then into the home network of the next house.

The easiest solution would be to set the new upstream power budget such that the signal would be sufficiently attenuated by the path described above so that it would not be a problem. This solution becomes harder when the customers are in a multiple-dwelling unit (MDU) where the coax drops may be shorter.

Worst case, in-line filters would have to be applied in-line with the drop cables of the adjacent home or within the adjacent home. Another approach is to put filters into the tap plate that serves an upgraded home and its adjacent homes. This would prevent the upgraded home from impacting the adjacent homes.

Thus, tap plates would only have to be replaced as part of a new deployment so the overall cost would be lower than having to replace them all at once. This assumes that the additional upstream path attenuation between taps on separate enclosures is sufficient.

As far as potential tuner sensitivity, the upstream spectrum could skip the frequencies from 41 to 47 MHz. This can be done, but it is a loss of 6 MHz of spectrum. The better plan is to make sure that the attenuation of the upstream signals into the downstream is sufficient that even 41 to 47 MHz is fine.

Summary

In summary, an external filter may not be needed. The HGW can be used to protect the upgraded home, although it has to be wired in line. The adjacent home should have enough attenuation from the drop cables and tap assembly. More caution may be needed in MDUs. An external in-line filter should be made available to fix the exception condition. Filtered taps may be good for dense situations such as MDUs.

3.3.5 Legacy OOB

Problem Definition

The out-of-band (OOB) channel is used on legacy STB to provide information to the STB and get information back. The OOB channel was used prior to the development of DOCSIS Set-top Gateway (DSG).

The downstream carrier is 1 MHz wide for SCTE 55-2 (Cisco) and approx 1.7 MHz wide for SCTE 55-1 (Motorola). Typical placement of center frequency is between 73.25 and 75.25 MHz as there is a gap between channels 4 and 5. The older "Jerrold" pilot (prior to Motorola/GI) was at 114 MHz. By spec [25], the STB must be able to tune up between 70 MHz and 130 MHz.

There is an upstream OOB carrier as well that is usually placed below 20 MHz.

CableCards are one-way and typically use only a downstream OOB channel.

There are no compatibility issues with the STB OOB channel and low-split or top-split. For mid-split, if the OOB channel can be placed above 108 MHz in the downstream spectrum then the problem is solved. This should work except for very old STBs that are fixed frequency. These STBs would have to be upgraded.

For high-split, this is probably the biggest issue. The 200+ MHz target cutoff for high-split is well above the 130 MHz upper end of the OOB tuner range.

Solutions

This is primarily a North American issue. In the rest of the world where legacy STB penetration with OOB is much lower or non-existent, and may not be a significant issue.

Of the STBs deployed in North America, many of the newer ones can actually tune to a frequency greater than 130 MHz because it was just as cheap to use a full spectrum tuner. Cisco estimates that > 70% of the Tier 1 installed base of Cisco STBs in 2015 would have this capability. (Further research is required. Software upgrades may be required.).

Then there is DSG. DSG is basically OOB over DOCSIS. Many of the deployed STBs have DSG built in but the DSG has not been enabled. Cablevision is an exception who has 100% DSG deployed, as does South Korea. So, DSG is proven to work.

It turns out there was a financial hitch with DSG. The original plan was add the STBs to an existing DOCSIS upstream channel. These upstream channels are engineered to be transmitted from the CMs on a home run cable. The STBs in the home have more attenuation, as they are deeper into the home coax network, so they are not always able to transmit onto an existing DOCSIS channel.

The solution is to use a separate QPSK DOCSIS channel. If this channel were the same modulation and power level as the existing OOB channel - which it would be - then if the OOB upstream worked, the

DOCSIS OOB upstream would also work. The problem is that this requires a dedicated carrier in the CMTS. This might be additional expense or the CMTS may not have the extra capacity. With newer CMTSs, there will be more upstream carriers available, so dedicating one carrier per port to DSG is a very reasonable solution.

It is also reasonable that any home that gets upgraded to a new high-split CM could also have their STBs upgraded to DSG compatible STB.

The OOB CableCard is easily replaceable and can migrate to DSG.

So that leaves STBs in North America, in non-upgraded homes, that are over 10 years old (by 2015), that can't tune above 130 MHz, that are non-DSG, and are not CableCards. That is really not a lot of STB. It could be around 0% to 10% of the STB population rather than the originally estimated 100%.

There is a motivation to replace these old STBs. They are beyond their capital write-down period. Further, these STB usually do not have the CPU or memory capacity required to run new applications. This means that new services cannot be sold to these customers.

Just to be on the safe side, there is a solution that does not require upgrading the old STB. That solution would be to put an inexpensive LMA behind legacy STB that provided an OOB channel. These LMAs would go inline with legacy STB. They would be cheap enough that they could be mailed out to customers who complain or are known to have specific legacy STBs.

If that does not work, only then a truck roll might be needed.

Summary

At first pass, the loss of the OOB channel seems like a major problem. However, by the time the next generation of DOCSIS is deployed, and with the variety of solutions, it is not really a problem at all.

Bear in mind that before the first high-split CM can be used in the new spectrum, the plant needs to be upgraded. But after the plant is upgraded, homes can be upgraded on a per home basis. This helps keep costs contained.   Also, in a phased approach to upstream bandwidth expansion, a mid-split architecture may buy yet more time to eliminate or actively retire the older STBs.

This is a far better proposition than if all legacy STBs had to be replaced prior to upgrading the plant.

3.4 The Legacy Mediation Adapter (LMA)

In several of the plans to deal with legacy, there is a back-up plan that involves an in-line device that we will refer to as a legacy mediation adapter (LMA).

- The LMA could be used for generating and receiving an OOB signals.
- The LMA could be used for blocking upstream energy from entering the downstream.
- The LMA could be used to isolate the ingress originating from the home when the home no longer needs a return path internal to the home.
- The LMA could even be used to generate an FM signal for European deployments.

There are at least two primary ways of designing this LMA. The first way uses a

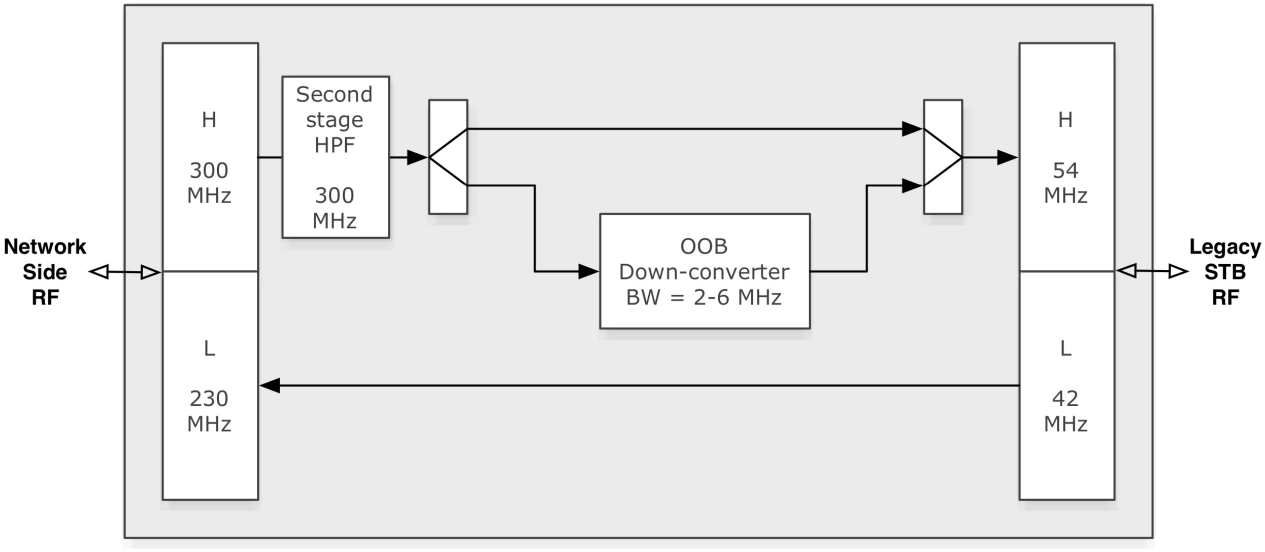


Figure 4 – LMA with Down-Conversion

simple down-conversion method. The second way uses an embedded circuit.

Another interest aspect of the LMA is that it interfaces between the new and old HFC spectrum plans. On the network side of the LMA, it interfaces into the high-split, 200 MHz (for example) plant. On the subscriber side of the LMA, it interfaces into the legacy sub-split 42 MHz plant.

3.4.1 LMA with Down-Conversion

In this approach, the headend would generate two OOB downstream carriers. The first one would be the standard downstream OOB carrier. This first carrier might be at 75 MHz for example.

The headend then generates a second OOB carrier at a frequency that is in the available downstream spectrum that is above the upstream cut-off frequency. This second carrier might be at 750 MHz for example.

This second carrier would fit into a 6 MHz or 8 MHz TV channel slot. This channel would be wide enough that multiple carriers could be fit. That way, any plants that are dual-carry with two STB manufacturers on it could be accommodated.

If necessary, the bandwidth could be expanded to allow for the FM band to be placed at a higher frequency as well.

The first carrier at the lower frequency would be received by legacy STB on areas of the plant that have not been upgraded. The second carrier would be received by the LMA that has been placed behind the legacy equipment.

The use of two carriers at different frequencies presumes a scenario where the LMAs are distributed over a period of time prior to the HFC plant upgrade. Thus, during the transition period, there would be legacy devices on both carriers.

A block diagram of the down-converting LMA is shown in Figure 4. Starting at the network side, the RF signal is separated with a diplexer into downstream and upstream frequency paths. The

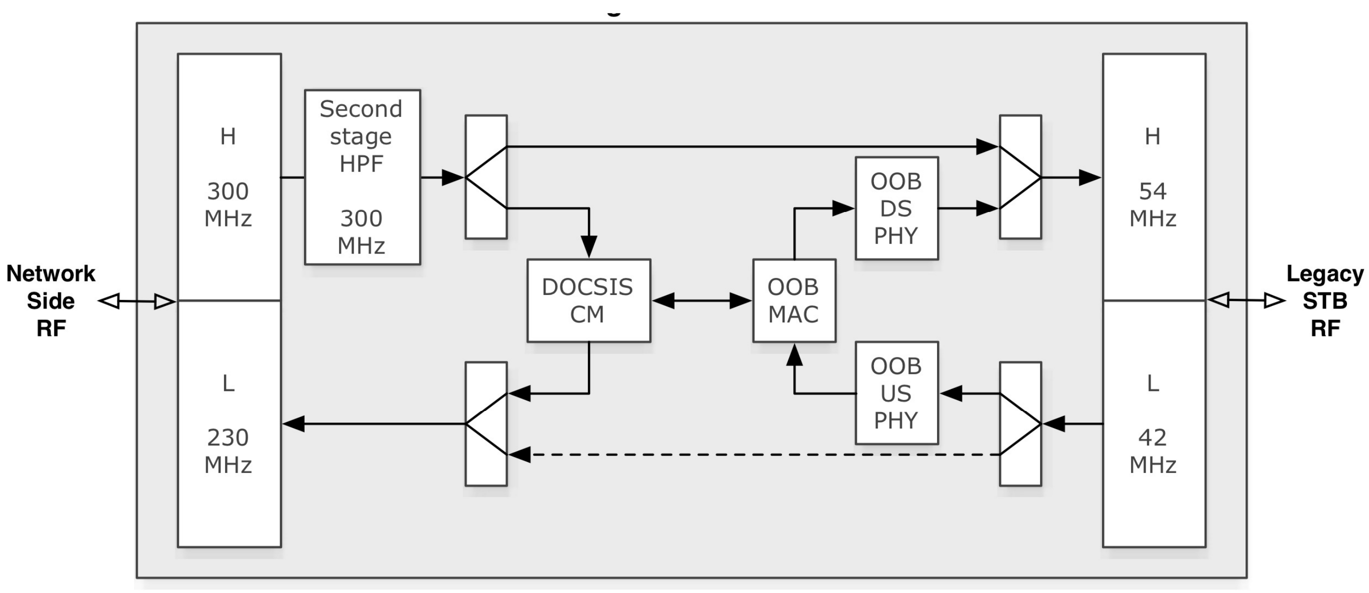


Figure 5 – LMA with CM

downstream path may require further filtering to remove any upstream energy.

The higher frequency OOB carrier is tapped off and passed to a down-converter. In the example used here, the down-converter would down convert from 750 MHz to 75 MHz. This carrier is then combined back into the downstream spectrum and then passed to the legacy STB.

To further reduce the cost of the LMA, the upper frequency that is used for the OOB carrier could be standardized through CableLabs. The LMA would then be a fixed frequency device and would not require any configuration.

The return path is left intact as the legacy STB will need to send an OOB carrier back to the headend.

3.4.2 LMA with DOCSIS CM

This approach achieves similar goals but with a different method. In this method, a DOCSIS CM is used to communicate the OOB information over IP from the headend to a local OOB circuit. This design would be good for operators who are using DSG as a baseline to control their network or for a scenario where the LMA needs to be configured.

DSG can be used on the network side in the downstream. Alternatively, a basic IP tunnel can be used to transport the raw OOB channel. An IP tunnel will have to be defined for the upstream that carries the upstream OOB information to the headend. This can be done at CableLabs.

The LMA has an entire two-way OOB MAC and PHY. This circuit generates a local OOB circuit. A clever implementation could implement both the SCTE 55-1 and SCTE 55-2 OOB standards. Otherwise, there would need to be two separate LMAs.

This design could use a DOCSIS 1.1 CM as part of a reduced cost implementation as only single carrier implementations are needed.

The return path from the home to the network could be disabled so that the LMA would isolate the ingress from the home from getting to the network.

3.5 Downstream Concerns

The downstream frequency band above 1 GHz will have a few challenges as well. In addition to the higher attenuation and micro-reflections, there are some frequency bands to be careful of. Here are two of the more common spectrum usages to be aware of.

3.5.1 MoCA®

MoCA is a technology that allows peer to peer communication across coax in a home environment. It typically is used for communicating between set-top boxes.

The concern would be that new frequencies on the cable plant above 1 GHz could interfere with MOCA in homes that are both upgraded to DOCSIS NG that don't isolate the HFC plant from the home and homes that are legacy.

MoCA 1.1 defines a 100 Mbps data channel that consumes 50 MHz of spectrum that can be located anywhere in between 1125 MHz and 1525 MHz.

MoCA 2.0 defines a 500 Mbps data channel that consumes 100 MHz of spectrum that can be located anywhere in between 500 MHz and 1650 MHz. MOCA 2.0 also has a special 1 Gbps data channel that is bonded across two 100 MHz channels.

A key observation is that MOCA does not occupy the entire operating frequency range. The large frequency range allows multiple MoCA system to coexist.

The most probably solution is to set aside some amount of downstream spectrum, say 200 MHz, for use by MoCA, and let MoCA find it.

3.5.2 GPS

GPS L3 (1381.05 MHz) is an encoded alarm signal broadcast worldwide by the GPS constellation. It is used by part of the US DOD Nuclear Detection System (NDS) package aboard GPS satellites (NDS description [29]). Encoding is robust and is intended for receipt by military ground-based earth stations. These installations are not susceptible to terrestrial signal interference (i.e. skyward-looking antennas).

Despite being so, large scale, wide area leakage into L3 (as from a distributed cable plant) would not be looked upon favorably by either the US or Canadian governments, or by radio astronomy organizations, who already suffer from GPS L3 signals corrupting “their” skyward-looking receive bands near 1381 MHz. [30]

In contrast, L1 (1575.42 MHz) and L2 (1227.60 MHz) are susceptible to terrestrial interference, despite CDMA encoding. This is due to the low-cost nature of the patch antennas and receivers used to detect them in consumer applications. Unlike the military receive systems and precision GPS packages used in commercial navigation (aviation and shipping), which are robust in the presence of terrestrial interference, consumer GPS are not so. Consumer GPS (including auto and trucking) navigation systems rely upon a wide-pattern patch antenna with a low-noise, high-gain preamplifier.

Such a configuration has no discrimination against terrestrial signals. The low level of received signal at the preamp creates a condition ideal for “blanking” of L1 and L2 should a terrestrial signal of sufficient spectral power density –

particularly from overhead cable plant – be present.

Finally, new applications of the latest civilian GPS frequency, L5 (1176.45 MHz), are currently emerging. Despite being CDMA encoded with FEC, it is not possible to predict how consumer receivers for this latest band will perform in the presence of broad-area interference.

It is of some interest to note that the target application for L5 is “life safety”, see [31]. To get a feel for a L1, L2, and L3 receiver architectures, see the following overview paper on civilian GPS receiver parameters, [32].

3.6 Summary

While initially there were many concerns about the logistics of implementing high-split, there are good mediation strategies. Analog video can be removed or remapped. Adjacent device interference should not be a general problem, and a filter LMA or tap plate filter can manage exception cases. Even the OOB channel is quite manageable with DSG or with an LMA.

This LMA can be multi-purpose and include OOB support and downstream high-split filtering. There may be other functions such as FM radio support that may also be interesting to consider.

The LMA has two different implementations. One is a down-conversion. The advantage is low cost, no ASIC needed, and re-use of OOB headend equipment. The second design could be low-cost if done right, requires ASIC integration, and is better suited to a DSG environment.

More research is needed on the impact to the aeronautical band and to the adjacent tuners below 54 MHz.

It is clear, however, that there are no logistical show stoppers in the deployment of a mid-split or high-split system.

4 COAXIAL NETWORK COMPONENTS AND TOPOLOGY ANALYSIS

The goal of any cable operator is a drop in upgrade to add spectrum capacity when needed. This saves time and money in resizing the network such as node and amplifier location and spacing. Adding network elements or changing network element locations will impact cost for electrical powering requirements. [35]

Ideally, the upgrade would touch the minimum number of network elements to reduce cost and time to market. In the section, the technologies, systems and architecture options are explored. The analysis will examine some of the pros and cons of several technologies and architectures, which could be used to provide additional capacity.

The analysis considered the capabilities of a "Drop in Upgrade" to determine the viability and impact for upstream spectrum expansion as a starting point. [35]

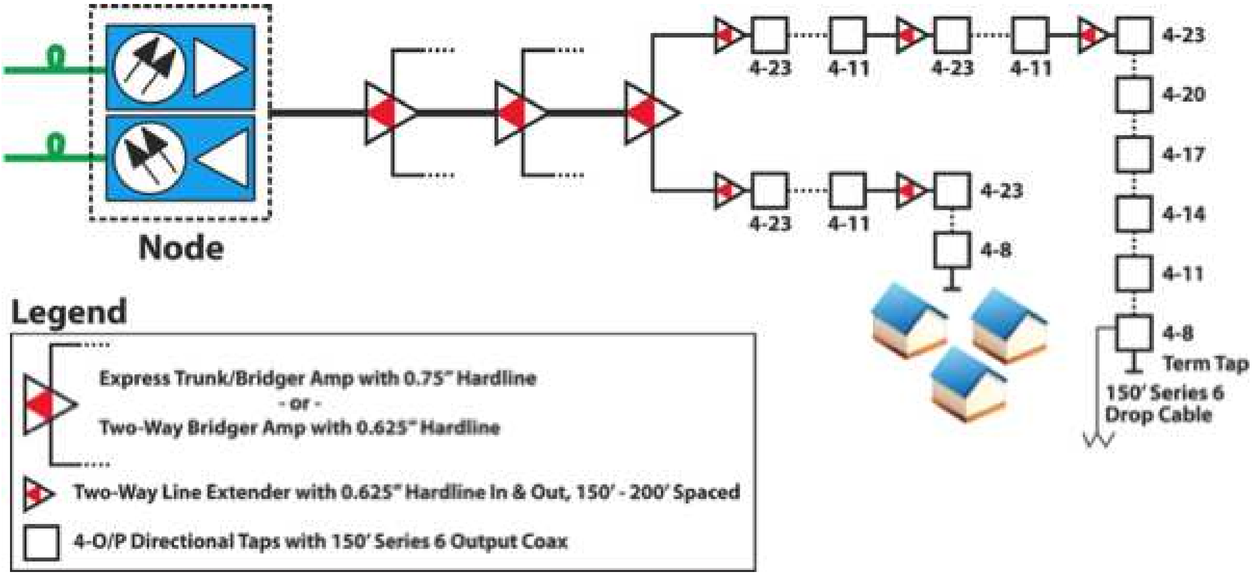
- Target starting point is a "Typical" 500 HHP Node Service Group
- Typical number of actives (30) and passives (200)
- Existing spacing, cabling types and distance (see Figure 6)

4.1 Overview of Important Considerations and Assumptions

This report has highlighted some important areas for network planners to consider while making the decisions for the next generation cable access network.

4.1.1 Avoidance of Small Node Service Groups or FTTLA

The analysis and conclusions found in this report indicates that the need for smaller node groups with few actives and passives such as Node +3 or even Fiber to the Last



4-23
4-11
4-23
4-11
4-23
4-20
4-17
4-14
4-11
4-8
Term Tap
150' Series 6
Drop Cable
4-23
4-11
4-23
4-8
Node
Legend
Express Trunk/Bridger Amp with 0.75" Hardline
- or -
Two-Way Bridger Amp with 0.625" Hardline
Two-Way Line Extender with 0.625" Hardline In & Out, 150' - 200' Spaced
4-O/P Directional Taps with 150' Series 6 Output Coax

Figure 6 – Coaxial Network Assumptions

Active (FTTLA) is not required to meet capacity, service tier predictions or network architecture requirements for this decade and beyond.

4.1.2 500 HHP Node Long-Term Viability

Our analysis finds that upstream and downstream bandwidth needs may be met while leveraging a 500 HHP node service group for a majority of this decade and even beyond. The maintaining of a 500 HHP service group is of immense value to the MSOs. The ability to solve capacity changes while maintaining the node size and spacing enables an option for a drop-in capacity upgrade.

If the goal is to achieve 1 Gbps capacity upstream this may be achieved using a typical 500 HHP node service group with 30 actives and 200 passives, and over 6 miles of coax plant in the service area as fully described later in this analysis, see Table 5.

The existing 500 HHP node has long-term viability in 750 MHz or higher systems providing enough downstream capacity to last nearly the entire decade. In the upstream a 500 HHP node is predicted to last until mid-decade when the sub-split spectrum may reach capacity and then a choice of node split, node segment or add spectrum like mid-split to maintain the 500 HHP service group are options.

The physical 500 HHP node service group may remain in place with High-split (238) beyond this decade providing 999 Mbps or 1 Gbps of MAC layer capacity. The Top-split 900-1050 with Sub-split has more capacity than Mid-split and will last through the decade.

4.1.3 1 GHz (plus) Passives - A Critical Consideration for the Future

The industry will be considering several spectrum splits and special consideration should be made to the most numerous network elements in the outside plant, the passives. Avoiding or delaying modification to the existing passives will be a significant cost savings to the MSO. Below are key factors about the 1 GHz passives:

1. Introduced in 1990 and were rapidly adopted as the standard
2. This was prior to many major rebuilds of the mid-late 90s and early 2000s
3. Prior even to the entry of 750 MHz optical transport and RF amplifiers/ products in the market place
4. Deployment of 1 GHz passives that would have more capacity than the electronics would have for nearly 15 years
5. Passives are the most numerous network element in the Outside Plant (OSP)
6. Volumes are astounding perhaps as many as 180-220 behind every 500 HHP Node or about 30 per every plant mile (perhaps 40-50 Million in the U.S. alone)
7. 1 GHz Passives may account for 85% of all passives in service today
8. Vendor performance of the 1 GHz Passives will vary and some support less than 1 GHz
9. Our internal measurements indicate that most will support up to 1050 MHz
10. Taps in cascade may affect capacity, thus additional testing is required

4.1.3.1 Assessment of the Passives

The authors believe that special consideration should be given to solutions that leverage the existing passive. This will avoid upgrades that may not be needed until the 2020 era when the MSOs may pursue spectrum above 1 GHz.

If the 1 GHz passives are considered and the desired use is over 1 GHz we believe that 1050 MHz is obtainable. There will be challenges with AC power choke resonances, which may impact the use of passives greater than 1050 MHz with predictably.

4.2 Characterization of RF Components

The network components that most affect signals carried above 1 GHz are the coaxial cable, connectors, and taps. The characteristics of these components are critical, since the major goal in a next generation cable access network is to leverage as much of the existing network as possible.

Before getting into the specifics about the RF characterization and performance requirements, it is worthwhile to establish the quality of signals carried above 1 GHz and below 200 MHz. The bottom line is that while return path signals can be carried above 1 GHz, they cannot be carried with as high order modulation as is possible at lower frequencies.

For example, if the goal is to meet similar return path data capacity the signal carriage above 1 GHz is possible using QPSK for 300 MHz of RF spectrum (47 channels of 6.4 MHz each). Whereas below 200 MHz 256-QAM is possible (due to lower coaxial cable loss) and only 24 channels occupying about 180 MHz spectrum are required, using rough estimates.

Additionally, the over 1.2 GHz solutions will require a 125 HHP service group to support QPSK, where as the High-split 200 solution may use a 500 HHP service group, this is a key contributing factor to the cost deltas of the split options.

4.3 Path Loss and SNR

In a typical HFC Node + N architecture, the return path has many more sources for extraneous inputs, “noise” than the forward path. This includes noise from all the home gateways, in addition to all the return path amplifiers that combine signals onto a single return path (for a non-segmented node).

For now we will ignore the gateway noise, since in principle it could be made zero, or at least negligible, by only having the modem return RF amplifier turned on when the modem is allowed to “talk”.

The RF return path amplifier noise funneling effect is the main noise source that must be confronted; and it cannot be turned off! This analysis is independent of the frequency band chosen for the “New Return Band” (e.g., Mid-split 5-85 MHz; High-split 5-200 MHz; or Top-split with UHF return), although the return path loss that must be overcome is dependent on the highest frequency of signals carried. For a first cut at the analysis, it suffices to calculate the transmitted level from the gateway required to see if the levels are even possible with readily available active devices.

The obvious way to dramatically reduce the funneling noise and increase return path capacity is to segment the Node. That is not considered here to assess how long the network remains viable with a 4x1 configuration, a 500 HHP node service group.

The thermal mean-square noise voltage in 1 Hz bandwidth is kT, where k is the Stefan-Boltzmann constant, 1.38x10^-23

J/deg-K, and T is absolute temperature in degrees Kelvin. From this we have a thermal noise floor limit of -173.83 dBm/Hz. For a bandwidth of 6.4 MHz and 75-ohm system, this gives -57.0 dBmV per 6.4 MHz channel as the thermal noise floor. With one 7 dB noise figure amplifier in the chain, we would have a thermal noise floor of -50 dBmV/6.4 MHz channel.

Two amplifiers cascaded would give 3 dB worse; four amplifiers cascaded give 6 dB worse than one. And since the system is balanced to operate with unity gain, any amplifiers that collect to the same point also increase the noise floor by 10*log(N) dB, where N is the total number of amplifiers in the return path segment.

For a typical number of 32 distribution amplifiers serviced by one node, this is five doubles, or 15 dB above the noise from one RF Amplifier, or -35 dBmV/6.4 MHz bandwidth. The funneling effect must be considered in the analysis for the NG Cable Access Network.

If the return path signal level at the node from the Cable Modem (CM) is +15 dBmV, it is clear that the Signal-to-Noise Ratio (SNR) in a 6.4 MHz bandwidth is 50 dB; very adequate for 256-QAM or even higher complexity modulation. But if the Return path level at the node port is 0 dBmV, the SNR is 35 dB; this makes 256-QAM theoretically possible, but usually at least 6 dB of operating margin is desired.

Table 4 – Legacy Modulation and C/N Performance Targets

Modulation Type	Uncoded Theoretical C/N dB	Operator Desired C/N Target
QPSK	16	22
8-QAM	19	25
16-QAM	22	28
32-QAM	25	31
64-QAM	28	34
128-QAM	31	37

Theoretical SNRs Uncoded with BER of 10^-8
Practical C/N is chosen to give 6 dB headroom above Uncoded

If only -10 dBmV is available at the node return input, the SNR is 25 dB; and so even the use of 16-QAM is uncertain. This illustrates (Table 4) the very high dynamic range of “Pure RF” (about 15 dB higher than when an electrical-to-optical conversion is involved).

Table 5 documents many important assumptions and assumed node configuration conditions. An important assumption is the CM maximum power output level of +65 dBmV into 75 ohms.

What this means is that if many channels are bonded (to increase the amount of data transmitted), the level of each carrier must be decreased to conform to the CM maximum power output constraint. Two channels bonded must be 3 dB lower each; four channels must be 6 dB lower than the Pout(max).

Since the channel power levels follow a 10*log(M) rule, where M is the number of channels bonded to form a wider bandwidth group. For 16 channels bonded, each carrier must be 12 dB lower than the Pout(max).

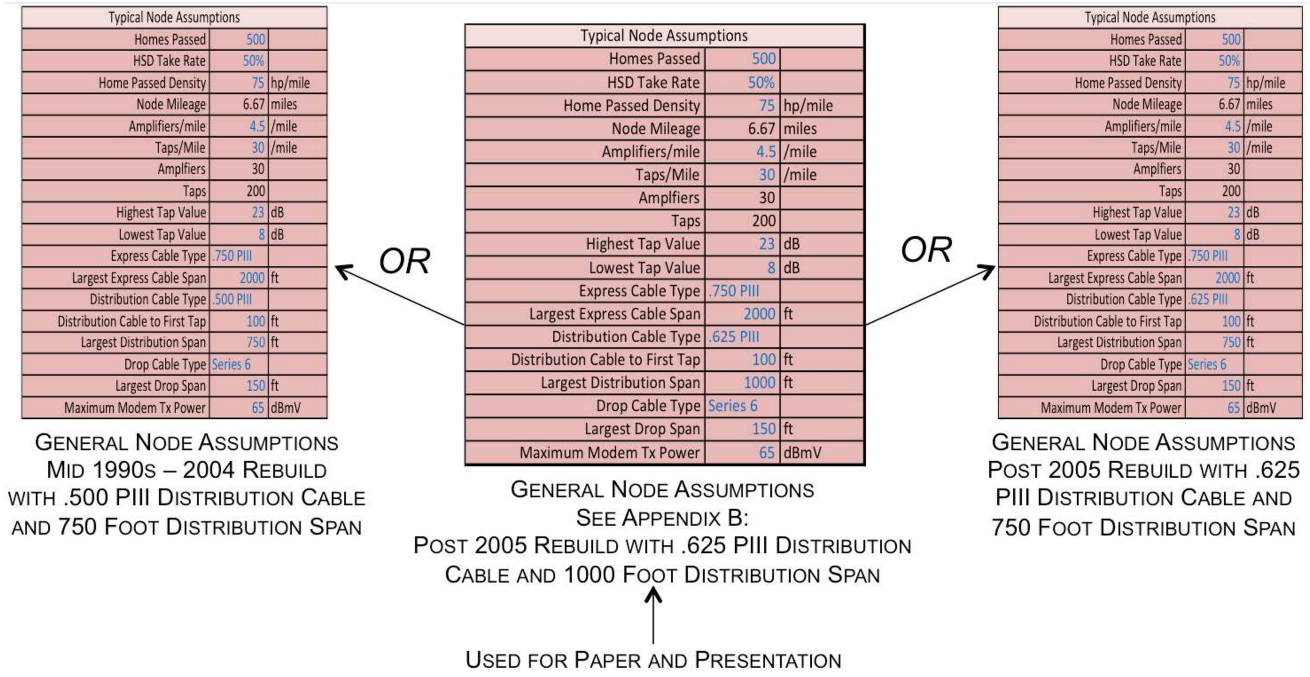
For 48 channels bonded, each must be 16.8 dB lower than the Pout(max). So for 48-bonded channels, the level per channel is at most 65 dBmV -17 dB = +48 dBmV. If there is more than 48 dB of loss in the return path to the node return input, the level is <0 dBmV and 64-QAM or lower modulation is required. The node and system configuration assumptions are as follows.

4.4 Cable Loss Assessment

Two different lengths of 1/2” diameter hardline coax were tested for Insertion Loss and Return Loss (RL). The loss versus frequency in dB varied about as the square root of frequency. But as can be seen below, the loss at 2 GHz is about 5% higher than expected by the simple sq-rt(f) rule. The graph below illustrates a slightly more than twice the loss at 2 GHz compared to 500 MHz, see Figure 7.

In the plot of Figure 8, the coax Return Loss (RL) did not vary as expected above 1200 MHz. This appears due to an internal low-pass matching structure in the hardline-to-75N connectors (apparently for optimizing the 1-1.2 GHz response). The connectors are an important element to return loss with signals above 1 GHz.

Table 5 – Node and Coaxial Network Assumptions Typical of U.S based MSOs



Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.500 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	750	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

General Node Assumptions
Mid 1990s – 2004 Rebuild
with .500 PIII Distribution Cable
and 750 Foot Distribution Span

Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	1000	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

General Node Assumptions
See Appendix B:
Post 2005 Rebuild with .625 PIII Distribution
Cable and 1000 Foot Distribution Span

Used for Paper and Presentation

Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	750	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

General Node Assumptions
Post 2005 Rebuild with .625
PIII Distribution Cable and
750 Foot Distribution Span

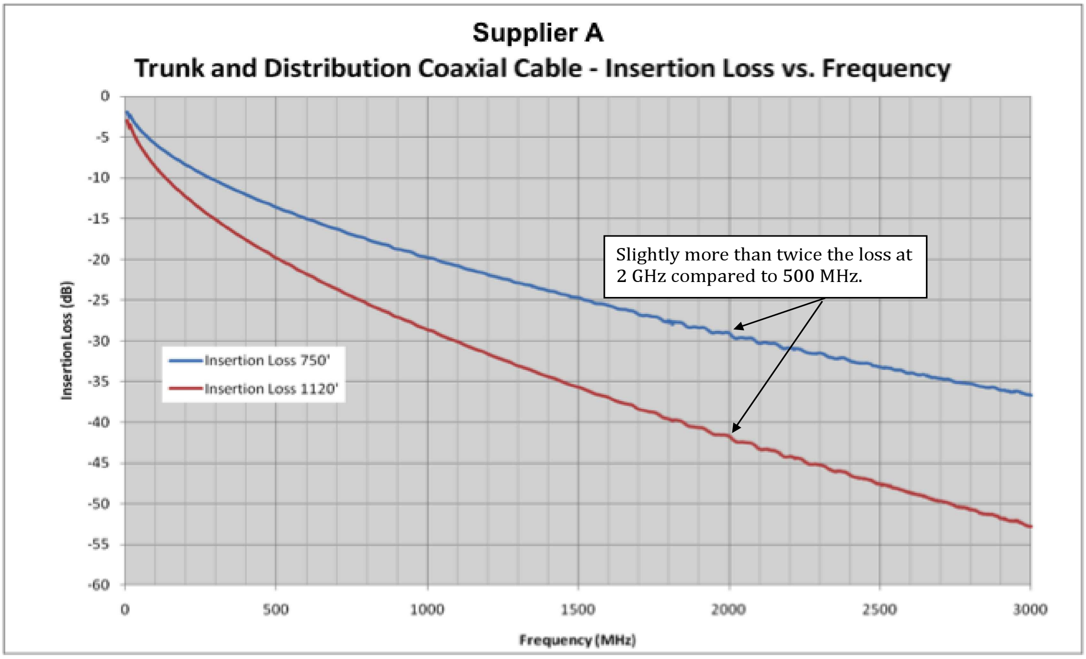


Figure 7 – Distribution Coaxial Cable – Insertion Loss vs. Frequency

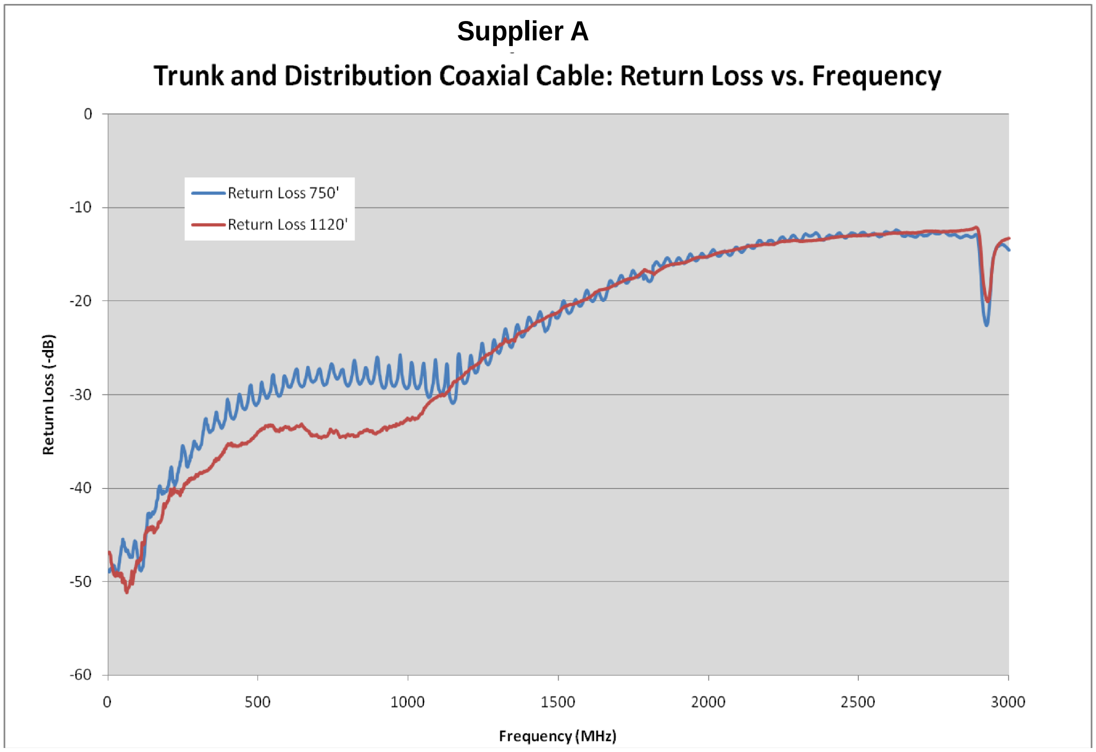


Figure 8 – Distribution Coaxial Cable – Return Loss vs. Frequency

4.5 Tap Component Analysis

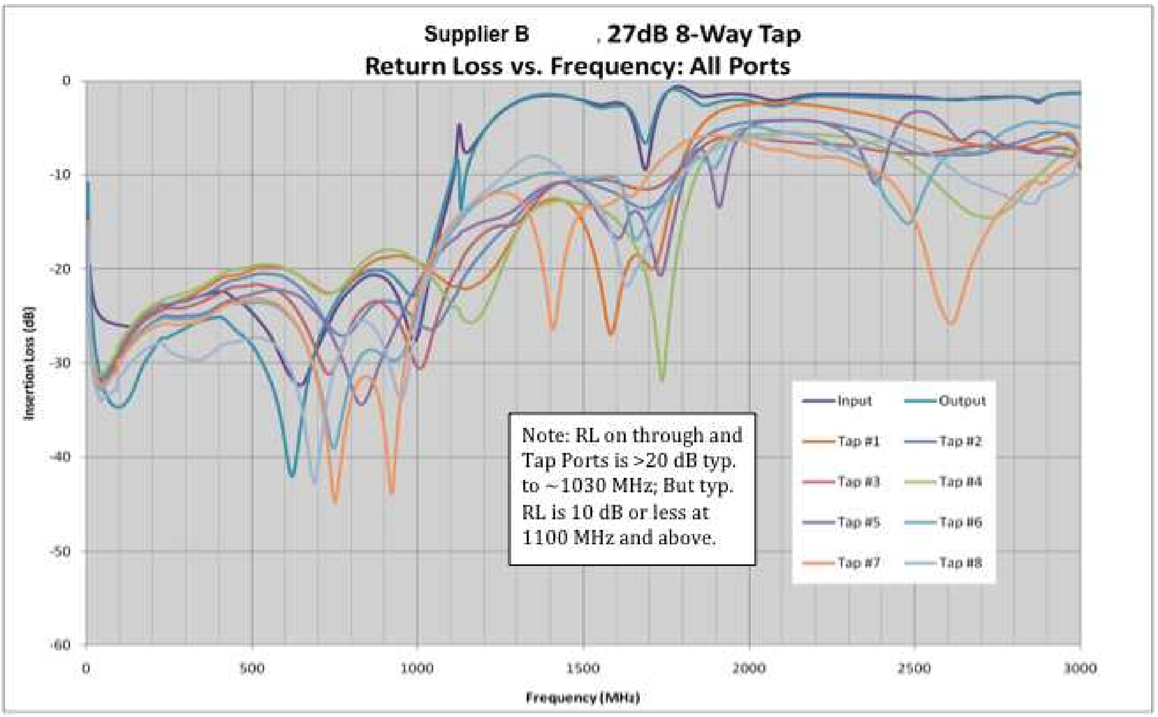
Taps are the components with the most variability in passband characteristics, because there are so many different manufacturers, values, and number of outputs. Most were designed more than ten years ago, well before >1 GHz bandwidth systems were considered. One of the serious limitations of power passing taps is the AC power choke resonance.

This typically is around 1100 MHz, although the “notch” frequency changes with temperature. Tap response resonances are typical from ~1050 to 1400 MHz. A limitation of power passing taps is the AC power choke resonance. This is an important finding when leveraging the existing passives; therefore the use above 1050 MHz may not be predictable or even possible.

Even the newer, extended bandwidth taps, with passband specified 1.8 GHz or 3 GHz, the taps usually have power choke resonances (or other resonances, e.g., inadequate RF cover grounding) resonances in the 1050 MHz to 1300 MHz range. Especially on the tap coupled port. However, most Taps work well to ~1050 MHz.

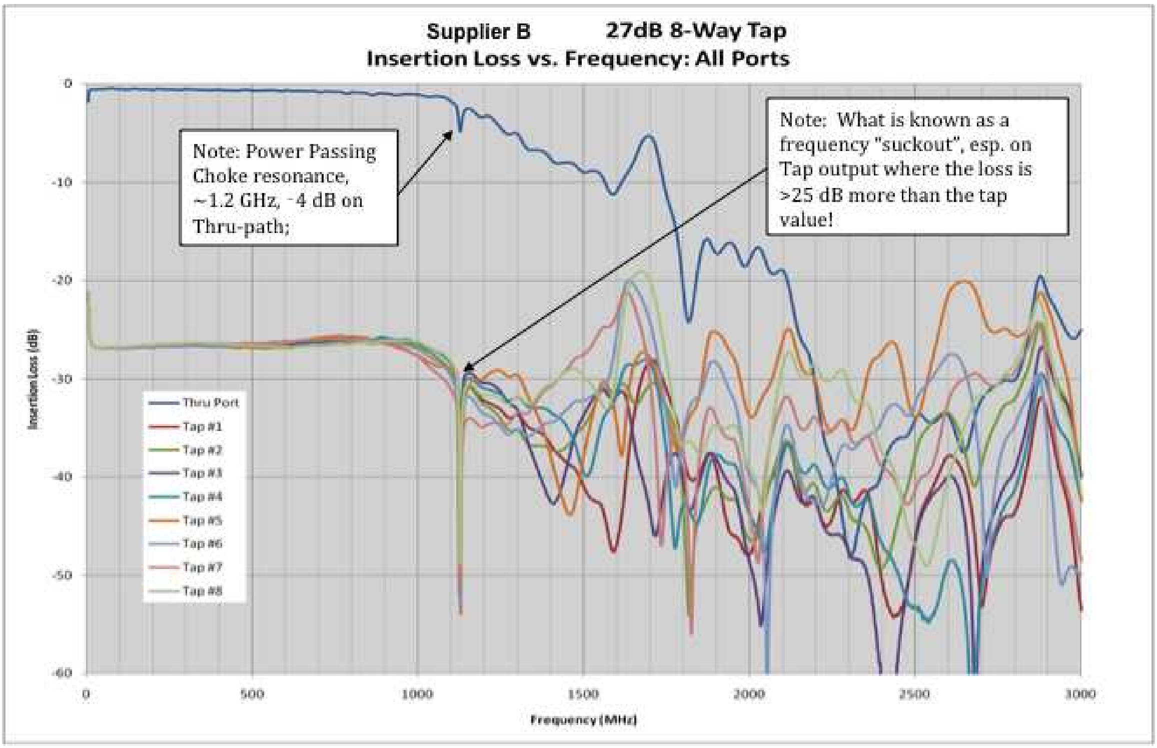
Nearly all taps exhibit poor RL characteristics on all ports above 1400 MHz. Some are marginal for RL (~12 dB), even at 1 GHz. Therefore tap cascades must be tested and over temperature to verify the actual pass band response due to close-by tap reflections.

Figure 9 to Figure 11 show examples of the variability of key RF parameters for an array of Taps evaluated.



Note: RL on through and Tap Ports is >20 dB typ. to ~1030 MHz; But typ. RL is 10 dB or less at 1100 MHz and above.

Figure 9 – 27 dB x 8 Tap - Return Loss vs. Frequency: All Ports


Supplier B 27dB 8-Way Tap
Insertion Loss vs. Frequency: All Ports

Note: Power Passing Choke resonance, ~1.2 GHz, -4 dB on Thru-path;

Note: What is known as a frequency "suckout", esp. on Tap output where the loss is >25 dB more than the tap value!

Figure 10 – 27 dB x 8 Tap - Insertion Loss vs. Frequency: All Ports

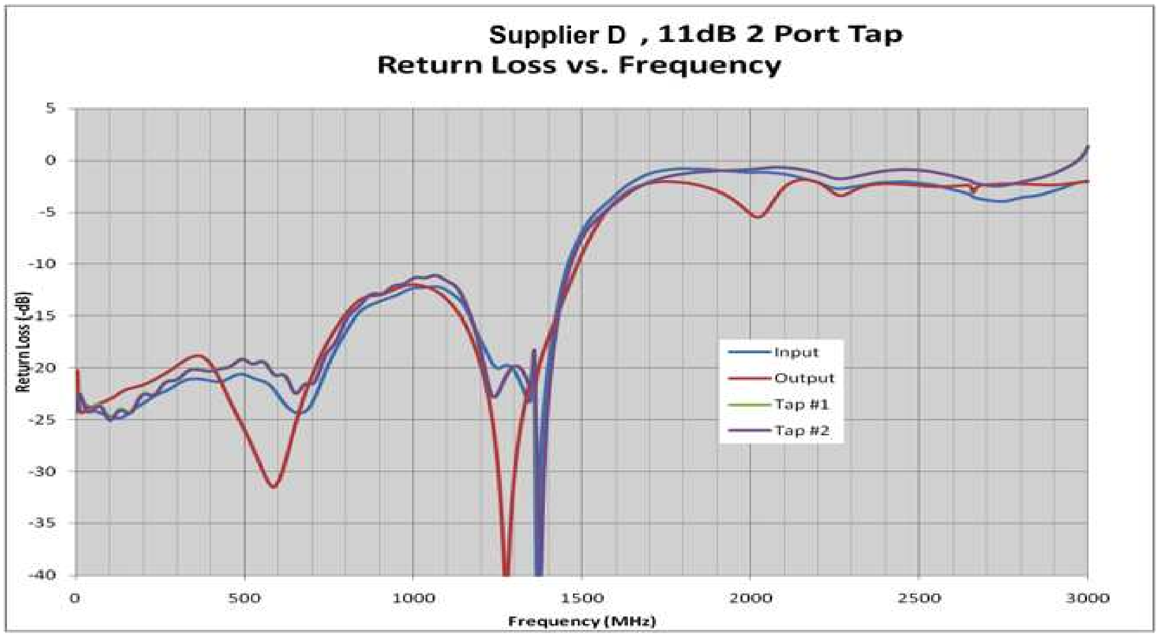

Supplier D , 11dB 2 Port Tap
Return Loss vs. Frequency

Figure 11 – 11 dB x 2 Tap - Return Loss vs. Frequency

4.6 Field Performance – Passive Coax Above 1 GHz

Let's pull together what we have discussed around taps and passives, the analysis of Section 4.2 and summarize how these components behave together in the context of recent field characterizations performed for the AMP initiative.

As discussed above, coaxial cable and even some current 1 GHz taps are indeed capable of supporting useful bandwidth above 1 GHz [4]. However, the frequency dependence of cable loss (see Figure 7) quickly attenuates signals above 1 GHz when we consider its use relative to attenuation characteristics of a low band upstream. The combination of drop cables, trunk cable, and taps add up to significant losses to the first active.

We can anticipate almost twice the loss (in dB) extending the return band to 200 MHz, such as in the high-split architecture introduced. However, above 1 GHz, the loss may increase by roughly a factor of five (in dB, dependent on Top-Split case chosen) compared to legacy return for such a span. CPE devices must make up for that loss to maintain equivalent performance, all else the same. As we observed in analyzing the case with an increasing amount of channel bonding, they also must generate additional total power associated with the wider bandwidth they would occupy to enable peak rates of a Gbps, relative to today's maximum of 6.4 MHz single or 2-4x bonded channel power.

This is not your father's cable modem – an L-Band, wideband, high power linear transmitter. It is a significantly more complex RF device. It is not a technology challenge, but it will come at a cost premium relative to retail CPE today. Quantifiably, the result is that very high CPE transmit power becomes necessary to close a bandwidth efficient link budget.

Conversely, for a given maximum transmit power, such as 65 dBmV chosen previously, we can favorably assume it is the same transmit power number for low split or for top split frequencies. The additional top-split loss translates to lower SNR at the first active, and every subsequent one if a cascade is in place. This impacts composite SNR formed by the combination of RF funneling and optical link performance.

The end result is that potential bps/Hz of top split is inherently lower for top split, and to achieve an equivalent modulation efficiency, the top split must be deployed over smaller service groups to reduce the noise contributions associated with the lower inherent SNR created by the loss. We will quantify this in further detail in Section 9.6.

However, Motorola performed field measurements as part of the AMP initiative, and the conclusions provide insight into the nature of this issue. We illustrate with a simple, and best case (N+0) example from field characterization done exactly for this purpose. Figure 12 shows field characterized loss [4] [5] of an RF leg of recently-built underground plant, measured from the end of a 300 ft coaxial drop from the final tap of a five-tap string on an otherwise typical suburban architecture.

The five taps, manufactured by Javelin Innovations, where extended bandwidth models, utilizing modified faceplates installed within existing tap housing to extend the RF passband of the network.

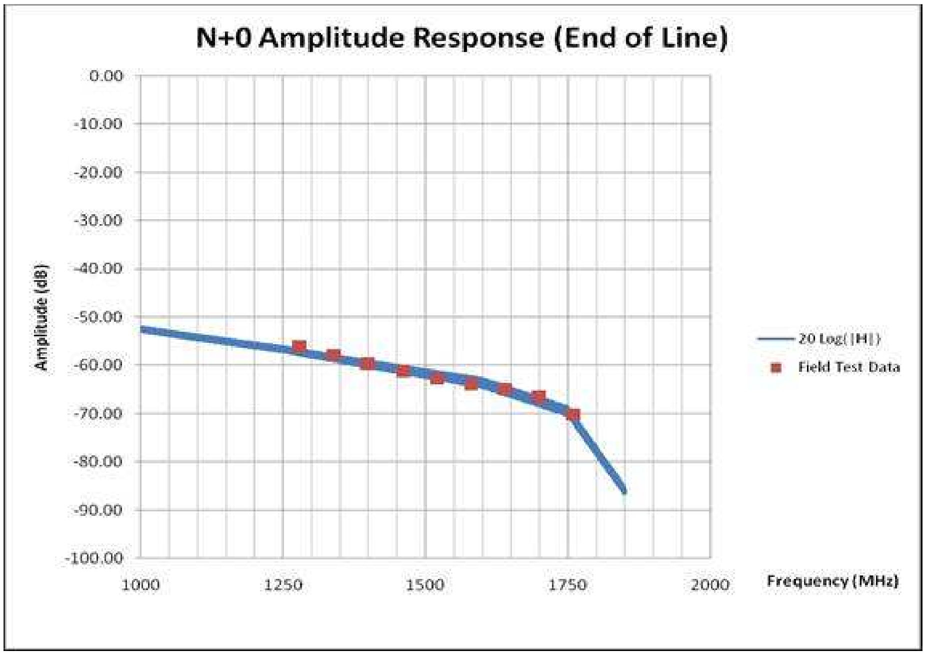


Figure 12 – Top Split Loss Characterization vs Model

Losses from 50-70 dB are observed, with measured data points highlighted in Figure 12. While the drop length represents an extended length scenario, the lack of any home connection removes any effects of additional splitters commonly found inside the home and outside the reach of the MSO until there is a problem in the home.

Let's take a look at the lowest, least attenuation part of the band, 1-1.2 GHz. A reasonable case can be made for a bandwidth efficient link budget for a remote PHY termination, as transmitters that increase the transmit power level over today's requirements to support 65 dBmV will reach the first active with solid SNR.

Mathematically, consider the following:

- Thermal Noise Floor: -65 dBmV/MHz
- Signal BW: 200 MHz
- Total Noise: -42 dBmV/200 MHz
- Active NF+Loss: 8 dB (est)
- Rx Noise Power, Plant Terminated: -34 dBmV

Using the 55 dB of loss observed at the low end of the band for the first 200 MHz, a 58 dBmV transmitter will leave us with an SNR of 37 dB. This is in the neighborhood of the SNR required, with margin, for 1024-QAM if advanced FEC is assumed. In Table 4,1024-QAM is quantified as SNR = 39 dB without FEC using typical HFC upstream optics. Higher orders would become challenging. A 65 dBmV capability would more ably support a higher modulation profile.

Based on the attenuation slope in Figure 12 above 1200 MHz, this gets more challenging as higher bands are considered. Note that the tap performance of the extended band units is very good, but there is simply unavoidable attention associated with deployed coaxial infrastructure that becomes the dominant SNR characteristic of the link.

Now, consider that the above characterization included the following favorable conditions:

- Faceplate tap replacements
- N+0
- Pristine, unused plant
- Extra transmit power assumed in a much higher frequency band
- No connected users
- No home losses

We can easily remove the first of these assumptions for most practical networks. Without the investment in tap faceplate change-outs, typical 1 GHz taps in the band directly above their specified maximum have more loss than these specially designed faceplates.

The additional loss observed is up to 9 dB for the cascade of taps at the end of the usable band, in this case characterized as 1160 MHz [5](worse above that, less below). More loss comes directly off of the SNR as the signal power is dropped into the noise floor.

Thus, in current tap architectures, under N+0 conditions, and constrained to the lowest end of "top-split," in good plant conditions, we are already seeing pressure on SNR for bandwidth efficient modulation profiles as the SNR drops to 30 dB or less. The sensitivity of QAM profile to SNR loss in Table 4 – Legacy Modulation and C/N Performance Targets shows that 2-3 modulation profiles, and the associated capacity, become compromised.

Now, to remove another assumption, if we instead think of the actives as amplifiers, and cascade them on the way to a node with equivalent degradation and potentially combining noise impacts at the node a described in Section 4.3, we find that a bandwidth efficient link budget becomes even more difficult to achieve.

Thus, top-split, while potentially within technology and investment reach, is off to a very difficult start as a viable alternative. The potential bps/Hz efficiency metric is inherently lower, and to achieve an equivalent modulation efficiency, the top-split must be deployed over smaller service groups to reduce the noise contributions associated with the lower inherent SNR created by the loss. This has been shown to be the case analytically as well as in field characterization in a better-than-typical environment.

4.7 Using "Top-Split" Spectrum for New Forward Path Capacity

While the challenges on the upstream above the forward band are significant obstacles to practical deployment, this is not necessarily so on the downstream. This is important, because as the upstream side of the HFC diplex extends, it intrudes on downstream bandwidth and thus removes available downstream capacity. We believe that use of new coaxial spectrum will be required in the evolution of HFC and of DOCSIS, and that both should be part of cable's migration plan. However, in the case of new spectrum above 1 GHz, we believe that is best utilized for new forward capacity.

We have discussed the possibility of a phased architecture. While forward bandwidth loss is relatively modest for an 85 MHz split, if the band extends further, such as to 200-300 MHz, then a significant chunk of downstream capacity is lost. Today, this band may be only carrying analog services, and thus is not reducing the actual deployed downstream capacity, but it is reducing the available capacity for future growth – i.e. it is assumed that at some point analog services will be removed in favor of digital capacity.

With this loss of downstream bandwidth, it then becomes important to uncover new downstream bandwidth, and the logical place to find this is directly above today's forward band. If the architecture is 750 MHz or 870 MHz, then of course there is already technology in place to exploit out to 1 GHz. Beyond 1 GHz, there is very little outdoor gear designed to operate in this band, and no CPE designed to work in this band (just as is the case for upstream).

We can identify at least three compelling advantages to considering use of the band over the end of the defined tap bandwidth for forward services, as opposed to reverse:

1) High Fidelity Forward Path – The fundamental characteristics of the forward path have always been to around a high SNR, low distortion environment to ably support analog video. As we know, the reverse path was not originally architected with high fidelity in mind. Over time, technology has been introduced to enable a high-speed data channel, but the low noise and high linearity architected into the forward path is orders of magnitude above the return path. This difference translates to a much more straightforward exploitation of bandwidth with high performance on the downstream.

2) Broadband RF Power – The forward path levels are designed for RF path losses out to 1 GHz. Because of this, the parasitic losses above 1 GHz of the coax, and the minimal additional attenuation, are not a stretch to achieve when extending the forward path. It is an entirely different case in the return, where the architecture has relied on the low loss end of the band, which increases only modestly as it is extended to 85 MHz or even 200 MHz. This issue was highlighted in Sections 4.3 and 1.1.

3) Cost of New RF BW – Forward path RF systems already extend to the 1 GHz range, so are designed with the expectation of the loss implications. There has therefore been continuing investment in broadband RF hybrids driving higher levels over increasing forward bandwidths, still based on supporting a full analog and digital multiplex. As a result, the output levels of these hybrids and nonlinear characteristics have continued to improve. However, investment in these premium devices for the forward path is spread over the number of homes serviced by the actives. The HFC downstream delivers high linearity and high levels over multiple octaves, and the hybrids are shared, spreading the investment across a subscriber pool. In the reverse path, each home needs a high power, linear transmitter (though less than an octave), and also in a much higher frequency band that would likely require a higher cost technology implementation.

4) The use of spectrum above the forward band implies a new guard band. Since guard bands are a percentage of edge frequency, the lost spectrum is sizable, cost significantly lost capacity. The eliminated spectrum will remove prime forward path digital bandwidth from use, costing on the order of 1 Gbps for DOCSIS NG technology, in order to enable *less* capable upstream bandwidth above 1 GHz.

Without question, HFC will need to mine new bandwidth to enable new capacity for continued traffic growth. Today's coax remains unexploited above 1 GHz in all cases, and above 750 MHz and 870 MHz in other cases in North America. Current forward path technology is already within striking distance and readily capable of being extended to take advantage of latent coaxial capacity above wherever the forward path ends today [6]. And, while this spectrum is non-ideal in the forward path as well, it will benefit from the introduction of OFDM for NG DOCSIS, but without the spectrum loss and RF power implications of use as upstream band.

Based on the above reasoning, our recommendation is to enable additional coaxial capacity above today's forward band, and to exploit this spectrum for downstream purposes exclusively. We will quantify this band for downstream use in subsequent sections derving data capacity, network performance, and lifespan.

In Section 0, we will estimate the available data capacity of the forward path under various implementations of an extended forward band.

Then, in Sections 10.2.1 and 10.2.2, we will quantify available network capacity and discuss the implications to forward path lifespan.

Finally, in Sections 10.2.3 and 10.2.4, we will describe how this bandwidth could be managed within the system engineering of downstream HFC, implemented within linear optics and RF (not an RF overlay).

5 HFC OPTICAL TRANSPORT TECHNOLOGY OPTIONS

The optical layer will be examined in this section. We will look at two technologies of optical transport return, analog return path and digital return, which may commonly be referred to as Broadband Digital Return (BDR), or simply Digital Return. First, we will review the forward path. [36]

5.1 Overview - Analog Forward Path Transport

Analog Forward path is currently the only economical method for the transmission of cable signals downstream. The advances in analog forward laser technologies enable transmission of the 54-1002 MHz of spectrum this is over 150 channels, each 6 MHz wide. This is approximately 6 Gbps of data capacity assuming the PHY layer transmission utilizing 256-QAM (8 bits per Hz BW efficiency, excluding overhead).

The forward path is a layer 1 media-converter style architecture. The optical transmission may be shared with multiple HFC nodes. There are two network architectures for the forward: Full Spectrum as illustrated in Figure 13; and another called QAM Narrowcast Overlay, or simply Narrowcast Overlay, as in Figure 14.

The MSO serving area between headend and node will be in most cases is less than 40 km. Therefore this will be easily

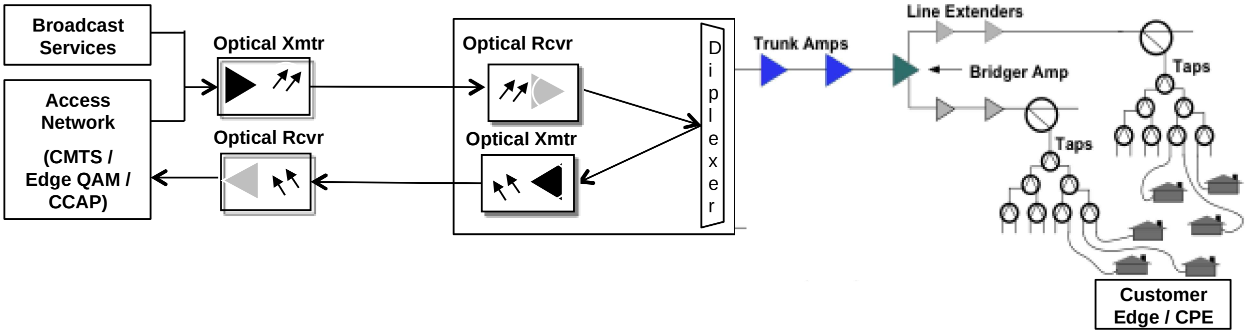


Figure 13 – Hybrid Fiber Coax (HFC) with Full Spectrum and Node +N

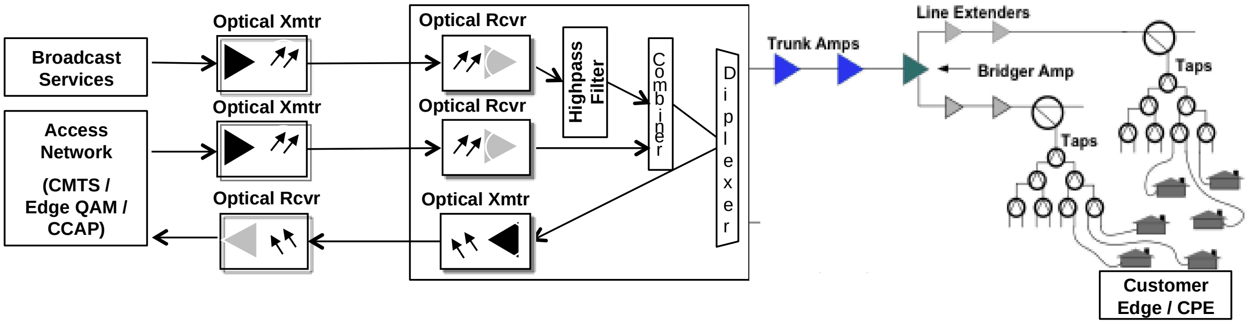


Figure 14 – Hybrid Fiber Coax (HFC) with QAM Narrowcast Overlay and Node +N

supported with an HFC architecture. The support for extremely long distance to and from the node may be a factor for the HFC. The optical capabilities of HFC simply have lots of dependencies, variables, and trade-offs to determine the HFC optical link distance.

We will use round numbers and generalities to discuss some the capabilities of HFC optical transport when considering long distances. So, we will use an example of HFC analog optical transmission of full spectrum, no analog video, and 150 QAM channels, we will assume a 100 km optical reach is achievable in most cases.

In a narrowcast overlay architecture, we assume as many as 40 wavelengths / lambdas per fiber, 80 QAMs of narrowcast spectrum, and a reach of approximately 100 km to the node. HFC optical distance will vary based on many factors, including narrowcast channel loading, the number of analog video channels, and many other factors. We could assume that a greater distance is achievable with an HFC Digital Forward, as well as DFC (Digital Fiber Coax) style optical transport, compared with HFC analog forward optics without the use of EDFAs (erbium-doped fiber amplifier).

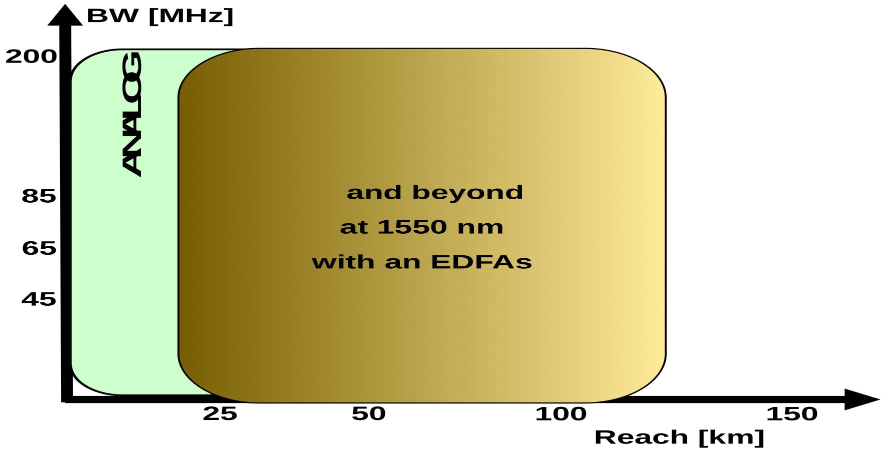


Figure 15 – Return Analog Optical bandwidth and Reach

In some cases, fiber count is insufficient, regardless of the distance. Therefore, to avoid over lashing new fiber to service groups, separate wavelengths are placed on the fiber. The use of HFC analog optics today supports far fewer optical wavelengths than that which is supported using optical Ethernet technology. This may be a challenge for HFC style architectures.

5.2 Overview - Analog Return Path Transport

Analog return path transport is now mostly done with a Distributed Feedback (DFB) laser located in the node housing and an analog receiver located in the headend or hub. Analog return path transport is considered as a viable option for Mid-split, High-split, and Top-split returns. Supporting short to moderate return path distances of 0-50 km with full spectrum High-split is achievable. If the wavelength is changed to 1550 nm with an EDFA, then greater distances are possible. This is shown in Figure 15.

The analog optical return path transport presently supports up to 200 MHz loading; but typically only 5-42 MHz or 5-65 MHz is carried, depending on the distribution diplex filter split. The major benefit with analog optical return is its simplicity and flexibility, when compared with HFC style digital optical transmission. Distance is the chief challenge of analog optical transport. Refer to the Figure 15 and Figure 16.

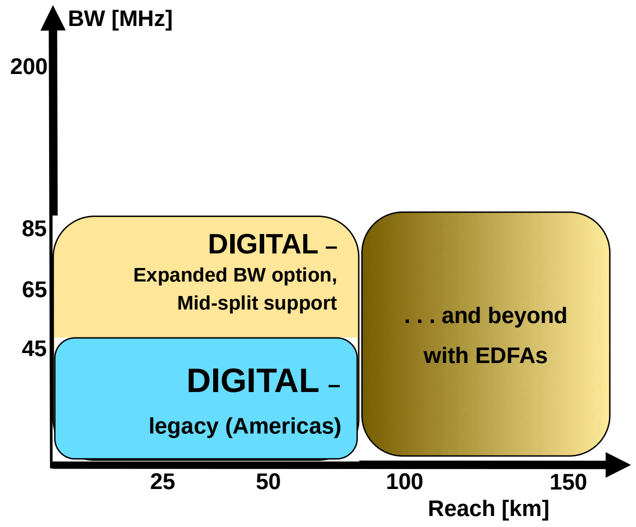


Figure 16 – Return Optical bandwidth and Reach

Pros

The chief advantage of analog return is its cost effectiveness and flexibility. If analog return optics are in use in the field today, there is a good chance that they will perform adequately at 85 MHz; and even 200 MHz loading may be possible, if required in the future. This would allow an operator to fully amortize the investment made in this technology over the decade.

Cons

There are drawbacks to using analog optics. Analog DFB's have demanding setup procedures. RF levels at the optical receiver are dependent on optical modulation index and the received optical power level. This means that each link must be set up carefully to produce the desired RF output at the receiver (when the expected RF level is present at the input of the transmitter). Any change in the optical link budget will have a dramatic impact on the output RF level at the receiver, unless receivers with link gain control are used.

Also, as with any analog technology, the performance of the link is distance dependent. The longer the link, the lower the input to the receiver, which delivers a lower C/N performance. The practical distance over which an operator can expect to deliver 256-QAM payload on analog return optics is limited.

Assessment

The analog return transmitter will work well for the low and high frequency return. Analog return path options should be available for the higher frequency return options at 900-1050 MHz and 1200-1500 MHz. However the cost vs. performance at these frequencies when compared to digital alternatives may make them less attractive. There will be distance limitations and EDFAs will impact the overall system performance noise budgets. The distance of 0-50 km are reasonable and longer distance would be supported with an EDFA.

5.3 Overview – Digital Return Path

Digital return path technology is commonly referred to as broadband digital return (BDR). The digital return approach is “unaware” of the traffic that may be flowing over the spectrum band of interest. It simply samples the entire band and performs an analog to digital conversion continuously, even if no traffic is present. The sampled bits are delivered over a serial digital link to a receiver in the headend or hub, where digital to analog conversion is performed and the sampled analog spectrum is recreated.

The parameters of analog to digital conversion will need to be considered when determining the Digital Return optical transport requirements. There are two important factors in the A-to-D conversion:

1. Sampling Rate and
2. Bit Resolution (number of bits of resolution).

Sampling Rate

- Inverse of the time interval of which samples of the analog signal are taken.
  - Referred to as Samples per Second or Sampling Frequency.
- Nyquist Sampling Theorem governs the minimum sampling rate.
  - Minimum sampling frequency must be at least twice the frequency width of the signal to be digitized.
  - Example: Return band from 5 – 42 MHz must be sampled at 84 MHz (at least). For practical filter realization, the sampling rate should be at least 10-20% greater.

Bit Resolution

- Number of bits to represent the amplitude for each sample taken.
- Each bit can be “1” or “0” only, but multiple bits can be strung together as “words” of “n” number of bits.
- Number of amplitude levels can be calculated as 2^n, where “n” is the number of bits of resolution. Example: 8 bits leads to 2^8 = 256 levels.

Pros

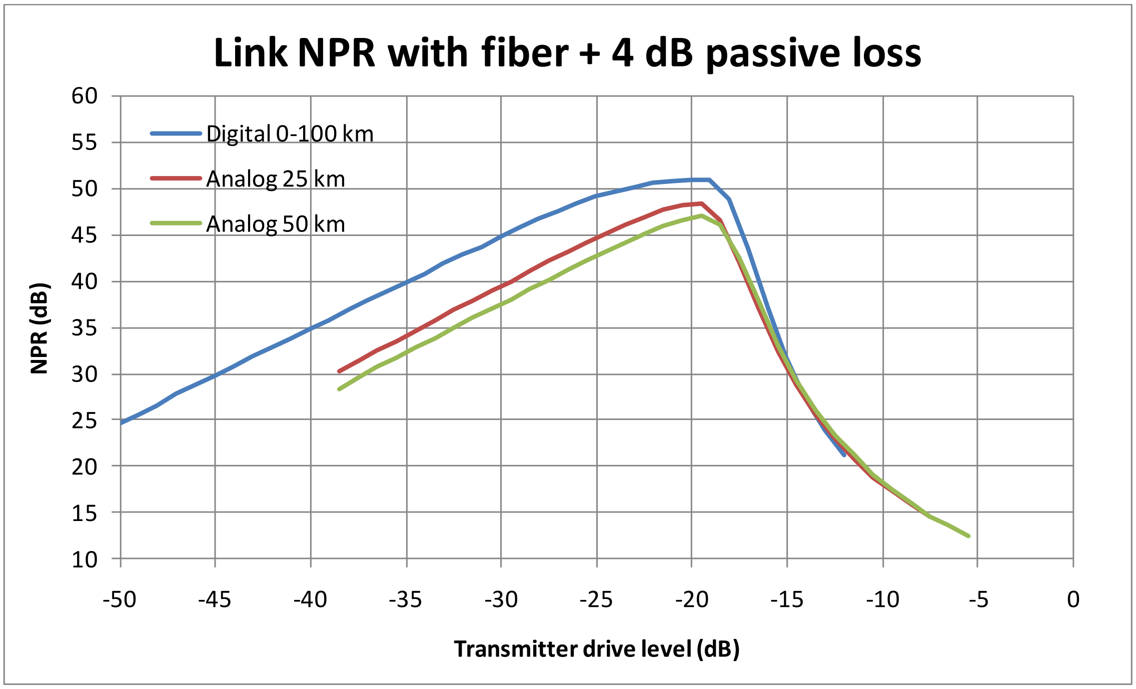


Figure 17 – Analog & Digital Return NPR

There are a number of advantages to the digital return approach. The output of the receiver is no longer dependent on optical input power, which allows the operator to make modifications to the optical multiplexing and de-multiplexing without fear of altering RF levels. The link performance is distance independent – same MER (Modulation Error Ratio) for 0 km as for 100 km, and even beyond as Figure 17 illustrates. The number of wavelengths used is not a factor since on/off keyed digital modulation only requires ~20dB of SNR; thus fiber cross-talk effects do not play a role in limiting performance in access-length links (<160 km)

The RF performance of a digital return link is determined by the quality of the digital sampling, rather than the optical input to the receiver; so consistent link performance is obtained regardless of optical budget. The total optical budget capability is dramatically improved since the optical transport is digital. This type of transport is totally agnostic to the type of traffic that flows over it.

Multiple traffic classes (status monitoring, set top return, DOCSIS, etc) can be carried simultaneously. Figure 17 below is an illustration of performance and distance when examining the analog and digital optical transport methods. With regards to the link noise power ratio (NPR) with fiber and 4 dB optical passives loss, the digital return used 1470 – 1610 nm; analog 25 km used 1310 nm, while the analog 50 km used 1550 nm. The optical output power of each transmitter was 2 mW (+3 dBm).

The Digital Return main drivers are as follow:

- “Set it and forget it” – technician and maintenance friendly
- Signal to noise performance does not degrade with distance
- Supports redundancy over uneven lengths/longer lengths
- Pairs well with “fiber deep” architectures, enables “service group aggregation”
- Pluggable optics for less costly inventory

Cons

The chief drawback to digital return is the fact that nearly all equipment produced to date is designed to work up to 42 MHz. Analog receivers are not useable with digital return transmissions. Further, the analog-to-digital converters and digital return receivers aren’t easily converted to new passbands. It requires “forklift upgrades” (remove and replace) of these optics when moving to 85 MHz and 200 MHz return frequencies. There is currently no standardization on the digital return modulation and demodulation schemes, or even transport clock rates.

Another chief drawback to digital return is the Nyquist sampling theorem. It requires a minimum sampling rate, f_s >2B for a uniformly sampled signal of bandwidth, B Hz. For n-bit resolution, this requires a Transport Clock frequency >2nB. It is assumed that the higher the transport clock, the more costly it is. And with higher clock speed, there is more fiber dispersion, which sets an upper limit on transport rate! This causes some practical limitations as to how high the return spectrum can cost effectively reach when considering digital return.

The key points about Nyquist Sampling are captured below. This may be a major driver for the use of analog optics when modest distances are possible and also a major reason to move away from HFC style architectures to a Digital Fiber Coax (DFC) class of architecture when distance is a challenge.

Nyquist Sampling Theorem governs the minimum sampling rate

- Minimum sampling frequency must be at least twice the frequency width of the signal to be digitized

Nyquist Theorem causes some practical limitations

- A 6 MHz baseband signal requires a sampling frequency of 12 MHz minimum
- A 42 MHz return band requires 84 MHz minimum (at least)
- To digitize the entire forward band, we would need to sample at 1.1 GHz (550MHz system) to 2.0 GHz (1GHz system)
- Higher speed A/D converters typically have less Effective Number of Bits (ENOB), translating to decreasing performance at increasing clock speeds for a fixed number of bits.

The total data rate for any given digitized signal can be calculated as follows:

- Determine the minimum sampling rate. As discussed, this is always at least 2X the frequency width of the signal to be digitized (at least). Multiply by the number of resolution bits desired, n, to get the minimum transport clock. And add overhead bits for error correction and framing.

Example: Digital Return

- Typical Return band is 5-42 MHz
- Minimum Sampling frequency is 84 MHz (2*42 MHz) (at least for practical filter realization the sampling rate may be at least 10-20% greater to allow for an anti-aliasing filter.)
- For simple math, we will use 100 MHz or 100 Million samples/second
- Determine the bit resolution will be largely dependent on the SNR required
- For simple math we will use 10-bit resolution or 10 bits/sample
- Multiply bit resolution and sampling rate
  - 100 Million samples/second * 10 bits per sample = 1,000,000,000 bits/second
  - Approximately 1 Gb/s required to digitize the return band

Key Summary:

- >1 Gbps of optical transport was required to transport the 5-42 MHz of spectrum / data capacity
- Estimate of 4 Gbps plus of optical transport was required to transport the 5-250 MHz of spectrum / data capacity at 10 bits per sample (490 Million samples/second * 10 bits per sample = 4,900,000,000 bits/second. This is an estimate only)

Example: Digital Forward

- How about a 550 MHz forward band requiring 52 dB SNR?
- >1.1 Giga samples/second * 10 bits per sample = 11.0 Gb/s!!!

Assessment

It is more difficult and therefore more costly to manufacture digital return products. This may be a driver to use Analog DFB products for the new return applications. The selection of digital return products may be driven by distance and performance requirements. Another driver to move to digital return will be when there is near cost parity with DFB. This may be the case in the future with the new spectrum returns.

5.4 HFC Return Path Analysis and Model

Analog return path transmitters used in HFC applications need to be examined to determine their capability to transmit higher orders of modulation or additional channel loading while maintaining adequate performance. Operating conditions such as the optical link budget, actual channel loading, and desired operational headroom are all contributing factors with respect to performance of these transmitters. Here, operational headroom can be defined as the amount of dynamic range required to provide sufficient margin against the effects of temperature variation, variation from system components (transmitter, receiver, CM/CMTS, etc…), and ingress noise.

In optical networking, the amount of dynamic range for a given modulation format needs to be considered to ensure proper operation of the transmitter under fielded conditions. Typically, 12dB of operational headroom has been recommended for robust operation. However, there may be opportunities in the future to reduce the operational headroom by up to 3dB (perhaps to 9dB). In the future, smaller node sizes and shorter cascades may reduce the amount of ingress noise and the impact of temperature can be lessened with the use of analog DWDM lasers, which are tightly controlled over temperature.

Testing conducted on a standard,

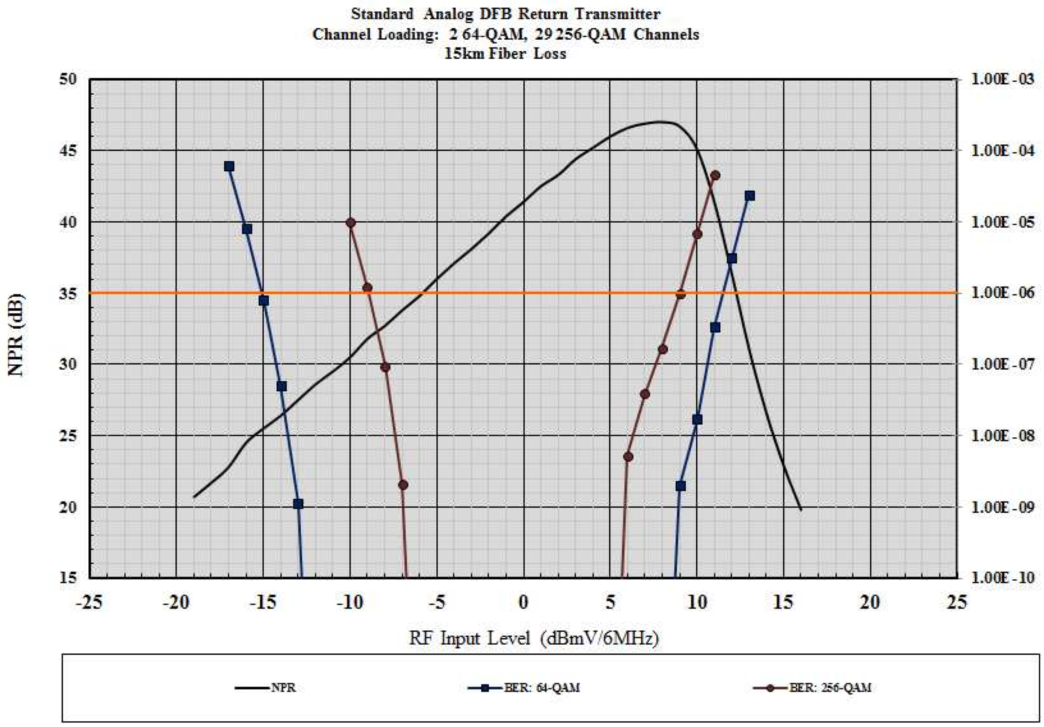


Figure 18 – High-split Standard Analog DFB Return Transmitter

analog DFB return transmitter (+3dBm) and an analog DWDM return transmitter, under “high split” loading conditions yielded acceptable dynamic range for 256 QAM operation. Figure 18 provides the results of the +3dBm analog DFB return transmitter. This test was conducted over a 15km link budget with a received power of -3dBm. The RF channel loading consisted of 31 QAM channels upstream containing two 64 QAM channels and twenty-nine 256 QAM channels. The measured dynamic range for a BER< 1E-06 for the 256 QAM channels is 18dB, which provides adequate operational headroom.

Figure 19 and Figure 20 provide data, taken at three frequency splits (low, mid, and high) using 64 QAM and 256 QAM channel loading, for an analog DWDM return transmitter, operating at +8dBm output power over a 16dB optical link (40km of fiber plus 8dB of passive loss). In the “high split” case, this transmitter provides 13dB of dynamic range (1E-06) for 256 QAM, adequate both for present day scenarios where 12dB of operational headroom may be required and for future scenarios where reduced operational headroom is sufficient.

.

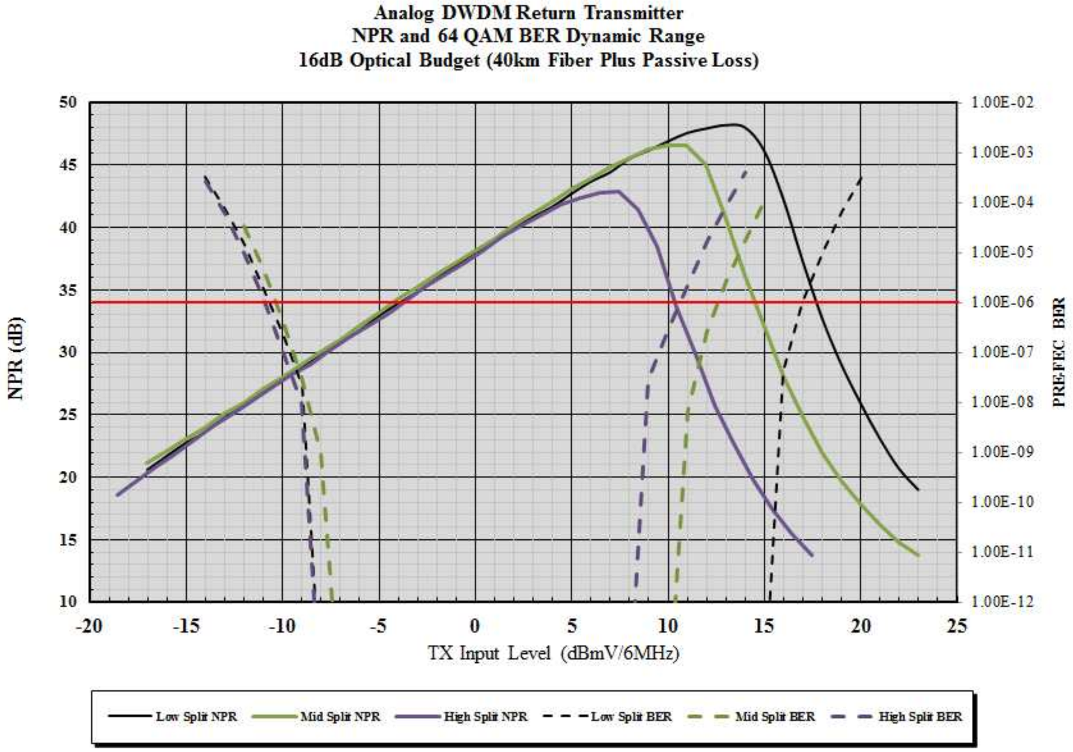


Figure 19 – Analog DWDM Transmitter: 64 QAM (Low/Mid/High Split)

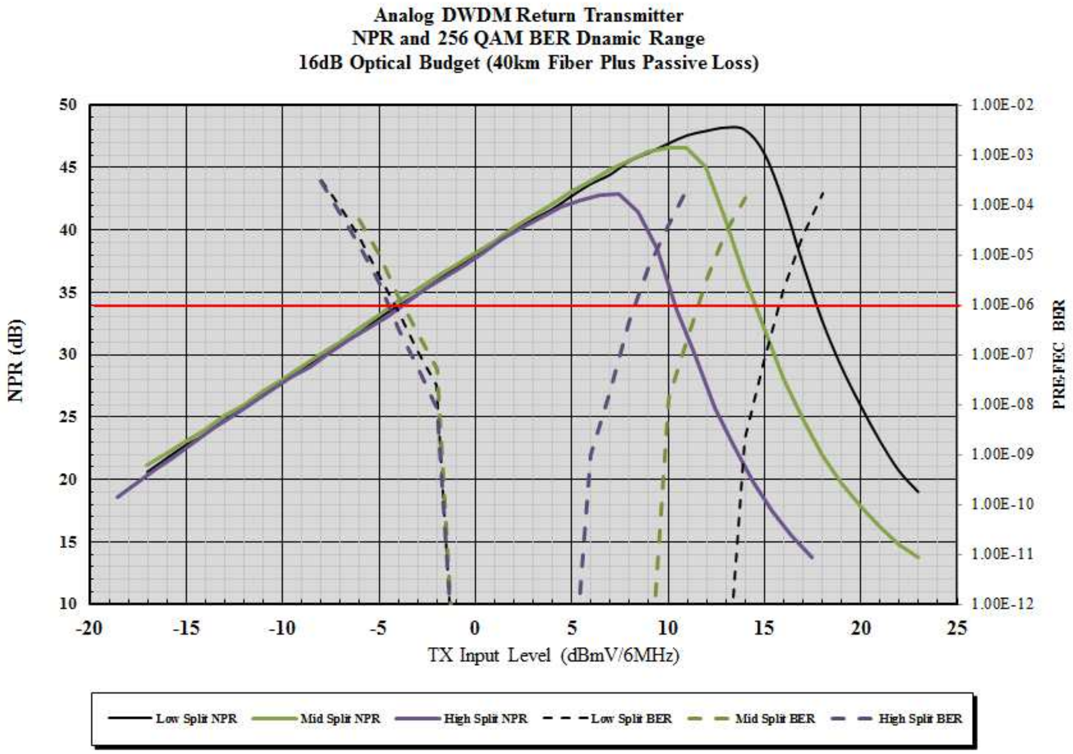


Figure 20 – Analog DWDM Transmitter: 256 QAM (Low/Mid/High Split)

6 SUMMARIES FOR HFC NETWORK COMPONENTS AND TOPOLOGY ANALYSIS

The analyses of the coaxial and optical network, the Hybrid Fiber Cox (HFC) network and the issues that need to be considered that may impact performance are summarized in Table 6. The spectrum selection will play a major role in terms of data capacity and network architecture.

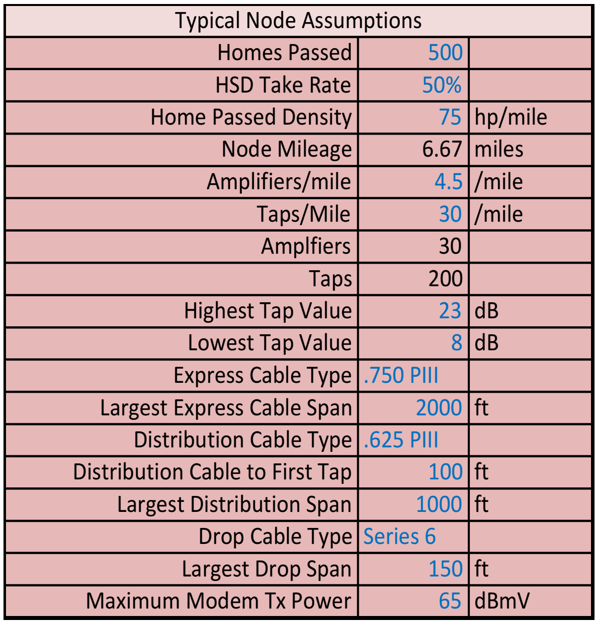
6.1 Major Considerations for Coaxial Network Performance

- **First Major Consideration:** Spectrum Selection
- **Second Major Consideration:** Path Loss or Attenuation
    - Overall System loss progressively increases as frequency increases, thus a major factor when considering higher frequency return.
    - Path Loss from the Last Tap including: Tap Insertion, Tap Port, Cable Loss Hardline, Cable Loss Drop, In Home Passive Loss to Modem/Gateway (these impact Top-splits)
- **Third Major Consideration:** Transmit Power Constraints
    - Modem maximum power output composite not to exceed +65 dBmV (to minimize power and cost, and maintain acceptable distortion)
- **Fourth Major Consideration:** Noise Funneling Effect
    - The effects of large number of return path amplifiers. This is not a factor at low frequency because the cable loss is low enough that a cable modem can provide adequate power level to maintain high C/N.
- **Fifth Major Consideration:** Optical CNR Contribution
- **Sixth Major Consideration:** Error Correction Technology

6.2 Analysis

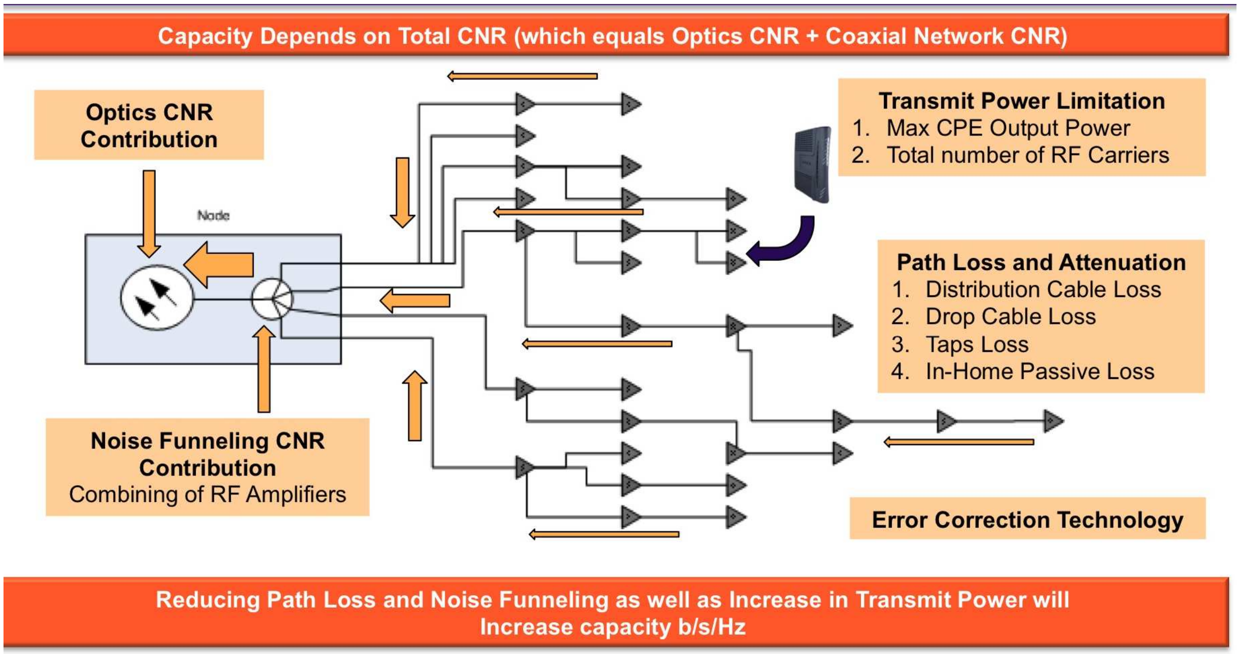
An analysis will be performed on the network in Figure 21 and described by Table 6

Table 6 – Node Service Group and Coaxial Network Assumptions



Typical Node Assumptions		
Homes Passed	500	
HSD Take Rate	50%	
Home Passed Density	75	hp/mile
Node Mileage	6.67	miles
Amplifiers/mile	4.5	/mile
Taps/Mile	30	/mile
Amplfiers	30	
Taps	200	
Highest Tap Value	23	dB
Lowest Tap Value	8	dB
Express Cable Type	.750 PIII	
Largest Express Cable Span	2000	ft
Distribution Cable Type	.625 PIII	
Distribution Cable to First Tap	100	ft
Largest Distribution Span	1000	ft
Drop Cable Type	Series 6	
Largest Drop Span	150	ft
Maximum Modem Tx Power	65	dBmV

For this analysis, 0.75" PIII class cable was assumed for express amplifier spans and 0.625" PIII class cable was assumed for tapped feeder spans. Table 7 shows what the gain requirements would be for an upstream express amplifier at the ranges of Figure 21.


Capacity Depends on Total CNR (which equals Optics CNR + Coaxial Network CNR)

Optics CNR Contribution

Transmit Power Limitation
1. Max CPE Output Power
2. Total number of RF Carriers

Path Loss and Attenuation
1. Distribution Cable Loss
2. Drop Cable Loss
3. Taps Loss
4. In-Home Passive Loss

Noise Funneling CNR Contribution
Combining of RF Amplifiers

Error Correction Technology

Reducing Path Loss and Noise Funneling as well as Increase in Transmit Power will Increase capacity b/s/Hz

Figure 21 – Major Considerations for Coaxial Network Performance

It is worth noting that the Sub-split, Mid-split and High-split gain requirements can be satisfied with commonly available components that are currently used in amplifier designs today and would likely involve no cost premium. However, the Top-Split options would likely require multistage high gain amplifiers to overcome predicted losses, which would be more costly.

It is also important to note that thermal control would likely become a major issue in the Top-split designs. Table 7 shows seasonal temperature swings of 5 to 6 dB loss change per amplifier span would be likely in the top-split solutions.

Reverse RF AGC systems do not exist today, and could be complex and problematic to design. Thermal equalization would be sufficient to control the expected level changes at 200 MHz and below, but it is not certain that thermal equalization alone will provide the required control above 750MHz. This needs more study.

Table 8 is a summary of path loss comparisons from home to the input of the first amplifier, which will ultimately determine the system operation point. It is interesting to note that as soon as the upper frequency is moved beyond the Sub-split limit, the maximum loss path tends toward the last tap in cascade as opposed to the first tap. There is a moderate increase in expected loss from 42 to 200 MHz, and a very large loss profile at 1000 MHz and above. The expected system performance can be calculated for each scenario.

Table 7 shows the compared performance calculations for the 500 home passed node outlined in Figure 21 and Table 6. The desired performance target is 256-QAM for each scenario; if it can be achieved, the throughput per subscriber will be maximized.

Table 7 – Express" (untapped) Segment Characterization

"Express" (untapped) Segment Characterization		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Typical Maximum Cable Loss (Amp to Amp 70 deg F)	dB	6.5	9.2	14.6	24.8	36.9	45.4	60.3
Additional Gain Required for Thermal Control (0 to 140 deg F)	+/-dB	0.5	0.6	1.0	1.7	2.6	3.2	4.2
Total Reverse Amplifier Gain Required	dB	**6.9**	**9.8**	**15.7**	**26.5**	**39.5**	**48.5**	**64.5**

For each approach, it is assumed that a CPE device is available with upstream bonding capability that can use the entire spectrum available at a reasonable cost. The number of bonded carriers transmitting must not exceed the maximum allowable modem transmit level, so the maximum power per carrier is calculated not to exceed 65 dBmV total transmitted power.

The maximum power, along with the worst-case path loss, yields the input level to the reverse amplifiers in the HFC Network. If the return level was greater than 15 dBmV, it was assumed that it would be attenuated to 15 dBmV.

Armed with the input level and station noise figure, the single station amplifier C/N is calculated and then funneled through the total number of distribution amplifiers serving the node to yield the C/N performance expected at the input of the node.

The HFC return optical links considered in the model are the analog DFB lasers or broadband digital return (BDR) systems. The selection DFB option was selected for the low frequency returns up to the High-split of 238 MHz. However, High-split 500 was modeled with Digital HFC Return. All the Top-split spectrum options used the Digital HFC Return optics as well.

In the model used to determine the performance of the optical link at several we used the following inputs for the various spectrum options and as well as optical link types, see the Table 9 below.

Table 8 – "Distribution" (tapped) Segment Characterization

"Distribution" (tapped) Segment Characterization		Sub-Split	Mid-Split	High-Split 238	High-Split 500	Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Worst Case Path Loss	dB	**29.0**	**30.0**	**34.5**	**43.1**	**67.0**	**75.3**	**80.0**
Path Loss from First Tap	dB	*29.0*	*30.0*	*32.2*	*35.4*	*44.2*	*43.2*	*50.1*
Distribution Cable Loss	dB	0.4	0.6	0.9	1.5	2.2	2.7	3.6
Tap Port Loss	dB	23.0	23.0	23.0	23.0	27.0	23.0	24.0
Drop Cable Loss	dB	2.1	2.9	4.7	7.4	10.4	12.8	17.0
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5
Path Loss from Last Tap	dB	*25.5*	*28.0*	*34.5*	*43.1*	*67.0*	*75.3*	*80.0*
Distribution Cable Loss	dB	4.0	5.7	9.1	15.0	22.0	27.0	35.9
Tap Insertion Loss	dB	7.9	7.9	9.2	9.2	18.0	21.8	12.6
Tap Port Loss	dB	8.0	8.0	8.0	8.0	12.0	9.0	9.0
Drop Cable Loss	dB	2.1	2.9	4.7	7.4	10.4	12.8	17.0
In Home Passive Loss to Modem	dB	3.5	3.5	3.5	3.5	4.6	4.7	5.5

Table 9 – Optical Segment Characterization Assumed per Spectrum Split

						Top-Split	Top-Split	Top Split
				High-Split	High-Split	(900-1125)	(1250-1700)	(2000-3000)
Optical Segment Characterization		Sub-Split	Mid-Split	238	500	Plus Sub-split	Plus Sub-split	Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	48	48

The inputs and results in Table 9 show following:

- 5 - 238 MHz have sufficient performance to support 256-QAM modulation at a 500 HHP node.
- 5 - 500 MHz have sufficient performance to support 128QAM modulation at a 500 HHP node.
- The top-split options suffer from cable loss, not to exceed +65 dBmV, and noise funneling.
  - The Top-split (900-1125) may operate at QPSK modulation with only 24 carriers at 6.4 widths.
  - The Top-split (1250-1700) may operate at QPSK modulation with only 3 carriers at 6.4 widths.
  - The Top-split (2000-3000) may operate at QPSK modulation with only 1 carrier at 6.4 widths. .

Further analysis of the Top-split options as shown in Table 10 through Table 13 concludes that reducing the node size, and thereby the funneled noise in the serving group could yield higher modulation capability. In these tables are red arrows, which highlight the key service group size and performance.

The comparison of low spectrum return options like that of Sub-split, Mid-split, and High-split versus the Top-split spectrum choices are measured in the following tables.

These table show that spectrum selection is one of the most important choices the cable operators could make for expanding the upstream. The spectrum options have vastly different performance capabilities when compared in the same cable topology. The Top-split option "MUST" reduce the noise funneling level, which requires smaller service group to increasing loading. Top-split allows only low order modulation and few carries will operate.

All of these assumptions are based on the use of single carrier QAM based systems using Reed-Solomon codes. Section 7 "DOCSIS PHY Technologies" describes the use of different error correction technologies and improvement that may be achieved in operating conditions and use of higher order modulation.

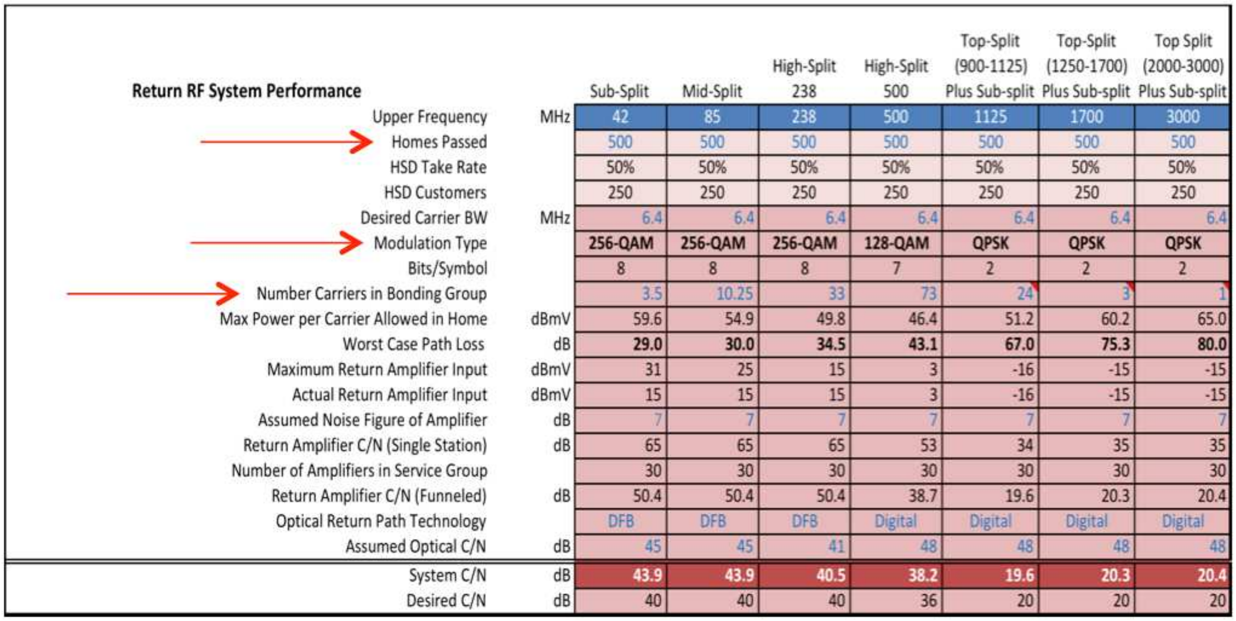
The use of Top-split frequencies will drive higher costs for additional node segmentation, nodes splits, and even running fiber deeper in the network.

The existing passive have an AC power choke resonances, which varies between 1050 - 1400 MHz making portions unusable or predictable. The recommendation on the low side is not to exceed 1050 MHz and high side 1125 MHz. Some passives may not even reach 1 GHz in cascade, so test your passives.

Plan to use low frequency return (Mid-split and High-split) and allow the downstream to use 1 GHz plus, like 1125 MHz or as high as the cascade of existing taps will allow.

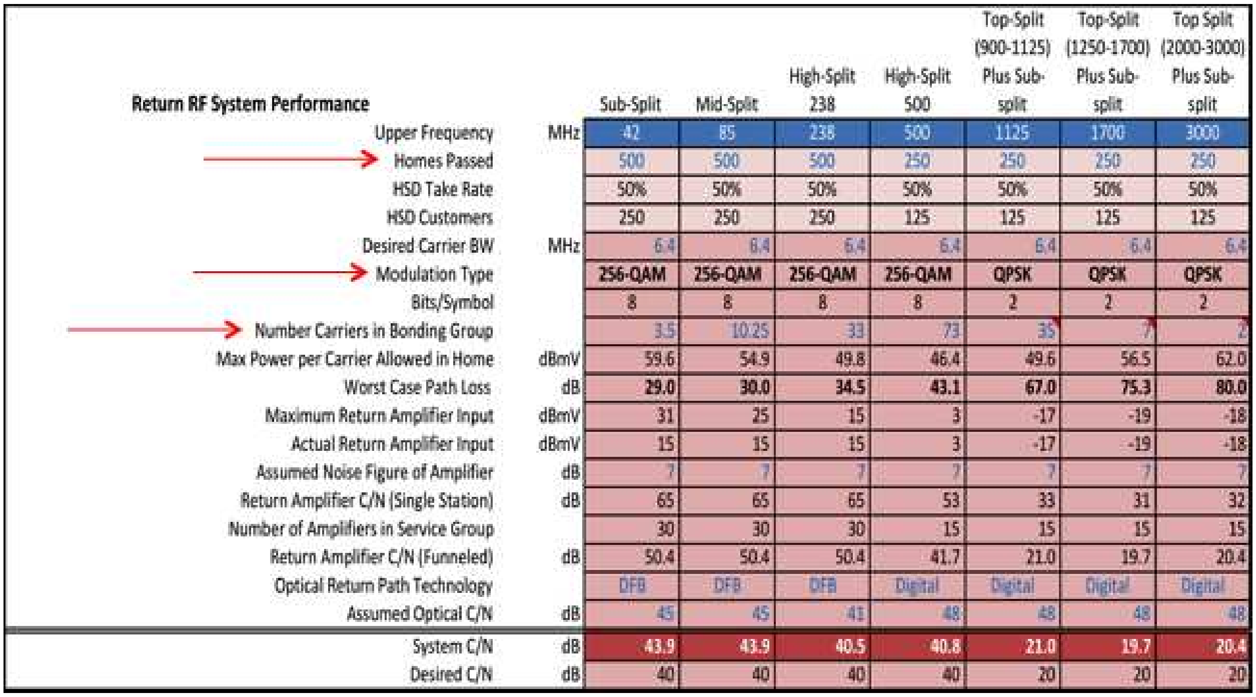
Consider touching the taps as a last resort.

Table 10 – Network Performance of a 500 HHP Optical Service Group



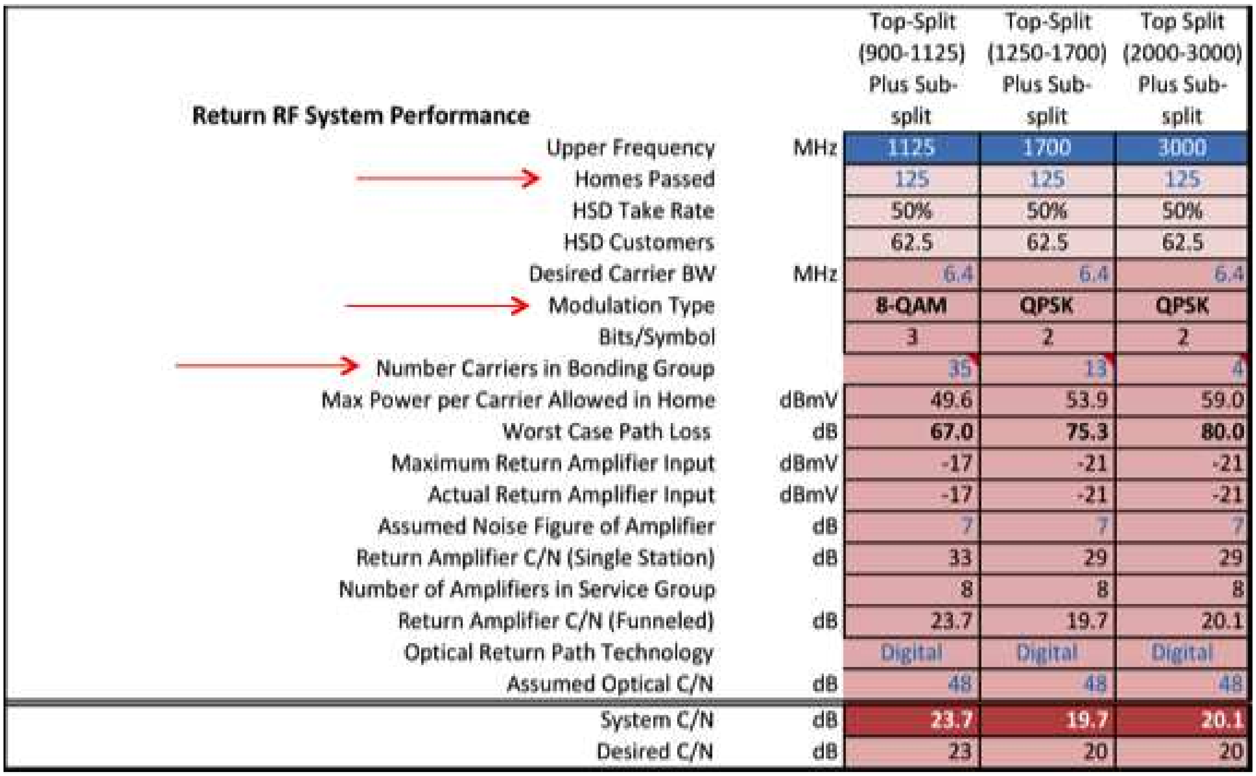
						Top-Split	Top-Split	Top Split
				High-Split	High-Split	(900-1125)	(1250-1700)	(2000-3000)
Return RF System Performance		Sub-Split	Mid-Split	238	500	Plus Sub-split	Plus Sub-split	Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Homes Passed		500	500	500	500	500	500	500
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	250	250	250	250
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4
Modulation Type		**256-QAM**	**256-QAM**	**256-QAM**	**128-QAM**	**QPSK**	**QPSK**	**QPSK**
Bits/Symbol		8	8	8	7	2	2	2
Number Carriers in Bonding Group		3.5	10.25	33	73	24	3	1
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	49.8	46.4	51.2	60.2	65.0
Worst Case Path Loss	dB	**29.0**	**30.0**	**34.5**	**43.1**	**67.0**	**75.3**	**80.0**
Maximum Return Amplifier Input	dBmV	31	25	15	3	-16	-15	-15
Actual Return Amplifier Input	dBmV	15	15	15	3	-16	-15	-15
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7
Return Amplifier C/N (Single Station)	dB	65	65	65	53	34	35	35
Number of Amplifiers in Service Group		30	30	30	30	30	30	30
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	38.7	19.6	20.3	20.4
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	48	48
System C/N	dB	**43.9**	**43.9**	**40.5**	**38.2**	**19.6**	**20.3**	**20.4**
Desired C/N	dB	40	40	40	36	20	20	20

Table 11 – 250 HHP Optical SG High-split 500 & Top-split Options



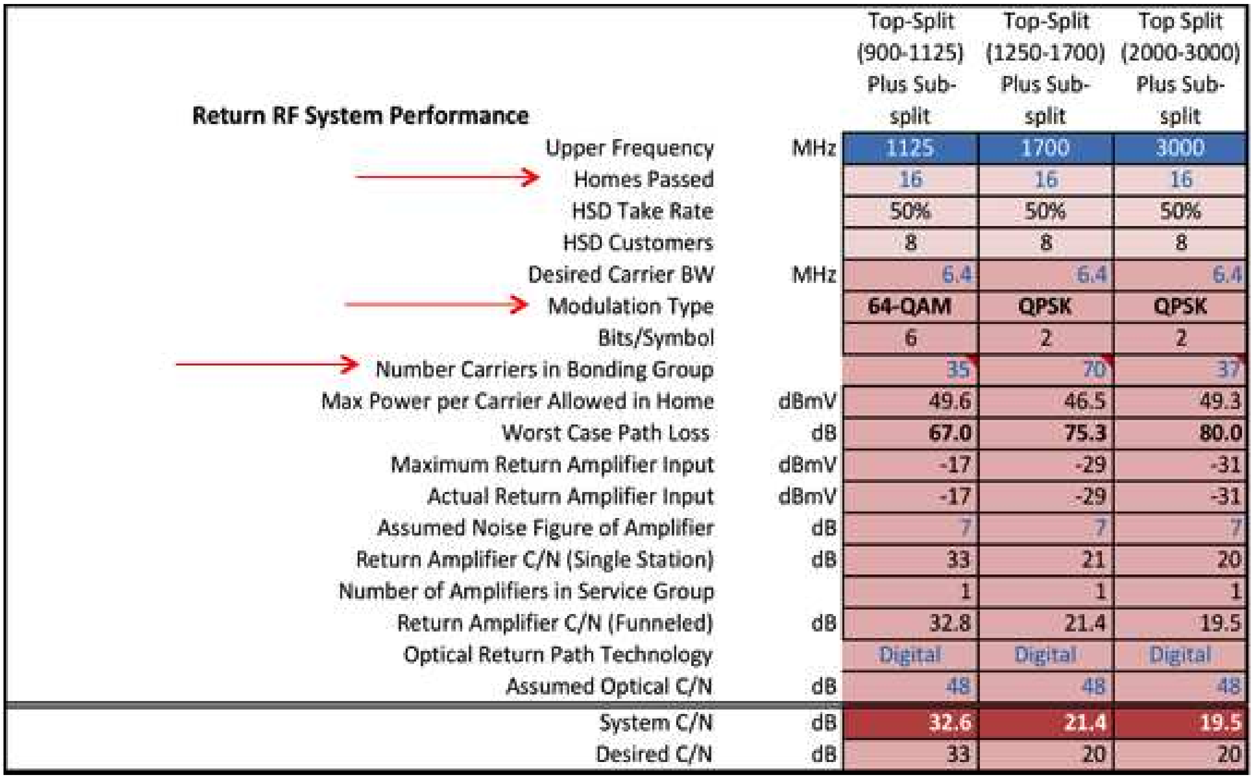
						Top-Split	Top-Split	Top Split
						(900-1125)	(1250-1700)	(2000-3000)
				High-Split	High-Split	Plus Sub-	Plus Sub-	Plus Sub-
Return RF System Performance		Sub-Split	Mid-Split	238	500	split	split	split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Homes Passed		500	500	500	250	250	250	250
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	125	125	125	125
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4
Modulation Type		**256-QAM**	**256-QAM**	**256-QAM**	**256-QAM**	**QPSK**	**QPSK**	**QPSK**
Bits/Symbol		8	8	8	8	2	2	2
Number Carriers in Bonding Group		3.5	10.25	33	73	35	7	2
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	49.8	46.4	49.6	56.5	62.0
Worst Case Path Loss	dB	**29.0**	**30.0**	**34.5**	**43.1**	**67.0**	**75.3**	**80.0**
Maximum Return Amplifier Input	dBmV	31	25	15	3	-17	-19	-18
Actual Return Amplifier Input	dBmV	15	15	15	3	-17	-19	-18
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7
Return Amplifier C/N (Single Station)	dB	65	65	65	53	33	31	32
Number of Amplifiers in Service Group		30	30	30	15	15	15	15
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	41.7	21.0	19.7	20.4
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	48	48
System C/N	dB	**43.9**	**43.9**	**40.5**	**40.8**	**21.0**	**19.7**	**20.4**
Desired C/N	dB	40	40	40	40	20	20	20

Table 12 – 125 HHP Optical SG Top-split Options



Return RF System Performance		Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	1125	1700	3000
Homes Passed		125	125	125
HSD Take Rate		50%	50%	50%
HSD Customers		62.5	62.5	62.5
Desired Carrier BW	MHz	6.4	6.4	6.4
Modulation Type		8-QAM	QPSK	QPSK
Bits/Symbol		3	2	2
Number Carriers in Bonding Group		35	13	4
Max Power per Carrier Allowed in Home	dBmV	49.6	53.9	59.0
Worst Case Path Loss	dB	67.0	75.3	80.0
Maximum Return Amplifier Input	dBmV	-17	-21	-21
Actual Return Amplifier Input	dBmV	-17	-21	-21
Assumed Noise Figure of Amplifier	dB	7	7	7
Return Amplifier C/N (Single Station)	dB	33	29	29
Number of Amplifiers in Service Group		8	8	8
Return Amplifier C/N (Funneled)	dB	23.7	19.7	20.1
Optical Return Path Technology		Digital	Digital	Digital
Assumed Optical C/N	dB	48	48	48
System C/N	dB	23.7	19.7	20.1
Desired C/N	dB	23	20	20

Table 13 – 16 HHP Optical SG Top-split Options



Return RF System Performance		Top-Split (900-1125) Plus Sub-split	Top-Split (1250-1700) Plus Sub-split	Top Split (2000-3000) Plus Sub-split
Upper Frequency	MHz	1125	1700	3000
Homes Passed		16	16	16
HSD Take Rate		50%	50%	50%
HSD Customers		8	8	8
Desired Carrier BW	MHz	6.4	6.4	6.4
Modulation Type		64-QAM	QPSK	QPSK
Bits/Symbol		6	2	2
Number Carriers in Bonding Group		35	70	37
Max Power per Carrier Allowed in Home	dBmV	49.6	46.5	49.3
Worst Case Path Loss	dB	67.0	75.3	80.0
Maximum Return Amplifier Input	dBmV	-17	-29	-31
Actual Return Amplifier Input	dBmV	-17	-29	-31
Assumed Noise Figure of Amplifier	dB	7	7	7
Return Amplifier C/N (Single Station)	dB	33	21	20
Number of Amplifiers in Service Group		1	1	1
Return Amplifier C/N (Funneled)	dB	32.8	21.4	19.5
Optical Return Path Technology		Digital	Digital	Digital
Assumed Optical C/N	dB	48	48	48
System C/N	dB	32.6	21.4	19.5
Desired C/N	dB	33	20	20

7 DOCSIS PHY TECHNOLOGIES

7.1 ATDMA & J.83 (Single Carrier QAM)

7.1.1 Potential for Higher Symbol Rate A-TDMA

With the increasing deployment of wideband (6.4 MHz) 64-QAM upstream channels and in some cases bonding of upstream channels, operators are beginning to take advantage of the most powerful set of DOCSIS 2.0 and DOCSIS 3.0 tools available for maximizing capacity of a given channel and delivering higher peak service rates.

Nonetheless, as these advancements have matured – they are 11 years and 6 years since initial release, respectively – the pace of bandwidth consumption and market demand for higher rate service has continued. While it has slowed in the upstream relative to the downstream, it has nonetheless marched forward such that we speak of 10 Mbps and 20 Mbps upstream service tiers today, with an eye towards 100 Mbps in the near future.

The nature of reasonable traffic asymmetry ratios for efficient operation of DOCSIS may pull 100 Mbps along as well as the downstream heads towards a 1 Gbps. Certainly, for DOCSIS-based business subscribers – already outfitted with CMs, for example, or without convenient access to a fiber strand – 100 Mbps is often not just an objective but a requirement.

It is also likely one that operators can derive increased revenue from and consider SLA management options to deliver higher-end services.

7.1.1.1 100 Mbps Residential Upstream

For residential services, while a need for a 1 Gbps service appears far off into the next decade, a 100 Mbps offering is a reasonable target for the near term, and projects as the CAGR-based requirement in 4-6 years for 20 Mbps services today using traffic doubling periods of every two years (approximately 40%) or every three years (approximately 25%).

Unfortunately, today, only through bonding four 64-QAM carriers can 100 Mbps service rate, accounting for overhead loss to net throughput, be provided. The addition of 256-QAM as a modulation profile, to be described in the next section, helps to alleviate this somewhat by enabling a 100 Mbps rate to be offered over three bonded upstreams.

In either case, however, the added complexity of latency of bonding is required to achieve what is expected to be a fundamental service rate target to likely be implemented in bulk. Latency in particular has become a topic generating much interest because of the impact packet processing delay can have on gaming.

While relatively low average bandwidth, high quality gaming demands instantaneous treatment for the fairness and QoE of the gaming audience. Performance has been quantified against latency and packet loss by game type [1], and the variations in performance have led to solution variation exploiting the video architecture, managing server locations, and using potential QoS or priority mapping schemes. While bonding is not the dominant network constraint, elimination of

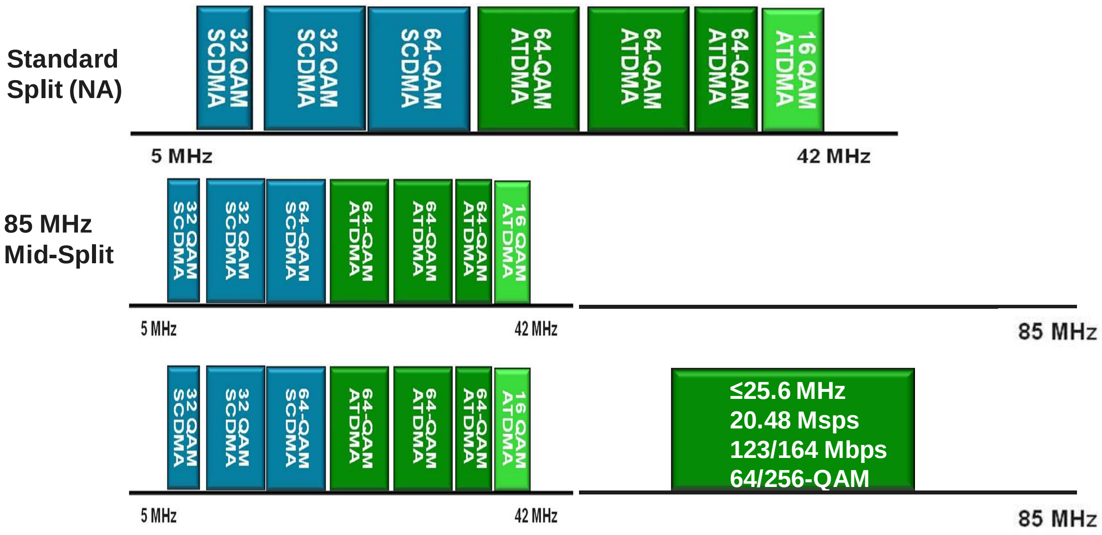


Figure 22 – Higher Symbol Rates Applied Over an 85-MHz Mid-Split Architecture

bonding is favorable for improving processing latency for gaming and other latency-sensitive applications that may arise in the future.

There is also a concern that upstream bonding capability will be limited to a maximum of 8 carriers, due to the increasing complexity associated with the tracking of packets and scheduling operation to process the payload across PHY channels. While operators are not ready to bond even four channels today, if this eight-channel limit were indeed the case, then peak upstream speeds could never exceed 240 Mbps at the PHY transport rate, or 320 Mbps under a 256-QAM assumption.

So, while 1 Gbps of capacity or service rate is likely not a near-term concern, a path to achieve that within the HSD infrastructure should be made available for the long-term health and competitiveness of the network.

Both concerns – 1 Gbps and the bonding implementation for 100 Mbps services – are addressed by a straightforward, integer-scale widening of the symbol rate of today's robust, single-carrier architecture. This approach is shown in Figure 22, where it is displayed as it might be implemented with an 85 MHz Mid-Split architecture. While not obvious from Figure 23, because of the full legacy band, two wider symbol rate channels could be operated within an 85 MHz architecture.

With an excess bandwidth (α) of 15%, there would be a reduced relative bandwidth overhead over today's α = .25. This represents a savings of over 2 MHz of excess bandwidth at 20.48 Msps symbol rates, and two channels would consume less than 48 MHz of spectrum. This leaves plenty of additional spectrum for legacy carriers in a clean part of the lower half of the upstream.

By increasing the maximum symbol rate by a factor of four, from 5.12 Msps to 20.48 Msps, a basic unit of single-carrier operation now is capable of being a 100 Mbps net throughput channel, and simple delivery of this key peak speed service rate is achieved.

7.1.1.2 Achieving 1 Gbps

By bonding eight such carriers together, coupled with the introduction of 256-QAM,

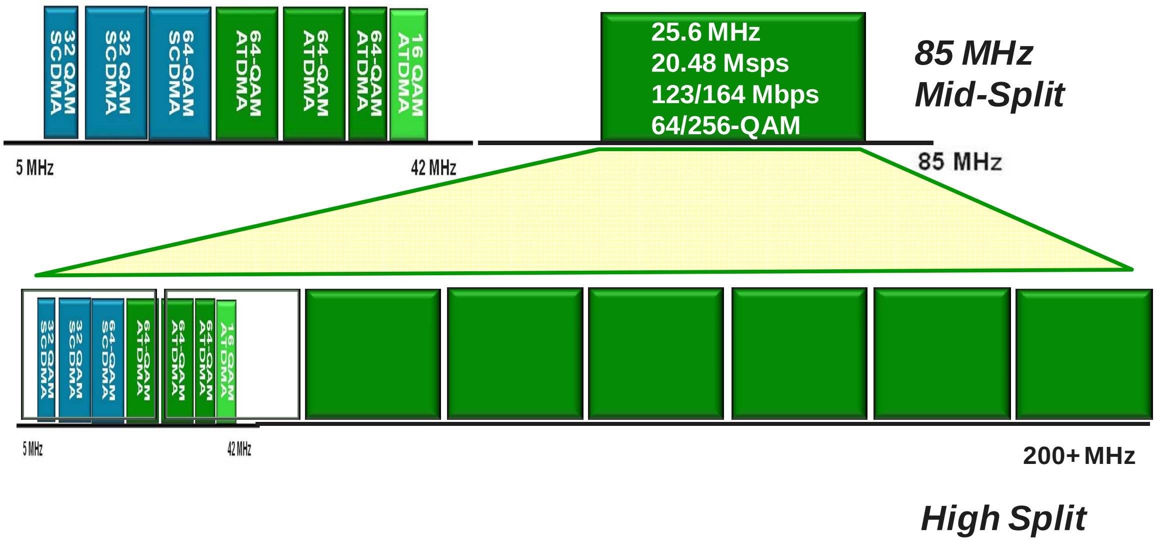


Figure 23 – 8x Bonded Higher Symbol Rates Over a "High-Split" Architecture

an aggregate throughput of over 1 Gbps can also now be enabled with a 4x symbol rate approach, when required. While it is not clear yet if there is an 8-bonded upstream limit, this technique takes that potential risk off of the table. This scenario is shown in Figure 23. In principle, these eight carriers can fit within 200 MHz of spectrum, making the approach comfortably compatible, even with the minimum bandwidth "high-split" spectrum architecture.

In practice, given that legacy services already populate the return path and will only grow between now and any new evolution of the channel or architecture, a high-split based upon a 250 MHz or 300 MHz upstream band is the more likely deployment scenario, with the possibility that it could increase further over time. A flexible FDD implementation would allow the traffic asymmetry to be managed as an operator sees fit based upon need.

7.1.1.3 Wider Band Channel Implications

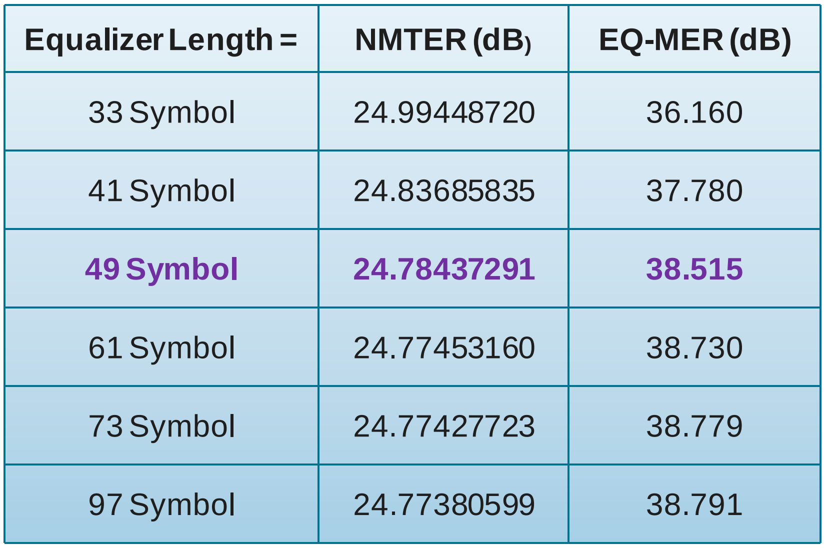
The complexity of DOCSIS 2.0's wideband 64-QAM is largely around the ability to equalize the signal under frequency response distortions. The 24-Tap architecture evolved from the 8-Tap structure of DOCSIS 1.0, providing a very powerful tool for both ISI mediation as well as for plant characterization and diagnostics through the use of the pre-equalization (pre-EQ) functionality.

Every individual CM has its RF channel effectively characterized for reflection content and frequency response distortions, such as roll-off and group delay distortion. Use of pre-EQ has become an immensely powerful tool for MSOs in optimizing their return and efficiently diagnosing and zeroing in on problem locations. Optimization of use has matured and MSOs have learned how best to make use of this powerful tool as wideband 64-QAM has become a critical component of the upstream strategy.

Today's equalizer architecture is also, therefore, quite mature, and the ability to provide real-time processing of burst upstream signals has advanced considerably in the intervening years per Moore's Law as it pertains to processing power. This is important to consider as we ponder higher symbol rates.

Higher symbol rates translate directly to wider channel bandwidths, and thus the equalizer is impacted by this technique. For the T-Spaced implementation of DOCSIS 3.0, if the symbol rate increases by a factor of four, then time span of an equalizer using the same number of taps has *shrunk* by a factor of one-quarter. In other words, the equalizer length must be increased by a factor of four to provide the same span of compensation for micro-reflections, for example.

Table 14 – Post-EQ MER as a Function of Tap Span



Equalizer Length =	NMTER (dB)	EQ-MER (dB)
33 Symbol	24.99448720	36.160
41 Symbol	24.83685835	37.780
49 Symbol	**24.78437291**	**38.515**
61 Symbol	24.77453160	38.730
73 Symbol	24.77427723	38.779
97 Symbol	24.77380599	38.791

Since equalizer taps are a complex multiply operation, it means 16x as many calculations take place in the equivalent algorithm. While this sounds imposing, considering that the 24-Tap structure is over ten years old, a 16x increase in processing is actually well below the “Moore’s Law” rate of compute power capability growth.

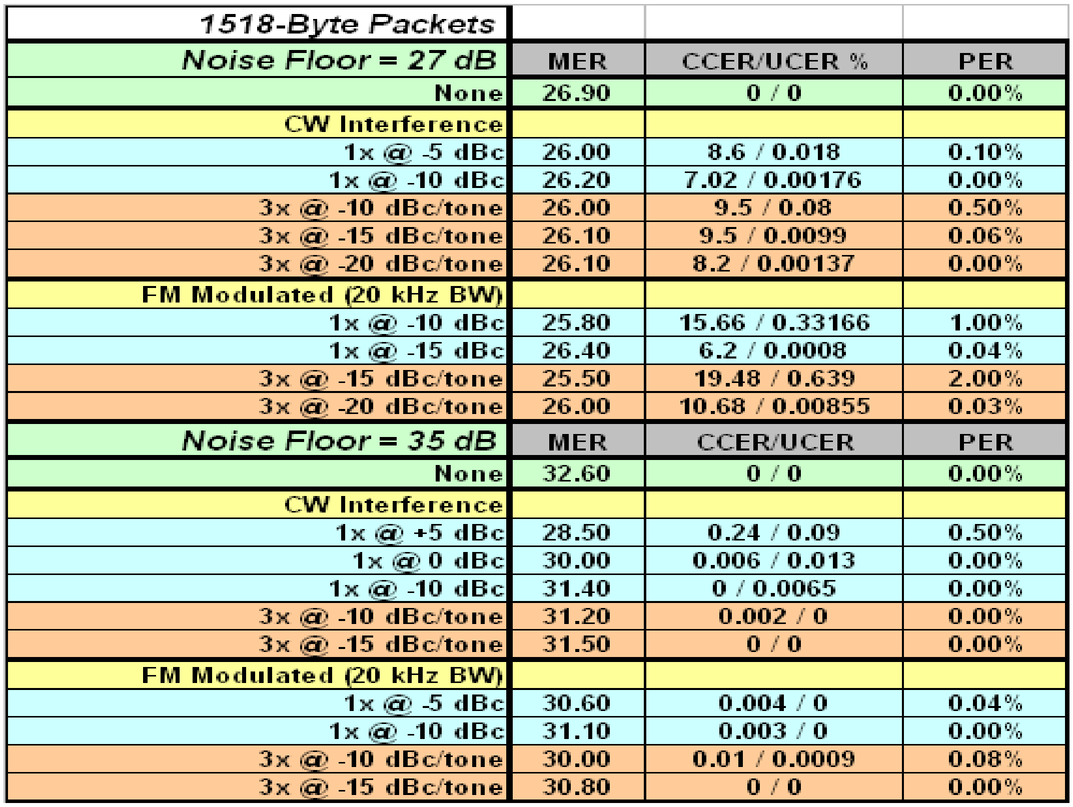
For example, at a doubling of capability even every two years, this would project out to more than 32x the processing power available today than was available when the current equalizer was *deployed*, much less designed. The technology capability to achieve a 96-Tap structure does not appear to be an obstacle, although its fit within modest variations to existing silicon is an important consideration.

There is some evidence that the 4x symbol rate may be a reasonable extension for today’s equalizer architecture to handle. Recent characterization of wideband channels in the > 1 GHz band has shown that the dithering on the last few taps in the equalizer may be minimal for short cascades.

In these environments, spectral roll-off caused by many filters in cascade is limited, as is the group delay impact of this roll-off. Also, fewer connected homes means fewer opportunities for poor RF terminations and the micro-reflections they cause.

Table 14 quantifies test results for a 4x symbol width in an unspecified part of the coaxial band at 1.5 GHz through a cascade of taps in the passive leg of the plant. The frequency response above 1 GHz is generally not specified today. However, this characterization was done with taps with faceplates installed to extend their bandwidth to about 1.7 GHz.

Table 15 – A-TDMA Narrowband Interference Suppression Capability



1518-Byte Packets			
Noise Floor = 27 dB	MER	CCER/UCER %	PER
None	26.90	0 / 0	0.00%
CW Interference			
1x @ -5 dBc	26.00	8.6 / 0.018	0.10%
1x @ -10 dBc	26.20	7.02 / 0.00176	0.00%
3x @ -10 dBc/tone	26.00	9.5 / 0.08	0.50%
3x @ -15 dBc/tone	26.10	9.5 / 0.0099	0.06%
3x @ -20 dBc/tone	26.10	8.2 / 0.00137	0.00%
FM Modulated (20 kHz BW)			
1x @ -10 dBc	25.80	15.66 / 0.33166	1.00%
1x @ -15 dBc	26.40	6.2 / 0.0008	0.04%
3x @ -15 dBc/tone	25.50	19.48 / 0.639	2.00%
3x @ -20 dBc/tone	26.00	10.68 / 0.00855	0.03%
Noise Floor = 35 dB	MER	CCER/UCER	PER
None	32.60	0 / 0	0.00%
CW Interference			
1x @ +5 dBc	28.50	0.24 / 0.09	0.50%
1x @ 0 dBc	30.00	0.006 / 0.013	0.00%
1x @ -10 dBc	31.40	0 / 0.0065	0.00%
3x @ -10 dBc/tone	31.20	0.002 / 0	0.00%
3x @ -15 dBc/tone	31.50	0 / 0	0.00%
FM Modulated (20 kHz BW)			
1x @ -5 dBc	30.60	0.004 / 0	0.04%
1x @ -10 dBc	31.10	0.003 / 0	0.00%
3x @ -10 dBc/tone	30.00	0.01 / 0.0009	0.08%
3x @ -15 dBc/tone	30.80	0 / 0	0.00%

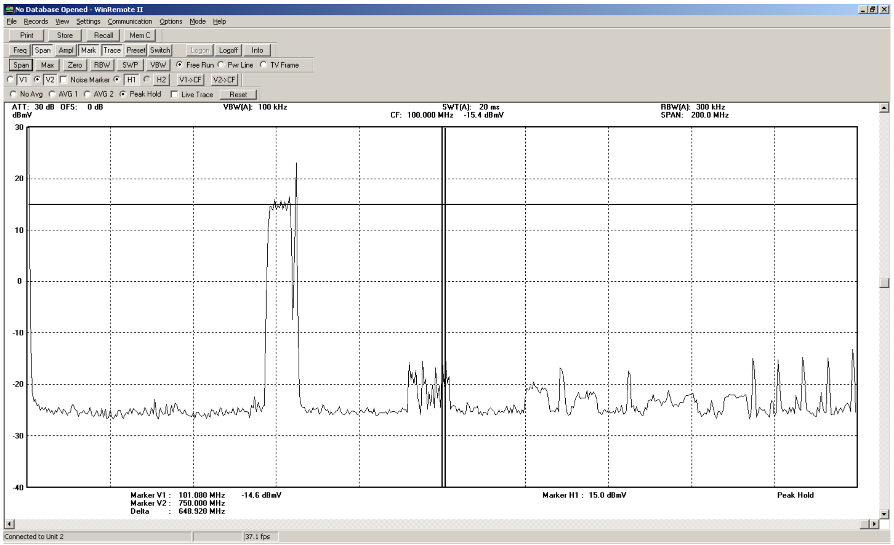
Evident in this essentially “N+0” segment is that the MER after equalization improves only incrementally as we include more taps up to about T = 49 symbols. The T=48 symbols would, of course, mean a doubling of the Tap span for a quadrupling of the symbol rate.

As cascades reduce and new, cleaner upstream bands are used to exploit more capacity, favorable channel condition with respect to frequency response are likely to result. This data certainly is favorable to the thought that even above 1 GHz, where little has been defined for CATV, a 4x symbol rate can be accommodated for the downstream.

Now, switching to the upstream, the spectrum expected to be exploited is in fact well-defined – return loss requirements and all – and will benefit from the same architectural migration shifts to shorter cascades and passive coax architectures. Because of this, the potential complexity increase of a 96-Tap equalizer and the corresponding time span that it supports may not be necessary to effectively use an extended upstream with 4x symbol rate transport. This may be valuable news to silicon implementers who may then be able to allocate silicon real estate and MIPS to other receiver processing functions.

7.1.1.4 Narrowband Interference

Another concern associated with increased symbol rates is the increased likelihood by a factor of four on average (slightly less with less excess bandwidth, of course) that narrowband interference will fall in-band and degrade the transmission. Unlike multi-carrier techniques, which can drop sub-channels out that can become



No Database Opened - WinRemote II
File Records View Settings Communication Options Mode Help
Print Store Recall Mem C
Freq Span Ampl Mark Trace Preset Switch Logon Logoff Info
Span Max Zero RBW SWP VBW
● Free Run ○ Pwr Line ○ TV Frame
○ V1 ● V2 □ Noise Marker ● H1 ○ H2 V1->CF V2->CF
○ No Avg ○ AVG 1 ○ AVG 2 ● Peak Hold □ Live Trace Reset
30
20
10
0
-10
-20
-30
-40

ATT: 30 dB OFS: 0 dB
dBmV
VBW(A): 100 kHz
SWT(A): 20 ms
CF: 100.000 MHz -15.4 dBmV
RBW(A): 300 kHz
SPAN: 200.0 MHz

Marker V1 : 101.080 MHz -14.6 dBmV
Marker V2 : 750.000 MHz
Delta : 648.920 MHz
Marker H1 : 15.0 dBmV
Peak Hold

Connected to Unit 2
37.1 fps

Figure 24 – Observed FM Band Interference on Deliberately Poor CM RF Interface

impaired by such interference (at the expense of throughput), a single carrier system must find a way to suppress the interference and reconstruct the symbol without it.

Fortunately, such techniques have matured, and today's ingress cancellation technology is very powerful in delivering full throughput performance in the face of strong narrowband interference. These processing algorithms sense ingress and adapt the rejection to the location and level of detected interference.

Table 15 quantifies the measured robustness under controlled testing of the DOCSIS 3.0 narrowband interference mechanism in suppressing interference [8]

It is readily apparent that today's DOCSIS 3.0 narrowband incision capability handles in-band interference very effectively over a range of much-worse-than-typical SNR, impulse, and interference conditions.

For example, at an SNR of 27 dB, which represents the return path quality of very old Fabry-Perot return paths long since replaced in most cases (DOCSIS minimum being 25 dB), it takes three tones of 20 kHz bandwidth a piece and adding up to about a 10 dB C/I to register a PER that might be considered objectionable (2%) from a user QoE perspective.

A borderline 1% PER occurs at C/I = 10 dB for a single interferer. These C/I values represent very high levels of plant interference in practice, although not completely uncommon, especially at the low end, shortwave area of the return band.

At SNRs closer to what is expected today (35 dB), no static interference case has PER of any consequence, even with C/I taken to 5 dB (modulated) and -5 dB (unmodulated) tones. This data suggests that wider symbols in the ever-cleaner part of the spectrum are likely to comfortably operate, quite robustly.

As the high-split architecture is deployed, interference levels over the air bands – particularly FM radio in North America, as discussed in Section 3.3.2– become important to understand. Figure 24 shows a field test with a diplex split extended above the 85 MHz Mid-Split for purposes of quantifying the potential for such interference.

In what was a very harsh metropolitan environment, with older plant cabling and nearby FM towers, a deliberately loose fitted CM resulted in relatively modest. However, because it is a wideband spectrum of channels, it would not be able to be compensated for by receiver ingress suppression. The roughly 30 dB of SNR would still yield high throughput, though because the interference effect may have non-Gaussian qualities, the uncorrected error rates may be higher.

However, it is expected this would be well within FEC capability to yield error-free output. Similar C/I's resulted with various arrangements of splitters, modems and deliberately radially and longitudinally damaged cables. While only one example, given the ground conditions, this trial was highly encouraging with respect to the high split running well in the region of spectrum occupied by FM radio over the air.

Note that the ingress-only performance shown in Table 16 in fact identifies a potential *advantage* of the single carrier approach to interference suppression relative to OFDM – there is no loss of available data rate; there is instead an overhead increase for channel knowledge. In OFDM, the C/I on a single sub-channel and closest neighbors, must be removed or have their modulation profile decreased at the cost of available data rate. If the C/I environment worsens however, OFDM can gracefully degrade where SC has threshold behavior.

7.1.1.5 Joint Impairment Thresholds

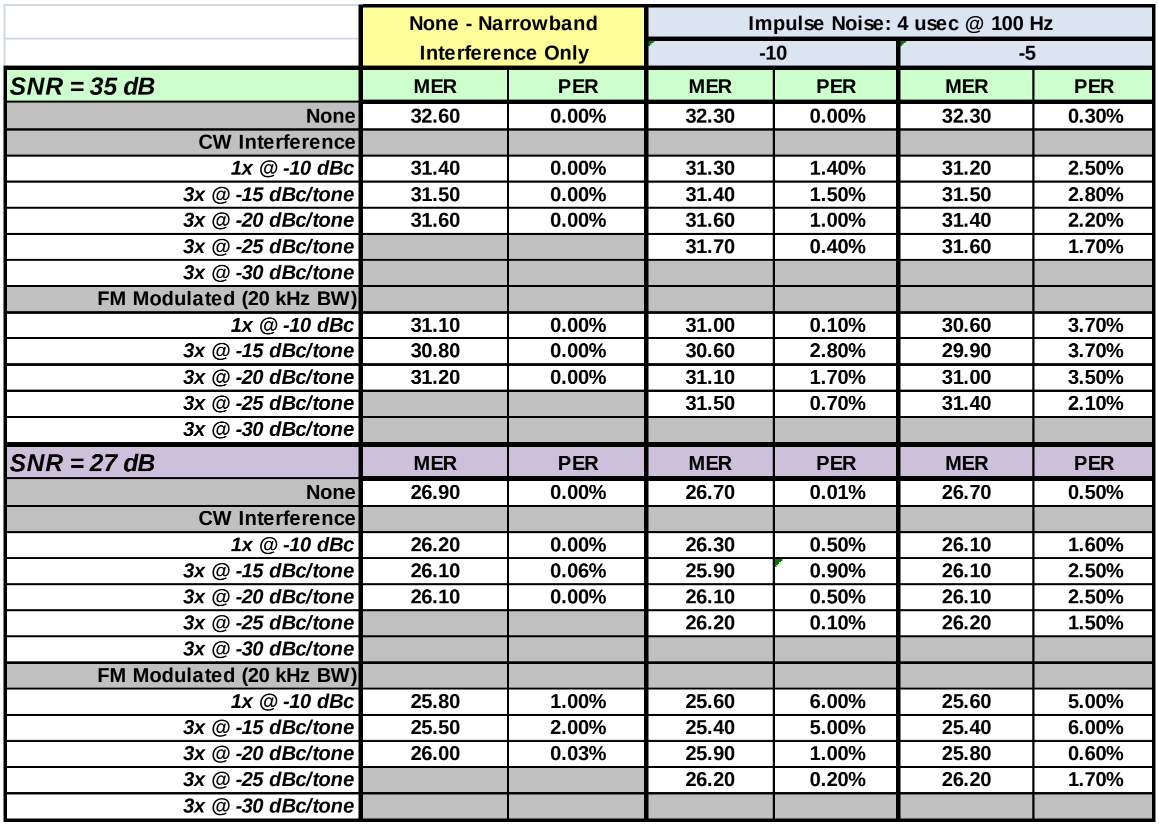
When impulse noise is added as a joint impairment, we can then begin to count more cases of potentially objectionable PER from a user QoE perspective. However, it is quite clear from the comparison that the error rate is being dictated by the very impulse noise component. This is indeed an area where OFDM would have benefits, much like will be seen with S-CDMA, through the use of longer symbol times to outlast the impulse events.

Of course, impulse noise tends to be restricted to the low end of the return band. Above about 20 MHz, there is little evidence that the joint impairment scenario occurs in a meaningful way to degrade A-TDMA performance. Indeed, where A-TDMA is the most vulnerable is relative to impulse noise. It is left to defend itself only with FEC today, and this has been proven to be sufficient in the vast majority of 64-QAM deployments implemented in the middle to high end of the 42 MHz upstream spectrum.

7.1.1.6 Summary

DOCSIS is currently a predominantly A-TDMA system, and exclusively so in the vast majority of deployment worldwide. As such, a natural and simple extension, with perhaps only minor impact on silicon development, is the increase the symbol rate of the already existing protocol to be better aligned with service on the near-term horizon, but also compatible with the direction of data services requirements for the long term.

Table 16 – A-TDMA Performance with Interference and Impulse Noise



	None - Narrowband Interference Only		Impulse Noise: 4 usec @ 100 Hz			
			-10		-5	
SNR = 35 dB	MER	PER	MER	PER	MER	PER
None	32.60	0.00%	32.30	0.00%	32.30	0.30%
CW Interference						
1x @ -10 dBc	31.40	0.00%	31.30	1.40%	31.20	2.50%
3x @ -15 dBc/tone	31.50	0.00%	31.40	1.50%	31.50	2.80%
3x @ -20 dBc/tone	31.60	0.00%	31.60	1.00%	31.40	2.20%
3x @ -25 dBc/tone			31.70	0.40%	31.60	1.70%
3x @ -30 dBc/tone						
FM Modulated (20 kHz BW)						
1x @ -10 dBc	31.10	0.00%	31.00	0.10%	30.60	3.70%
3x @ -15 dBc/tone	30.80	0.00%	30.60	2.80%	29.90	3.70%
3x @ -20 dBc/tone	31.20	0.00%	31.10	1.70%	31.00	3.50%
3x @ -25 dBc/tone			31.50	0.70%	31.40	2.10%
3x @ -30 dBc/tone						
SNR = 27 dB	MER	PER	MER	PER	MER	PER
None	26.90	0.00%	26.70	0.01%	26.70	0.50%
CW Interference						
1x @ -10 dBc	26.20	0.00%	26.30	0.50%	26.10	1.60%
3x @ -15 dBc/tone	26.10	0.06%	25.90	0.90%	26.10	2.50%
3x @ -20 dBc/tone	26.10	0.00%	26.10	0.50%	26.10	2.50%
3x @ -25 dBc/tone			26.20	0.10%	26.20	1.50%
3x @ -30 dBc/tone						
FM Modulated (20 kHz BW)						
1x @ -10 dBc	25.80	1.00%	25.60	6.00%	25.60	5.00%
3x @ -15 dBc/tone	25.50	2.00%	25.40	5.00%	25.40	6.00%
3x @ -20 dBc/tone	26.00	0.03%	25.90	1.00%	25.80	0.60%
3x @ -25 dBc/tone			26.20	0.20%	26.20	1.70%
3x @ -30 dBc/tone						

While many advances in PHY technology have occurred, the existing signal flow, knowledge base, silicon maturity, and understanding of management of the single carrier approach all favorably weigh in towards working to tweak something that doesn’t need outright fixing. Couple this maturity with the ability of single carrier tools to handle the upstream channel environment across the vast majority of the spectrum, creating a higher symbol rate of 4x, as described here, represents a logical, incremental, low-risk step for the transmission system portion of the PHY.

7.1.2 256-QAM Upstream

With the introduction of DOCSIS, cable operators created a specification for high speed data services that was built around the architecture and technology realities of the time – large serving groups of subscribers funneled through deep cascades of amplifiers and onto into a single laser transmitter – typically of the low-cost, low quality, Fabry-Perot variety – and with the anticipation of a lot of unwanted interference coming along for the ride.

The resulting requirements spelled out ensured robust operation under the condition of a 25 dB SNR assumption, among other impairments defined. Robust performance was assured through the use of relatively narrowband, robust modulation formats (QPSK and 16-QAM), a limited number of channels competing for spectrum power , and the ability to use powerful forward error correction.

Now, of course, many of the characteristics that defined the return have changed significantly, and DOCSIS 2.0 took advantage of many of them by calling for support of a 64-QAM modulation profile of up to twice the bandwidth if conditions allowed it.

It was not the case everywhere that it could be supported, but all phases of evolution were trending towards the ability to squeeze more and more capacity out of the return. Better, Distributed Feedback (DFB), analog optics became cost effective, digital return optics came on the scene, cascades shortened as serving groups shrunk during node splitting operations, and lessons learned over the years brought improvements in return path alignment and maintenance practices.

These same lessons brought about the introduction of S-CDMA, based on a better understanding of the characteristics of the low end of the return spectrum.

DOCSIS 2.0 itself is now over ten years old. DOCSIS 3.0 subsequently added channel bonding for higher peak speeds, as well as calling our support for return path extension in frequency up to 85 MHz.

Fortunately, the HFC architecture and supporting technology has continued to evolve favorably towards more upstream bandwidth, used more efficiently. In Section 2, the case was made for the use of the 85 MHz mid-split as an excellent first step for cable operators looking to add essential new bandwidth for upstream services. In this section, we will show how today’s return paths, extended to 85 MHz, are now capable of exploiting this band while also increasing the modulation profile to 256-QAM. It is within the capability of the upstream and demonstrably proven in the field that a 256-QAM modulation profile can be supported, and over a wider band than the legacy 42 MHz bandwidth in North America and the 65 MHz Euro split.

7.1.2.1 Upstream Link Analysis

While early generation CMTS equipment was designed to support 16-QAM as the maximum modulation profile, vendors generally provided enough margin in their systems to enable 64-QAM once networks evolved towards better HFC optics. 64-QAM was subsequently embraced in DOCSIS 2.0.

In Figure 17 through Figure 20 in Section 5, we introduced noise power ratio (NPR) curves to characterize return path optical technologies. NPR curves have the desirable feature of representing a worst-case (no TDMA operating) fully loaded return link from a signal stimulus standpoint while simultaneously quantifying the SNR and S/(N+D) on a single curve.

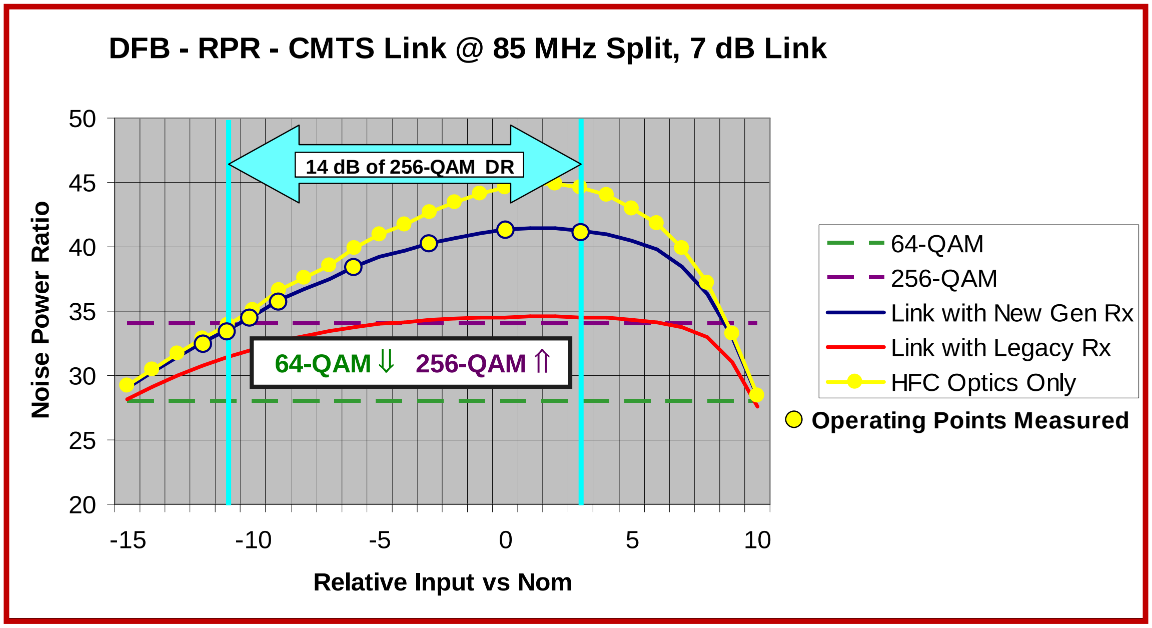


Figure 25 – HFC DOCSIS System Performance

In the NPR curves shown in this section, the optical performance will be augmented with other contributors to the link SNR – in particular RF contributions in the form of noise funneling previously discussed, and receiver noise figures associated with receivers, such as DOCSIS CMTS front ends. We will consider "legacy" DOCSIS receiver – designed originally for 16-QAM maximum profiles, and modern receivers aimed at higher sensitivity for better modulation efficiency.

Consider Figure 25. The red curve marks the performance characteristics of and HFC+CMTS link for legacy-type receivers optimized for 16-QAM and a DFB-RPR link of nominal length under an assumption of 85 MHz of spectrum loading. Clearly, it shows margin over and above the (green) 64-QAM threshold (chosen at 28 dB – an uncorrected 1e-8 error rate objective).

DFB HFC optics plus most of today's CMTS receivers comfortably support 64-QAM with sufficient, practical, operating dynamic range. This lesson is being proven everywhere DOCSIS 3.0 is being deployed. In some cases newer, high quality FP lasers can support 64-QAM as well. While DFBs are recommended for upstream as new channels are added and profiles enabled, it is comforting to realize that newer FPs can get 64-QAM started while the large task of exchanging lasers methodically takes place.

Though legacy receiver exceeded their original design requirements in being extended to 64-QAM (with the help of plant upgrades), enabling 256-QAM design margin – an additional 12 dB of performance over 16-QAM – was not cost effective to consider in early stages of DOCSIS.

As a result, there is zero margin to run 256-QAM (purple), as shown in Figure 25, or otherwise insufficient margin if we aid the factor in more power-per-Hz by limiting the bandwidth to the 65 MHz Euro split by comparison (about 1 dB higher peak) or the 42 MHz split (about 3 dB higher peak).

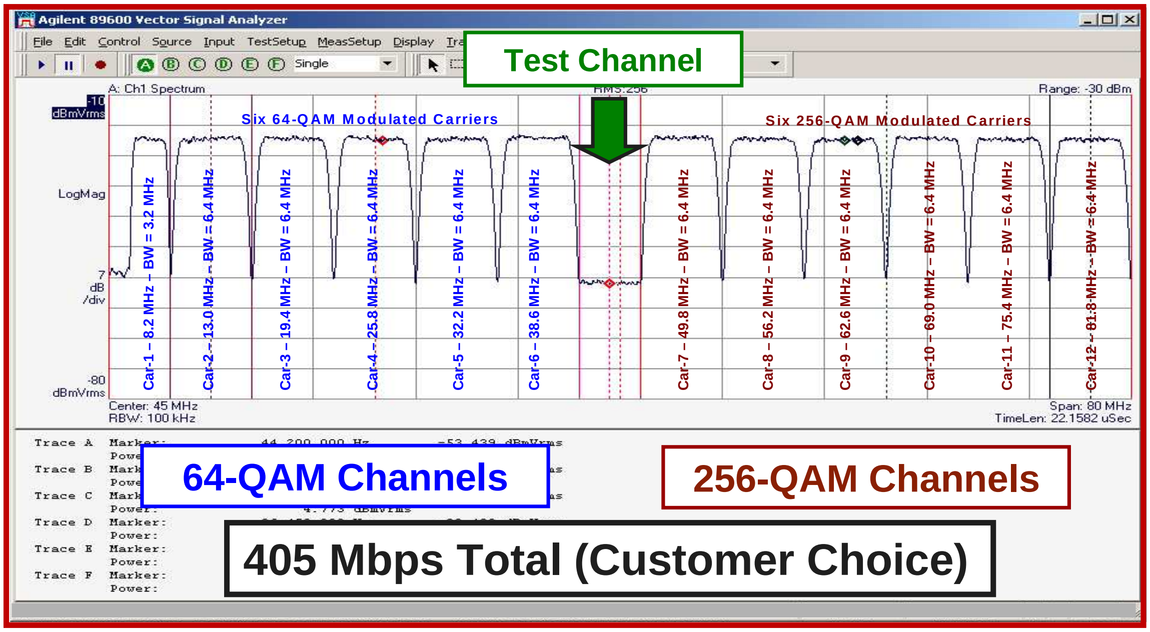

Agilent 89600 Vector Signal Analyzer
File Edit Control Source Input TestSetup MeasSetup Display Tra
Single
Test Channel
A: Ch1 Spectrum
RMS:256
Range: -30 dBm
-10 dBmVrms
LogMag
7 dB /div
-80 dBmVrms
Six 64-QAM Modulated Carriers
Six 256-QAM Modulated Carriers
Car-1 – 8.2 MHz – BW = 3.2 MHz
Car-2 – 13.0 MHz – BW = 6.4 MHz
Car-3 – 19.4 MHz – BW = 6.4 MHz
Car-4 – 25.8 MHz – BW = 6.4 MHz
Car-5 – 32.2 MHz – BW = 6.4 MHz
Car-6 – 38.6 MHz – BW = 6.4 MHz
Car-7 – 49.8 MHz – BW = 6.4 MHz
Car-8 – 56.2 MHz – BW = 6.4 MHz
Car-9 – 62.6 MHz – BW = 6.4 MHz
Car-10 – 69.0 MHz – BW = 6.4 MHz
Car-11 – 75.4 MHz – BW = 6.4 MHz
Car-12 – 81.8 MHz – BW = 6.4 MHz
Center: 45 MHz
RBW: 100 kHz
Span: 80 MHz
TimeLen: 22.1582 uSec
Trace A Marker: 44 200 000 Hz -53 439 dBmVrms
Trace B
Trace C Power: 4.773 dBmVrms
Trace D Marker: Power:
Trace E Marker: Power:
Trace F Marker: Power:
64-QAM Channels
256-QAM Channels
405 Mbps Total (Customer Choice)

Figure 26 – Mid-Split Channel Loading

New receivers, however, provide a higher fidelity upstream termination in order to support 64-QAM with margin and S-CDMA synchronization. Because of these requirements and the continued advances in performance of DFB return optics (higher power laser transmitters), 256-QAM can now be comfortably supported.

The performance of the combined HFC+CMTS link for modern receivers is shown in the blue curve of Figure 25. DOCSIS does not yet call out 256-QAM, although this is a change currently in process.

However, much of the existing silicon base already supports this mode. Note that the yellow points on the blue curve represent points measured in the field that achieved low end-of-line packet error rate performance, as a way of verifying the predicted dynamic range on a real HFC link (NPR would be an intrusive measurement).

Note also that the dynamic range supported for 256-QAM is nearly the same dynamic range that existing receivers provide for 64-QAM – an indication of the robustness potential for 256-QAM links.

Finally, comparing the HFC (yellow) NPR trace to the HFC+CMTS (blue) trace, it is apparent also how little loss of NPR is incurred by new high fidelity CMTS receivers.

Figure 26 shows a snapshot of a recent trial of an Mid-Split architecture, where the upper half of the band was used to support 256-QAM channels, but with all signals at the same power level except for the lowest frequency (narrower) channel. A mid-band test channel was left unoccupied for monitoring the most probable location of maximum distortion build-up as dynamic range was exercised.

Evident from Figure 26 is the high available SNR delivered by the HFC link using existing analog DFB return optics at nominal input drive. The available SNR as measured at the input to the CMTS receiver is about 45 dB. In this case, the tested link was an N+3 architecture.

Table 17 shows a full 85 MHz optimization, using 12 carriers of both S-CDMA and A-TDMA, employing modulations from 32-QAM to 256-QAM across the band. The results indicate a maximum of nearly 400 Mbps of Ethernet throughput under the packetized traffic conditions used.

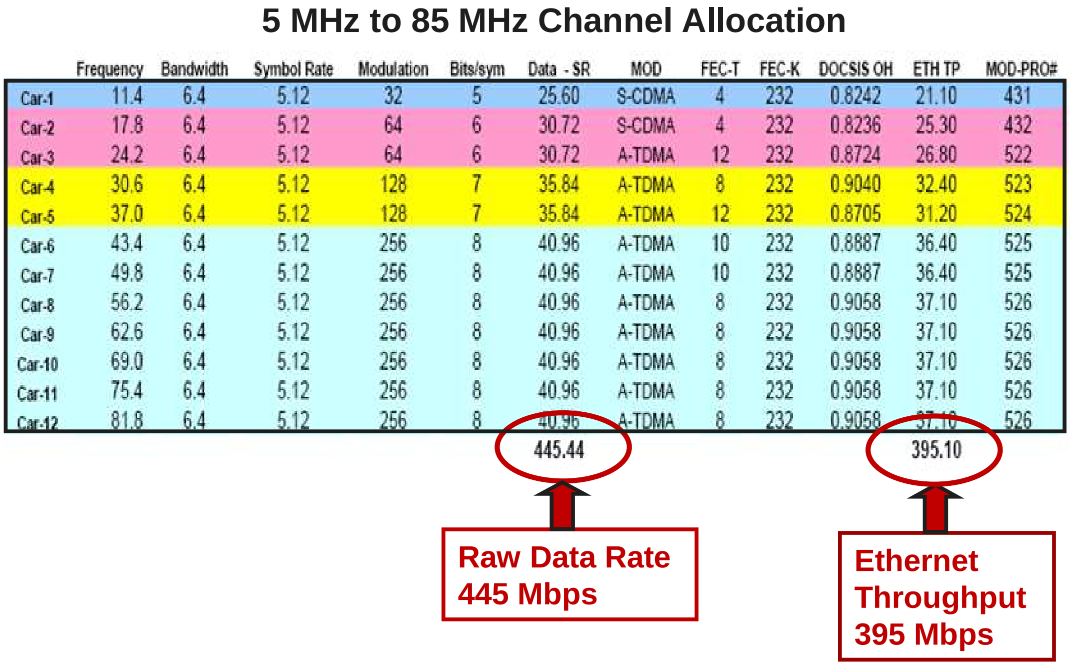
7.1.2.2 Extended HFC Performance

To show the robustness potential of 256-QAM upstream, we can extend the performance calculations in Figure 25 to include longer HFC links and the contribution of potentially long RF cascades summed together, resulting in the "noise funnel" aggregation of amplifier noise figures.

The cases shown in Figure 27 assumes a deep cascade (N+6) in a 4-port node, and thus 24 amplifiers summed, and optical links of 7 dB and 10 dB. While the yellow curve still represents 7 dB optics only, both 7 dB and 10 dB links are shown with the RF cascade included (dashed), and then each of the same with the CMTS receiver contribution included (solid).

The loss due to an analog optical link length is very predictable, as the optical receiver SNR drops as input light level drops. The RF cascade can be shown to create the effect of pushing the performance peak down, reflecting the SNR contribution of amplifier noise to the optical link. However, its effect on the dynamic range for supporting 256-QAM is negligible.

The stronger dynamic range effect is the extended optical link of 10 dB, which ultimately reduces 256-QAM dynamic range by about 2 dB, but with the dynamic range still showing a healthy 11 dB of robust wiggle room.



5 MHz to 85 MHz Channel Allocation

	Frequency	Bandwidth	Symbol Rate	Modulation	Bits/sym	Data - SR	MOD	FEC-T	FEC-K	DOCSIS OH	ETH TP	MOD-PRO#
Car-1	11.4	6.4	5.12	32	5	25.60	S-CDMA	4	232	0.8242	21.10	431
Car-2	17.8	6.4	5.12	64	6	30.72	S-CDMA	4	232	0.8236	25.30	432
Car-3	24.2	6.4	5.12	64	6	30.72	A-TDMA	12	232	0.8724	26.80	522
Car-4	30.6	6.4	5.12	128	7	35.84	A-TDMA	8	232	0.9040	32.40	523
Car-5	37.0	6.4	5.12	128	7	35.84	A-TDMA	12	232	0.8705	31.20	524
Car-6	43.4	6.4	5.12	256	8	40.96	A-TDMA	10	232	0.8887	36.40	525
Car-7	49.8	6.4	5.12	256	8	40.96	A-TDMA	10	232	0.8887	36.40	525
Car-8	56.2	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-9	62.6	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-10	69.0	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-11	75.4	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
Car-12	81.8	6.4	5.12	256	8	40.96	A-TDMA	8	232	0.9058	37.10	526
						445.44					395.10	

Raw Data Rate 445 Mbps

Ethernet Throughput 395 Mbps

Table 17 – Optimized 85 MHz Mid-Split Channel Loading

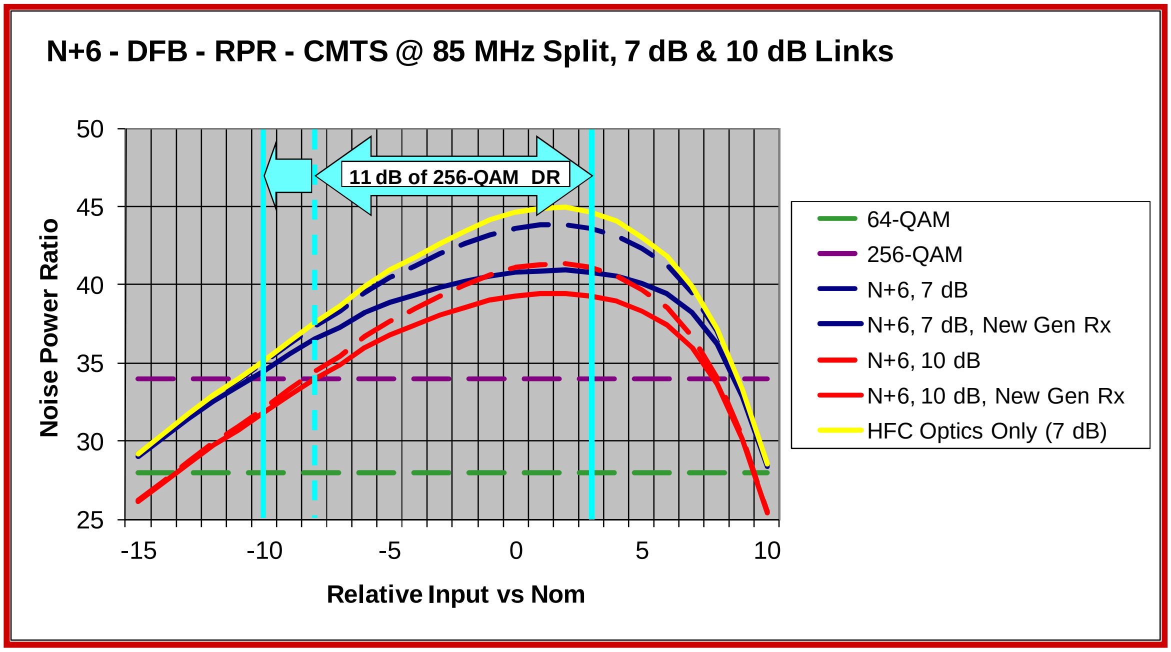


Figure 27 – HFC DOCSIS System Performance for Longer RF Cascades

7.1.2.3 *Extended "High-Split" Bandwidth Projection*

A 1 Gbps capacity threshold upstream requires the split to move to 200 MHz or higher. The 5-200 MHz bandwidth itself supports well over 1 Gbps of theoretical capacity, but legacy use may not make the full spectrum available for higher efficiency, and overhead loss will decrease transport capacity to a lower net throughput.

A higher spectrum diplex will likely therefore be required. However, we quantify the 200 MHz case because of its potential compatibility with current equipment outfitted with 200 MHz RF hybrids, or with minor modifications thereof.

Figure 28 is the analogous figure to Figure 25 for 85 MHz Mid-Split, showing, in this case, projected performance on a 200 MHz "high" split when factoring in an "equivalently performing" CMTS receiver (DOCSIS does not extend to 200 MHz) and DFB optics performing at today's noise density (adjusted only for power loading).

As would be expected, with the receiver performance equivalent to legacy CMTS receivers, inherently not equipped for 256-QAM, performance does not even breach the threshold. However, with a new generation of high fidelity receivers, system analysis projects that there exists 10 dB of dynamic range to 256-QAM performance over a fully loaded 200 MHz return path.

This would see degradation when RF amplifiers are included, but again to minor effect on dynamic range. Conversely, it is anticipated that by the time the need for high split is required, very small serving groups have already been established, leading to a much less significant noise funnel.

While dynamic range (10 dB) is still relatively high, there is observable loss of peak above the 256-QAM threshold, meaning much of the dynamic range exists over a relatively low steady-state operating margin. This could make the link more

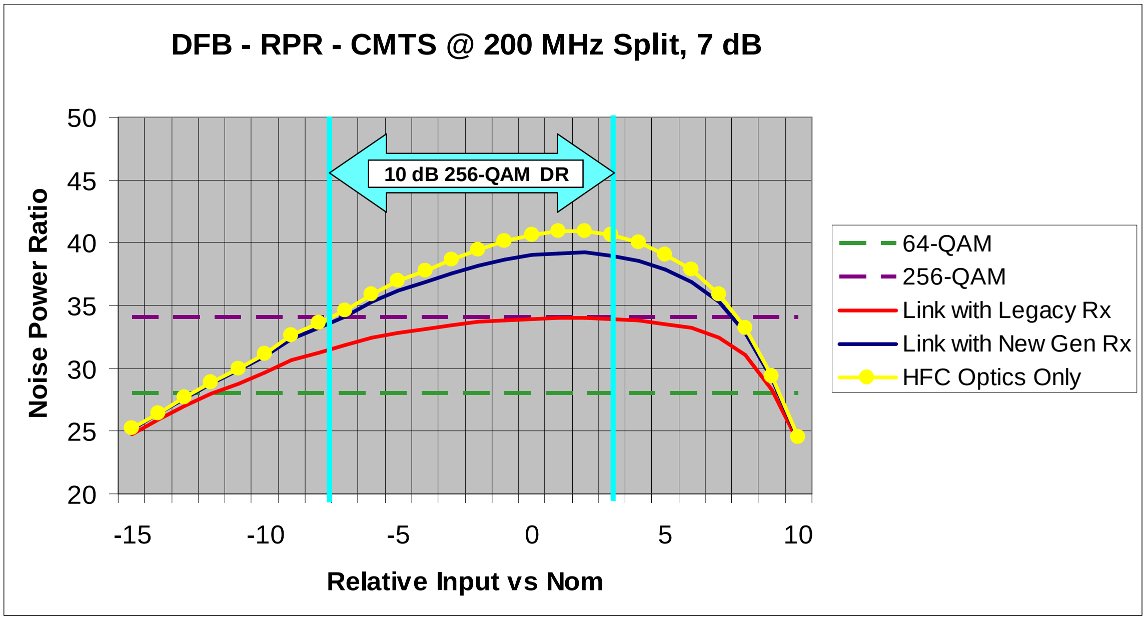


Figure 28 – HFC-DOCSIS System Performance using 200MHz "High Split"

susceptible to moderate transients, drift, temperature extremes, or misalignment, and thus require more regular maintenance.

As such, Figure 28 points out the near term potential for high split operation over HFC optics, but also indicates that performance improvements over time will be welcome to ensure robust operations. Also, note that measured performance for a high split return to 185 MHz, shown in Figure 20, is similar to the analysis in Figure 28. In fact, measured performance of the 1550 nm DWDM return in Figure 20 is slightly better (by about 1.5 dB) than the extrapolated performance in Figure 28 using a standard 1310 nm DFB, pointing out additional margin for the high split case already existing today.

7.1.2.4 *Modem Performance Characterization Findings*

Recent results [17] have evaluated 256-QAM transmission in the presence of narrowband interference to assess the capability of the ingress suppression capability for the higher order of modulation. Table 18 quantifies these results in terms of Codeword Errors (CCER, UCER) and Packet Errors (PER) as are calculated and made available in the DOCSIS MIB.

Results for 64-QAM were shared, along with results for 256-QAM, in [16]. However, Table 18 updates the results for 256-QAM with a more robust performance assessment using higher performance recovers for the proper SNR baseline. This is simply mirroring what was already described and identified in Figure 25 – legacy DOCSIS receivers do not have acceptable margin to run a robust 256-QAM profile.

Nonetheless, it is difficult to make apples-to-apples ingress suppression comparisons, as the SNR margin for 64-QAM offers inherently 6 dB more room for the ingress cancellation to operate under than 256-QAM.

The DFB-RPR link in Table 18 was setup to provide higher SNR than the 64-QAM case in [16] in order than a very low

Table 18 – 256-QAM Interference Performance Low PER Thresholds

	256-QAM				
	Level (dB, dBc)	UNCORR%	CORR%	PER%	MER (dB)
Baseline - AWGN	**36**	0.000%	0.000%	0.000%	37
Single Ingressor Case					
QPSK 12kHz 0.5%	3	0.254%	0.435%	1.060%	34
QPSK 12kHz 1.0%	1	0.447%	0.944%	2.300%	34
FSK 320ksym/s 0.5%	29	0.278%	0.032%	0.110%	35
FSK 320ksym/s 1.0%	27	0.633%	0.230%	0.810%	35
FM 20kHz 0.5%	2	0.128%	0.295%	0.750%	34
FM 20kHz 1.0%	1	0.187%	0.554%	1.260%	34
Three Ingressor Case					
CPD 0.5%	28	0.297%	0.041%	0.190%	34
CPD 1.0%	27	0.698%	0.144%	0.750%	33

BER threshold in each was a baseline. However, it was not the same absolute margin of the M-QAM to the SNR of the link (5 dB vs 2 dB). It did lead to a very important conclusion, however.

With this low BER steady state case in [8] for 256-QAM, for nearly equivalent relative performance (6 dB difference) for nominal single-interference cases was observed. However, for multiple interferers and for wideband (100's of kHz) there was still substantially more robustness in the case of 64-QAM. Refer to [8] for full details.

Overall, proof of the functionality of ingress cancellation was achieved for 256-QAM, but with degraded performance when the channel is at its noisiest. Of course, the strategy for deploying 256-QAM is to place in the clean part of the upstream, where it can be supported – above 25 MHz. And, certainly consider it to extract capacity in the 85 MHz Mid-Split case above 42 MHz.

This is the approach used to "optimize" the 85 MHz band and shown in Table 18 – a mixture of 256-QAM, 128-QAM, 64-QAM, and S-CDMA based 64-QAM and 32-QAM.

This is the upstream line-up that led to the 445 Mbps transport rate proof of concept reported in [12].

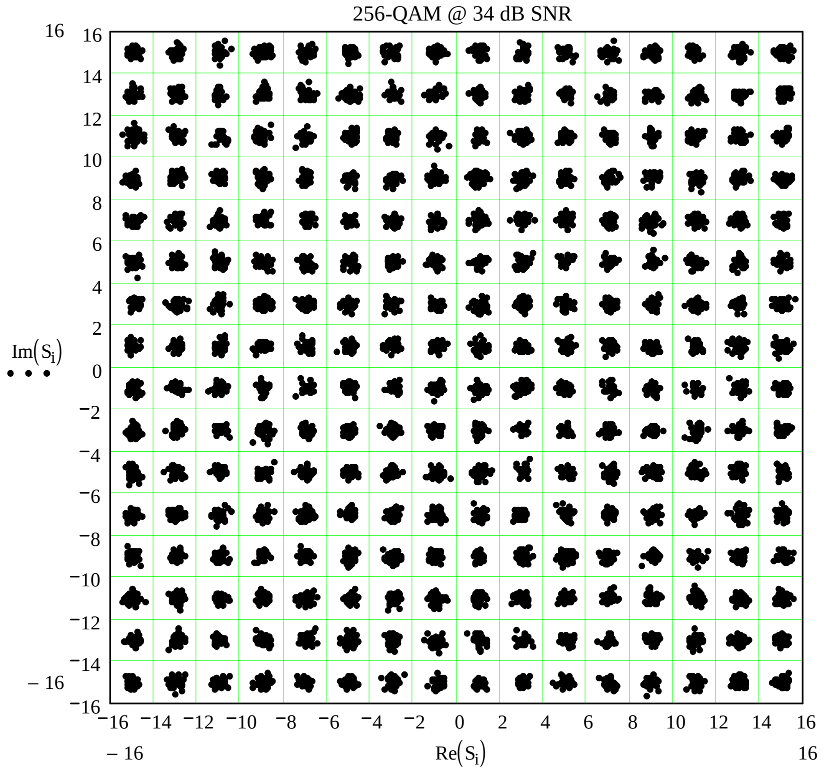


Figure 29 – 256-QAM @ 34 dB SNR

7.1.3 1024-QAM Downstream

In Section 9.5 "Downstream Capacity", we will calculate the downstream capacity for a fully digitized forward band, multiplying the number of 6 MHz slots by the modulation profile allowed by DOCSIS (256-QAM) to arrive at data capacities for 750 MHz, 870 MHz, and 1 GHz networks. We then calculated the case for a Next Generation PHY using LDPC and OFDM, making the reasonable assumption that by updating the FEC, we can achieve two QAM orders of modulation higher in bandwidth efficiency, which effectively suggests 6 dB can be gained.

However, not all of this may be in the FEC (depending on code rate). Some incremental link budget dB may be obtained through some of the business-as-usual operations of fiber deeper and cascade reduction, which reduces noise and distortion accumulation, and through the conversion of analog carriers to digital, which reduced (2x analog + digital) composite carrier-to-noise (CCN) distortion effects. Lastly, newer STBs in the field tend to higher sensitivity (lower noise figure).

Because of this, the FEC is not left to make up all of the dB between 256-QAM and 1024-QAM. And, in fact, it is now possible to make a case based only on these HFC changes that 1024-QAM may be possible in evolved architectures today, even without the addition of new FEC on silicon that can support this QAM mode. This offers the potential for 25% more bandwidth efficiency. This section quantifies this potential.

Let's begin the discussion with the use of QAM over HFC for downstream video as it has evolved to date.

The cable plant has kept up with the bandwidth consumption by adding RF bandwidth and using efficient digital modulations to mine the capacity effectively and with robustness. What started as 64-QAM digital signals became yet more bandwidth efficient with the deployment of 256-QAM downstream, the dominant QAM approach today. The ability to successfully deploy such schemes is due to the very high SNR and very low distortion downstream.

This was to ensure proper conditions for supporting much less robust analog video. In addition to high linearity and low noise, the downstream channel has a flat frequency response on a per-channel basis, minimizing both amplitude and phase distortion, although it can be prone to reflection energy.

As a simple example of the possibilities, the theoretical capacity of a 6 MHz channel with a 40 dB SNR is approximately 80 Mbps. Yet, for J.83-based 256-QAM, the transmission rate is only about 40 Mbps. When accounting for overhead, there is even less throughput.

The next higher order, square-constellation, modulation is 1024-QAM. This technique achieves an efficiency of 10 bits/symbol, or another 25% efficiency over 256-QAM, and an impressive 67% improvement relative to 64-QAM. To support 1024-QAM, a more stringent set of specifications must be met.

Analysis was performed to identify implications to the plant and its performance requirements for robust downstream transmission [1]. The analysis quantified SNR, beat distortion interference, and phase noise, and interpreted the results. We summarize the problem statement here and describe the conclusions.

7.1.3.1 SNR

Let's consider the implications of 1024-

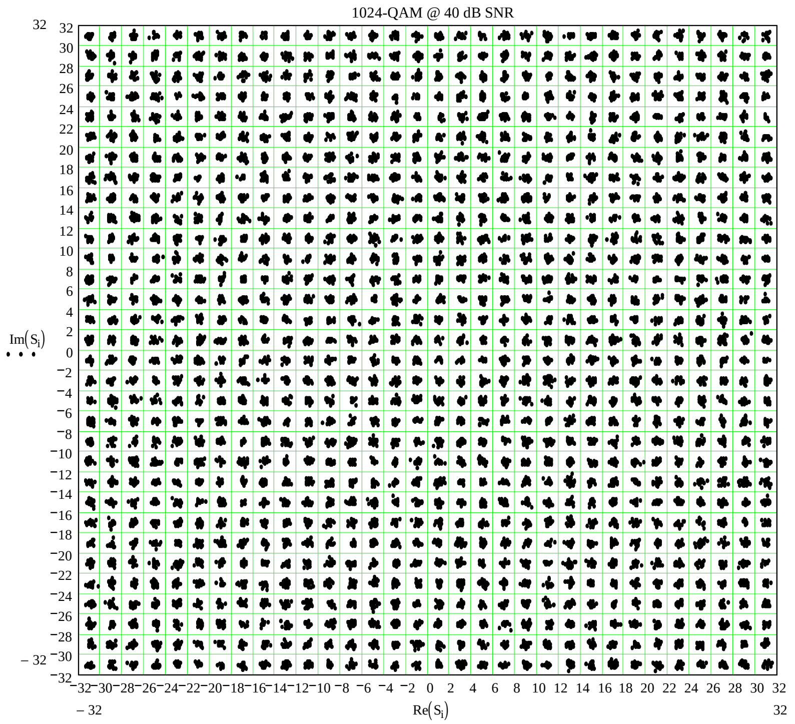


Figure 30 – 1024-QAM @ 40 dB SNR

Table 19 – Power Loading Effects of Analog Reclamation - 870 MHz

	Channel Uptilt @ 870 MHz					
	Flat		12 dB		14 dB	
	Delta Ref	QAM Increase	Delta Ref	QAM Increase	Delta Ref	QAM Increase
79 Analog	Ref Load	---	Ref Load	---	Ref Load	---
59 Analog	-0.7	2.5	-1.0	1.5	-0.9	1.5
39 Analog	-1.6	3.5	-1.7	2.5	-1.6	2.0
30 Analog	-2.1	4.0	-2.0	2.5	-1.9	2.5
All Digital	-4.5	4.5	-2.8	3.0	-2.5	2.5

QAM. Figure 29 and Figure 30 show constellation diagrams of 256-QAM @ 34 dB SNR and 1024-QAM @ 40 dB SNR. Being 6 dB apart, these are equivalent uncorrected error rate cases (@1E-8). The congested look of the 1024-QAM diagram, emphasized by the small symbol decision regions, signals the sensitivity this scheme has to disturbances.

Now consider what 40 dB means in terms of use on the plant. For an end-of-line 46 dB of plant (analog) CNR, QAM SNR becomes 40 dB when backed off by 6 dB. We've thus removed virtually all link available margin under an objective of 1E-8, and are now into a region of measurable errors, relying on FEC to finish the job under even the most benign circumstance of thermal noise only.

On the STB side, there is similar margin-challenged mathematics. For a STB noise figure of 10 dB, and for QAM signals arriving at the STB at the low end of the power range, some simple math shows the following:

- Residual Thermal Noise Floor: -58 dBmV/5 MHz
- STB Noise Figure, NF = 10 dB: -48 dBmV/5 MHz
- Analog Level into STB: 0 dBmV
- Digital Level into STB: - 6 dBmV
- STB SNR contribution: -6 -(-48) = 42 dB

Note that NF = 10 dB is not a technically difficult performance requirement. However, in practice, given the cost sensitivity of CPE equipment and without a historical need to have better RF sensitivity, 10 dB and higher is quite normal.

The combined link delivers an SNR of about 38 dB. This simple example leads to the conclusion that existing conditions and existing deployment scenarios create concerns for a seamless 1024-QAM roll-out under a "J.83"-type PHY situation. It reveals the necessity of at least 2 dB of coding gain to ensure robust link closure.

Improving the noise performance of CPE is of course one option to enable more bandwidth efficient link budgets, particularly as yet more advanced modulation profiles beyond 1024-QAM are considered. The sensitivity of CPE cost and the existing deployment of 1024-QAM capable receivers and current noise performance, however, leads to a desire to remain conservative in the expectation of CPE performance assumptions.

7.1.3.2 Favorable Evolution Trends

A couple of favorable trends are occurring in HFC migration that potentially free up some dB towards higher SNR of the

Table 20 – Noise and Distortion @ 550 MHz vs Analog Channel Count

	CCN		CTB		CSO	
Analog Channels	**N+6**	**N+0**	**N+6**	**N+0**	**N+6**	**N+0**
79	48	51	58	70	56	64
59	48	52	60	70	59	65
30	48	52	68	74	67	70

QAM channels – analog reclamation and cascade shortening.

Table 19 shows the potential for higher SNR by taking advantage of the RF power load when compared to a reference of 79 analog channels for 870 MHz of forward bandwidth. In the table, the left hand column for each case – Flat, 12 dB tilt, 14 dB tilt – represents the decrease in total RF load compared to the 79-analog channel reference. The right column for each case represents how much more power could be allocated to each digital carrier in order to maintain the same total RF power load. This is the potential available theoretical SNR gain.

The flat case represents the effect on the optical loading of the analog reclamation process. There is headroom that can be exploited in the optical link and RF cascade by increasing the total power of the analog plus digital multiplex, gaining SNR for all channels and offering potential mediation against the 6 dB increased SNR requirement.

The SNR discussion above refers only to the improvement relative to the thermal noise floor. The additional distortion component (composite inter-modulation noise or CIN) and practical RF frequency response means not all of the theoretical dB will be realized (refer to [1] for details).

Now consider Table 20 quantifying modeled performance for a sample HFC link under different assumptions of line-up and cascade. The data underscores the impact on noise and distortion of decreasing analog channel loads and shorter RF cascades. CCN represents Composite Carrier-to-Noise – a combination of the CNR or SNR and digital distortion products.

Moving across rows, noise and distortion improvements associated with the elimination of the RF cascade (N+6 to N+0) is clear. Moving down columns, the benefits of doing analog reclamation also becomes clear. Both activities enable the network to more ably support higher order modulation SNR performance requirements.

From the perspective of noise (CCN), shortening of the cascade reduces the accumulation of amplifier noise, freeing up 3-4 dB additional SNR available relative to a typical line-up and cascade depth of today. When coupled with possible loading adjustments with the larger digital tier and new headroom available – a few dB here and a few dB there approach – we can come close to 6 dB of new SNR as we evolve the network and use the gains to our benefit. This is, of course, the amount of increased SNR sensitivity of 1024-QAM compared to 256-QAM.

Table 21 – Inner (5/6) LDPC Coded M-QAM Throughput and Comparison to J.83 [2]

Mode	Efficiency (bits/symbol)	Representative Symbol Rate (Msps)	Representative Inner Code Bit Rate (Mbps)	TOV Es/No (dB)	Delta from Capacity* (dB)
Proposed 64QAM	5.333	5.056	26.96	18.02	0.52
Proposed 256QAM	7.333	5.361	39.31	24.27	0.46
Proposed 1024QAM	9.333	5.361	50.03	30.42	0.50
J83.B 64QAM	5.337	5.056	26.97	20.75	3.25
J83.B 256QAM	7.244	5.361	38.84	26.90	3.44
"J83.B" 1024QAM	9.150	5.361	49.05	33.03	3.80

*Note that "Capacity" in this case is an abbreviation for *Constrained* Capacity, as opposed to Shannon Capacity. For this example, the constraint is a symbol set of uniformly distributed QAM symbols. Please refer to above text and [2] for details.

7.1.3.3 Modern FEC

So far, we have considered only existing FEC with 1024-QAM, relying on HFC migration phases to extract additional dB from the plant to create sufficient operational margin. Fortunately, we are not limited to legacy error corrections schemes. While powerful in its day, concatenated Reed-Solomon FEC used in J.83 is now roughly 15 years old – an eternity in information theory technology development. While J.83 leaves us several dB from theoretical PHY performance, modern FEC, typically built around Low Density Parity Check (LDPC) codes – also concatenated to avoid error flooring – achieves performance within fractions of dB of theoretical.

A proposal made during DOCSIS 3.0 discussions [2] quantified additional gains available using LDPC for current 64-QAM and 256-QAM systems, as well as for potential 1024-QAM use. Table 21 summarizes some of the core findings of that system design. The analysis references a common Threshold of Visibility (TOV) threshold for video of 3e-6 and compares constrained capacity (limited to QAM signal sets) of the various profiles. This constraint has an inherent offset from Shannon capacity that grows as a function of SNR.

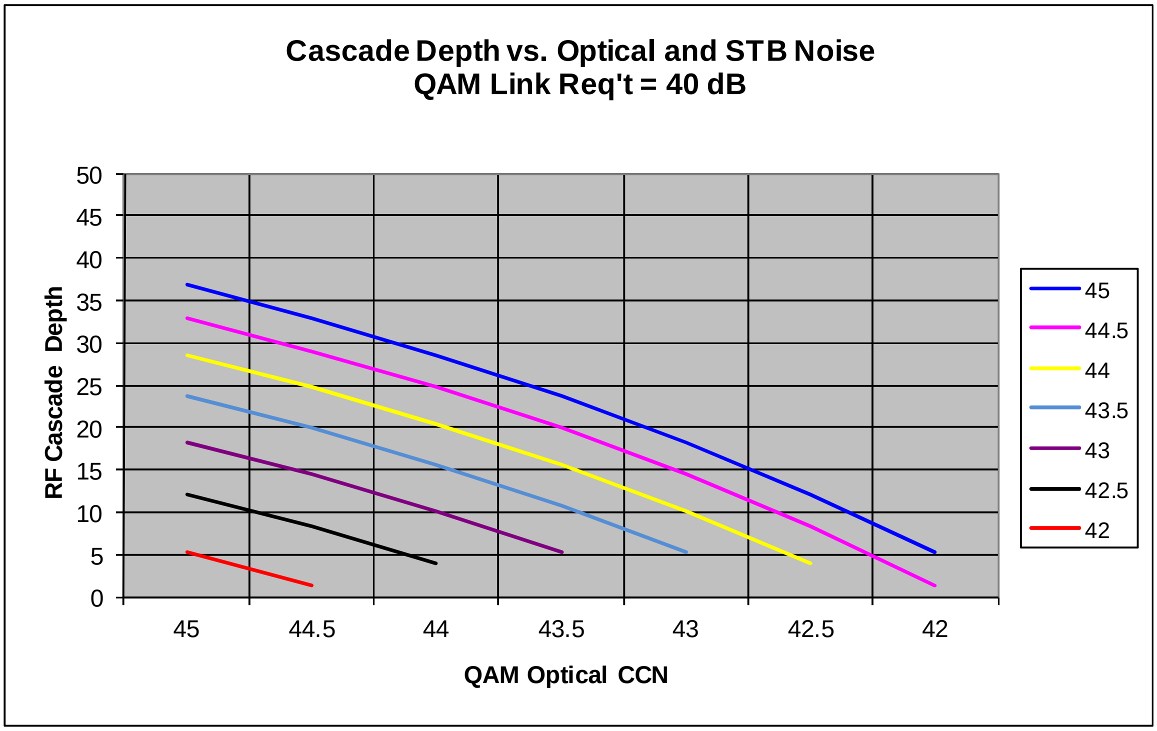


Figure 31 – 1024-QAM, Noise, and Cascade Depth – 40 dB Link Requirement

With the recognition of another 3.3 dB of coding gain, the proposal pointed out the accessibility of 1024-QAM for the downstream channel in a legacy 6 MHz format. This constraint (6 MHz) can also be removed for wider band channels, leading to more flexibility in code design and thus more available coding gain. However, we will see below that even just assuming a modest 3 dB more coding gain provides very meaningful SNR margin for robust 1024-QAM.

We can now execute architecture trade-offs of noise contributions and the depth of the RF cascade to evaluate support for 1024-QAM. HFC cascade thresholds are shown in Figure 31 and Figure 32, as a function of STB noise figure and optical link CCN, as a function of a pre-defined overall SNR link objective (40 dB or 37 dB). Each curve represents a different value of SNR as set by the STB alone, associated with the noise figure and digital level (de-rated from analog) at its input.

Note from the figures that there is a wide range of SNR combinations that essentially offer no practical limit to RF cascade depth as it relates to noise degradation. Clearly, tolerating a 37 dB link requirement is exactly this scenario, and this is quite a reasonable requirement under the capability of new FEC. It provides a very comfortable range of operation, even for poor performing optical links with respect to noise.

However, the 40 dB range includes conditions that could lead to a sharp reduction in the cascade acceptable. From a sensitivity analysis standpoint, such conditions hinge on small dBs and even fractions thereof. This makes it more valuable to be able to earn back, for example, just 1-2 dB SNR in the analog reclamation process.

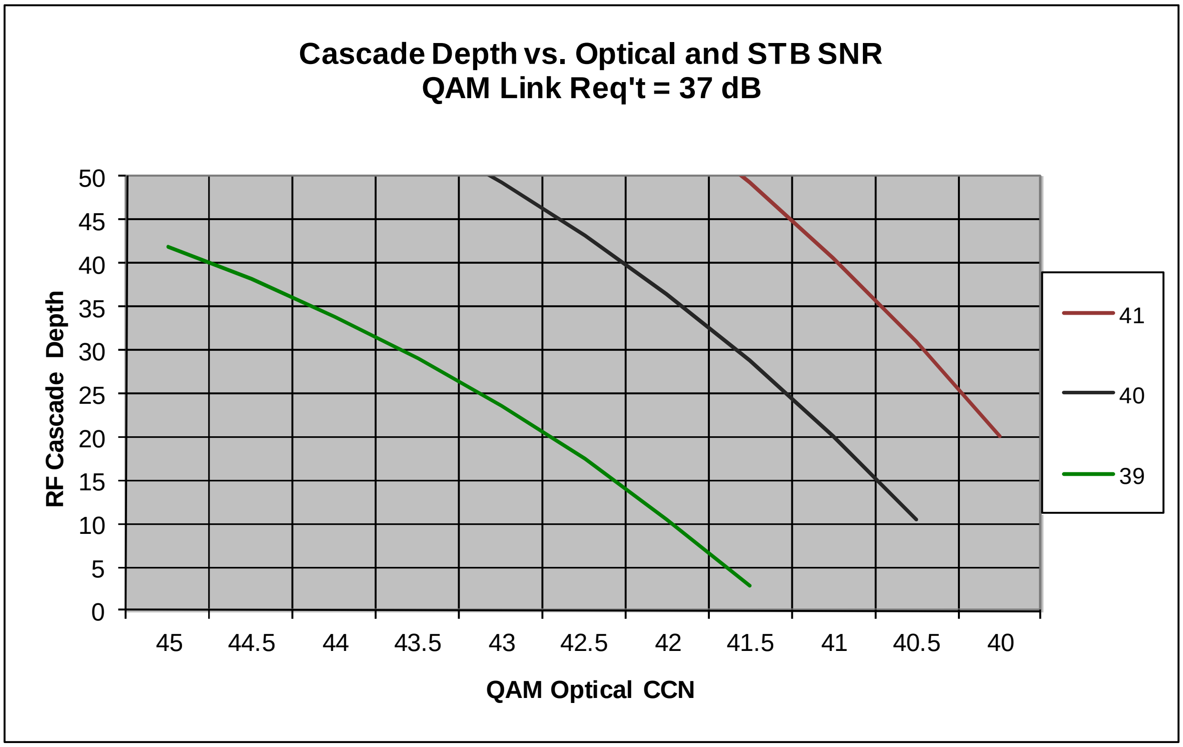


Figure 32 – 1024-QAM, Noise, Cascade Depth – 37 dB Link Requirement (Improved FEC)

Finally, note specifically the SNR = 42 dB at Optical CCN = 45 point on the bottom left of Figure 31. For a quite typical 51 dB Optical CNR requirement, a digital CCN of 45 dB would occur under 6 dB back-off. These conditions yield a cascade depth of five (N+5) as tolerable. Note, however, that 42 dB was a NF = 10 CPE, and, as previously identified, higher NF's (10-14 dB) may be the case.

This points out simply that STB clients of higher NF than 10 dB, under nominal optical link performance and deeper cascades may struggle to achieve the 40 dB requirement for 1024-QAM. FEC may save the link from a QoE perspective, but this example points out how relatively nominal conditions of legacy plant add up to make 1024-QAM a challenge. It also emphasizes the value of the dB available in migration, and especially the value of new FEC, most readily observable in Figure 32.

7.1.3.4 Distortion

As observed in Table 20, in addition to its positive effects on digital SNR, analog reclamation offers benefits in the distortion domain as well. Table 20 results are arrived at through tools such as shown in Figure 33 – a sample of a distortion beat map for 79 analog channels on a 12 dB tilt to 870 MHz. Such analysis is used to calculate the impact of varying channel line-ups on relative distortion level. Coupled with the sensitivity of 1024-QAM under CTB/CSO impairment, we can then evaluate the ability of an HFC cascade to support 1024-QAM.

The performance thresholds for CTB were taken from laboratory evaluation of error-free or nearly error-free 1024-QAM with actual live-video CTB generated as the impairment source [1]. It is interesting to note in that testing how pre-FEC and post FEC results are related, indicative of CTB as a "slow" disturbance relative to the symbol rate, and thus a burst error mechanism that challenges FEC decoding.

A result of the use of these CTB thresholds to find HFC architecture limitations is shown in Figure 34. It plots cascade depth thresholds over a range of given RF amplifier CTBs, specified at typical RF output levels, and varying analog channel counts used using a CTB threshold of 58 dBc [1].

It is clear to see that analog reclamation to 30 channels enables virtually any practical RF cascade depth. However, it also becomes clear how for 79-channel systems and 59-channel systems, some limitations may appear.

Prior analysis had investigated the effects of analog beat distortions on 256-QAM, developing relationships for the comparative performance of 64-QAM and 256-QAM [3]. It was observed that 10-12 dB difference existed in susceptibility to a single, static, in-band narrowband interferer at the main CTB offset frequency. Under the assumption that ingress mediation performance can achieve equivalent rejection relative to the M-QAM SNR (potentially an aggressive assumption), this relationship might be assumed hold between 256-QAM and 1024-QAM for narrowband interference.

7.1.3.5 Phase Noise

Untracked phase error leads to angular symbol spreading of the constellation diagram as shown in Figure 35 for 1024-QAM with .25° rms of Gaussian-distributed untracked phase error imposed. This non-uniform impact on symbols is critical to understand to explain phase noise sensitivities for increasing M in M-QAM. It was observed in [1] that .25° rms represents a loss due to phase noise of about 1 dB, assuming low error rate conditions, and with no practical phase noise-induced BER floor.

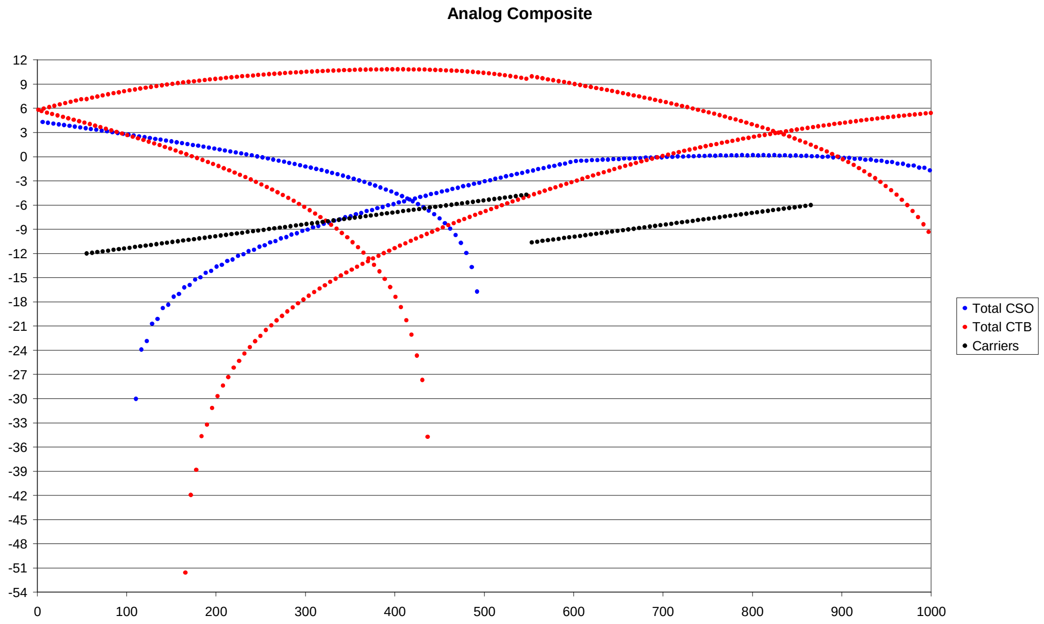


Figure 33 – Distortion Map - 79 Analog Channels, 12 dB Tilt

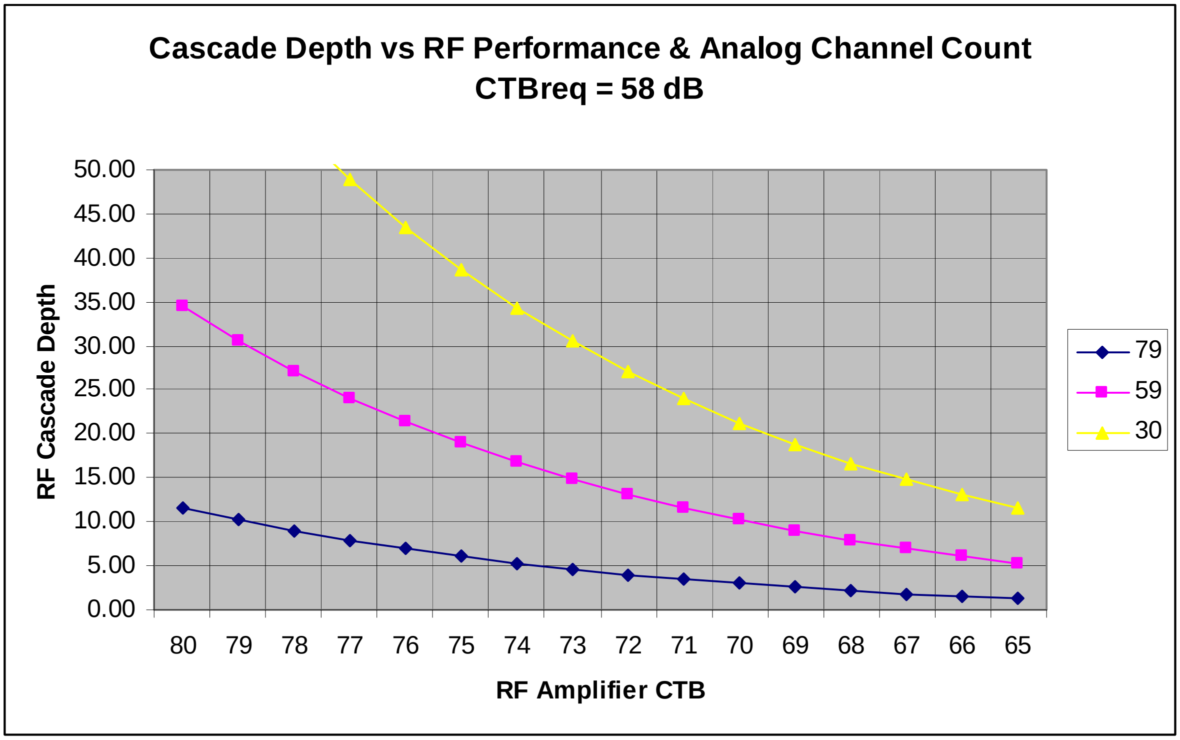


Figure 34 – 1024-QAM, CTB, and Cascade Depth, Thresh = 58 dB

A floor in the 1E-8 or 1E-9 region will be induced at roughly 50% more jitter, or .375 deg rms. Measurements of phase noise showed that for high RF carrier frequencies, typically associated with higher total phase noise, wideband carrier tracking still left about .33 deg rms of untracked error, enough to cause a BER floor to emerge at very high SNR.

The use of degrees rms is more easily understood when expressed as signal-to-phase noise in dB. Note that 1° rms is equivalent to 35 dBc signal-to-phase noise. Doubling or halving entails 6 dB relationships. Thus, we have the following conversions:

1 deg rms = 35 dBc SNRφ

.5 deg rms = 41 dBc SNRφ

.25 deg rms = 47 dBc SNRφ

The values .33 deg rms and .375 deg rms represent 44.6 dBc and 43.5 dBc, respectively. This is instructive to compare to the SNR under AWGN only (40 dB used above), as it illustrates the nature of the phase noise impairment on M-QAM with high M.

Error rate measurements [1] show that error flooring appears to be occurring as measured by pre-FEC errors, suggesting that there have not been significant enough tuning (historically analog, now full-band capture) noise improvements or carrier recovery system changes to mitigate this effect.

However, although phase noise is a slow random process that challenges burst correcting FEC, the combination of the interleaver, Reed-Solomon, and the relatively low floor, has been seen to result in zero post-FEC errors. Note that the phase noise alone is requiring the FEC to work to clean up the output data, and is thus consuming some FEC "budget" in the process.

Phase noise can be improved through design as well, almost without limit, but as strong function of cost for broadband performance. Current performance appears sufficient, although perhaps coming at the expense of increased sensitivity to other impairments that may also require FEC help.

These observations are likely a harbinger of issues to come as M increases further in search of higher bandwidth efficiency, such as 4096-QAM.

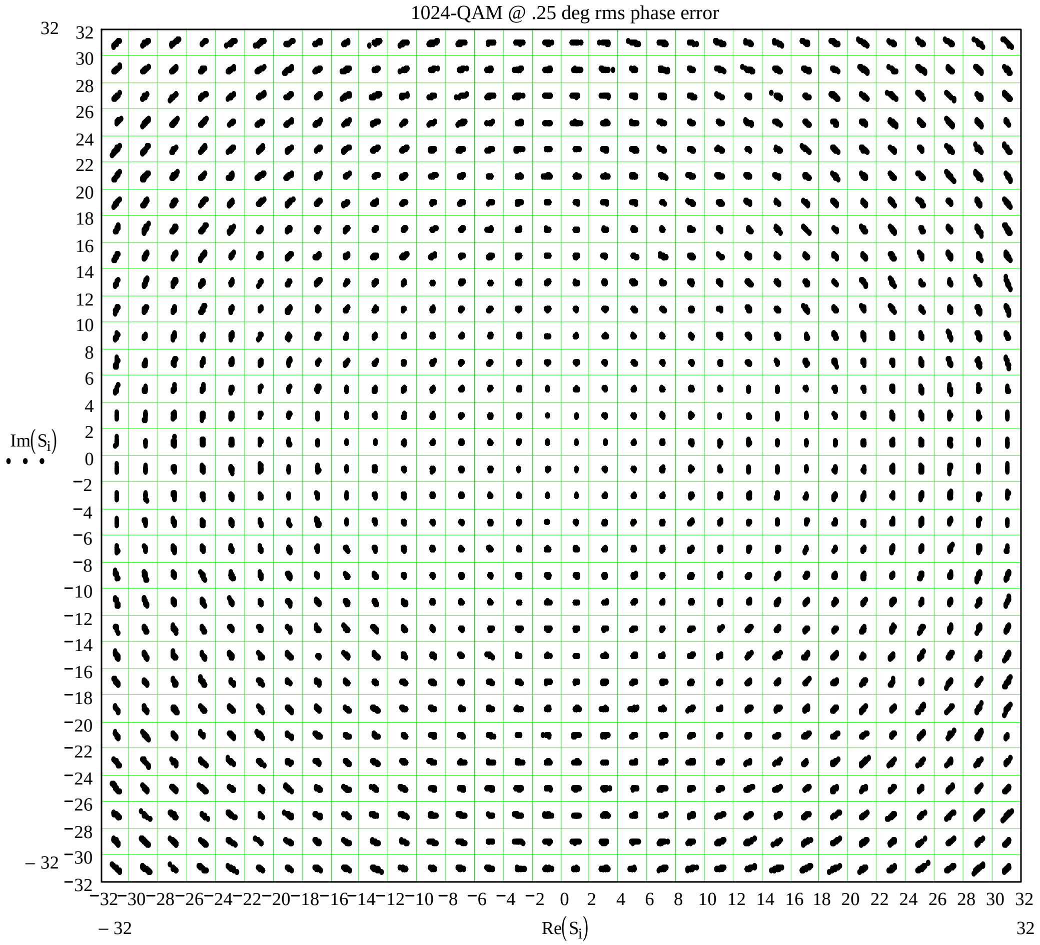


Figure 35 – 1024-QAM with .25° RMS Phase Noise

7.2 S-CDMA

Leveraging S-CDMA has many benefits, including reclamation of regions of upstream spectrum considered previously unusable with TDMA, lower overhead for FEC, and even feasibility of higher-order constellations. Some frequency regions are, of course, readily accessed leveraging Advanced Time Division Multiple Access (A-TDMA).

A-TDMA can be made very robust to a broad set of impairments including noise, distortion, and interference when it's coupled powerful tools such as Forward Error Correction (FEC), Equalization, and Ingress Cancellation. Problems arise when impairments exceed the performance limits of what A-TDMA can mitigate, resulting in objectionable codeword errors and packet loss.

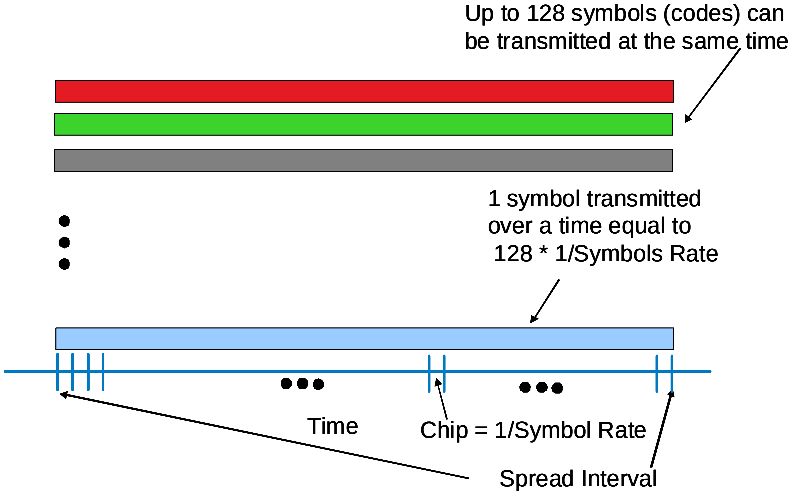

Up to 128 symbols (codes) can be transmitted at the same time
1 symbol transmitted over a time equal to 128 * 1/Symbols Rate
Time
Chip = 1/Symbol Rate
Spread Interval

Figure 36 – S-CDMA Parallel Symbol Transmission

Fortunately, DOCSIS 2.0 and later includes Synchronous Code Division Multiple Access (S-CDMA), which offers additional robustness against impairments, and in particular against impulse noise. This robustness against impulse noise exceeds that of A-TDMA by a factor of 100 times or more [14].

As powerful as DOCSIS 2.0 S-CDMA has been proven to be in field trials, DOCSIS 3.0 has S-CDMA features that further enhance robustness against impairments. These techniques were standardized to create a very high-performance, sophisticated PHY for cable, capable of supporting high data rates in the most difficult of environment.

The latest features include Selectable Active Codes (SAC) Mode 2, Trellis Coded Modulation (TCM), Code Hopping, and Maximum Scheduled Codes (MSC). Despite these advances aimed at adding more capability to the upstream, most of the DOCSIS 3.0 features remain largely unused, and DOCSIS 2.0 deployments are minor in scale in North America.

Let's take a look at what is available in DOCSIS to maximize the throughput of the upstream band, and discuss how today's PHY toolsets complement one another. First, let's understand what S-CDMA does best – high throughput performance under difficult channel conditions.

7.2.1 Impulse Noise Benefits of S-CDMA

There are several benefits to S-CDMA, but the most important by far is its burst protection capability. The ingredient that makes the robustness to impulse noise possible is the spreading out of the symbols by as much as 128 times in the time domain, which directly translates to stronger protection against impulse noise.

This spreading operation is pictured in Figure 36. Noise bursts that may wipe out many QAM symbols of an A-TDMA carrier must be two orders of magnitude longer in duration to have the same effect on S-CDMA, which is very unlikely. It is the spread signaling approach itself, without even considering FEC settings, that enables S CDMA to withstand much longer impulsive events.

There is no reduction in throughput as a result of this spreading, of course, because the slower symbols are transmitted simultaneously. S-CDMA has similarities conceptually to OFDM in this manner, with the difference being S-CDMA's use of the orthogonality in the code domain versus OFDM's use of orthogonality in the frequency domain.

Now consider Figure 37, which illustrates how S-CDMA's primary benefit translates to return path bandwidth access. Through its effectiveness against impulse noise, S-CDMA facilitates efficient use of what is otherwise very challenging spectrum for A-TDMA. It is a critical tool for squeezing every last bit-per-second possible out of return spectrum.

Additionally, the lower the diplex split used in the system, the more important S-CDMA becomes. It has become well-understood that the most consistently troublesome spectrum is at the low end of the band, typically 5-20 MHz.

This region is where S CDMA shines in comparison to A-TDMA. As such, S-CDMA matters more for maximizing use of 42 MHz than it does to 65 MHz (Euro Split) or 85 MHz (Mid-Split) because of the percentage of questionable spectrum.

Purely in terms of spectrum availability then, S-CDMA is most valuable to the North American market, where upstream spectrum is the scarcest and use of DOCSIS services is high. Depending on the upstream conditions, about 35-50% of extra capacity can be made available using S-CDMA.

Nonetheless, S-CDMA's benefits have been largely unused in practice by operators, despite its availability in DOCSIS 2.0 and DOCSIS 3.0 certified equipment.

7.2.2 Quantifying Performance

Again, by far S-CDMA's most compelling advantage is its ability to perform in harsh impulse noise environments. Impulse noise is, by definition, a transient event – interference of finite duration and often periodic or with

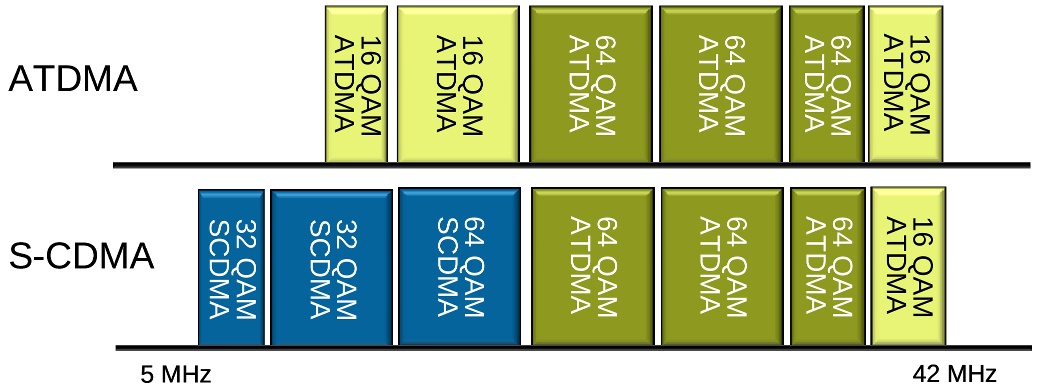


Figure 37 – Maximizing 5-42 MHz Throughput Using S-CDMA

repetitive frequency of occurrence.

Characterization of impulse noise includes duration, rate, and amplitude. It is generated in a variety of ways. When noisy devices such as dimmer switches, hair dryers, garage door openers, power tools, automobile ignition circuits – the list goes on – are in close proximity to the cable network, impulse noise may enter into upstream. The majority of impulse noise originates in and around the home.

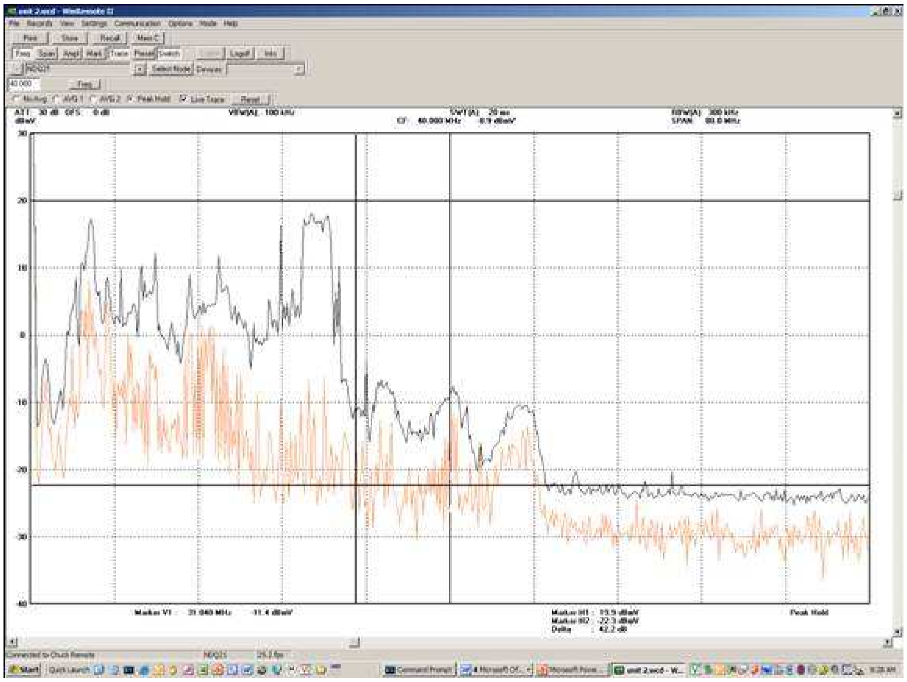
Figure 38 is a spectral snapshot of impulse noise, where a noticeable wideband burst above the noise average (red) is very likely interfering with DOCSIS signaling by creating a temporary condition whereby the SNR is only about 18 dB.

The impact of such a burst on a discrete set of QAM symbols is to cause the symbols to jump decision boundaries, or increase the probability that they will do so, resulting in codeword errors, as shown in Figure 39. Note the wideband nature of the degradation in the frequency domain of short duration impulse noise.

Consider just the DOCSIS-described scenario of duration 10µs and rate 1 kHz. A 10 usec burst will corrupt 52 symbol at 5.12Msps, which translates to 39 bytes of data for 64-QAM. This is beyond the capability of the Reed-Solomon FEC, with a maximum burst protection of t = 16 bytes.

For this scenario, the FEC cannot be effective without assistance of interleaving. An interleaver, in theory, could be used to break-up clusters of impacted bytes so that they span multiple codewords, allowing FEC to be more effective. However, byte interleaving requires longer packets for adequate shuffling of the bytes. Minimum packet lengths of 2x the designated codeword length are necessary, and the longer the better.

Unfortunately, of course, most upstream packets tend to be short and not suited to effective interleaving.


File Records View Settings Communication Options Mode Help
Print
Store
Recall
Main C
Freq Span Ampl Mark Trace Preset Switch
Logoff
Info
Select Node
Freq
Peak Hold
Live Trace
Reset
30
20
10
0
-10
-20
-30
-40

ATT 30 dB OFS 0 dB
dBmV
VBW(A) 100 kHz
SWT(A) 20 ms
CF 40.000 MHz -6.9 dBmV
RBW(A) 300 kHz
SPAN 80.0 MHz

Marker V1 : 31.040 MHz -11.4 dBmV
Marker H1 : 19.9 dBmV
Marker H2 : -22.3 dBmV
Delta : 42.2 dB
Peak Hold

Connected to Chuck Remote

Figure 38 – Impulse Noise Illustration, -18dBc

Such situations are where S-CDMA is the best choice for achieving high throughput. S-CDMA has greater ability to recover transmissions through long noise bursts, and is not sensitive to packet size the way interleaving is in a burst environment.

A most recent head-to-head comparison under simultaneous RF impairments of impulse noise and interference is shown in Table 22.

Three impulse noise sources were used:

1. Duration = 10µs, Rate = 1kHz (per DOCSIS specification)
2. Duration = 20µs, Rate = 4kHz
3. Duration = 40µs, Rate = 4kHz

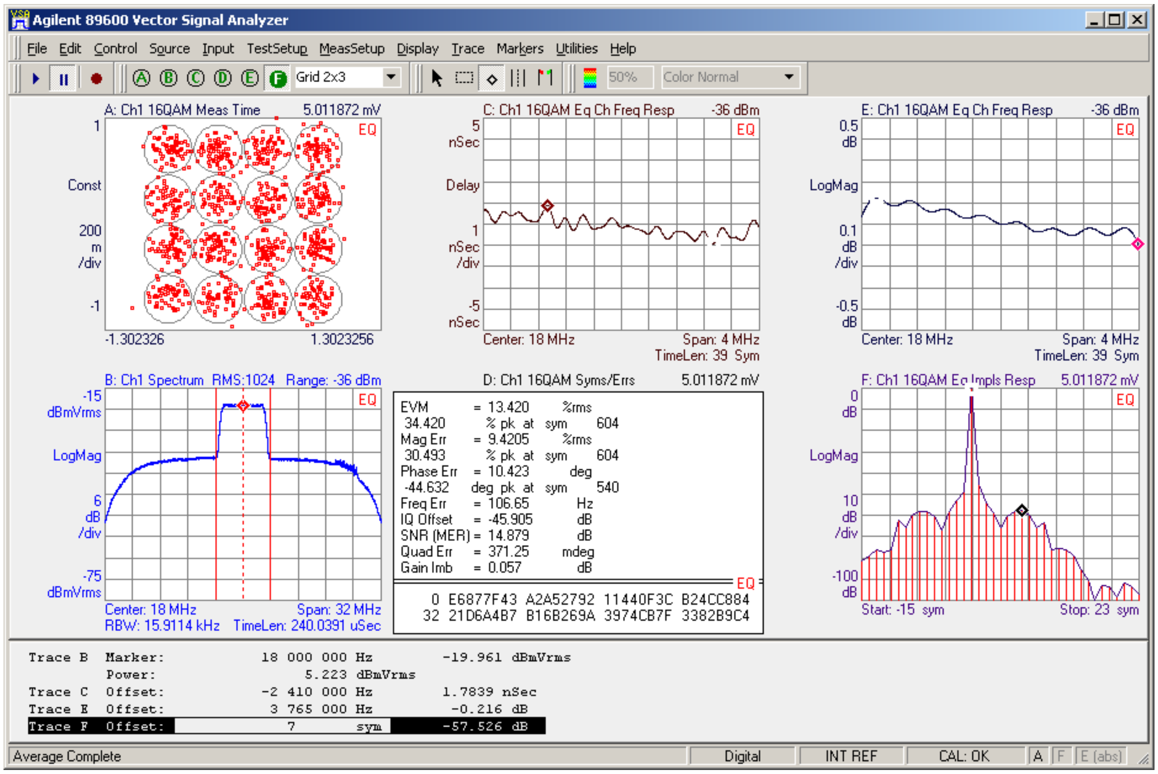
Three interference patterns used, centered around the signal center frequency:

- A. 4x π/4-DQPSK Carriers @16ksym/s, Spacing = 400kHz
- B. 2x π/4-DQPSK Carriers @16ksym/s, Spacing = 1600kHz
- C. 1 π/4-DQPSK Carriers @16ksym/s

The interference was modulated in order to randomize it and give it some spectral width, which makes ingress cancellation more challenging.

Table 22 shows the comparative results, with S-CDMA clearly and significantly outperforming A-TDMA under the dual impairment conditions. A-TDMA FEC is working much harder in each of the cases evaluated, primarily because of the impulse noise.

Uncorrected Codeword Error Rate (UCER) and packet error rate (PER) for A-TDMA under each of the impairment conditions shows performance that would



```
EVM         = 13.420      %rms
 34.420       % pk  at  sym     604
Mag Err     = 9.4205      %rms
 30.493       % pk  at  sym     604
Phase Err   = 10.423       deg
 -44.632    deg  pk  at  sym     540
Freq Err    = 106.65        Hz
IQ Offset   = -45.905       dB
SNR (MER)   = 14.879        dB
Quad Err    = 371.25      mdeg
Gain Imb    = 0.057         dB

 0  E6877F43  A2A52792  11440F3C  B24CC884
32  21D6A4B7  B16B269A  3974CB7F  3382B9C4
```

```
Trace B  Marker:          18 000 000 Hz        -19.961 dBmVrms
         Power:                5.223 dBmVrms
Trace C  Offset:          -2 410 000 Hz        1.7839 nSec
Trace E  Offset:           3 765 000 Hz         -0.216 dB
Trace F  Offset:                   7      sym  -57.526 dB
```

Figure 39 – Impulse Noise Impaired 16-QAM

likely noticeably degrade the customer experience.

Not only is the S-CDMA FEC not working as hard as A-TDMA FEC, there is also less S-CDMA FEC applied. FEC for A-TDMA was at its maximum setting of t=16, and k=219, whereas field trial results previously published [1] resulted in lower FEC for S-CDMA of t=6, and k=239.

As previously discussed, FEC operating requirements can be lowered for S-CDMA because the robustness of the spreading function itself.

Clearly, for equal or even more strenuous impairment scenarios than the A-TDMA cases, S-CDMA offers error-free UCER and PER with no impact to the customer experience.

Additionally, proactive monitoring of Corrected Codeword Error Rate (CCER) with S-CDMA could better facilitate impulse noise problem diagnostics, whereas A-TDMA links would not.

Additional testing in the field on live plants has confirmed the advantage that S-CDMA delivers in the poorer part of the upstream spectrum. A result from a comparison of S-CDMA and A-TDMA on the same return path channel using logical channel operation, centered in a noisy portion of the upstream (about 13 MHz), is shown in Figure 40.

Apparent from Figure 40 is that A-TDMA is taking errors in transmission at a nearly 20% clip, while S-CDMA is taking

Table 22 – S-CDMA & TDMA Performance against Impulse Noise + Interference

16-QAM. 6.4MHz						
1518-Byte Packets	S-CDMA			ATDMA		
Noise Floor = 35dB	MER	CCER/UCER	PER	MER	CCER/UCER	PER
Interference Characteristics	Impulse Noise Characteristics: Duration = 10us, Rate = 1kHz, Level = -11dBc					
Pattern A @ -20dBc	33.1	3.2653%/0.0000%	0.00%	32.2	9.8190%/0.3643%	1.82%
Pattern B @ -18dBc	33.3	2.2164%/0.0004%	0.00%	30.4	9.4996%/0.4362%	1.84%
Pattern C @ -16dBc	33.6	6.0938%/0.0000%	0.00%	30.5	9.1357%/0.9920%	4.86%
Interference Characteristics	Impulse Noise Characteristics: Duration = 20us, Rate = 4kHz, Level = -13dBc					
Pattern A @ -22dBc	29.0	6.2512%/0.0000%	0.00%	29.6	39.7214%/0.2657%	1.46%
Pattern B @ -22dBc	23.0	6.4386%/0.0000%	0.00%	28.2	36.8949%/0.0730%	0.39%
Pattern C @ -20dBc	33.5	5.3450%/0.0000%	0.00%	25.6	36.5901%/1.1087%	4.61%
Interference Characteristics	Impulse Noise Characteristics: Duration = 40us, Rate = 4kHz, Level = -14dBc					
Pattern A @ -22dBc	17.3	13.1082%/0.0000%	0.00%	26.6	39.7623%/0.0639%	0.40%
Pattern B @ -22dBc	26.1	13.8848%/0.0000%	0.00%	20.3	35.1569%/0.0079%	0.05%
Pattern C @ -13dBc	34.2	7.6259%/0.0000%	0.00%	28.0	38.3802%/1.7060%	6.91%

16-QAM. 3.2MHz						
1518-Byte Packets	S-CDMA			ATDMA		
Noise Floor = 35dB	MER	CCER/UCER	PER	MER	CCER/UCER	PER
Interference Characteristics	Impulse Noise Characteristics: Duration = 10us, Rate = 1kHz, Level = -7dBc					
Pattern A @ -22dBc	32.2	6.9036%/0.0000%	0.00%	33.5	18.1515%/2.7396%	14.87%
Pattern B @ -26dBc	21.1	4.0558%/0.0000%	0.00%	28.9	19.2957%/0.7367%	3.99%
Pattern C @ -11dBc	33.1	3.6618%/0.0000%	0.00%	34.0	16.8403%/5.2196%	22.86%
Interference Characteristics	Impulse Noise Characteristics: Duration = 20us, Rate = 4kHz, Level = -10dBc					
Pattern A @ -23dBc	25.6	8.1255%/0.0005%	0.00%	26.2	79.9084%/4.3388%	22.07%
Pattern B @ -24dBc	19.5	17.1071%/0.0000%	0.00%	24.8	81.1037%/0.1378%	0.85%
Pattern C @ -12dBc	32.6	13.3983%/0.0000%	0.00%	18.0	65.1727%/20.9625%	65.44%
Interference Characteristics	Impulse Noise Characteristics: Duration = 40us, Rate = 4kHz, Level = -12dBc					
Pattern A @ -20dBc	22.9	15.8017%/0.0000%	0.00%	18.6	85.0225%/2.8658%	13.41%
Pattern B @ -23dBc	31.3	16.5487%/0.0000%	0.00%	20.0	83.8348%/0.4118%	2.01%
Pattern C @ -13dBc	31.6	24.5632%/0.0000%	0.00%	23.0	71.7126%/17.3259%	56.71%

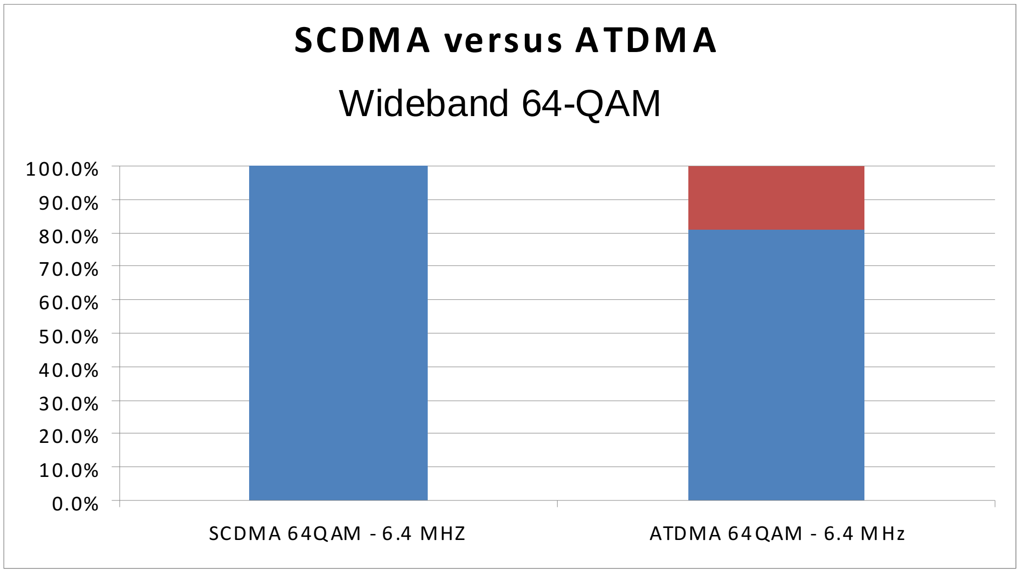


Figure 40 – Corrected Error Statistics

none. In this case, FEC settings for A TDMA are again t=16, while for S-CDMA, they are set to just t=2. S-CDMA inherently takes advantage of its impulse immunity properties rather than relying on FEC.

It is worth noting that, for A-TDMA, impulse noise can also wreak havoc on adaptive processes such as equalization and ingress cancellation, resulting in appreciable variation in cancellation estimates. For example, Figure 41 shows Non-Main Tap to Total Energy Ratio (NMTER) for a population of eight cable modems where impulse noise caused significant variation in equalizer correction.

NMTER is useful as a Figure of Merit to describe the linear distortion level of the upstream path. Here, it is indicating that the frequency response correction process is being significantly disturbed, resulting in a period of increased ISI until the impulse noise subsides and the taps updated.

Even should FEC be able to handle the impulse duration, this increase in ISI can degrade performance because of the increased susceptibility to detection errors at the slicer. The FEC budget may be required to deal with both ISI and burst correction, and is therefore more likely to be overwhelmed until the next tap update can be processed.

7.2.3 More Capability Remains

S-CDMA's impulse noise robustness has been demonstrated, but there is still more that can be leveraged to take advantage of all of the DOCSIS 3.0 features of S-CDMA.

Additional features include Selectable Active Codes (SAC) Mode 2, Trellis Coded Modulation (TCM), Code Hopping, and Maximum Scheduled Codes (MSC). These features provide more flexibility and capability for extracting bandwidth from noisy, limited spectrum, and yet remained largely unused despite more being standardized for many years.

Briefly, these features provide the following:

SAC Mode 2 – Allows for customization of the active codes. Instead of fixed active codes (SAC Mode 1) codes

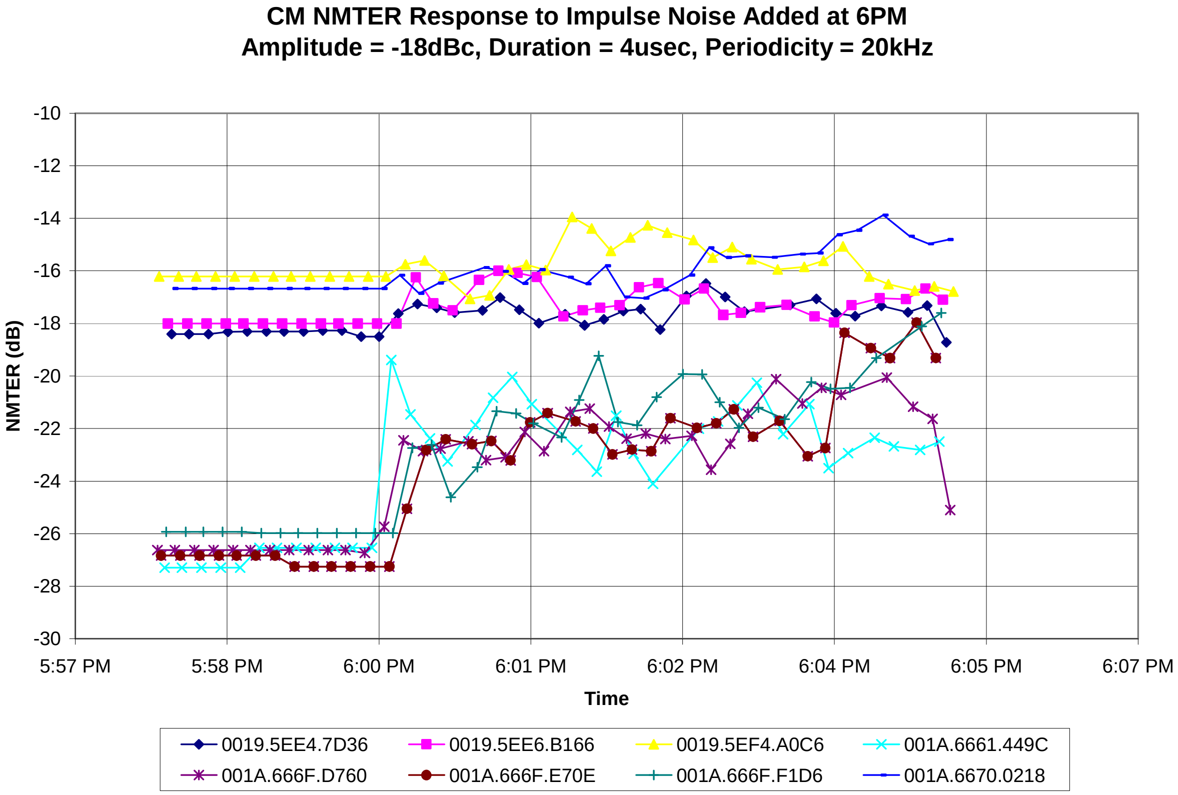


Figure 41 – NMTER vs. Time Impaired by Impulse Noise

may now be optimally allocated between spreading and ingress cancellation.

Trellis-Coded Modulation (TCM) – The well-known technique for optimizing coding structure through integration with symbol mapping, adding gain without adding bandwidth overhead to do so.

Code Hopping – Provides cyclic shifts of the active code set at each spreading interval, further randomizing code allocation to achieve a uniformity of robustness of performance

Maximum Scheduled Codes (MSC) – Offers the flexibility to trade-off between the power allocated per-code and the number of codes turned on. For example, if 128 codes are on transmitting at Pmax, each code is allocated Pmax/128. If only 64 codes are used, each code is allocated Pmax/64, or 3 dB more power per code. This comes at the expense of throughput, but offers some choices to the operator that may be better than an equivalent A-TDMA alternative.

7.2.4 Summary

S-CDMA delivers proven, substantial gains in impulse noise robustness – performance verified in detailed lab testing and in the field, around the world.

It clearly outperforms A-TDMA on difficult channels, enables high-throughput access to the otherwise abandoned lower portion of the return spectrum, and has been shown to operate robustly on channels where A-TDMA will not operate at all.

Many available, but as yet unused, features of S-CDMA, including SAC Mode

2, MSC, Code Hopping, and TCM, provide further capability against upstream impairments. Nonetheless, while a long-standardized tool in DOCSIS, operators have not widely deployed S-CDMA.

In low-diplex architectures, where DOCSIS extensions may be the most straightforward, low-complexity way to light up new spectrum, S-CDMA already exists to support the delivery of high throughput on difficult low-end spectrum. It is capable of providing the same benefits as in any new spectrum deployed for upstream that becomes prone to high interference and noise levels.

The combination of updated A-TDMA with the full features of S-CDMA may, in fact, be a sufficient PHY toolset for upstream growth and lifespan extension, eliminating the need to develop a third upstream PHY, such as an OFDM-based system.

7.3 OFDMA, OFDM & LDPC (A Proposal for a New PHY)

7.3.1 Problem Statement

Once it is acknowledged that current DOCSIS 3.0 MAC provides all the necessary capabilities to extend DOCSIS service to future gigabit rates, the challenge becomes optimizing the PHY layer.

Before choosing the technology for that new PHY, key selection criteria need to be established. These criteria apply to both upstream and downstream.

1. Bandwidth capacity maximization
2. Transparency toward the existing D3.0 MAC
3. Robustness to interference
4. Robustness to unknown plant conditions
5. Throughput scalability with plant condition (SNR)
6. Implementation complexity and silicon cost
7. Time to market
8. PAPR considerations
9. Frequency agility

7.3.1.1 Bandwidth Capacity Maximization

According to Shannon theorem the maximum achievable throughput capacity for a communication system is a function of signal to noise ratio and bandwidth. Both of these resources, the signal power relative to an unavoidable noise and the useful bandwidth of the coaxial part, are limited in an HFC plant.

An upgrade of the HFC plant is costly, and therefore before (or in parallel with) this upgrade, the available SNR and bandwidth utilization can, and must be maximized using state-of-the-art modulation and coding techniques.

7.3.1.2 Transparency Toward the Existing D3.0 MAC

One of the extremely useful features of the D3.0 MAC is the physical channel bonding. This feature allows trafficking of logical flows on information through multiple and different physical channels. Apart from the lower level convergence layer features, the DOCSIS 3.0 MAC is not aware what type of Physical channel(s) the information is flowing through, be it 256-QAM or 64-QAM in downstream, or ATDMA or SCDMA in upstream.

Allowing the new PHY to follow the same transparency will allow the products introduced to the market migrate gradually from using the old PHY to using the new PHY by utilizing (rather than giving up) throughput from existing legacy channels, until these are gradually replaced with new ones. For example, there are CMs deployed in the field with eight downstream channels. Until all these CMs are replaced, those eight channels will continue to occupy the shared spectrum. A transition period product will be able to make use of both the legacy PHY and the new PHY through channel bonding; and hence will maximize the data throughput as illustrated in Figure 42 and Figure 49.

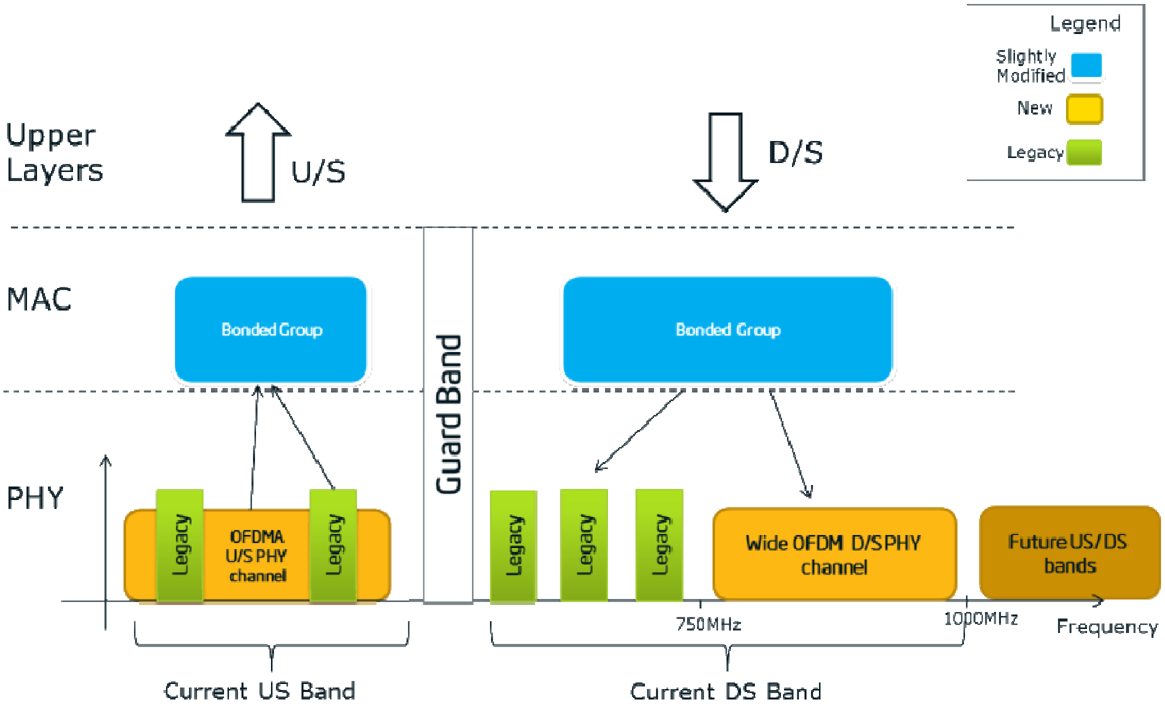


Figure 42 – Illustration of bonding the legacy and the new PHY channels

As a comparison, a non-DOCSIS technology will not be able to benefit from the bandwidth occupied by legacy.

7.3.1.3 Robustness to Interference

As the home and business environment becomes flooded with electronic equipment, the level of interference becomes a significant limiting factor of bandwidth usage in some regions of the HFC spectrum, particularly in the upstream. A modulation scheme of choice should be designed to minimize the effect of interference on the achievable throughput.

7.3.1.4 Robustness to Unknown Plant Conditions

The new PHY should be well equipped to be deployed in spectrum that is currently unused for cable systems, such as spectrum beyond 1GHz. Also, it should be equipped to maximize throughput given unknown parameters in the existing installation, as these differ significantly by region, type of installation, countries, etc. Planning for the worst case adds inefficiency and cost, hence agility to optimize capacity per given condition is required.

7.3.1.5 Throughput Scalability with Plant Condition (SNR)

As mentioned above, SNR sets the maximum achievable capacity over a given bandwidth. Ability to scale the throughput accordingly with the SNR available to the modem will allow squeezing the maximum throughput possible per given installation condition. Simply put, more bits/sec/Hz configurations are needed with finer granularity, spanning a wide SNR scale.

7.3.1.6 Implementation Complexity and Silicon Cost

Adding more throughput capability to the modem will result in more silicon complexity that translates to silicon cost. It is essential that the new PHY technology chosen is able to offer cheaper implementation in terms of dollars per bits/sec/Hz over other alternatives. As a side note, one thing worth noting is that process technology scaling (Moore's law) allows increasing the PHY complexity without breaking the cost limits.

7.3.1.7 Time to Market

It is important to isolate the proposed changes to specific system elements without affecting system concepts. Changing only the PHY channel, without any significant changes to the MAC minimizes the scope of impact of the change and allows quicker standardization and implementation of the change. Utilizing existing, proven, and well-studied technologies helps accelerate the standardization and the productization.

7.3.1.8 PAPR Considerations

Good (low) peak to average ratio properties of the modulation technique may help in squeezing more power out of the amplifiers in the system by moving deeper into the non-linear region. Hence, good PAPR properties are desirable, as these have system impact beyond the end equipment.

7.3.1.9 Frequency Agility

The ability of the new PHY channel to be deployed in any portion of the spectrum is a great advantage. This is especially useful during the transition period where various legacy services occupy specific frequencies and bands and cannot be moved.

Next we consider the alternatives of the PHY channels in light of the above-mentioned criteria, focusing on the parameters of the suggested proposal.

7.3.2 Solution Analysis

7.3.2.1 Channel Coding – Optimizing Spectral Efficiency

FEC has the most significant impact on spectral efficiency. Traditional error control codes such as J.83 Annex B are concatenations of Trellis and Reed-Solomon block codes. Modern coding techniques such as LDPC and Turbo use iterative message passing algorithms for decoding, thereby yielding significant coding gains over traditional techniques. LDPC has been shown to out-perform Turbo codes at relatively large block sizes. LDPC also has the parallelism needed to achieve high throughputs.

Figure 43 shows a comparison of different coding schemes used in Cable technologies[1]. 256-QAM modulation is taken as baseline for comparison. The horizontal axis is the code rate and the vertical axis shows the SNR required to achieve a BER of 1e-8. The two DVB-C2 LDPC codes are shown, the long code with a block size of 64800 bits and the short code with a block size of 16200 bits. [28]

As expected, the code with the longer block size does provide better performance

[1] Although code rate 0.8 is not present in current J83 specification, the system was simulated with RS codes (204, 164) for J.83 Annex A and (128, 108) for J.83 Annex B to get the effective performance of these codes at a rate of 0.8.

although the difference is very small (0.2 dB) for high code rates needed for cable applications. The two DVB-C2 LDPC codes do include a weak BCH code to assist with the removal of the error floor.

The graph in Figure 43 shows that the DVB-C2 LDPC offers about 3 dB more coding gain over J.83 Annex B code for a code rate of 0.9 implying an increase of capacity of 1 bit/s/Hz, i.e. a 12.5% increase with respect to 256-QAM. The increase in coding gain and hence the capacity is much higher (about 5 dB) with respect to just the RS code used in J.83 Annex A, i.e. DVB-C.

Note that since existing coding schemes are compared, the code word lengths are not the same, implying an advantage to longer code words. Theoretically, if the J.83 Annex B FEC is extended to a longer code word, the difference will be less than 3 dB, but the DVB-C2 code will still give the better performance.

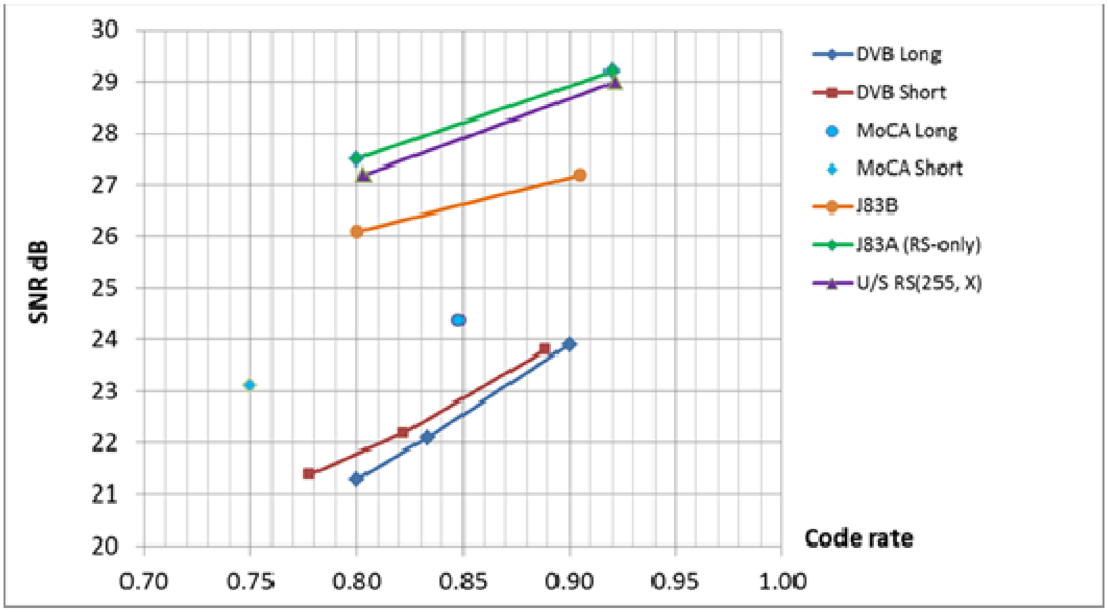


Figure 43 – FEC Comparison for 256-QAM Modulation

To enable efficient stuffing of upstream bursts with code words, two types of codes with different code word length are necessary. A short code word for short bursts, and a long code word for long bursts are recommended. Since the ambitious throughput requirements are usually on the long bursts (streaming data, rather than maintenance messages), no system throughput loss is expected due to usage of shorter code word.

7.3.2.2 Modulation Scheme

The options considered for the modulation scheme of the next gen PHY are as follows:

1. Legacy modulation, narrow Single Carrier QAM, 6/6.4 MHz channels
2. A new, wide Single-Carrier channel modulation, e.g. Single Carrier QAM 24 MHz channel
3. A new, wide Multi-Carrier OFDM channel modulation

A comparison of these options is discussed next against the established criteria.

7.3.2.3 Implementation Complexity

To contain the total complexity increase due to scaling to gigabit throughputs, both PHY layer implementation itself, and its effect on the MAC layer need to be considered.

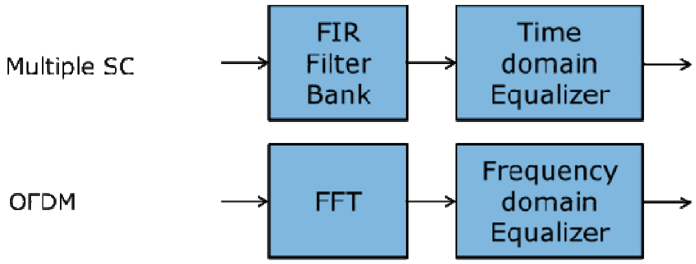


Figure 44 – Signal Processing Block for Computational Complexity analysis

If narrow channels are used to attain the high throughputs, a large number of such channels will be required, which may lead to a non-linear increase in the MAC complexity. Hence, there is a benefit of using wide channels to reduce the total number of bonded channels.

However, only OFDM out of the three options considered can give a computational benefit given a wide channel, due to differences in channel processing and equalization. The channelization for OFDM is based on FFT, which is computationally more efficient than the multiple sharp channel filters required for single carrier. Also the frequency domain equalization in OFDM is much lighter computationally than time domain equalization required with SC-QAM. Figure 44 and Table 23 show the processing power analysis of the options based on the number of real multiplication per second required. A clear advantage of OFDM is observed.

Another thing worth noting in favor of wide channels is that since today's analog front end technology for cable is based on direct digital-to-analog and analog-to-digital conversion, having wide channels does not pose a new implementation challenge for the analog front end design. All the channelization and up/down frequency conversion can be done digitally.

7.3.2.4 Channel Equalization

A common assumption for OFDM modulation is that the guard interval (GI) needs to be of a length equal to or higher than the longest reflection in the channel. However, this does not have to be the case. The reflection that is not completely covered by the GI affects only small part of the

Table 23 – Number of Multiplications per sec (real*real) for different modulations schemes

Function	32x6 MHz SC	8x24 MHz SC	16K OFDM
Modulation	1024-QAM	1024-QAM	1024-QAM
Channelization	32 FIR (sym.) filters 6.9e9 (40-tap)	8 FIR (sym.) filters 6.9e9 (40-tap)	16K FFT: 2.6e9
Equalizer	32e9 (40-tap)	125e9 (160-tap) 100e9 (128-tap) 75e9 (96-tap)	5.0e9

symbol, reducing the power of the inter-symbol-interference (ISI) on the entire symbol accordingly (approximately by 10log(T_interference_overlap/T_symbol) on top of the already weak power of the long echoes).

The result is extra gain in throughput of OFDM symbol, due to the GI being shorter than the longest anticipated reflection. To illustrate this, a simulation result of a 16K FFT OFDM system with 200 MHz channel bandwidth and DVB-C2 LDPC code with rate 8/9 is depicted in Figure 45. SCTE-40 reflection profile (SCTE-40) is simulated, as well as AWGN.

The 4.5 us SCTE-40 echo (-30 dB) is outside the 3.33 us cyclic prefix guard interval. However, the loss with respect to the 5 us guard interval is only 0.15 dB because the ICI/ISI noise floor due to echo outside guard is at -42 dB.

An OFDM scheme can have multiple options for guard intervals without any silicon cost penalty, whilst the SC time equalizer approach needs to be designed for the worst case. As DOCSIS moves into new spectrum, this additional flexibility gives OFDM an advantage over SC.

7.3.2.5 Robustness to Interference

In OFDM, narrow interference typically affects only a small number of carriers, causing only a minor loss in capacity. If the locations of the interferences are known, it is possible not to transmit at those carriers or reduce the modulation order of transmission for those carriers only. Also, since the LDPC decoding is done based on SNR estimation per carrier, the error contribution of the noisy carriers will be minimized by the LDPC decoder even if the location of the interference is not known.

Robustness to interference of wide single carrier channels would be based on the same ingress cancellation techniques

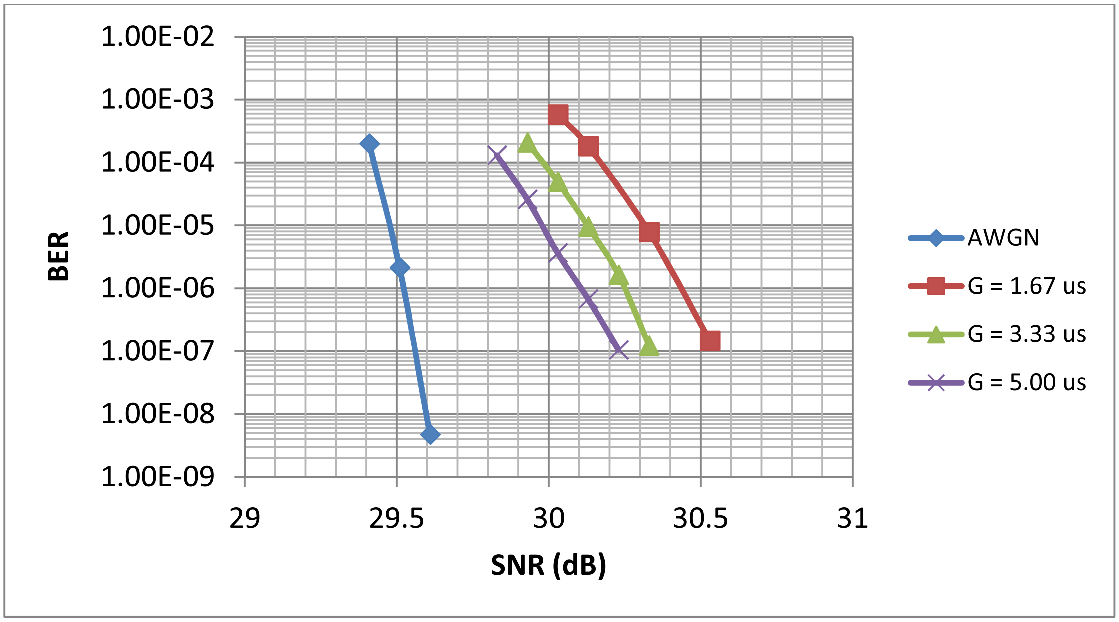


Figure 45 – OFDM/LDPC system performance in presence of SCTE-40 channel echoes

currently used in downstream and upstream receivers. However, for wider channels, these functions could become more challenging because of the increased probability of multiple interferers. This would result in inferior performance compared to today's single carrier in spectral regions beset by interference, or an increase in complexity to achieve the same performance.

In general, OFDM offers particular, understood simplicity and flexibility advantages for dealing with the narrowband interference environment. These could benefit DOCSIS, particularly as previously unused, unpredictable bands become used.

7.3.2.6 Throughput Scalability with SNR

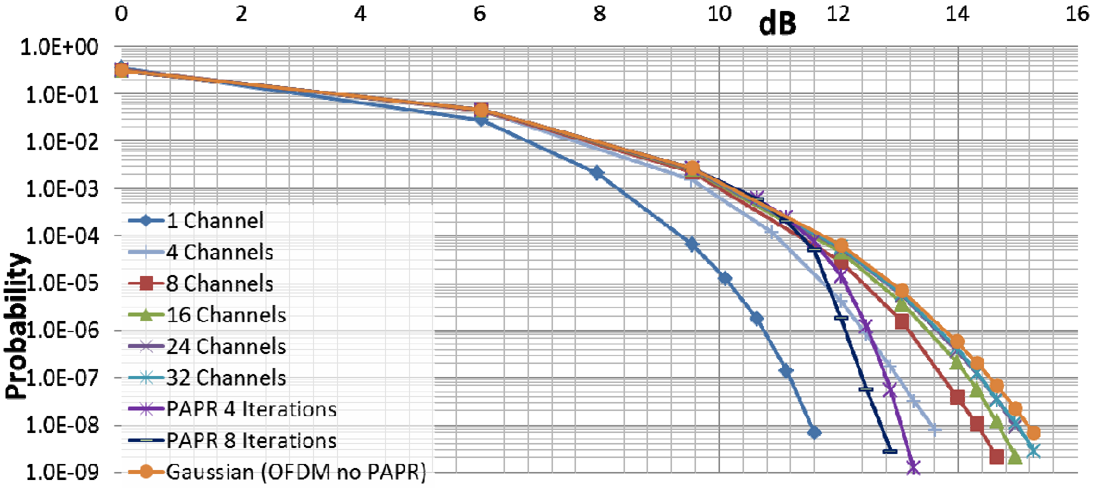
Another useful feature of OFDM modulation is that it enables use of different QAM constellations per carrier (also known as "bit loading"). This allows keeping all the benefits of a wide channel, while having the ability to fit modulation per the existing SNR at a narrow portion of spectrum. This enables maximizing throughput when the SNR is not constant within the channel band.

The non-flat SNR case is especially relevant for spectrum beyond 1 GHz, where signal attenuation falls sharply with frequency, or above the forward band of sub-1 GHz systems. Using a wide single carrier channel in this case would mean a compromise on throughput, and using a narrow singe carrier channel would require a myriad of channels.

7.3.2.7 Peak to Average Power Ratio (PAPR)

Peak to Average Ratio of OFDM modulation is frequently considered as its disadvantage due to the fact that OFDM symbol has Gaussian amplitude distribution (that's because of its multicarrier nature). It is true, but mainly in comparison to a single channel or a small number of channels.

DOCSIS 3.0 systems have at least 4 upstream channels, and this number will continue going up as long as single carrier channels are used to reach higher rates.



Notes: 6 MHz channels with 0.15 alpha and wide OFDM with Peak to Average Reduction Algorithm.

Figure 46 – Probability of Clipping as a Function of Peak to RMS Ratio

Figure 46 shows the PAPR profiles for OFDM and different numbers of single-carrier channels. The vertical axis is the clipping probability for the clipping threshold given in the horizontal axis.

The Gaussian profile is for OFDM with no PAPR reduction. Graphs for different numbers (1, 4, 8, 16, 24 and 32) of single-carrier channels are also shown (each with 0.15 RRC roll-off). It is seen that even when the number of single carrier channels is as low as four, the PAPR is not too different from Gaussian.

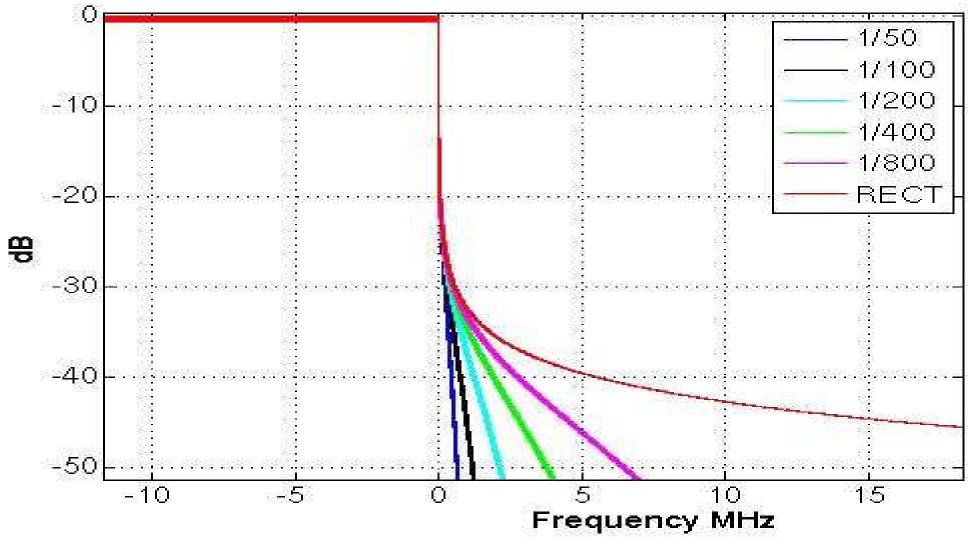
However, unlike single-carrier, OFDM offers ways of reducing peak-to-average power. One such method illustrated using this graph is called tone reservation. In this method a few (< 1%) of the tones are reserved to reduce the high amplitudes in an OFDM FFT. The results shown have been obtained by simulating the specific method given in the DVB-T2 specification. It is seen that the peak power of OFDM can be made to be less than four single-carrier channels at clipping probabilities of interest to cable applications.

Hence, as far as next gen DOCSIS PHY is concerned, OFDM actually has an advantage over bonded single carrier modulation of four channels or greater in terms of PAPR.

7.3.2.8 Frequency Agility

All options considered for downstream have width of multiples of 6 MHz or 8 MHz, for compatibility with the existing downstream grid.

A wide OFDM channel allows creating a frequency “hole” in its spectrum to enable legacy channels inside it, should there be a frequency planning constraint (as graphically shown in Figure 42. With this feature, OFDM retains the frequency agility of a narrow channel, while keeping all the benefits of a wide channel. A wide single carrier channel will be at a disadvantage in that respect.



Notes: lengths (ratio to total number of carriers)

Figure 47 – 16K symbol frequency response with different pulse shaping

To reduce the interference of OFDM

channel to the QAM channel inside it, an OFDM symbol shaping (windowing) can be employed as shown on Figure 47. This windowing makes the OFDM symbol length longer which implies a reduction in the bit rate. Nevertheless, as seen from the figure, windowing significantly sharpens the edge of the OFDM spectrum. This allows data carriers to be inserted until very close to the edge of the available bandwidth. So we have a capacity loss seen from the time domain representation and a capacity gain seen from the frequency domain representation. The net effect is a significant capacity gain and the optimum excess time for windowing has been found (for 12.5 KHz carrier separation) to be 1% of the useful OFDM symbol period (black line in Figure 47).

7.3.2.9 Upstream Multiple Access Considerations

Allowing simultaneous access of multiple CMs is essential for containing latency and for ease of CM management. OFDM modulation can be extended into an OFDMA (Orthogonal Frequency Division Multiple Access) modulation where several modems can transmit on different carriers at the same time.

The good news is that the DOCSIS 3.0 MAC convergence layer already supports that type of access for a case of SCDMA modulation in DOCSIS 3.0. The same concepts can be adopted with minor adjustments for OFDMA convergence layer. The concept of minislots that serves as an access sharing grid for the upstream transmission opportunities can be kept. The two dimensional minislot numbering used in SCDMA can also be kept for OFDMA. The contention, ranging and station maintenance arrangements can be kept.

In order to allow different bit loading per carrier, the minislots, if chosen as constant in time, may be different in size. That would be a change from constant size minislots in legacy DOCSIS, but this is an isolated change. Figure 48 shows an example of such access.

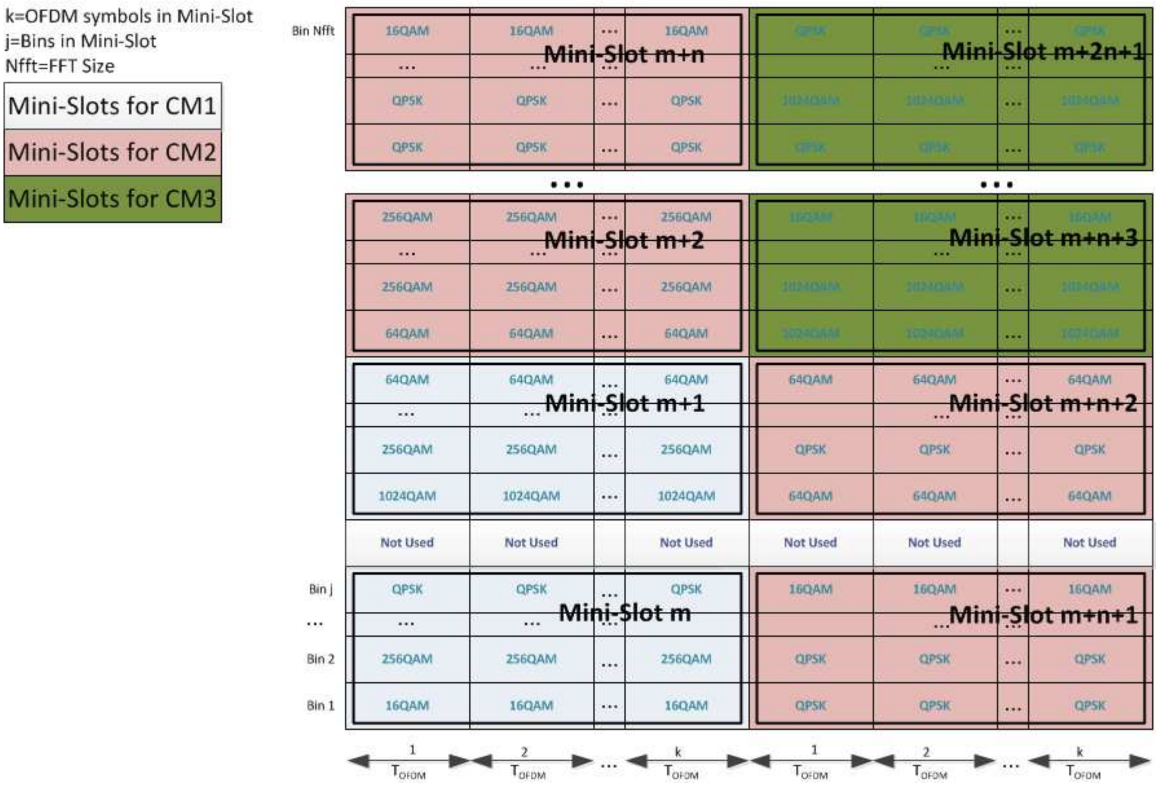


Figure 48 – Mini-slot based scheduling for OFDMA

7.3.3 OFDM Channel Parameter Examples

Table 24 – OFDM Channel Parameters for 192 MHz Wide Channel

Parameter	**Value**
Channel bandwidth	192 MHz
Useful bandwidth	190 MHz (-95 MHz to +95 MHz) -44 dB attenuation at 96 MHz band-edge
FFT size	16384
FFT sample rate	204.8 MHz (multiple of 10.24 MHz)
Useful symbol time	80 us
Carriers within 190 MHz	15200
Guard interval samples	683 (ratio=1/24; 3.33 us)
Symbol shaping samples	164 (ratio=1/100; 0.8 us)
Total symbol time	84.13us
Continuous pilots	128 (for synchronisation)
Scattered pilots	128 (for channel estimation)
PAPR pilots	128 (for PAPR reduction)
Useful data carriers per symbol	14816
QAM Constellations	4096-QAM, 1024, 256, 64, 16
Bit rate for 4096-QAM w/o FEC	2.11 Gbit/s (11.0 bits/s/Hz)
Bit rate for 1024-QAM w/o FEC	1.76 Gbit/s (9.17 bits/s/Hz)

Table 25 – OFDM Channel Parameters for 96 MHz Wide Channel

Parameter	Value
Channel bandwidth	96 MHz
Useful bandwidth	94 MHz (-47 MHz to +47 MHz) -44 dB attenuation at 48 MHz band-edge
FFT size	8192
FFT sample rate	102.4 MHz (multiple of 10.24 MHz)
Useful symbol time	80 us
Carriers within 94 MHz	7520
Guard interval samples	341 (ratio=1/24; 3.33us)
Symbol shaping samples	82 (ratio=1/100; 0.8 us)
Total symbol time	84.13us
Continuous pilots	64 (for synchronisation)
Scattered pilots	64 (for channel estimation)
PAPR pilots	64 (for PAPR reduction)
Useful data carriers per symbol	7328
QAM Constellations	4096-QAM, 1024, 256, 64, 16
Bit rate for 4096-QAM w/o FEC	1.05 Gbit/s (10.9 bits/s/Hz)
Bit rate for 1024-QAM w/o FEC	0.87 Gbit/s (9.07 bits/s/Hz)

Table 26 – OFDM Channel Parameters for 48 MHz Wide Channel

Parameter	Value
Channel bandwidth	48 MHz
Useful bandwidth	46 MHz (-23 MHz to +23 MHz) -44 dB attenuation at 24 MHz band-edge
FFT size	4096
FFT sample rate	51.2 MHz (multiple of 10.24 MHz)
Useful symbol time	80 us
Carriers within 46 MHz	3680
Guard interval samples	171 (ratio=1/24; 3.33us)
Symbol shaping samples	41 (ratio=1/100; 0.8 us)
Total symbol time	84.13us
Continuous pilots	32 (for synchronisation)
Scattered pilots	32 (for channel estimation)
PAPR pilots	32 (for PAPR reduction)
Useful data carriers per symbol	3584
QAM Constellations	4096-QAM, 1024, 256, 64, 16
Bit rate for 4096-QAM w/o FEC	0.51 Gbit/s (10.65 bits/s/Hz)
Bit rate for 1024-QAM w/o FEC	0.43 Gbit/s (8.88 bits/s/Hz)

Table 27 – OFDM Channel Parameter for 37 MHZ Wide Channel, Upstream NA Band

Parameter	Value
Channel bandwidth	37 MHz
Useful bandwidth	36 MHz (-18 MHz to +18 MHz) -40 dB attenuation at 18.5 MHz (TBC)
FFT size	2048
FFT sample rate	51.2 MHz
Sub-carrier spacing	25 KHz
Useful symbol time	40 us
Carriers within 36 MHz	1440
Guard interval samples	192(ratio=3/32; 3.75 us)
Symbol shaping samples	41 (ratio=1/50; 0.80 us)
Total symbol time	44.55us
Continuous pilots	16 (for synchronisation)
Scattered pilots	none (Channel est. via preamble)
PAPR pilots	16 (for PAPR reduction)
Useful data carriers per symbol	1408
QAM Constellations	1024-QAM, 256, 64, 16, QPSK
Bit rate (for 1024-QAM)	0.32 Gbit/s (8.56 bits/s/Hz)

7.3.3.1 Modulation Summary

DOCSIS 3.0 equipment, completed in 2006, is now seeing increasing field deployment. While deployed CM percentages are still modest, CMTS capabilities are being installed and spectrum plans have been put into place. It has been proven to be rugged and capable, and it is now timely to consider the next phase of DOCSIS evolution. And, as powerful as DOCSIS 3.0 may be, it most certainly can be enhanced by taking advantage of modern tools and the continued advancement in cost-effective, real-time processing power.

Two such approaches have been identified here – adding new symbol rates, similar to the DOCSIS 2.0 extension in 2002 that introduced 5.12 Msps, or introducing multi-carrier modulation, which has been embraced in standards bodies across industries. Table 28 summarizes various attributes of these PHY modulation alternatives relative to today’s available DOCSIS 3.0 baseline for the scaling of services to Gbps rates.

Table 28 – Relative Impact of Extensions to DOCSIS 3.0 for Gigabit Services

Attribute	Wide SC	Wide OFDM	Comments
Silicon Complexity (cost per bit)	-	+	Based on # of real-time multiplication operations
Transparency to existing D3.0 MAC	Same		OFDM: Minor mods to convergence layer
Field Technician Familiarity	+	-	
Robustness to interference	-	+	SC-QAM improved with SCDMA (upstream only)
Robustness to unknown plant (e.g. > 1 GHz operation)	-	+	
Throughput scalability per plant condition (SNR)	-	+	
Peak-to-Avg Power Ratio (PAPR)	Same		OFDM: better with PAPR reduction algorithms
Spectrum Allocation Flexibility	-	+	
New Requirements Definition	+	-	

<u>Notes</u>: Wide SC-QAM refers to 8x24 MHz. Wide OFDM refers to 16k IFFT 192 MHz.
"+" and "-" compare wide SC and wide OFDM to a 6.4 Mhz channel-bonded DOCSIS 3.0 baseline.

7.3.4 In Summary

By first stating the criteria, and then analyzing the available options against the criteria, it is suggested that the OFDM/OFDMA/LDPC wide channel is the best candidate for next generation gigabits capable DOCSIS PHY layer. This scheme is based on well-studied, widely adopted methods, allowing quick standardization turn around.

It enables to maximize the throughput with the available and future bandwidth and SNR resources. It is flexible enough to cope with new, less studied spectrum portions and interferences. It is more cost efficient than other alternatives for same throughputs (cost per bit). All these traits suggest that this PHY can optimally serve the DOCSIS evolution going into the gigabit rates, minimizing the investment needed by doing it “once and for all”.

8 DOCSIS MAC TECHNOLOGIES

8.1 DOCSIS Channel Bonding

DOCSIS Channel bonding may support full spectrum downstream. Additional DOCSIS channel bonding upstream may support higher upstream capabilities with targets to 1 Gbps. Achieving larger bonding group will require software, hardware and perhaps specification changes.

A future release of DOCSIS should enable bonding across legacy DOCSIS 3.0 and the new DOCSIS NG, even if they use dissimilar PHY technologies. The MAC layer and IP bonding will stitch the PHY systems together.

8.2 DOCSIS Scheduler Benefits

The DOCSIS protocol allows multiple users to "talk" or transmit at same moment in time and on the same channel, this was part of DOCSIS 2.0 introduction of SCDMA. The introduction of channel bonding allowed ATDMA based system to transmit at the same moment in time on differ frequencies while part of a channel bonding group.

Unlike DOCSIS, the EPON MAC allows "only one" subscriber to "talk" or transmit at any given moment in time. If we consider a single Home Gateway with multiple services and devices behind it, these will contend with each other and neighbors for time slots for transport of voice service, video conferencing, real-time data services, and even normal data and IPTV TCP acknowledgments.

Now, we must consider all the Home Gateways in a serving area domain competing for time slots allocated only on a "per home" basis, if the MSOs move to this style of architecture.

In many ways the EPON and EPOC MAC is most equivalent to a DOCSIS 1.1 MAC, of the 2000 era, because this supports multiple service flows, however allows only "one" user to talk or transmit at a time. The DOCSIS 2.0 and 3.0 specifications changed this limitation to accommodate for more devices, bandwidth, services, and concurrency of users and latency sensitivity; this is a powerful difference between the MAC standards.

The DOCSIS MAC designers knew that shared access meant contention for both bandwidth resources "and" time, this is why DOCSIS 2.0 and 3.0 support simultaneous transmission upstream enabling Quality of Service (QoS) and Quality of Experience (QoE).

There is another major factor with the DOCIS MAC, the development and feature set is controlled by the Cable Industry and not a third party standards organization, like the IEEE or ITU. This allows the MSO to make design request directly to systems vendors for continue innovation and support for new features that come along over time.

The DOCSIS MAC continues to change as the MSOs think of new service differentiation features and the flexible DOCISS MAC enable this support and creating a best in breed and cost effective MAC for the cable industry.

8.3 Services Enabled by DOCSIS

The DOCSIS technology can support virtually any service. DOCSIS technology may enable support for the full range of IP and Ethernet based services. The challenges for support for advanced layer 2 and layer 3 VPN services are not found in the DOCSIS access layer technology, but rather the network elements.

The DOCSIS CMTS will need to add support for desired layer 2 and layer 3 VPN services. The DOCSIS protocol with the use of the advanced MAC should support Ethernet Services types and Bandwidth Profiles defined by the Metro Ethernet Forum (MEF).

8.4 Importance of Backward Compatibility with DOCSIS 3.0 and Any Successor

The authors of this analysis believe that DOCSIS and any successor should consider the value of backwards compatibility especially across channel bonding groups. This assures previous and future investment may be applied to create a large IP based bandwidth network while not stranding previous capital investment and spectrum.

The use of channel bonding leverages every MHz, which is finite and not free, this is all towards an effort to create one large IP pipe to and from the home. The use of backwards compatibility has benefitted the cable industry as well as other industries which use technologies like IEEE Ethernet, WiFi, and EPON creating consumer investment protection, savings, and a smooth migration strategy.

The adoption of backward compatibility simply allows the MSOs to delay and perhaps avoid major investment to the network such as adding more data equipment, spectrum, node splits, and running fiber deeper.

The Data over Cable System Interface Specification (DOCSIS) began development in the late 1990's and has since had four versions released. DOCSIS standards include DOCSIS 1.0, 1.1, 2.0 and 3.0. The standards allowed for backwards compatibility and coexistence with previous versions of the standard.

As the needs of subscribers and providers continued to evolve, the DOCSIS standard was progressively upgraded to accommodate the change in services. The DOCSIS 2.0 standards increased upstream speeds and the DOCSIS 3.0 standard dramatically increased upstream and downstream bandwidth to accommodate higher speed data services.

These transitions capitalized on the availability of new technologies (ex: SCDMA) and the processing power of new silicon families (ex: Channel Bonding).

The authors of this analysis believe that DOCSIS and any successor should consider the value of backwards compatibility especially across channel bonding groups. This assures previous and future investment may be applied to create a large IP based bandwidth network while not stranding previous capital investment and spectrum.

The use of channel bonding leverages every MHz, which are finite and not free, this is all towards an effort to create one large IP pipe, to and from the home. The use of backwards compatibility has benefitted the cable industry as well as other industries which use technologies like IEEE Ethernet, WiFi, and EPON creating consumer investment protection, savings, and a smooth migration strategy.

The adoption of backward compatibility simply allows the MSOs to delay and

perhaps avoid major investment to the network such as adding more data equipment, spectrum, node splits, or running fiber deeper.

1. DOCSIS 3.0 QAM based and any successor should consider that every MHz should all share the same channel bonding group, this maximizes the use of existing spectrum and delays investment

2. Sharing channel bonding groups with DOCSIS 3.0 and Any Successor creates "one" IP Network (cap and grow networks hang around awhile)

3. Sharing the same bonding group assures previous and future investment may be applied in creating larger IP based bandwidth and not stranding previous capital investment

4. Backward Compatibility has benefitted industries like the IEEE Ethernet, WiFi, and EPON saving the entire eco-system money

5. Backward Compatibility simply allows the MSOs to delay and perhaps avoid major investment to the network such as adding more spectrum or running fiber deeper.

6. Avoids the MSO having a RF Data Simulcasting Tax (as discussed in this report)

7. All of our analysis in this report assumes backward compatibility with DOCSIS 3.0 QAM and any successor, like DOCSIS OFDM; thus creating a larger and larger IP bonding group with each year's investment. If this is not the case the investment in HFC upgrades will pull forward. It is uncertain of the exact level of financial impact but the total cost of ownership may be higher when deploying two separate IP based network technologies.

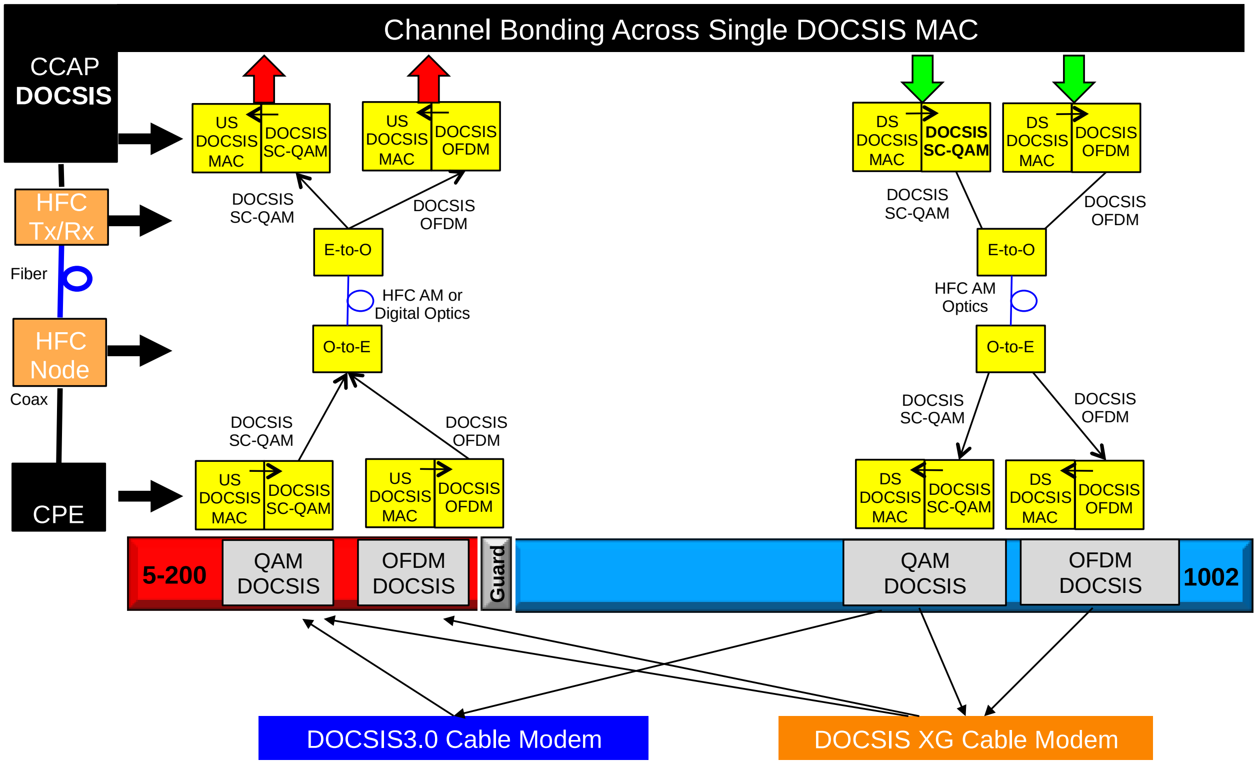


Figure 49 – Channel Bonded DOCSIS 3.0 and DOCSIS NG System

This is an illustration of channel bonding across a DOCSIS 3.0 and potential DOCSIS NG system. Figure 49 shows a DOCSIS 3.0 system coexisting with a DOCSIS NG system, then adding a DOCSIS NG system this platform could support legacy DOCSIS 3.0 SC-QAM, modulation and perhaps add 256-QAM upstream and 1024-QAM downstream, and RS and also supporting the new DOCSIS NG PHY. This will allow backward compatibility for the DOCSIS 3.0 cable modems and CMTS, while supporting the new PHY and likely in new spectrum.

Figure 50 is an illustration of the possible integration of HFC optics in the CCAP that will support DOCSIS 3.0 and DOCSIS NG. DOCSIS 3.0 and DOCSIS NG will likely be supported on the same card in the future without requiring HFC optical integration to the CCAP.

8.5 RF Data Simulcasting Tax

We would recommend strongly examining the history and impact of simulcasting services. If an alternative to DOCSIS is considered this will require new spectrum. The existing DOCSIS service and spectrum allocation may actually continue to grow during the initial introduction of the new data MAC/PHY technology, such as EPOC.

New spectrum that likely mirrors the size of DOCSIS would have to be found, so that at least the same services may be offered using an EPOC technology. The amount of new spectrum allocated by the MSO for DOCSIS and EPOC would begin the RF Data Simulcasting Tax Period.

It is true, that legacy networks tend to hang around for a long time. For example,

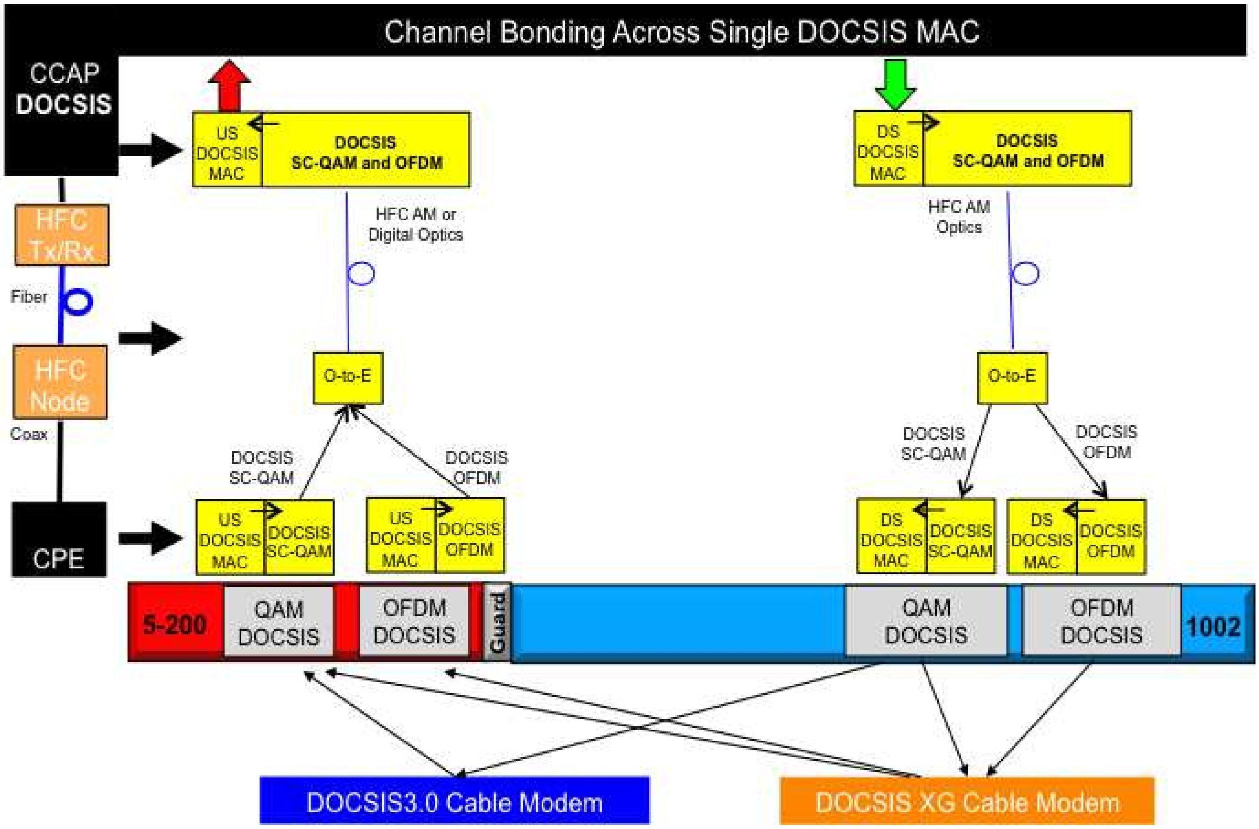


Figure 50 – Channel Bonded DOCSIS 3.0 and 4.0 System with CCAP Integrated HFC Optics

MSOs that deployed constant bit rate voice services, known as CBR voice, may still have these technologies occupying spectrum, even though they also have voice services using DOCSIS in the same network. The challenge is cost; the cost to reclaim spectrum is substantial, it requires new CPE and Headend systems, for no additional revenue.

The additional impact is finding new spectrum to offer what is a duplicate service using a different technology. It is fair to say that the cost for supporting parallel RF data networking technologies will have a capital and operational impact that will likely be more than expanding the current technology over the existing HFC network.

DOCSIS has the ability with each passing year investment to create larger and larger IP bonding groups, to enable higher speed service tiers and support traffic growth. Additionally, the DOCSIS CPEs may be channel bonded with legacy PHY and/or new PHY technologies, while all sharing the same MAC layer-bonding group.

Also, not a single DOCSIS CPE would be required to change to reclaim spectrum, because of backward compatibility or to eliminate the RF data simulcasting tax, as this network tax could be avoided with DOCSIS current and future systems.

This is a compelling feature of continuing to leverage DOCSIS 3.0 and why next generation DOCSIS needs to be backward compatible at the MAC layer, with different PHYs.

1. The amount of new spectrum allocated by the MSO for DOCSIS and EPOC would begin the RF Data Simulcasting Tax Period.
2. The existing DOCSIS service and spectrum allocation may actually continue to grow during an initial introduction of a new data MAC/PHY technology, such as EPOC.
3. Legacy networks tend to hang around for a long time, CBR Voice.
4. A challenge is the cost to reclaim spectrum is substantial; it requires new CPE and Headend systems, for likely no additional revenue.
5. The additional impact is finding new spectrum to offer what is a duplicate service offering using a different technology, to find capacity node splits, new node placement in the field, and/or spectrum expansion, new powering for the OSP equipment, and more are all impacts.
6. It is fair to say that the cost for supporting a parallel RF data networking technology will have a capital and operational impact.
7. The ability that DOCSIS has is that with each passing year spectrum is allocated creating larger and larger IP bonding groups, to enable higher speed service tiers and support traffic growth.
8. This is a compelling feature of continuing to leverage DOCSIS 3.0 and why next generation DOCSIS needs to be backward compatible.

9 NETWORK CAPACITY ANALYSIS

9.1 Intro

The network capacity of the cable access network is determined by the amount of spectrum available and the data rate possible within the spectrum. The modern cable network is incredibly flexible allowing the MSO to make targeted investments where and when needed to either incrementally or in some cases substantially increase network capacity depending on the capacity expansion method selected.

The use of capacity expansion methods may be applied across an entire network footprint or with laser beam focus to address capacity challenges. Figure 51 is an attempt to capture the various methods available to increase or improve capacity of the network. The diagram brings together methods and techniques used by various disciplines within the MSO, such as outside/inside plant, IP/Data, SDV, and Video Processing. The techniques will allow the MSO to transform their network from broadcast to unicast and from analog/digital to IP.

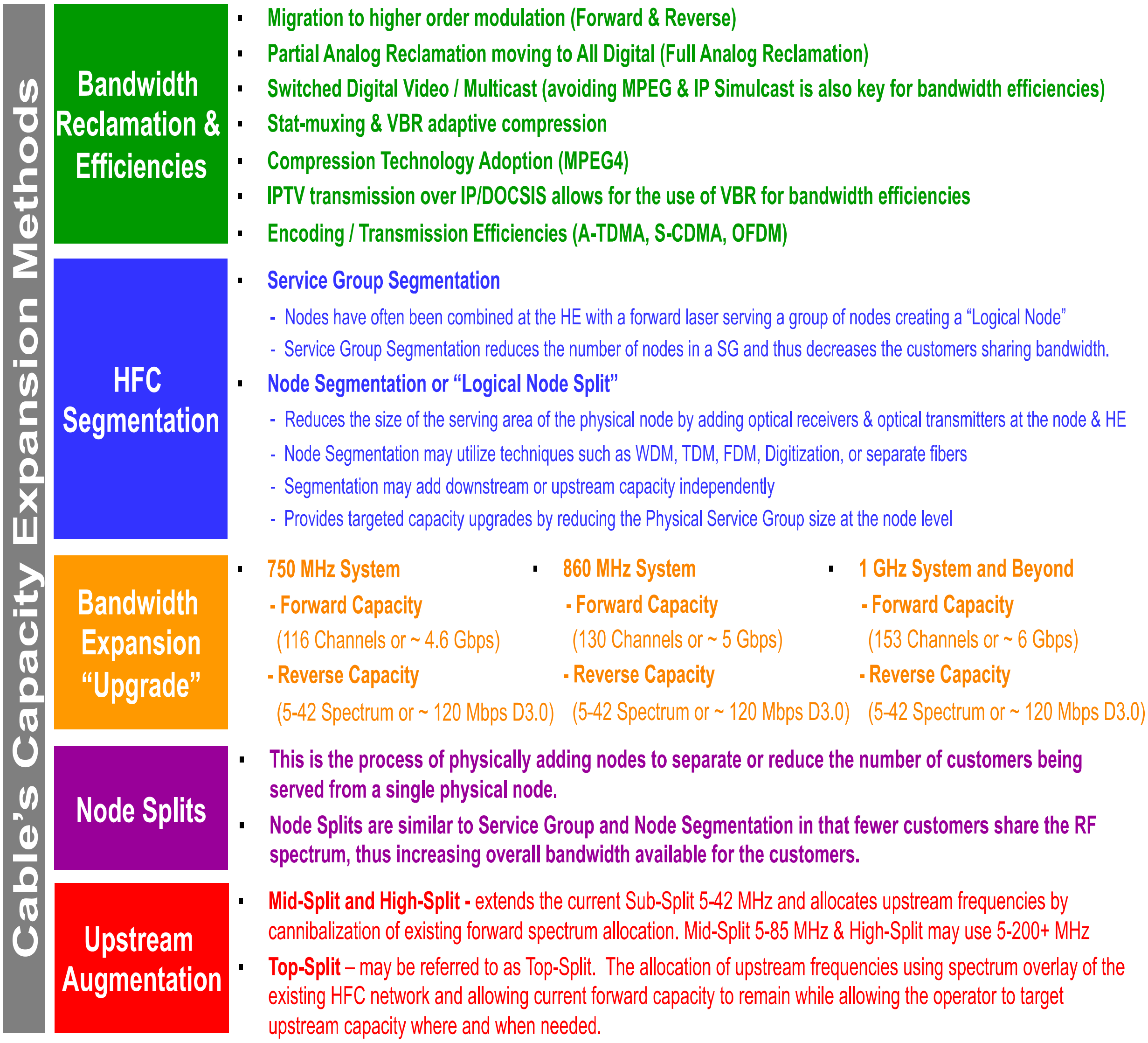
Today, in fact MSOs may use techniques to increase capacity without touching the outside plant; this is dramatically different than the approaches that were used for decades. The technique referred to as Bandwidth Reclamation and Efficiencies, as illustrated in the top of Figure 51 is becoming the primary method to address system wide capacity challenges. In most cases this technique may be implemented with equipment in the headend and home, thus not requiring conditioning of the outside plant or headend optics.

A technique recently put into practice by some cable operators is partial or even full analog reclamation. This enables the operator to transition the channels currently transmitted in analog and to transmit them only in digital format allowing greater bandwidth efficiencies by requiring the use of a digital terminal adapter (DTA) alongside televisions that may have only had analog services.

Another technique for Bandwidth Reclamation and Efficiencies is the use of Switch Digital Video (SDV). The use of SDV allows the cable operator to transmit in the network only the video streams that are being viewed by consumers. This allows the operator to increase the number of channels offered to consumers, in fact the actual channels offered to the consumers may exceed the throughput capabilities of the network but through careful traffic engineering and capacity planning this approach is an excellent way of adding additional capacity to the network.

This technique is a form of over-subscription and has been in practice for decades by the telecommunication industry. The items captured in Bandwidth Reclamation and Efficiencies are the modern methods to expand capacity. In many respects the Bandwidth Expansion “upgrade” approach as illustrated in Figure 51 whereby the entire network was upgraded to increase capacity, may be seldom used in the future. If used, this may be part of a joint plan to increase the spectrum allocation of the return path.

In the future, the use of IP for video delivery will provide even greater bandwidth efficiencies. IP used for digital video transmission and will also provide functionality similar to the techniques used in SDV. Another key advantage is that IP allows for the use of variable bitrate (VBR)



Cable's Capacity Expansion Methods

Bandwidth Reclamation & Efficiencies

- Migration to higher order modulation (Forward & Reverse)
- Partial Analog Reclamation moving to All Digital (Full Analog Reclamation)
- Switched Digital Video / Multicast (avoiding MPEG & IP Simulcast is also key for bandwidth efficiencies)
- Stat-muxing & VBR adaptive compression
- Compression Technology Adoption (MPEG4)
- IPTV transmission over IP/DOCSIS allows for the use of VBR for bandwidth efficiencies
- Encoding / Transmission Efficiencies (A-TDMA, S-CDMA, OFDM)

HFC Segmentation

- Service Group Segmentation
  - Nodes have often been combined at the HE with a forward laser serving a group of nodes creating a "Logical Node"
  - Service Group Segmentation reduces the number of nodes in a SG and thus decreases the customers sharing bandwidth.
- Node Segmentation or "Logical Node Split"
  - Reduces the size of the serving area of the physical node by adding optical receivers & optical transmitters at the node & HE
  - Node Segmentation may utilize techniques such as WDM, TDM, FDM, Digitization, or separate fibers
  - Segmentation may add downstream or upstream capacity independently
  - Provides targeted capacity upgrades by reducing the Physical Service Group size at the node level

Bandwidth Expansion "Upgrade"

- 750 MHz System
  - Forward Capacity (116 Channels or ~ 4.6 Gbps)
  - Reverse Capacity (5-42 Spectrum or ~ 120 Mbps D3.0)
- 860 MHz System
  - Forward Capacity (130 Channels or ~ 5 Gbps)
  - Reverse Capacity (5-42 Spectrum or ~ 120 Mbps D3.0)
- 1 GHz System and Beyond
  - Forward Capacity (153 Channels or ~ 6 Gbps)
  - Reverse Capacity (5-42 Spectrum or ~ 120 Mbps D3.0)

Node Splits

- This is the process of physically adding nodes to separate or reduce the number of customers being served from a single physical node.
- Node Splits are similar to Service Group and Node Segmentation in that fewer customers share the RF spectrum, thus increasing overall bandwidth available for the customers.

Upstream Augmentation

- Mid-Split and High-Split - extends the current Sub-Split 5-42 MHz and allocates upstream frequencies by cannibalization of existing forward spectrum allocation. Mid-Split 5-85 MHz & High-Split may use 5-200+ MHz
- Top-Split – may be referred to as Top-Split. The allocation of upstream frequencies using spectrum overlay of the existing HFC network and allowing current forward capacity to remain while allowing the operator to target upstream capacity where and when needed.

Figure 51 – Cable's Capacity Expansion Methods

encoding increasing the capacity of the network and the utilization of higher order compression techniques.

Cable operator's selection priority of the capacity expansion methods has and will continue to vary. The cable operators will eventually use all or nearly all of the capacity expansion methods in Figure 51

9.2 Importance of Error Correction Technologies

The paper by David J.C. MacKay and Edward A. Ratzer, titled "Gallager Codes for High Rate Applications", published January 7, 2003 [27], examines the improvements obtained by switching from Reed-Solomon codes to Gallager codes or Low Density

Parity-Check (LDPC) code. It is the opinion of this author, that the MacKay paper is one of the best comparisons of illustrating the benefits of switching to LDPC from Reed-Solomon. The paper initially released in 2003, suggests some modifications to Gallager codes to improve performance. The paper suggest about a 5 dB gain. The paper lists further ideas worth investigating that may improve performance.

The use of LDPC has expanded recently with the adoption by the IEEE WiMAX 802.16e, ITU-T G.hn. and the cable industry use for downstream transmission in DVB-C2. The use of LDPC may be used in any carrier modulation method, such as SC-QAM, OFDM, or Wavelet, and the expectation is the use of higher order modulation is achievable compared with Reed-Solomon based systems. It is reasonable to suggest a 6 dB gain is possible by switching from Reed-Solomon to LDPC and this will allow an increase in modulation by perhaps two orders, in other words perhaps one could move from 64-QAM to perhaps 256-QAM. In Table 29, the R-S using approximately 86-87% coding and LDPC using the inner code of 5/6 or 83% yields a 6 dB difference and will allow an increase of two orders of the modulation.

The key takeaway is the use of LDPC will improve network capacity or actual bit per second per Hertz over Reed-Solomon based systems, and this is achieved by enabling the use of higher order modulation with the same signal-to-noise ratio (SNR) condition. This allows operators to allocate less spectrum compared to Reed-Solomon based systems or have more network capacity in occupied spectrum.

The benefits of the cable industry's use can be seen in DVB-C2 systems. However, the use of LDPC for upstream cable data use is still under study as seen in this report. There are also other error correction technologies to consider that have been adopted by other standards groups.

This section will state the major differences and reasons why the use of modern error correction technology is key to increasing network capacity. The new error correction technology and the assumed two-order increase in modulation while operating in the same Signal to Noise Ratio (SNR) environment is the major reason there is an improvement in capacity.

Refer to Table 29 to Table 31 for the DOCSIS Single Carrier-QAM with Reed-Solomon system verse the performance estimates of a DOCSIS Multi-carrier OFDM with LDPC system and also refer to Table 32 to Table 34 for the analysis of these competing PHY layer technologies.

This section compares DOCSIS Single Carrier QAM and the current error correction technology with the proposed DOCSIS NG use of OFDM and the modern LDPC error correction technology.

9.3 DOCSIS 3.0 Single Carrier-QAM with Reed-Solomon

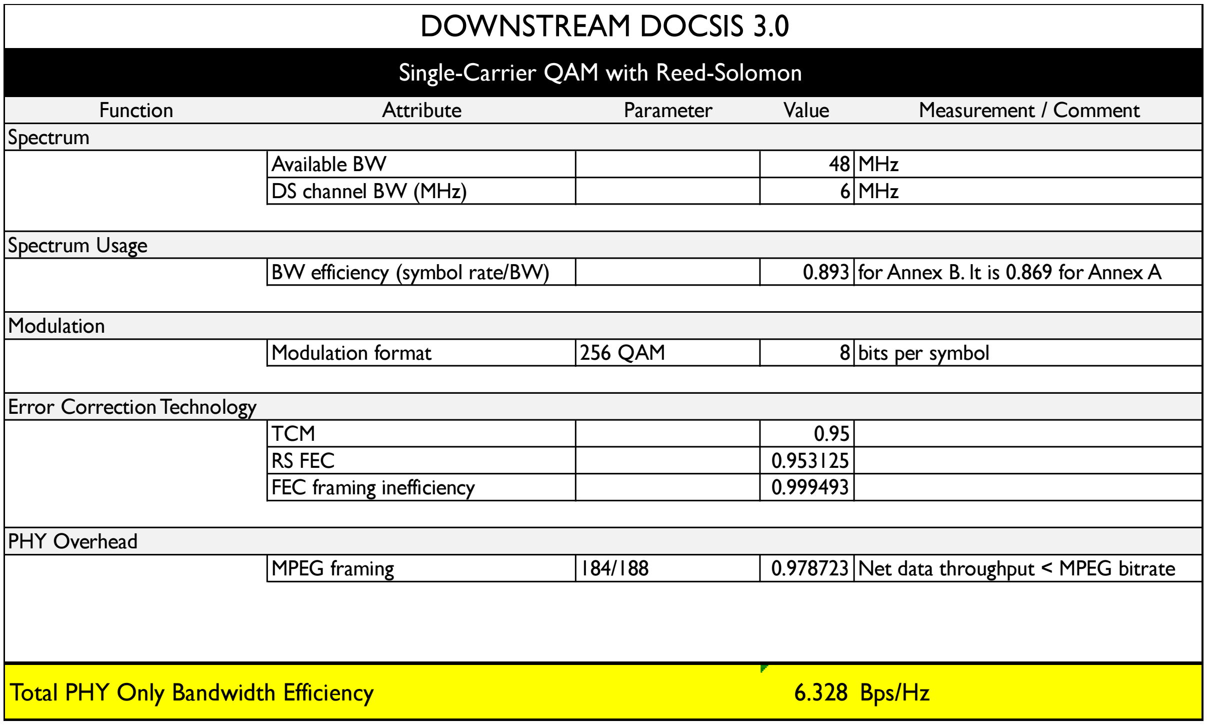
The DOCSIS SC-QAM 256-QAM downstream, as shown in Table 29 and the following two tables models the upstream using DOCSIS SC-QAM 64-QAM and DOCSIS 256-QAM. Each scenario assumes ATDMA.

These tables measure the PHY layer spectral efficiency of DOCSIS QAM based solutions. The channel coding for controlling errors in data transmission for the DOCSIS examples use Reed-Solomon forward error correction (RS-FEC) and Trellis Modulation or also known as Trellis Coded Modulation (TCM).

These are used to calculate the network capacity of the cable network considering several spectrum options found in the Network Capacity section.

A key take away is performance gap between 256-QAM PHY and 64-QAM layer efficiencies. The assumptions for 64-QAM at 4.1 bps/Hz would require 33% more spectrum and DOCSIS channels to maintain the equivalent PHY layer throughput. The use of DOCSIS 256-QAM for the upstream is not part of the DOCSIS standards. However some CMTS and CM products support this modulation profile in hardware.

Table 29 – Downstream DOCSIS 3.0 256-QAM with Reed-Solomon & TCM



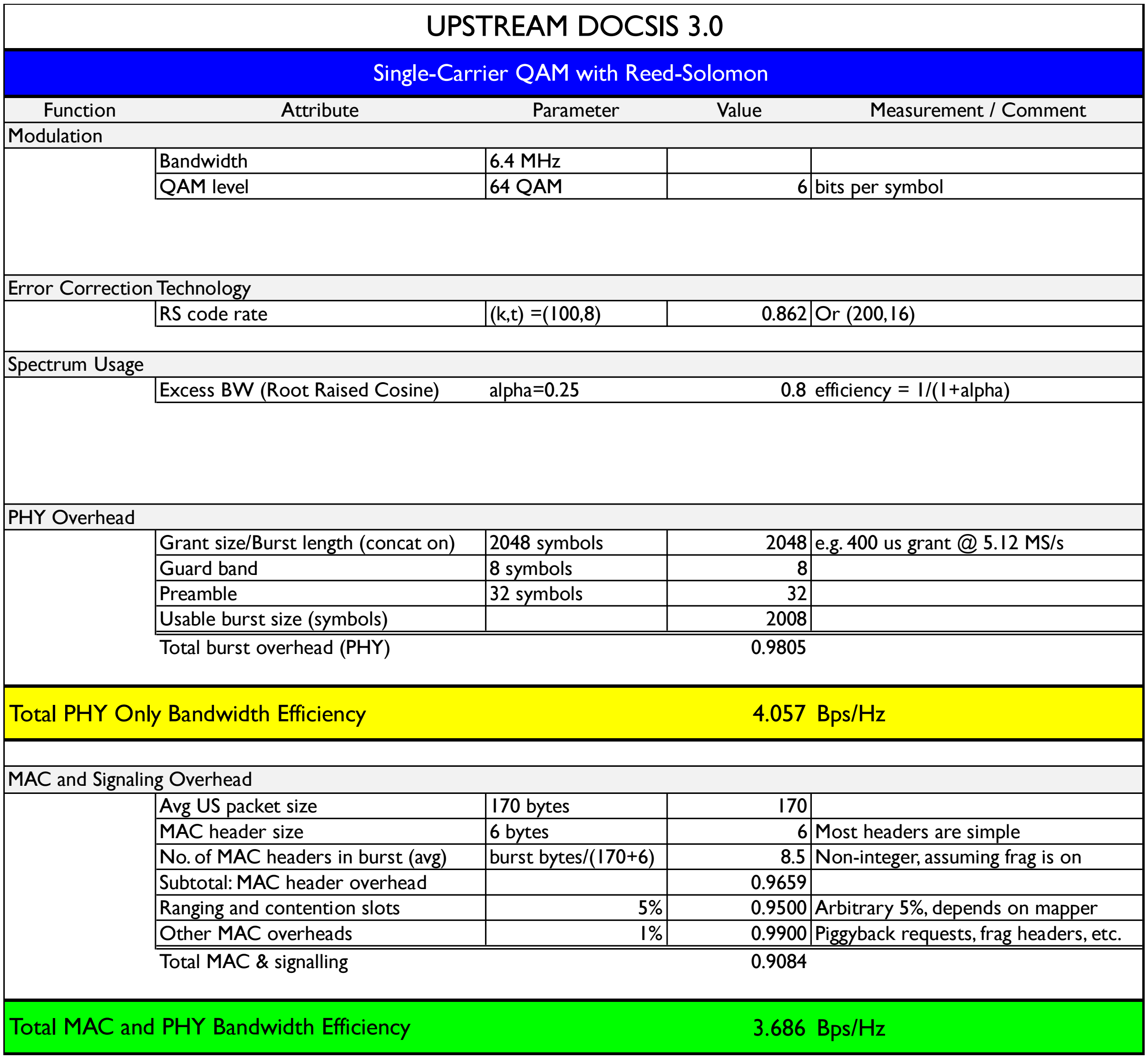
DOWNSTREAM DOCSIS 3.0

Single-Carrier QAM with Reed-Solomon

Function	Attribute	Parameter	Value	Measurement / Comment
Spectrum				
	Available BW		48	MHz
	DS channel BW (MHz)		6	MHz
Spectrum Usage				
	BW efficiency (symbol rate/BW)		0.893	for Annex B. It is 0.869 for Annex A
Modulation				
	Modulation format	256 QAM	8	bits per symbol
Error Correction Technology				
	TCM		0.95	
	RS FEC		0.953125	
	FEC framing inefficiency		0.999493	
PHY Overhead				
	MPEG framing	184/188	0.978723	Net data throughput < MPEG bitrate
Total PHY Only Bandwidth Efficiency			6.328	Bps/Hz

The DOCSIS specifications could be modified to include 256-QAM upstream as well as 1024-QAM in the upstream and downstream. However, the real major gains would be achieved by changing the error correction technology.

Table 30 – Upstream DOCSIS 3.0 64-QAM with Reed Solomon



UPSTREAM DOCSIS 3.0

Single-Carrier QAM with Reed-Solomon

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	6.4 MHz		
	QAM level	64 QAM	6	bits per symbol
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Excess BW (Root Raised Cosine)	alpha=0.25	0.8	efficiency = 1/(1+alpha)
PHY Overhead				
	Grant size/Burst length (concat on)	2048 symbols	2048	e.g. 400 us grant @ 5.12 MS/s
	Guard band	8 symbols	8	
	Preamble	32 symbols	32	
	Usable burst size (symbols)		2008	
	Total burst overhead (PHY)		0.9805	
Total PHY Only Bandwidth Efficiency			4.057	Bps/Hz
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	MAC header size	6 bytes	6	Most headers are simple
	No. of MAC headers in burst (avg)	burst bytes/(170+6)	8.5	Non-integer, assuming frag is on
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Piggyback requests, frag headers, etc.
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			3.686	Bps/Hz

Table 31 – Upstream DOCSIS 3.0 256-QAM with Reed Solomon

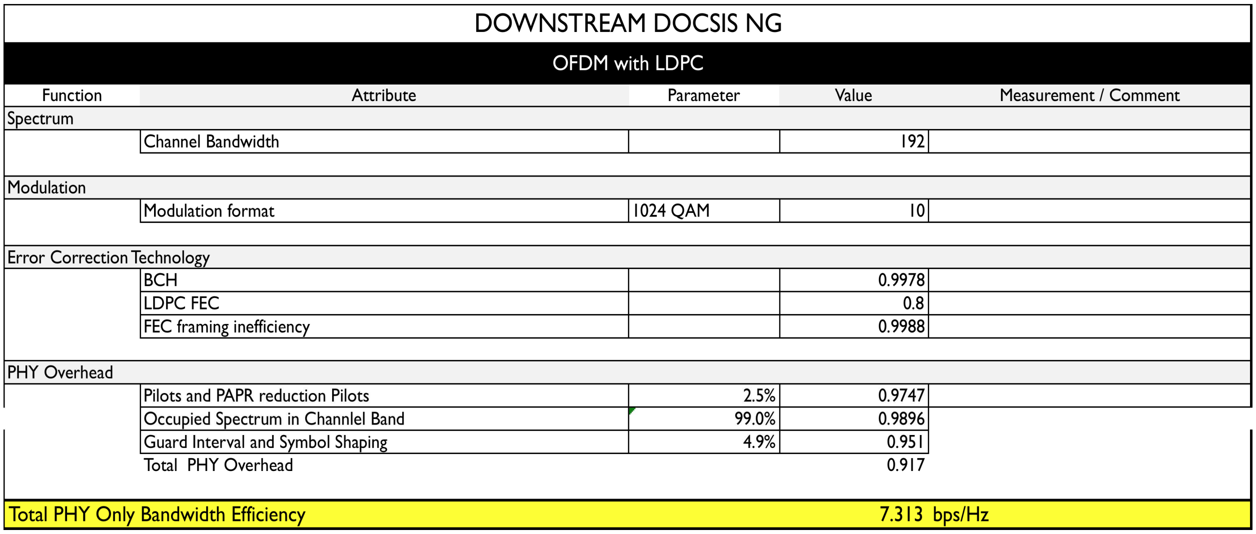
UPSTREAM DOCSIS 3.0				
Single-Carrier QAM with Reed-Solomon				
Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Bandwidth	6.4 MHz		
	QAM level	256 QAM	8	bits per symbol
Error Correction Technology				
	RS code rate	(k,t) =(100,8)	0.862	Or (200,16)
Spectrum Usage				
	Excess BW (Root Raised Cosine)	alpha=0.25	0.8	efficiency = 1/(1+alpha)
PHY Overhead				
	Grant size/Burst length (concat on)	2048 symbols	2048	e.g. 400 us grant @ 5.12 MS/s
	Guard band	8 symbols	8	
	Preamble	32 symbols	32	
	Usable burst size (symbols)		2008	
	Total burst overhead (PHY)		0.9805	
Total PHY Only Bandwidth Efficiency			5.409	Bps/Hz
MAC and Signaling Overhead				
	Avg US packet size	170 bytes	170	
	MAC header size	6 bytes	6	Most headers are simple
	No. of MAC headers in burst (avg)	burst bytes/(170+6)	11.4	Non-integer, assuming frag is on
	Subtotal: MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Piggyback requests, frag headers, etc.
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			4.914	Bps/Hz

9.4 DOCSIS NG Multi-carrier OFDM with Low Density Parity-Check (LDPC) code

The analysis in this section provides measurements using OFDM/OFDMA. Again OFDM is not part of the DOCSIS 3.0 standard. The channel coding for controlling errors in data transmission is assumed to use Low Density Parity-Check (LDPC) code also referred to as Gallager codes.

The analysis also uses values as described in Section 7.3.3 OFDM Channel Parameter Examples discuss in this paper. The target for these DOCSIS NG OFDM and LDPC estimates is to use an error correction amount referred to as 5/6 inner code rates or .833. The strong error correction used for the LDPC is modeled to achieve the Carrier to Noise target of 6 dB below Reed Solomon code rate of 86%. This will mean for the same modulation format R-S will yield greater b/s/Hz than LDPC using a stronger FEC in this effort to achieve a 6 dB decrease in C/N.

Table 32 – Downstream DOCSIS OFDM 1024-QAM with LDPC



DOWNSTREAM DOCSIS NG

OFDM with LDPC

Function	Attribute	Parameter	Value	Measurement / Comment
Spectrum				
	Channel Bandwidth		192	
Modulation				
	Modulation format	1024 QAM	10	
Error Correction Technology				
	BCH		0.9978	
	LDPC FEC		0.8	
	FEC framing inefficiency		0.9988	
PHY Overhead				
	Pilots and PAPR reduction Pilots	2.5%	0.9747	
	Occupied Spectrum in Channlel Band	99.0%	0.9896	
	Guard Interval and Symbol Shaping	4.9%	0.951	
	Total PHY Overhead		0.917	
Total PHY Only Bandwidth Efficiency			7.313	bps/Hz

The downstream DOCSIS OFDM 1024-QAM with LDPC system has about a 20% performance improvement of DOCSIS SC-QAM 256-QAM with Reed-Solomon. This is attributed primary to the FEC and not to the change in multi-carrier OFDM. The modern FEC will support greater Modulation QAM Format in the same SNR.

In the previous figures, 256-QAM was analyzed using estimates for PHY and MAC layer efficiency comparing DOCSIS single carrier 256-QAM and DOCSIS OFDM 256-QAM. The use of LDPC may allow higher upstream modulation schemes to be used compared with Reed-Solomon based approaches.

This could mean that 64-QAM Reed-Solomon system may actually be compared with an OFDM 256-QAM LDPC based system in the same Signal to Noise Ratio environment. Moreover, a 256-QAM Reed-Solomon system may actually be compared with a OFDM 1024-QAM LDPC based system in the same SNR environment.

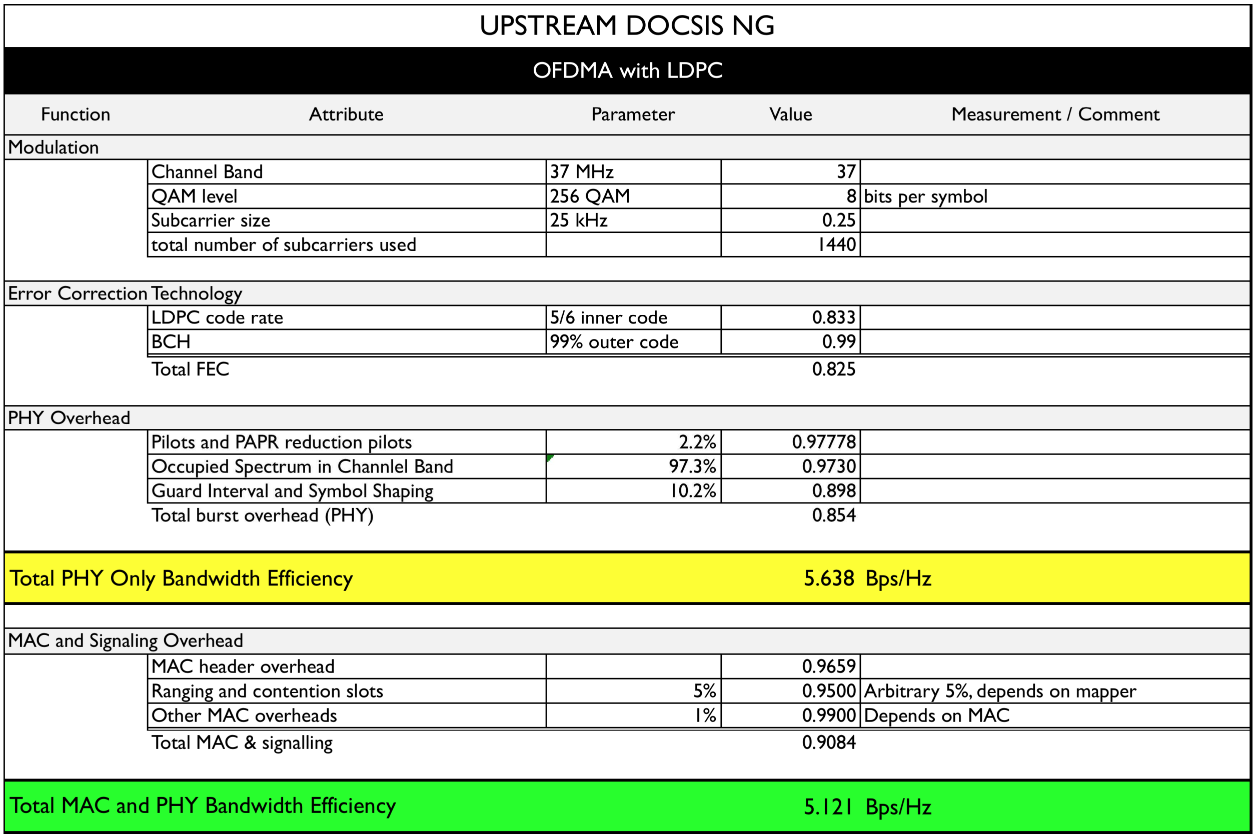
The goal to target the OFDM and LDPC system to operated in the same SNR environment and with two orders increase in

QAM level, required us to apply more error correction codes to LDPC.

Again, because we are assuming that LDPC will be capable of operating in the same SNR environment while using 2 orders higher modulation than a Reed Solomon system. This accounts for the added FEC overhead and lower performance when using the same QAM level.

The actual performance of either system in real-world HFC deployments is unknown. There are many attributes and assumptions than can be modified. We used an estimate that we considered to be fair for single carrier QAM and OFDM. These are subject to debate until systems are tested in a cable system.

Table 33 – Upstream DOCSIS OFDM 256-QAM with LDPC

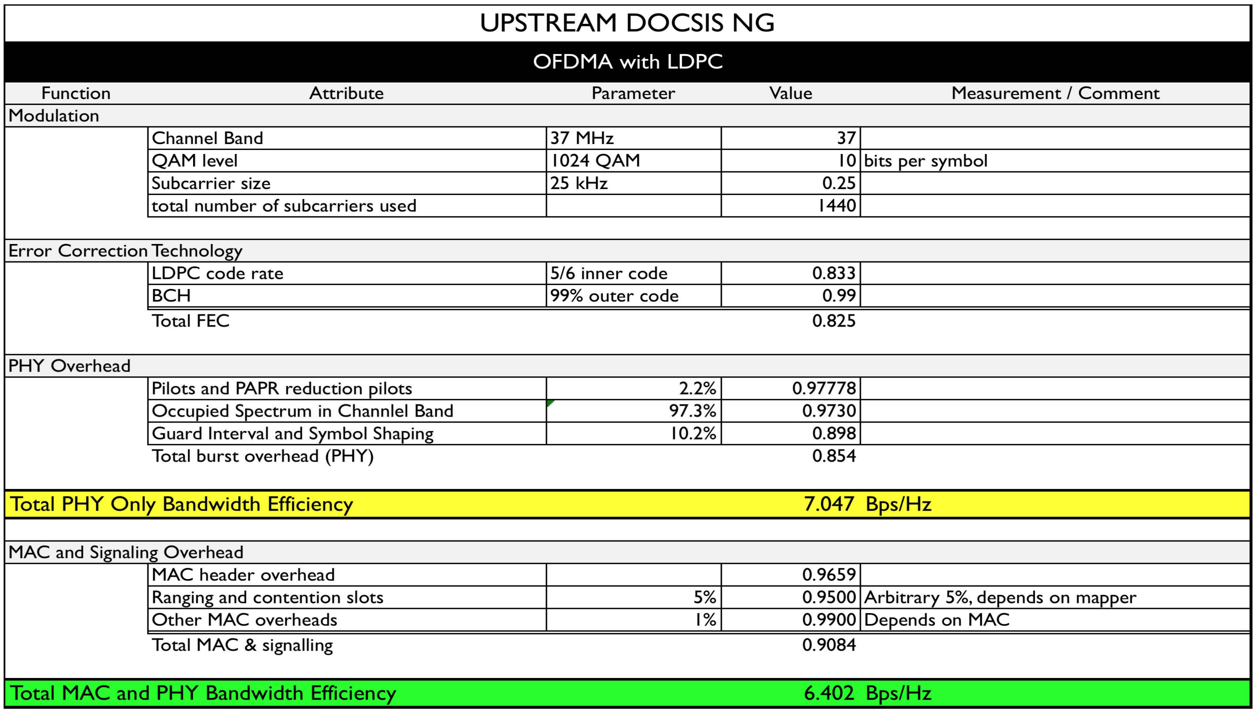


UPSTREAM DOCSIS NG

OFDMA with LDPC

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Channel Band	37 MHz	37	
	QAM level	256 QAM	8	bits per symbol
	Subcarrier size	25 kHz	0.25	
	total number of subcarriers used		1440	
Error Correction Technology				
	LDPC code rate	5/6 inner code	0.833	
	BCH	99% outer code	0.99	
	Total FEC		0.825	
PHY Overhead				
	Pilots and PAPR reduction pilots	2.2%	0.97778	
	Occupied Spectrum in Channlel Band	97.3%	0.9730	
	Guard Interval and Symbol Shaping	10.2%	0.898	
	Total burst overhead (PHY)		0.854	
Total PHY Only Bandwidth Efficiency			5.638	Bps/Hz
MAC and Signaling Overhead				
	MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Depends on MAC
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			5.121	Bps/Hz

Table 34 – Upstream DOCSIS OFDM 1024-QAM with LDPC



UPSTREAM DOCSIS NG

OFDMA with LDPC

Function	Attribute	Parameter	Value	Measurement / Comment
Modulation				
	Channel Band	37 MHz	37	
	QAM level	1024 QAM	10	bits per symbol
	Subcarrier size	25 kHz	0.25	
	total number of subcarriers used		1440	
Error Correction	Technology			
	LDPC code rate	5/6 inner code	0.833	
	BCH	99% outer code	0.99	
	Total FEC		0.825	
PHY Overhead				
	Pilots and PAPR reduction pilots	2.2%	0.97778	
	Occupied Spectrum in Channlel Band	97.3%	0.9730	
	Guard Interval and Symbol Shaping	10.2%	0.898	
	Total burst overhead (PHY)		0.854	
Total PHY Only Bandwidth Efficiency			7.047	Bps/Hz
MAC and Signaling Overhead				
	MAC header overhead		0.9659	
	Ranging and contention slots	5%	0.9500	Arbitrary 5%, depends on mapper
	Other MAC overheads	1%	0.9900	Depends on MAC
	Total MAC & signalling		0.9084	
Total MAC and PHY Bandwidth Efficiency			6.402	Bps/Hz

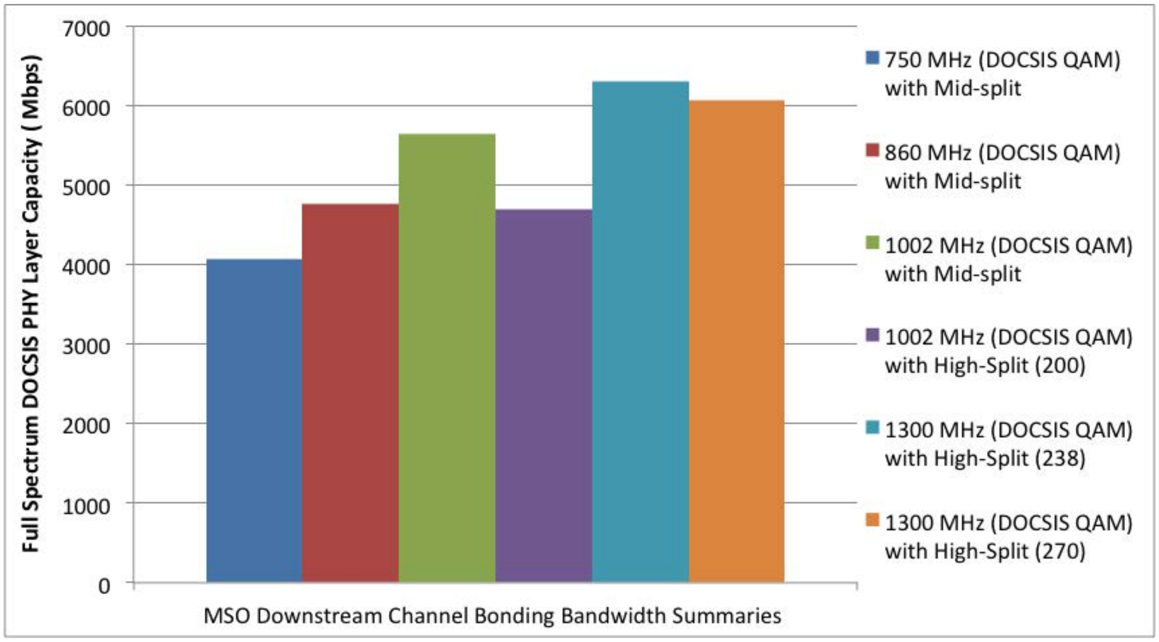


Figure 52 – 256 SC-QAM RS Codes PHY

9.5 Downstream Capacity

The most critical determination for the capacity of the network is the amount of spectrum available. The determination of the downstream capacity will assume the eventual migrations to an all IP based technology. The migration to all IP on the downstream which will optimize the capacity of the spectrum providing the versatility to use the network for any service type and provide the means to compete with PON and the flexibility to meet the needs of the future.

Table 35 provides capacity projections considering the upstream spectrum split and the use of DOCSIS Single Carrier QAM using several downstream spectrum allocations from 750 MHz to 1002 MHz. Certainly there are other spectrum options that could be considered such as moving the downstream above 1 GHz such as 1300 MHz as well as other spectrum options for the upstream. This table will calculate the estimated downstream PHY layer capacity using several spectrum options with limits of 256-QAM though higher modulations are possible.

Figure 52 shows different downstream spectrum allocations as well as the removal of upstream spectrum from the downstream. The downstream network capacity is illustrated using DOCSIS 256 SC-QAM Reed-Solomon Codes PHY or DOCSIS 1024-QAM OFDM LDPC capacity assuming full spectrum.

9.6 Upstream Capacity

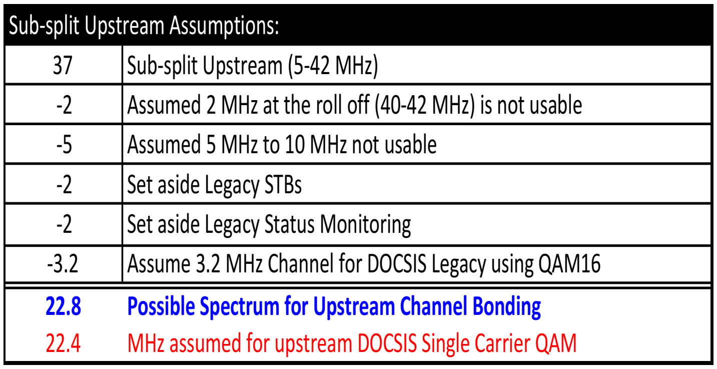
The upstream capacity measurements are more complicated and not as straightforward as the downstream capacity projections. In the Figure 53, many of the spectrum split options were evaluated considering several PHY layer options and modulation schemes within each spectrum split.

These are some key assumptions about the upstream capacity estimates:

Table 35 – 256 SC-QAM RS PHY or 1024-QAM OFDM LDPC Full Spectrum Capacity

Split Type	MSO Downstream Channel Bonding Bandwidth Summaries	Total Downstream Spectrum Available	DOCSIS QAM Usable Data Rate Per MHz 256 QAM)	DOCSIS OFDM Usable Data Rate Per MHz (Assuming 1024 QAM OFDM w/ LDPC)	Total Capacity Data Rate Usable (Mbps)
Downstream Capacity with Sub-split (5-42 MHz)	750 MHz (DOCSIS QAM) with Sub-split	696	6.328	7.313	4404
	750 MHz DOCSIS OFDM OFDM w/ LDPC with Sub-split	696	6.328	7.313	5090
	860 MHz (DOCSIS QAM) with Sub-split	806	6.328	7.313	5100
	860 MHz DOCSIS OFDM OFDM w/ LDPC with Sub-split	806	6.328	7.313	5894
	1002 MHz (DOCSIS QAM) with Sub-split	948	6.328	7.313	5999
	1002 MHz DOCSIS OFDM OFDM w/ LDPC with Sub-split	948	6.328	7.313	6933
Downstream Capacity with Mid-split	1002 MHz (DOCSIS QAM) with Mid-split	897	6.328	7.313	5676
	1002 MHz DOCSIS OFDM OFDM w/ LDPC with Mid-split	897	6.328	7.313	6560
Downstream Capacity with High-Split (238)	1050 MHz (DOCSIS QAM) with High-Split (238)	750	6.328	7.313	4746
	1050 MHz DOCSIS OFDM OFDM w/ LDPC with High-Split (238)	750	6.328	7.313	5485
	1300 MHz (DOCSIS QAM) with High-Split (238)	1000	6.328	7.313	6328
	1300 MHz DOCSIS OFDM OFDM w/ LDPC with High-Split (238)	1000	6.328	7.313	7313
Downstream Capacity with Top-split (900-1125)	750 MHz (DOCSIS QAM) with Top-split (900-1050)	696	6.328	7.313	4404
	750 MHz DOCSIS OFDM OFDM w/ LDPC with Top-split (900-1050)	696	6.328	7.313	5090
Downstream Capacity with Top-split (1250-1750)	1002 MHz (DOCSIS QAM) with Top-split (1250-1750)	948	6.328	7.313	5999
	1002 MHz DOCSIS OFDM OFDM w/ LDPC with Top-split (1250-1750)	948	6.328	7.313	6933

- Sub-split and/or Mid-split channel bonding spectrum was counted in capacity summaries with any new spectrum split (Figure 54 does illustrate Top-split spectrum options and the capacity. Note that Sub-split and Mid-split are add to these options)
- Included in the analysis are PHY layer efficiency estimates as well as MAC layer efficiency estimates. This will be labeled in each model



Sub-split Upstream Assumptions:	
37	Sub-split Upstream (5-42 MHz)
-2	Assumed 2 MHz at the roll off (40-42 MHz) is not usable
-5	Assumed 5 MHz to 10 MHz not usable
-2	Set aside Legacy STBs
-2	Set aside Legacy Status Monitoring
-3.2	Assume 3.2 MHz Channel for DOCSIS Legacy using QAM16
22.8	Possible Spectrum for Upstream Channel Bonding
22.4	MHz assumed for upstream DOCSIS Single Carrier QAM

Figure 53 – Sub-split Assumptions

An important assumption is that the upstream capacity measurements assume that spectrum blocks from the sub-split region and any new spectrum split will all share a common channeling bonding domain. This is essentially assuming that backwards compatibility is part of the upstream capacity projections.

The upstream capacity projections for each split will assume DOCSIS QAM – and if adopted in the future – DOCSIS OFDM based systems will all share the same channel-bonding group. This will allow for previous, current, and future investments made by the cable operator to be applied to a larger and larger bandwidth pipe or overall upstream capacity.

If backwards compatibility were not assumed, the spectrum options would have to allocate spectrum for

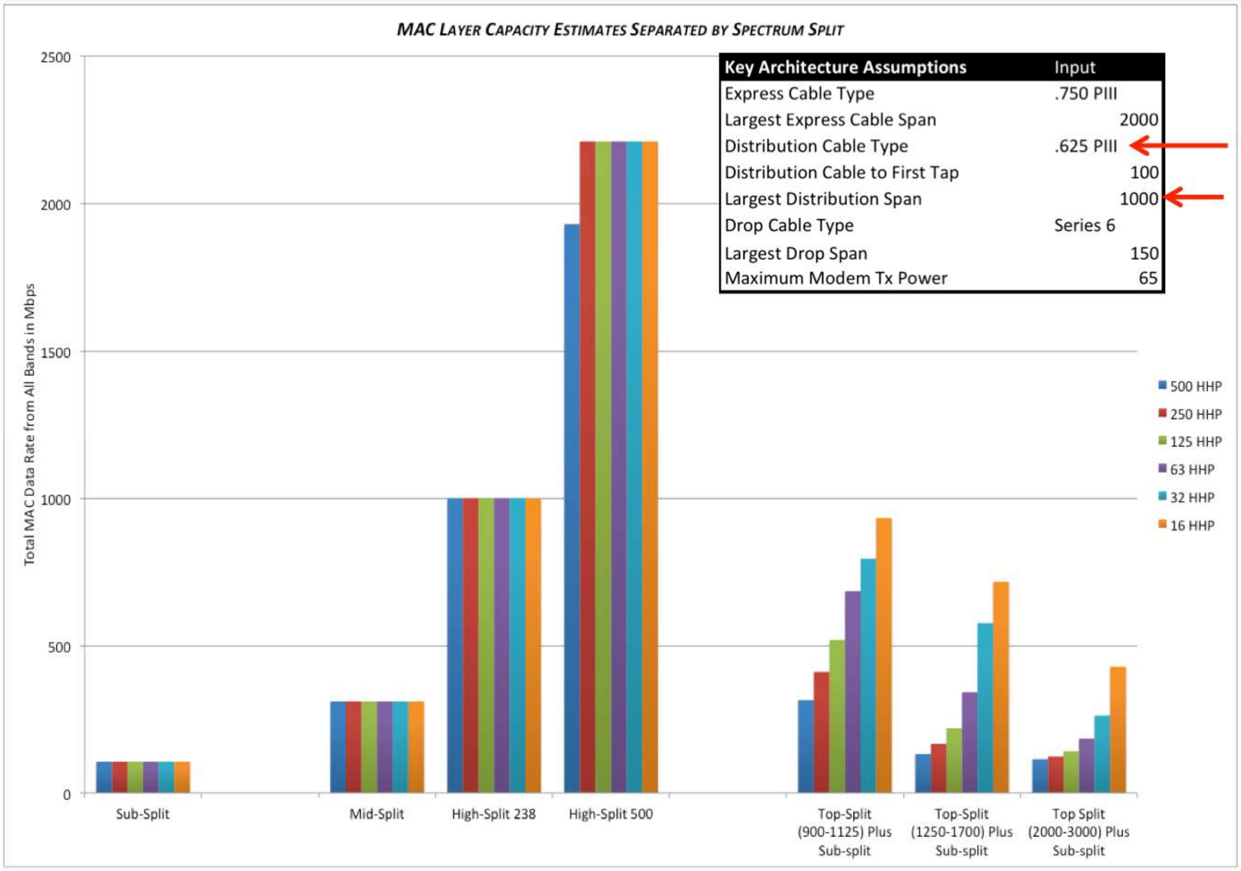
DOCSIS QAM and separate capacity for any successor technology, resulting in a lower capacity throughput for the same spectrum allocation. This would compress the duration of time that the same spectrum may be viable to meet the needs of the MSO.

9.6.1 Achieving 1 Gbps Symmetrical Services and Beyond with DOCSIS 3.0

A major interest of the cable operators is the understanding of the architecture requirements for each spectrum split option to achieve 1 Gbps MAC layer performance. The migration strategy to reach 1 Gbps may be of interest as well, so that an operator can make incremental investment if desired to meet the capacity needs over time, this is sort of a pay as you grow approach.

We have modeled the MAC layer capacity estimates for each node service group size starting at 500 HHP and splitting the service group size in half until reaching 16 HHP, equivalent of fiber to the last active (FTTLA). The model assumes .625 PIII distribution cable with the largest span of 1000 feet in the architecture calculations as shown in Figure 54.

The upstream capacity measurements found in Figure 54 compares various spectrum splits using DOCSIS single carrier QAM with Reed Solomon with a maximum of 256-QAM. The spectrum splits found in the table include Sub-split, Mid-split, High-split (238), High-split (500), Top-split (900-11125) with Sub-split, Top-split (1250-1700) with Sub-split, Top-split (2000-3000)



Key Architecture Assumptions	Input
Express Cable Type	.750 PIII
Largest Express Cable Span	2000
Distribution Cable Type	.625 PIII
Distribution Cable to First Tap	100
Largest Distribution Span	1000
Drop Cable Type	Series 6
Largest Drop Span	150
Maximum Modem Tx Power	65

Figure 54 – Upstream D3.0 MAC Layer Capacity Estimates over Dist. Cable .625 PIII at 1000′

with Sub-split.

The various spectrum splits, along with the overhead contributed from the current DOCSIS PHY, the MAC, the use of SC-QAM and the highest possible modulation type, are examined in Figure 54 to determine the Total MAC Channel Bond Capacity Usable. Traffic engineering and capacity planning should consider the headroom needed for peak periods.

Similar to the examination of the downstream capacity projections above, the upstream use of a new error correction technology such as LDPC will allow high order modulations to be used, thus increasing capacity compared to Reed-Solomon based systems. Higher order modulations will also mean less spectrum required for a desired data rate.

The actual gain for the upstream across an HFC network will need to be determined in the real-world deployments. All upstream capacity is limited to 256-QAM, all though higher order modulation may be possible under certain conditions. Figure 54 through Figure 56 are meant to show the vast difference in capacity and network architecture with upstream spectrum just for having different distribution cable and span of this section of the network. It is this layer of the cable network that is vastly different among MSOs and even within MSOs.

Figure 55 represents cable rebuilds or new builds after the year 2005. Figure 56 represents the Mid 1990s – 2004 Rebuild. Again maximum 256-QAM limitations are assumed as well other assumptions defined in the paper.

A ***major*** finding is that Top-split

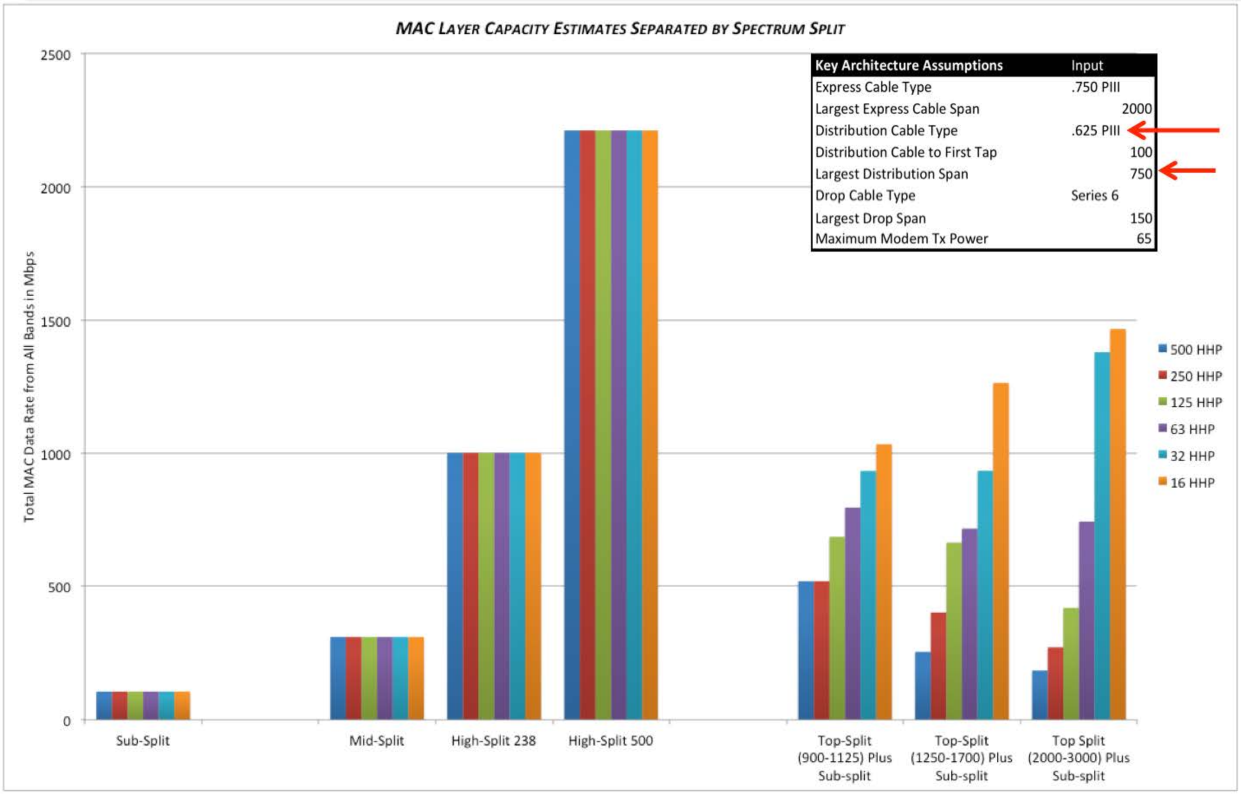


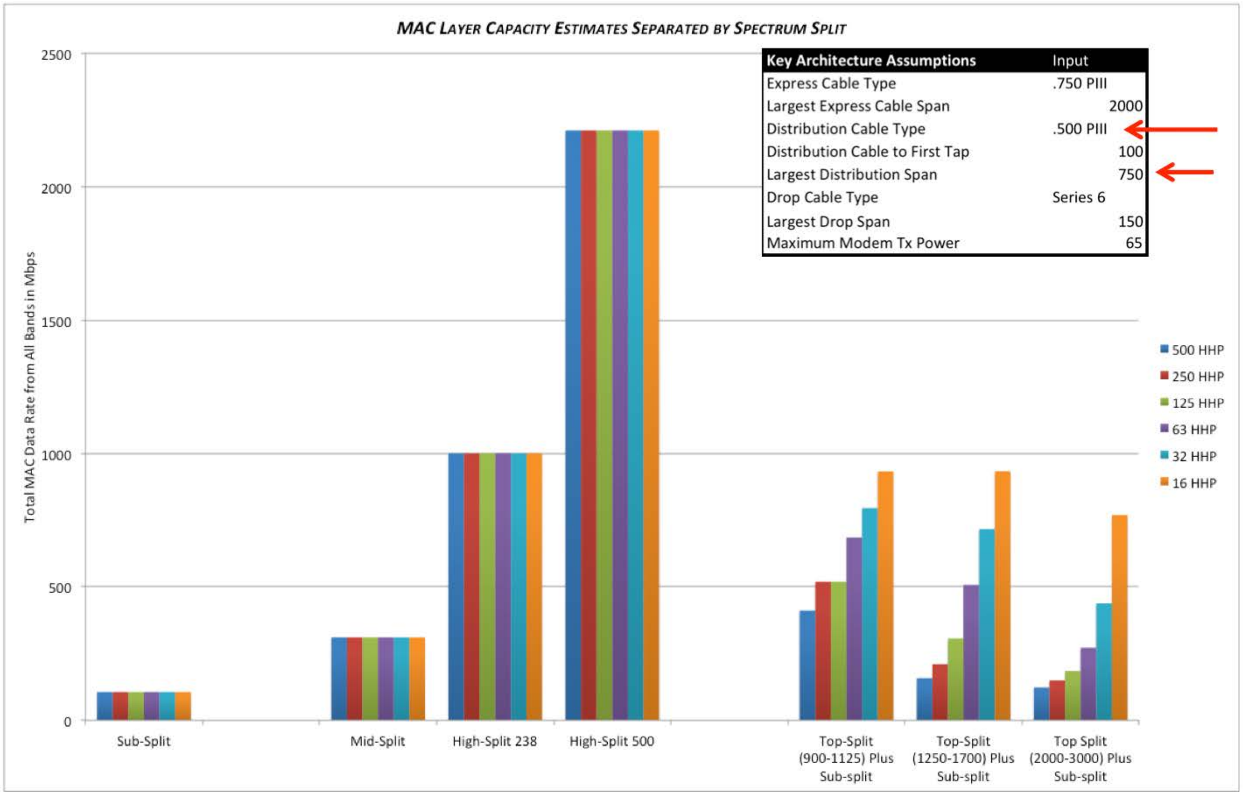
Figure 55 – Upstream D3.0 MAC Layer Capacity Estimates over Dist. Cable .625 PIII at 750'

options require Fiber to the Last Active (~16 HHP) and the placement of a node at each location to maximize the spectrum capacity. However, all Top-split options even if combined with the existing Sub-split will not reach the capacity any of the High-split option. If these two Top-split options are not combined with Sub and Mid-split achieving 1 Gbps MAC Layer performance is not possible, given the assumptions described in this analysis, .625 PIII at 1000 foot spans to last tap and other assumptions.

Another ***major*** finding is that even, given the assumption of the widely deployed cable architecture using .500 PIII distribution cable with 750 foot spans to the last tap, none of the Top-split with Sub-split reaches 1 Gbps with current DOCSIS PHY as shown in Figure 56. Only Top-split with .625 PIII at 750 foot spans to last tap will meet or exceed the 1 Gbps capacity.

Another very important point is that the network architecture and performance characteristics of the plant in the real world will determine the spectrum capacity to be used. The determination of the network architectures that may work at various spectrum splits, modulations, and number of carriers in different cable types and distance to the subscriber was a critical finding.

We have modeled the network architecture and performance assumptions to estimate the modulation and capacity possible for each spectrum split. This allowed us to determine the overall requirements and impacts to cost of the various split options and the ability for the spectrum split to meet the business needs of the MSO.



Key Architecture Assumptions	Input
Express Cable Type	.750 PIII
Largest Express Cable Span	2000
Distribution Cable Type	.500 PIII
Distribution Cable to First Tap	100
Largest Distribution Span	750
Drop Cable Type	Series 6
Largest Drop Span	150
Maximum Modem Tx Power	65

Figure 56 – Upstream D3.0 MAC Layer Capacity Estimates over Dist Cable .500 PIII at 750′

9.6.2 DOCSIS NG Network Capacity Estimates Upstream

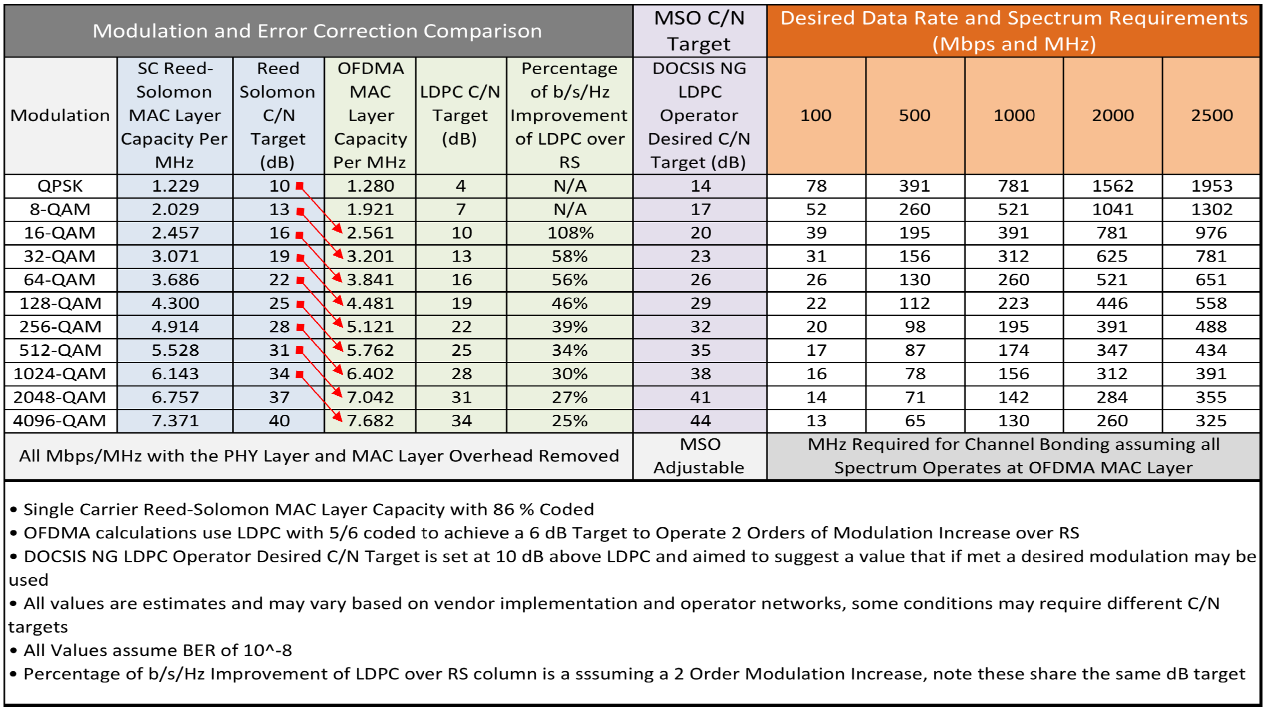
We have modeled the network architecture using several HFC coaxial network topologies using DOCSIS 3.0, however in this section DOCSIS NG will be compared. This section will provide a summary of the key methods and measurements to estimate sizing for DOCSIS NG.

The adoption of higher modulation formats in DOCSIS NG will increase b/s/Hz. A key finding is the use of DOCSIS 3.0 Single Carrier Reed Solomon verse OFDM using LDPC may allow two (2) orders of modulation increase. In Figure 57, DOCSIS 3.0 verse DOCSIS NG Modulation C/N and Capacity Estimates this summarize the major benefits of moving to DOCSIS NG.

Figure 57 illustrates that the use of Reed Solomon and LDPC with different code rates will have different b/s/Hz using the same modulation format. The major takeaway from the table is the use of a stronger error correction code will allow LDPC to operate in the same carrier to noise environment as Reed Solomon but LDPC may use two orders of modulation higher.

The table uses red arrows to illustrate the corresponding Reed Solomon modulation and C/N to the OFDMA LDPC modulation format, which shares the same C/N dB. The table will show that in the same modulation format Reed Solomon will have more b/s/Hz than LDPC and this is due to a higher code rate percentage applied to LDPC. The percentage of gain is measured using the SC Reed Solomon data rate for a given modulation and the used of two order of modulation increase using LDPC.

For example, in the table SC Reed Solomon b/s/Hz of QPSK is measured against OFDMA LDPC using 16-QAM, the percentage of gain in b/s/Hz 89%. As



Modulation and Error Correction Comparison						MSO C/N Target	Desired Data Rate and Spectrum Requirements (Mbps and MHz)				
Modulation	SC Reed-Solomon MAC Layer Capacity Per MHz	Reed Solomon C/N Target (dB)	OFDMA MAC Layer Capacity Per MHz	LDPC C/N Target (dB)	Percentage of b/s/Hz Improvement of LDPC over RS	DOCSIS NG LDPC Operator Desired C/N Target (dB)	100	500	1000	2000	2500
QPSK	1.229	10	1.280	4	N/A	14	78	391	781	1562	1953
8-QAM	2.029	13	1.921	7	N/A	17	52	260	521	1041	1302
16-QAM	2.457	16	2.561	10	108%	20	39	195	391	781	976
32-QAM	3.071	19	3.201	13	58%	23	31	156	312	625	781
64-QAM	3.686	22	3.841	16	56%	26	26	130	260	521	651
128-QAM	4.300	25	4.481	19	46%	29	22	112	223	446	558
256-QAM	4.914	28	5.121	22	39%	32	20	98	195	391	488
512-QAM	5.528	31	5.762	25	34%	35	17	87	174	347	434
1024-QAM	6.143	34	6.402	28	30%	38	16	78	156	312	391
2048-QAM	6.757	37	7.042	31	27%	41	14	71	142	284	355
4096-QAM	7.371	40	7.682	34	25%	44	13	65	130	260	325
All Mbps/MHz with the PHY Layer and MAC Layer Overhead Removed						MSO Adjustable	MHz Required for Channel Bonding assuming all Spectrum Operates at OFDMA MAC Layer				

- Single Carrier Reed-Solomon MAC Layer Capacity with 86 % Coded
- OFDMA calculations use LDPC with 5/6 coded to achieve a 6 dB Target to Operate 2 Orders of Modulation Increase over RS
- DOCSIS NG LDPC Operator Desired C/N Target is set at 10 dB above LDPC and aimed to suggest a value that if met a desired modulation may be used
- All values are estimates and may vary based on vendor implementation and operator networks, some conditions may require different C/N targets
- All Values assume BER of 10^-8
- Percentage of b/s/Hz Improvement of LDPC over RS column is a sssuming a 2 Order Modulation Increase, note these share the same dB target

Figure 57 – DOCSIS 3.0 versus DOCSIS NG Modulation C/N and Capacity Estimates

expected the percentage of gain will decrease as modulation increases, for example moving from 256-QAM to 1024-QAM is a smaller gain, than moving than the doubling of QPSK to 16-QAM.

The table estimates the use of OFDMA and the MAC layer bit rate in a given modulation as explained in the paper. The table calculated several desired MAC layer throughput capacities from 100 Mbps, 500 Mbps, 1,000 Mbps, 2,000 Mbps, and 2,500 Mbps and using the OFDMA estimated MAC layer data rate a required spectrum calculation and corresponding modulation format are aligned.

Table 36 – DOCSIS NG Modulation and C/N Performance Targets

Modulation Type	Uncoded Theoretical C/N dB	LDPC 5/6 Coded C/N dB	Operator Margin is Desired C/N Target
QPSK	16	4	14
8-QAM	19	7	17
16-QAM	22	10	20
32-QAM	25	13	23
64-QAM	28	16	26
128-QAM	31	19	29
256-QAM	34	22	32
512-QAM	37	25	35
1024-QAM	40	28	38
2048-QAM	43	31	41
4096-QAM	46	34	44

Theoretical SNRs Uncoded with BER of 10^-8
Practical C/N is chosen to give 10 dB headroom
Operator Margin above LDPC 5/6 coded

The MSO may require less upstream spectrum if a high modulation format may be used. The table illustrates a proposed Operator Desired C/N target for each Modulation format using LDPC, please note that the higher the modulation form the higher the C/N requirements but the lower percentage of gain in b/s/Hz.

In the past, our industry may have used The “Operating Margin” (OM) or Operator Desired carrier to noise target to be 6 dB above the theoretical uncoded C/N for a given BER, usually between 10E-6 or 10E-8, without any Forward Error Correction (FEC). The 6 dB of margin typically assumed a 500 HHP case; that is, for “Node +5” (or so), involving up to 30 return path RF amplifiers.

In the future perhaps we need to change the method by which we estimates the “Operating Margin” (OM) and perhaps we need to estimate the operating margin from the coded rate used for a given system and then add the Operating Margin, for the analysis below we used 10 dB above the LDPC dB value.

About the “Operating Margin” (OM) parameter, this is a variable (in dB) to account for the performance changes in the HFC return path system due to temperature variation and setup accuracy of the outside plant. This mainly involves RF level changes due to hardline and drop cable loss changes, Tap loss change, and RF Amplifier/Node Return RF drive path (Hybrid) gain changes, and Node passive loss changes with temperature. It also includes setup level tolerances (due to RF Testpoint accuracy and flatness over frequency) and laser optical power output changes over temperature.

Some of these changes are small or only occur in one place, while others are more significant as they occur at many places and in cascade (e.g., cable segments, RF Amplifiers, and Taps). With many amplifiers in a 500 HHP distribution sector (up to 30 for Node +5 sector), the number of cascaded Amplifiers is typically a maximum of 6. There typically will be 6 or more Taps used between each amplifier, so these elements contribute significantly.

About 2/3 of the 6 dB OM assumed in the calculation matrix is due to the cable part of the plant. The other 2 dB is due to the "optics" part; mainly for the Return laser. The laser is assumed a high quality uncooled CWDM analog laser, with 2 mW or higher optical output. The OM is added to the "Theoretical C/N" at 10E-6 BER (without encoding) to obtain a "Desired C/N" for determining the highest order modulation type allowed.

In the model that will estimate the use of DOCSIS NG and LDPC, we will use a 10 dB Operating Margin, on top of the coded value, please see Table 36 for the allocation.

In order to estimate the capacity of the different spectrum splits using DOCSIS NG we placed the values of the Operator Margin desired C/N target and the b/s/Hz estimates for DOCSIS NG. The model estimates the system C/N and in this case the model used .500 PIII distribution cable at 750 feet.

Please note the that model estimates that very high modulation format may be used in a 500 HHP node for the low frequency return while the Top-split spectrum selection is only capable of using substantially lower order modulation formats.

As seen in Table 37, 2048 QAM and 1024 QAM are possible in the upstream in a 500 HHP node with assumption defined in this table. This is an illustration of the modern DOCSIS PHY and the ability to maximize spectrum for the operator.

DOCSIS NG capacity is examined in Figure 58 considering several spectrum-split options. Please note the capacity of Sub-split, Mid-split, and the pair of High-split options. The MSOs may choose any of this spectrum split or others depending on the desired capacity. The estimates assume that the entire spectrum uses the highest modulation rate possible for a given spectrum selection.

9.6.3 DOCSIS 3.0 versus DOCSIS NG Side-by-Side Upstream Capacity Estimate

Table 37 – Upstream DOCSIS NG MAC Layer Capacity Estimates over Distribution Cable .500 PIII at 750 Feet

				High-Split	High-Split	Top-Split	Top-Split	Top Split
						(900-1125)	(1250-1700)	(2000-3000)
DOCSIS NG System Performance Estimates		Sub-Split	Mid-Split	238	500	Plus Sub-split	Plus Sub-split	Plus Sub-split
Upper Frequency	MHz	42	85	238	500	1125	1700	3000
Homes Passed		500	500	500	500	500	500	500
HSD Take Rate		50%	50%	50%	50%	50%	50%	50%
HSD Customers		250	250	250	250	250	250	250
Desired Carrier BW	MHz	6.4	6.4	6.4	6.4	6.4	6.4	6.4
Modulation Type		**2048-QAM**	**2048-QAM**	**1024-QAM**	**1024-QAM**	**8-QAM**	**QPSK**	**QPSK**
Bits/Symbol		11	11	10	10	3	2	2
Number Carriers in Bonding Group		3.5	10.25	33	73	35	22	9
Max Power per Carrier Allowed in Home	dBmV	59.6	54.9	49.8	46.4	49.6	51.6	55.5
Worst Case Path Loss	dB	**29.1**	**30.1**	**33.5**	**41.4**	**65.1**	**73.0**	**76.9**
Maximum Return Amplifier Input	dBmV	30	25	16	5	-16	-21	-21
Actual Return Amplifier Input	dBmV	15	15	15	5	-16	-21	-21
Assumed Noise Figure of Amplifier	dB	7	7	7	7	7	7	7
Return Amplifier C/N (Single Station)	dB	65	65	65	55	35	29	29
Number of Amplifiers in Service Group		30	30	30	30	30	30	30
Return Amplifier C/N (Funneled)	dB	50.4	50.4	50.4	40.4	19.9	14.0	14.0
Optical Return Path Technology		DFB	DFB	DFB	Digital	Digital	Digital	Digital
Assumed Optical C/N	dB	45	45	41	48	48	48	48
System C/N	dB	**43.9**	**43.9**	**40.5**	**39.7**	**19.9**	**14.0**	**14.0**
Desired C/N	dB	41	41	38	38	17	14	14

The paper has examines the downstream and upstream features of DOCSIS NG. The analysis has examined modulation profiles such as using LDPC with increased FEC to obtain a 6 dB gain over Reed Solomon in the same modulation format. Figure 59 examines the low frequency return spectrum options using DOCSIS 3.0 using 64 QAM against DOCSIS NG using the maximum modulation format possible given the assumptions and spectrum selection. Please note the much higher aggregate capacity of the DOCSIS NG system over current DOCSIS.

9.6.4 Summaries for Network Capacity

DOCSIS NG will greatly expand the capacity of the cable network and coupled with backward compatibility utilize spectrum efficiently

Downstream Capacity Expansion

1. DTA's & SDV will provide long term downstream plant capacity expansion
2. Reduced service group size enabling fewer customers to share bandwidth
3. Node segmentation and node splits will continue to be used in a targeted basis
4. Use of highest order modulation and channel bonding to increase throughput
5. Consider DOCSIS NG changes with modern error correction technology that allow the modulation rate to increased, given the same SNR, perhaps as much as two orders. For example, 256-QAM could be increased to 1024-QAM
6. Possible downstream bandwidth expansion along with upstream augmentation

Upstream Capacity Expansion

1. Use of highest order modulation and Channel Bonding to increase throughput

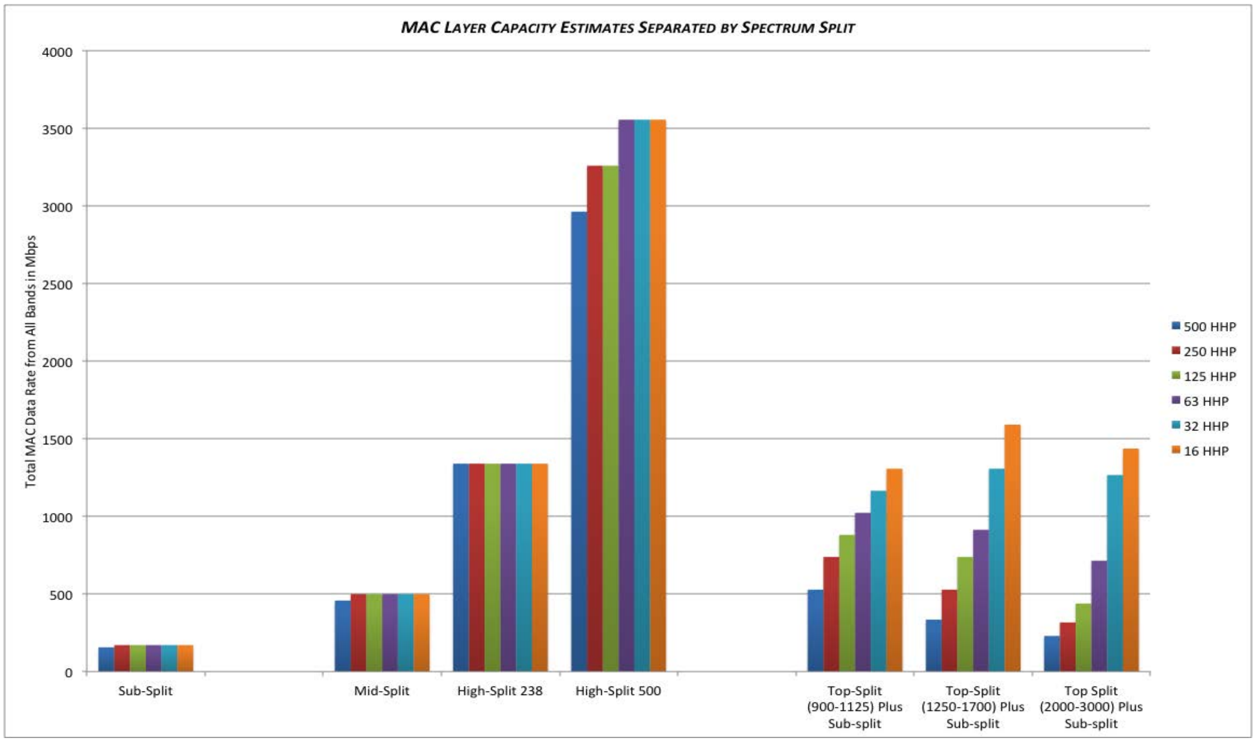


Figure 58 – Upstream DOCSIS NG MAC Layer Capacity Estimates over Dist Cable .500 PIII at 750'

2. Consider DOCSIS NG changes with modern error correction technology that allow the modulation rate to be increased, given the same SNR, perhaps as much as two orders. For example, 64-QAM to 256-QAM and perhaps 256-QAM to 1024-QAM

3. Progressively smaller upstream service groups
4. Ongoing node splits / segmentation
5. These incremental steps should last for a majority of the decade

Upstream augmentation expands upstream spectrum and bandwidth such as conversion to mid-split, high-split, or top-split options.

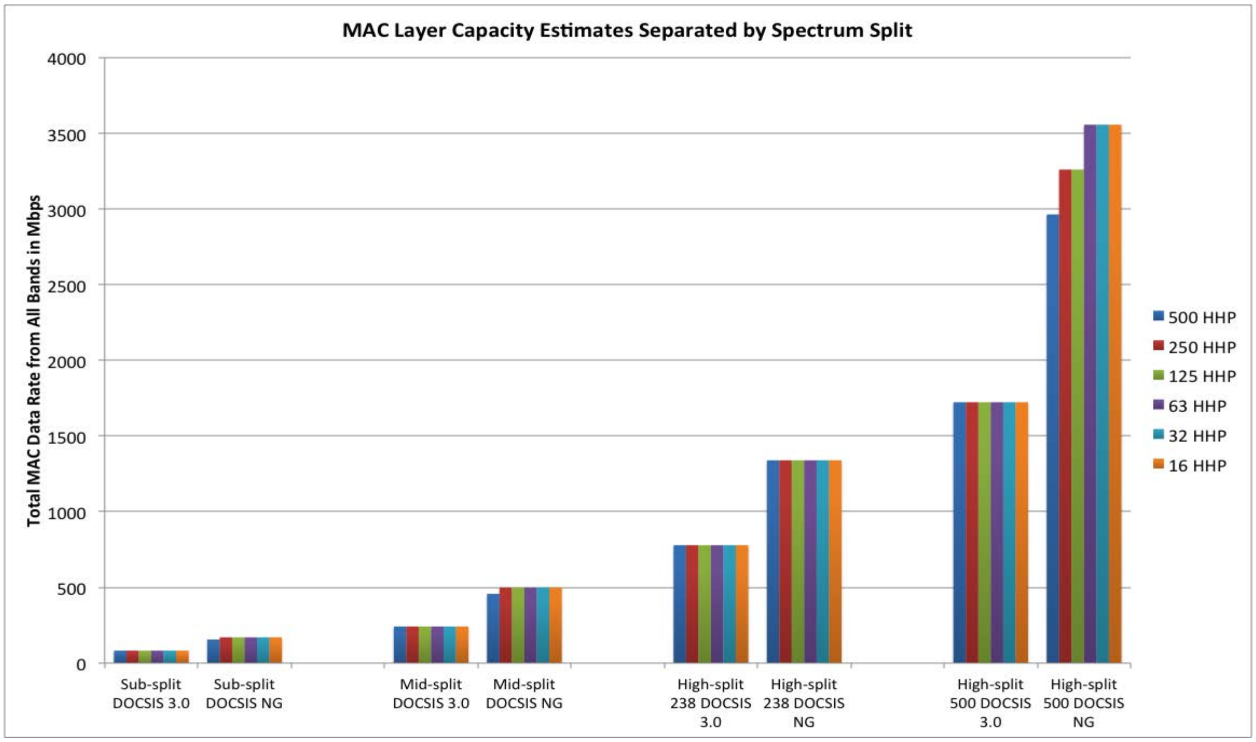


Figure 59 – DOCSIS 3.0 verse DOCSIS NG

10 NETWORK CAPACITY PROJECTION AND MIGRATION STRATEGIES

10.1 Upstream Migration Strategy

10.1.1 Phase 0: Sub-Split and Business as Usual

10.1.1.1 Sub-split Legacy Return Lifespan

Let's put our understanding of upstream data capacities to work in evaluating time-based migration strategies for the HFC upstream. Note that not every capacity number calculated in the paper to this point is represented on a chart in this section. We expect that the reader may have to extrapolate between displayed values in some case to draw conclusions from curves shown for some cases not explicitly plotted.

We introduced a version of an upstream lifespan analysis in Figure 2 of Section 2.6. A more traditional version is shown in Figure 60. Traffic models based on a compound annual growth (CAGR) methodology have been shown to represent historical traffic trends well. However, because of short-term fluctuations, particularly in the upstream, there is a need to engineer ahead of the curve to avoid being unprepared in the case of an unexpected step function in growth (a "Napster" moment).

We will use CAGR analysis such as this and Figure 2 as a guideline to understand the most fundamental of drivers for upstream evolution – the need to find more capacity,

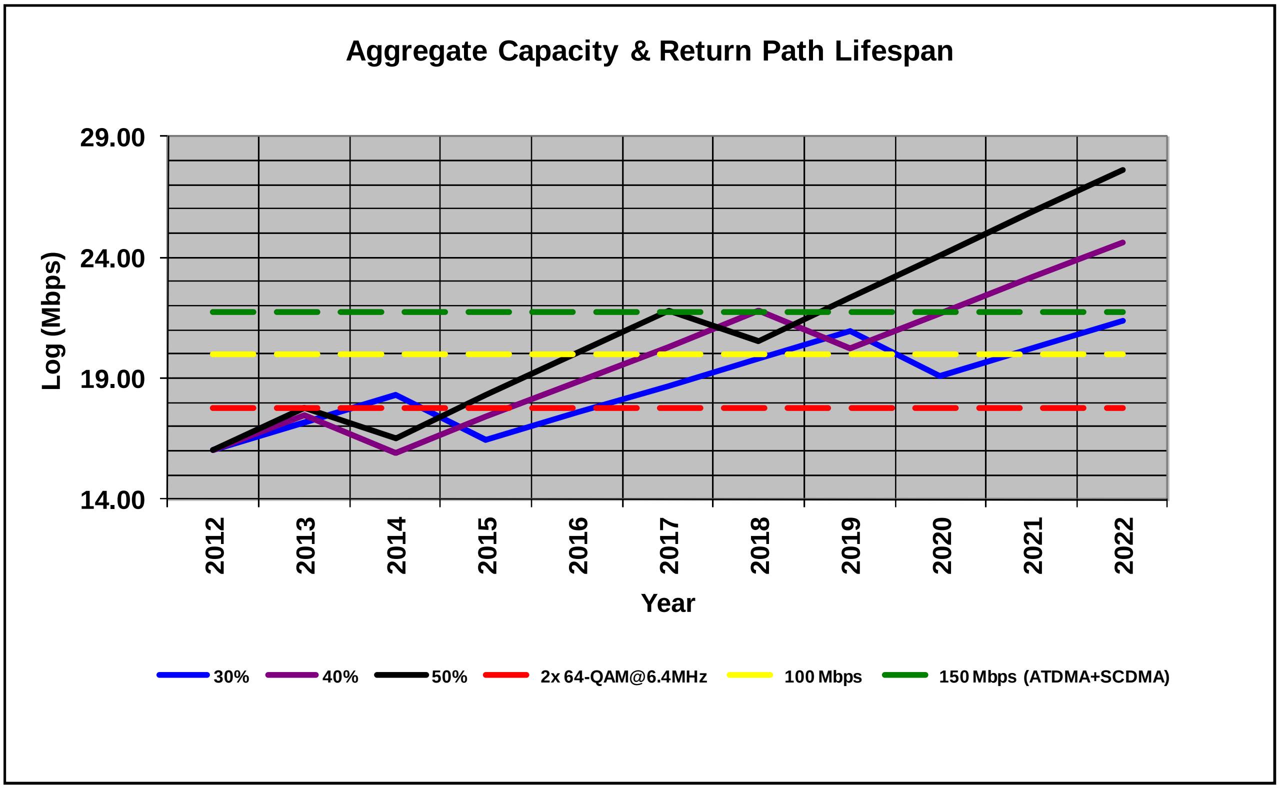


Figure 60 – Upstream CAGR vs. Available Capacity

coupled with a need to deliver competitive service rates, so that the upstream achieves a long and healthy lifespan.

Figure 60 shows this a CAGR approach for the upstream using three different assumptions – 30%, 40% and 50%. The three trajectories, representing a single aggregate service group, are interrupted by two breakpoints over the next ten years.

These represent node and/or service group splits – 3 dB (best case) offsets, or a doubling of average bandwidth per home. Note that the 3 dB is a step straight downward by 3 dB at implementation, so that by the time the next year comes around, some of that has been consumed.

These trajectories are plotted against three different HFC upstream capacity thresholds, using raw physical layer transport rate for simplicity and to remove the ambiguity around overhead of different configurations, packet sizes, and net throughputs. We will use raw transport rate for trajectories and thresholds throughout to simplify apples-to-apples comparisons.

- 60 Mbps – Approximately two 64-QAM DOCSIS channels at 5.2 Msps
- 100 Mbps – Approximate available bit rate in 5-42 MHz with only A-TDMA
- 150 Mbps – Approximately a fully utilized 5-42 MHz using both A-TDMA and S-CDMA

Using these, we can now estimate when various CAGRs exhaust the available upstream. Let’s assume 40 Mbps of upstream consumption at peak busy hour – 50% of 80 Mbps of deployed capacity, for example (2x 64-QAM + 16-QAM, all at 6.4 MHz).

Some key conclusions can be drawn from Figure 60. Clearly, a couple of 64-QAM DOCSIS channels get exhausted within a few years without a service group split. While node splits are costly and intrusive, they are well-understood business-as-usual (BAU) activities.

Most important to craft an evolution strategy is to estimate when 5-42 MHz itself gets exhausted, and when a more significant change must be considered. Referring again to Figure 60, note that a single split supports 4-6 years of growth considering 100 Mbps as the 5-42 MHz throughput boundary.

While further node splitting will provide more average bandwidth, the maximum service rate limit also come into play, where 100 Mbps upstream service rates require more total capacity to be achieved. Aside from merely keeping pace with upstream service rate growth, the service rate upstream should be somewhat aligned with 1 Gbps downstream rates from a timing perspective.

Finally, note that with S-CDMA the upstream could last through the decade for a very robust CAGR (40%).

Figure 60 is a useful guide for visualizing growth versus time. In Figure 61, as in Figure 2, we have displayed the same information differently, allowing us to understand the sensitivity of the exhaustion of the 5-42 MHz return path relative to the CAGR assumptions. Note that service group splits are instead represented by dashed traces for the 100 Mbps and 150 Mbps cases.

The three crosshairs on Figure 61 are positioned to help interpret between Figure 60 and Figure 61. For example, note the point at which a 50% CAGR exhausts a 150 Mbps maximum throughput threshold after one split in Figure 60. This occurs 5 years into the future. We can see this same point represented by the leftmost crosshair in Figure 61. Similarly, we can correlate between the crosshairs at 40% and 30% CAGR on Figure 61 and the corresponding breach of threshold in Figure 60.

We will use the format of Figure 61 in subsequent discussion because of the granularity and clarity it brings in an environment where CAGR tends to have more variation. This variation of CAGR points out why, for network planning decisions, upstream CAGR needs to be considered in the context of an average, long-term CAGR, rather than based on very high or very low periods of growth.

This is particularly true upstream, where there is not a set of knobs and levers at the operator's disposal to manage a spectrum congestion issue as there is in the downstream. In the downstream, while CAGR is consistent and generally higher, but there is more control over service delivery choices to manage spectrum. In the upstream, there is a hard bandwidth cap at 42 MHz in North America, for example, little control over the growth of Internet usage, and limited ability or authority to more actively manage traffic by type. As such, there are not any "easy" answers to creating more upstream capacity in the 42 MHz spectrum.

One area where there is some room to grow is in the low end of the return. A key problem for A-TDMA is its ability to operate efficiently or at all in this region. Some 30-40% of the 5 to 42 MHz return band is polluted by a combination of impulse noise emanating from homes and often times various narrowband interferes managing to get onto the cable in the short wave band.

However, it is the impulse noise that

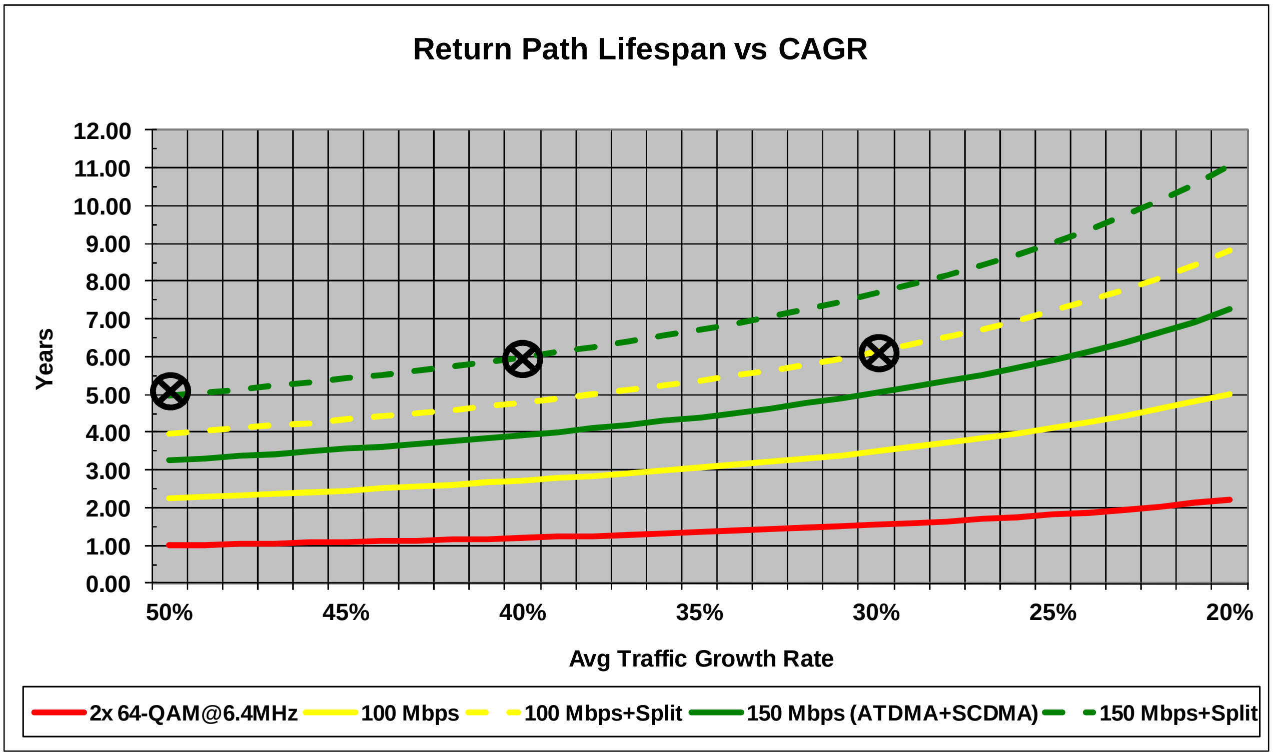


Figure 61 – Lifespan of 5-42 MHz vs CAGR

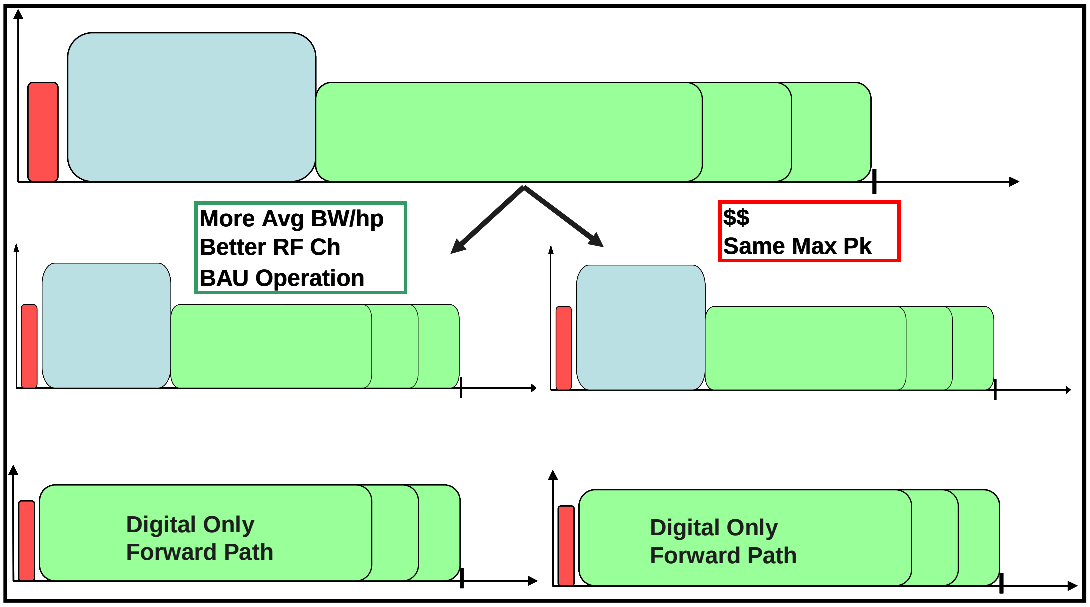


Figure 62 – Serving Group Segmentation

gives A-TDMA the most difficulty, even with powerful Reed-Solomon burst correction employed. To combat this, DOCSIS 2.0 introduced S-CDMA to the standard. By enabling use of the lower portion of the upstream spectrum, the total 5-42 MHz band improves in its total capacity by almost 50%, to about 150 Mbps. We will discussed S-CDMA in Section 7.2, and will use some of the results observed to add to the available capacity in 5-42 MHz to calculate the lifespan of a fully optimized 5-42 MHz.

10.1.1.2 Legacy Relief: Business-As-Usual Node Splitting

The classically deployed tool for improving average bandwidth per user is service group or node splitting. However, this does not enable service rate increases, and splitting nodes in the field runs into diminishing return because of the unbalanced nature of physical architectures.

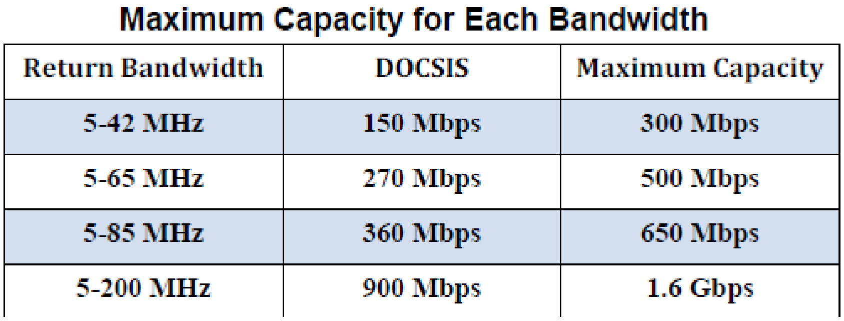
We observed in Figure 60 and Figure 61 how this lead to a longer lifespan for 5-42 MHz by simply sharing the fixed bandwidth among fewer users. The average bandwidth per user, often a good reflection of user QoE, will increase.

The most natural HFC methods to decreasing the service group size are the removal of combiners at the output of the return optical receivers that combine upstreams into a single port, or the splitting of nodes, either through a segmentable node or pulling fiber deeper.

Figure 62 illustrates this approach from a spectral allocation perspective, identifying also the pros and cons commonly associated with this well-understood tool.

The increased BW/user is an obvious benefit. Another key benefit of this straightforward approach is that, while heavy touch, it is a well-understood “business as usual” operation. In addition, reducing the serving group size can improve the RF channel in two ways.

Table 38 – Bandwidth, DOCSIS, and Theory @25 dB SNR



Maximum Capacity for Each Bandwidth

Return Bandwidth	DOCSIS	Maximum Capacity
5-42 MHz	150 Mbps	300 Mbps
5-65 MHz	270 Mbps	500 Mbps
5-85 MHz	360 Mbps	650 Mbps
5-200 MHz	900 Mbps	1.6 Gbps

First, fewer users means a lesser probability of interference and impulse from a troublesome subscriber. While the troublemaker has not gone away, he is now only inflicting his pain on half the number of users. Second, from a system engineering standpoint, the same funneling reduction that increases the probability of not having a troublemaker also reduces any amplifier noise aggregation effect, noticeable when deep RF cascades combine in multiport nodes, for example. All of this can lead to more efficient use of the existing spectrum than had existed prior to the split.

The primary performance disadvantage of only a segmentation strategy is that 5 to 42 MHz ultimately limits the maximum total bandwidth to around 100 Mbps. Under good conditions, a single 100 Mbps serving group may be all that can be obtained in an A-TDMA only system.

This limits the flexibility of this architecture to provide other services, such as mid-size business service tiers, and to support Nielsen's Law-based peak rate growth. And, peak rate offerings generally are topped out at some scale factor of the total available capacity for practical reasons.

Note that in Figure 62 we have added the "digital only" forward example. As we consume forward band for return applications, techniques that make more efficient use of the forward path also draw more focus. Digital only carriage (DTA deployments) is one of the key tools for extracting more from the downstream as upstream imposes on it, and for adding flexibility to the diplex split used in the architecture.

10.1.1.3 Delivering New DOCSIS Capacity

Because of the known limitations of return spectrum, the expectation that traffic growth in the upstream will continue to compound, and the anticipation that peak service rates will do the same, options to find new capacity are required.

There is consensus that new spectrum must eventually be mined for upstream use. The questions that remain are where do we find it and how much do we need. And, of course, at the core of the discussion, how much new capacity, for how long, and what are the practical implications of implementing such a change.

We will focus on the recommended evolution approach whereby cable maintains a diplex-only architecture for optimum bandwidth efficiency. We view a migration that has as a primary objective the most efficient long-term use of the cable spectrum to ensure the longest lifespan of the architecture, and preferably with the simplicity of implementation that cable enjoys today.

A diplex architecture achieves this. We view the selection of the actual frequency split as something that evolves with time, in an efficient way, and based on the traffic mix and projected services.

We note that it is possible that extracting the most bandwidth efficiency with flexibility theoretically involves a TDD implementation. However, the obstacles in place to enable TDD in the HFC environment are so great and will be so for so long, that it does not appear to be a sensible plan for typical HFC architectures.

However, with the very long observation window enabled by fiber deep migration and the recommendations made herein, it may at some point become a more practical consideration for cable if the need for increased flexibility of traffic allocation justifies the increase in complexity.

Table 38 illustrates the available DOCSIS transport rate for various low diplex-based frequency split architectures, and the theoretically available channel capacity at the DOCSIS-specified minimum of 25 dB.

While it is impractical to achieve theoretical capacity, the gap has indeed closed over time between practice and theory. This not a negative reflection on DOCSIS 1.0, only a reflection that its PHY basis is 15 years old – a very long time in technology evolution, and a period of extensive advances in communications theory and practice. For DOCSIS NG, we have already introduced the fact that a new FEC added to the PHY mix will enable a major step closer to capacity by enabling higher order profiles over the same SNR.

One simple conclusion of Table 38 is simply the power of the Shannon-defined proportional relationship between capacity and bandwidth for a fixed SNR. Indeed, for high SNR assumptions, capacity is directly proportional to both bandwidth available and SNR expressed in dB – the assumption being very relevant to the cable architecture. This leads to the inescapable conclusion that when discussing new actual upstream capacity, it is first about architecture and bandwidth, and not waveform.

As previously introduced, a straightforward and surprisingly powerful way to exploit new bandwidth and remain compatible with DOCSIS is use of the 85 MHz Mid-Split.

This band edge was wisely chosen to maximize clean low band return without overlapping the FM radio band and potential harmful effects of proximity to that band. Its advantages are numerous. First, however, let’s understand what new spectrum means in terms of that fundamental upstream problem – lifespan – that has us so concerned in the first place.

10.1.2 Phase 1: Deploy 85 MHz Mid-Split

10.1.2.1 Capacity and Lifespan

It was shown in Figure 2 how the 85 MHz Mid-Split delivers long-term new capacity to the HFC upstream. Consider **Error! Reference source not found.**, which adds the Mid-Split case to cases observed in Figure 61 for 42 MHz. The gap between the set of 5-42 MHz options and the maximized Mid-Split is readily apparent at 3.5-5.5 years at 30% CAGR, depending on whether S-CDMA is utilized or not.

The transition to Mid-Split pushes the lifespan of the return path to nearly a decade under a 256-QAM maximum assumption – a very comfortable chunk of next generation network planning time. This lifespan time frame is pushed beyond a decade for CAGRs of 35% and below if the Mid-Split is combined with one service group split, as shown in **Error! Reference source not found.**.

Though not apparent in an upstream analysis, it is straightforward to show that a ten-year lifecycle of growth aligns the upstream with what is also achievable in the downstream under similar assumptions about plant segmentation. Aligning these two in terms of physical plant segmentation has operational benefits.

Because of this result observed in **Error! Reference source not found.**, when combined with a service group split, Mid-Split (440 Mbps), in fact, represents a *long-term* solution, not merely an incremental one.

This is a very important, fundamental conclusion to recognize about the 85 MHz Mid-Split architecture, that is often not fully understood. The amount of lifespan afforded by 85 MHz with just a single split is nearly a decade – a technology eternity. If today's observed, low, CAGRs persist, it is even

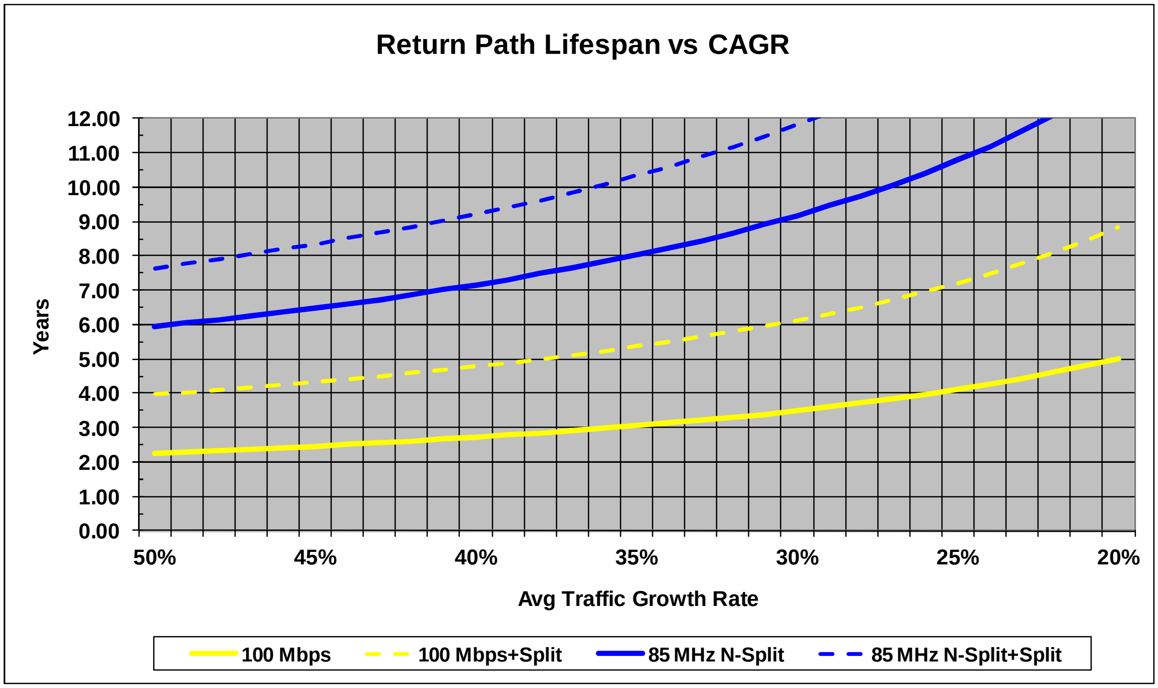


Figure 63 – 85 MHz Mid-Split vs. 42 MHz A-TDMA-Only with Segmentation

longer, and longer still if we assume that modulation profiles extend beyond the 256-QAM examples used for the Mid-Split analysis here. For example, 25% is a three year doubling period, so it offers 50% more lifespan than 40%. Similarly, 1024-QAM, which may become available with LDPC FEC, offers 25% more data capacity, pushing 400 Mbps of 85 MHz throughput to 500 Mbps available for growth.

The window of time to observe trends in traffic, applications, services, and technology, coupled with the runway for managing down legacy in an all-IP transition, is a very meaningful strategy component considering the low risk associated with implementation.

Even under an acceleration of CAGR, the architecture supports 100 Mbps services and an attractive long-term lifespan. A common traffic engineering assumption is to evaluate an increased CAGR resulting from the exploding number of devices looking for access to the upstream, using similar models for average application bandwidth of the access. The net effect for equivalent QoE is the potential requirement to adjust the oversubscription model.

In **Error! Reference source not found.**, we adjust this traffic engineering parameter by a factor of two to account for the increasing number of simultaneous users (devices) looking to access the upstream. Despite this acceleration, the Mid-Split architecture still achieves a decade of lifespan under two segmentations for common CAGR ranges.

Considering that a downstream CAGR analysis typically requires two splits over this same time period, there is the added opportunity to take advantage of this added lifespan to the upstream as well if necessary.

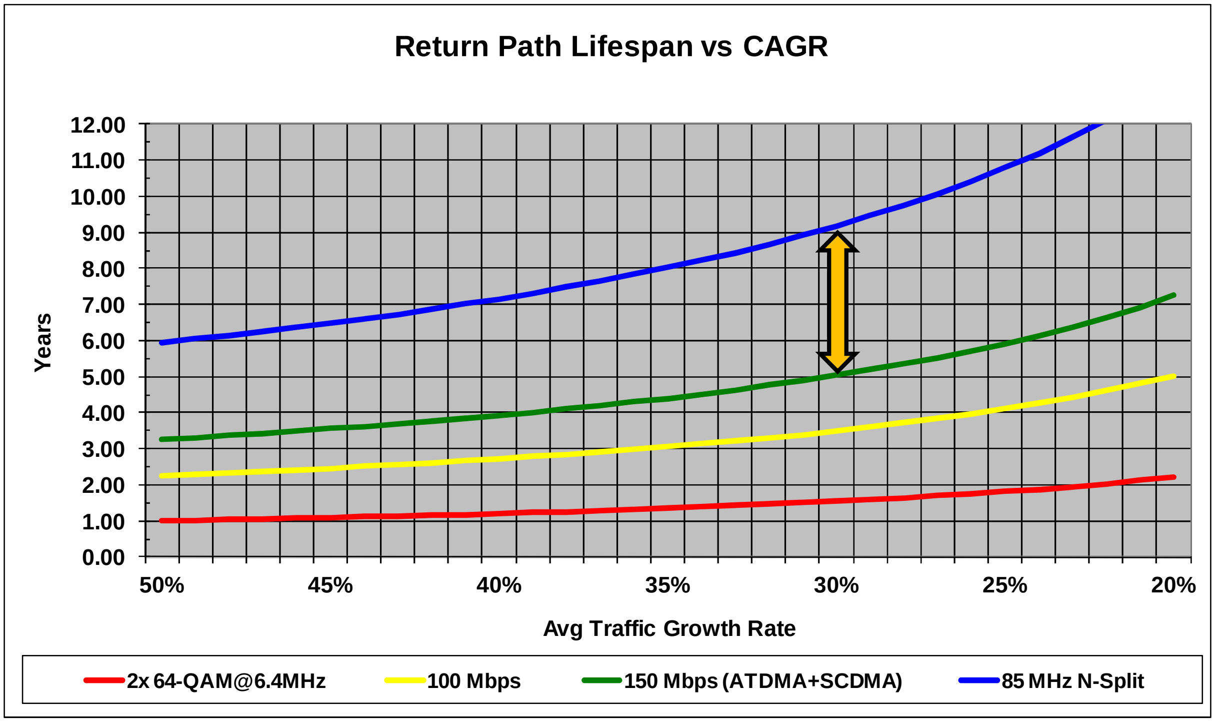


Figure 64 – 85 MHz Mid-Split Years of Growth vs. 5-42 MHz Use

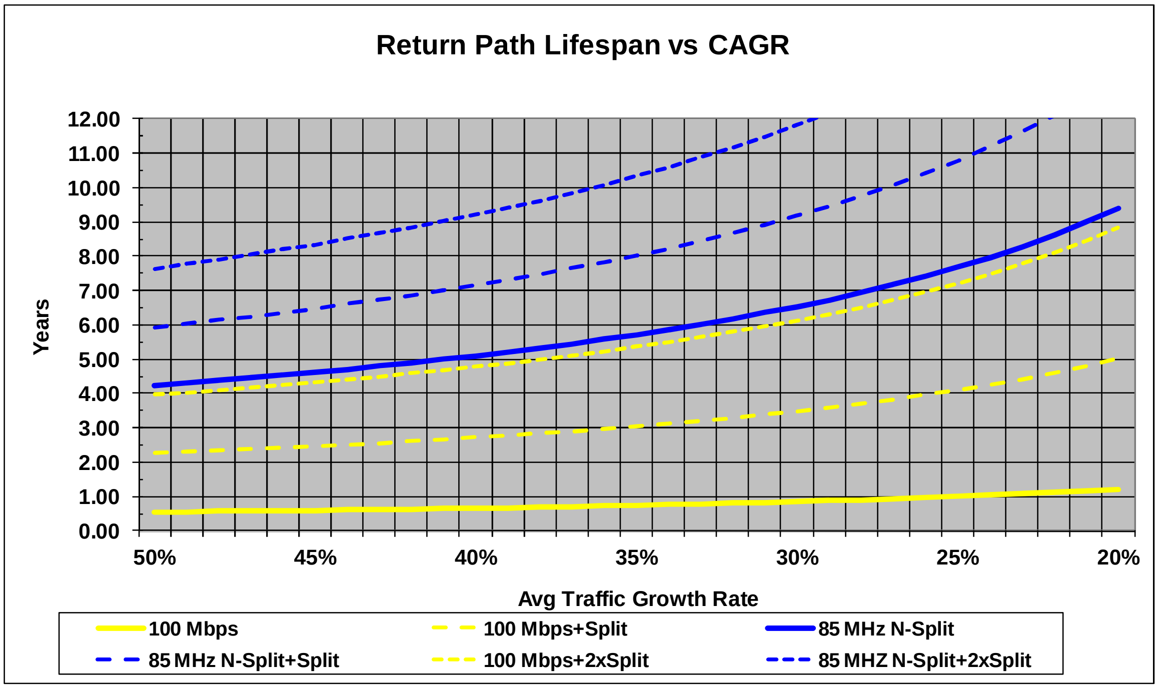


Figure 65 – Upstream Lifespan for Accelerated Usage Patterns

10.1.2.2 Architecture

We observed the clear relationship between available bandwidth and upstream capacity in Table 38. Unfortunately, there simply are no "easy" answers to adding new, real upstream capacity (as opposed to virtual, node splitting).

However, the 85 MHz Mid-Split looks to be the most compelling option in the near term in terms of implementation ease, availability, risk, compatibility, lifespan, and the strength of the value proposition, additional components of which are described in Section 2.1. We have seen in **Error! Reference source not found.** and **Error! Reference source not found.** and **Error! Reference source not found.**, that it also has perhaps unexpectedly powerful benefits.

The 85 MHz spectrum approach is diagrammed in Figure 66. Also shown is the combined case of the Mid-Split and a node split – clearly these are complementary tools.

This architecture has many very valuable and compelling advantages including the most important one of enabling a long upstream lifespan, while supporting key service expectations around data rate.

We summarize the 85 MHz Mid-Split benefits below:

- More than doubles the spectrum available, and more triple the available capacity compared to the use of 5-42 MHz today

- A decade of life OR MORE of upstream growth under aggressive assumptions for traffic growth using only an assumption of 256-QAM

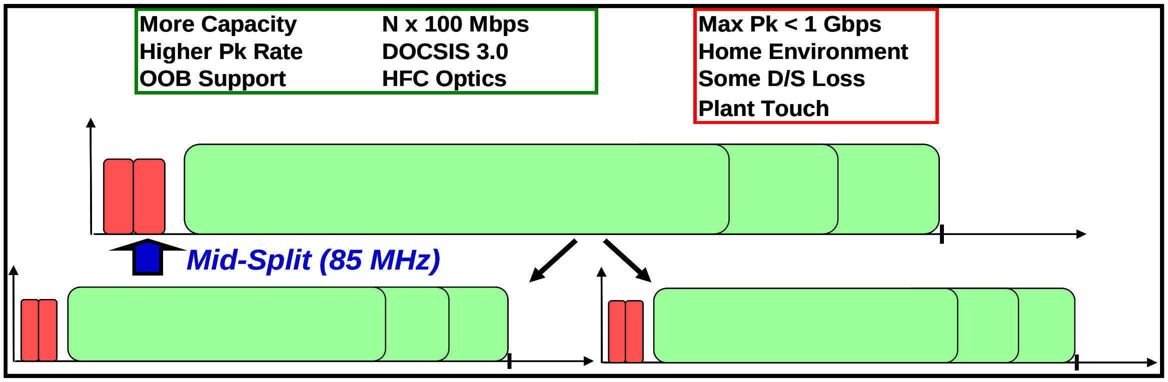


Figure 66 – Step 1: New Return Above the Old Return

- Accommodates multiple 100 Mbps peak rates. Accommodates higher peak rates if desired such as 150 Mbps or 200 Mbps. These may be important to run an effective 1 Gbps DOCSIS downstream service.

- Compatibility with DOCSIS 3.0. Current specification call out support of this extended spectrum. Equipment exists and has been proven for this band.

- Compatibility with standard downstream OOB carriers (70-130 MHz). Thus, no STB CPE using standard OOB is stranded (or at least the vast, vast majority, will not). Over time, as this older population of CPE is removed as part of an all-IP transition, even more flexibility for how to manage return spectrum become available.

- Can be implemented over standard HFC RF and linear optical returns, as well as digital returns. Products exist today for both.

- The new spectrum from 42-85 MHz tends to be cleaner, with less interference and impulse noise, and overall well behaved. This follows the characteristic of the current return that gets cleaner towards the higher end of the band.

- The Mid-Split architecture remains in the low-loss end of the HFC band. Combined with clean spectrum, the DOCSIS 3.0 implementation should have little if any differences, and any updated PHY approaches have the opportunity for even more bandwidth efficient modulation profiles.

- Entails minimal encroachment into the downstream bandwidth as a matter of capacity, and is even less significant when considered in the context of reclaiming the analog spectrum. In this case, it is basically the loss of one 6 MHz slot from a program count perspective – nine lost slots to cover the guard band

- Has similar cable loss versus frequency properties as legacy band – important for understanding CPE implications

- Very low risk, Proven in the field on a fully loaded upstream carrying 64-QAM and 256-QAM. Field trials using standard DFB lasers over typical link length and optical receivers have proven performance.

Note that the proven performance and link characterization for the Mid-Split architecture was discussed in detail in Section 7.1.2, where 256-QAM deployments for upstream were described

A few drawbacks are often cited for the Mid-Split, typically around cost and deployment obstacles. The primary concern is the need to touch actives throughout the plant. It is thus an imperative an upgrade activity be coupled with a segmentation operation and preferably with the ability to enable a Phase 2 of the evolution without requiring the same heavy touch.

Many potential solutions are available to ensure that an elegant transition from 85 MHz to a wider bandwidth in the future can be achieved. Unfortunately, as was originally stated for the upstream, there is no simple solution to more return spectrum.

Recognizing the intrusiveness of the work at hand to modify the frequency split, is commonly observed that the level of touch to the plant means that the "big" step to the 200+ MHz approach should be made.

However, in consulting with operators and suppliers, it is clear that the legacy CPE still requiring the downstream OOB channel for communications must be accommodated. The dynamics associated with this obstacle were detailed in Section 3.3.5. Also, the ability to absorb that amount of loss in the downstream is not tolerable at this phase of the IP migration, which currently might best be described as the "IP Simulcast Bubble" phase of evolution. Therefore, we recommend a phased approach.

Two key items must be recognized in implementing the change. First, it is intrusive, but it is also very low tech, very low risk, available and standardized today. Indeed, it has been proven in existing equipment. Second there is a perception that "just" going to 85 MHz with the effort involved is not enough.

In fact, as shown in the analysis of 85 MHz Mid-Split capacity and lifespan, this is not a band-aid, incremental upgrade, but one that delivers a powerful value proposition in the long term runway it enables, all the while maintaining the fundamental diplex architecture and simplicity of using the low-loss end of the spectrum for the return path.

The deployment challenge often arises out of concern for the home environment when an 85 MHz CM is installed. We described these dynamics in Section 3 and discussed strategies to deal with the challenge. For example, an installation may need to include a blocking filter for some STB CPE. Obviously, the risk here drops considerably if analog channels are removed, or if a Home Gateway architecture is adopted as part of an IP video transition. This is important to characterize and develop a sound operational model for, but is certainly not a technology challenge.

And, in Sections 2.6 we outlined the argument around the limitation often stated that that 85 MHz cannot achieve 1 Gbps of upstream. As was observed in Figure 2, with the time window made available by an Mid-Split upgrade, an extension of the Mid-Split is poised to deliver this capability when necessary and after legacy obstacles have had an opportunity to be addressed. The capacity requirements for residential 1 Gbps of capacity or service rate project well into the next decade on a CAGR basis.

10.1.2.3 Summary – Mid-split Migration Strategy

We recommend an 85 MHz Mid-Split upgrade for a near-term phase of spectrum expansion. Given the lifespan it will be shown to support over CAGRs much more aggressive than are observed today, the 85 MHz Mid-Split should be viewed as a long-term solution and not a temporary fix.

Key benefits are summarized as follows:

1. More than doubles the spectrum and triple the available capacity, providing a path to a decade of life OR MORE of upstream growth
2. Accommodates multiple 100 Mbps peak rates and higher.
3. Compatible with DOCSIS 3.0
4. Compatible with standard downstream OOB carriers (70-130 MHz)
5. Can be implemented over HFC RF and linear optical returns, as well as digital returns.
6. Cleaner spectrum from 42-85 MHz tends to be cleaner
7. Maintains use of the low-loss end of the HFC band. Any updated PHY approaches have the opportunity more bandwidth efficient modulation profiles, and CPE Tx power remains manageable.
8. Entails minimal encroachment into the downstream bandwidth as a matter of capacity
9. Very low risk, proven in the field on a fully loaded upstream carrying 64-QAM and 256-QAM using standard DFB lasers.

While we refer to Mid-Split as "Phase 1", it is a possibility that such a step becomes essentially a "forever" step from a business planning standpoint, on the way to some other long-term approach as greater than ten years of HFC migration is traversed.

Nonetheless, given the projected objectives for the upstream as we see them today, ensuring a path to 1 Gbps in the upstream within the context of HFC tools and technologies is a good long-term objective and a necessary part of long term planning.

Thus, a smooth transition plan beyond Mid-Split requires thinking through the aspects of the Phase 1 implementation that clears the way for this point in the distant future when 1 Gbps becomes a requirement. In this way, the best of multiple key objectives is achieved – many comforting years of immediately available lifespan, support for a long transition window of legacy services, and a strategy for effectively dealing with the continuous traffic growth to come with new bandwidth on-demand.

10.1.3 Phase 2: Deploy High-split – Enabling Gigabit Plus

10.1.3.1 High-Split Extension

Though there are many benefits to an 85 MHz extension, one aspect that cannot be accomplished is support of the 1 Gbps capacity or service rate. This is the case within the parameters of DOCSIS use of the band (360 Mbps), and also the case considering theoretical capacity under DOCSIS SNR assumptions of 25 dB (650 Mbps).

Interestingly, a theoretical 1 Gbps within the 85 MHz Mid-Split architecture would require a 38 dB return path SNR. While well above the DOCSIS requirement, this is, in fact, a relatively easily achievable optical link SNR today using modern DFB transmitters or digital returns. In addition, we can expect higher order modulation profiles enabled at lower SNRs because of the new FEC anticipated – such as 1024-QAM. This would increase data capacity by 25% over 256-QAM and 67% over 64-QAM.

In practice, a manageable operating dynamic range must be considered, as must the other factors that contribute to SNR degradation – RF cascade, user interference, CMTS receivers, and upstream combining, for example. And, though this may be

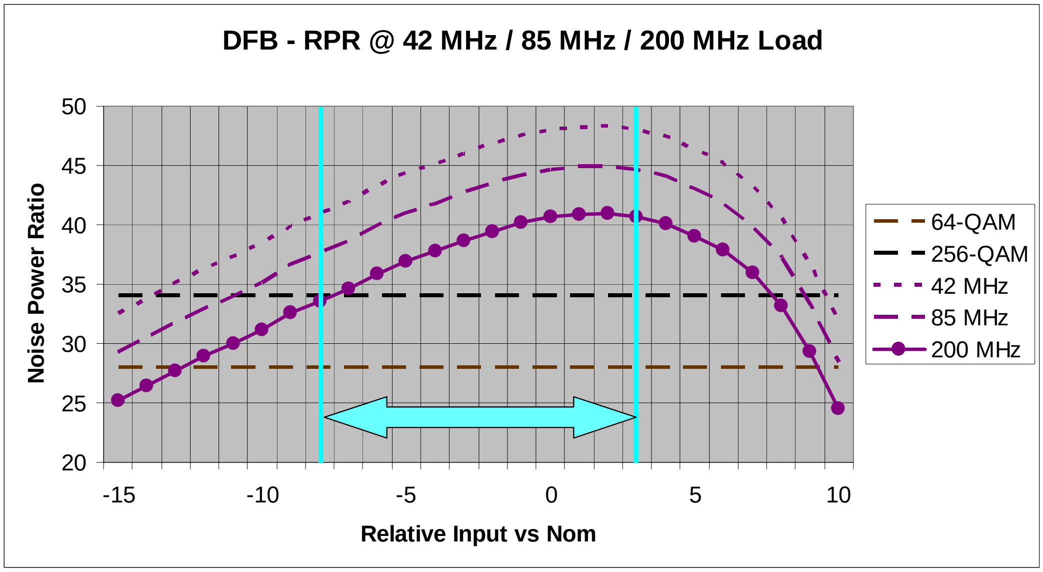


Figure 67 – Bandwidth Loading Effect, 42/85/200 MHz

possible in principle, there is likely to be legacy constraints to having the entire band available for a new, capacity-capable PHY to reach 1 Gbps.

However, this fact does point out that we are entering a new realm of possibilities on the return. Now, with de-combined Headends, 85 MHz of spectrum, modern HFC optics, and new CMTS receivers, and eventually new FEC, many new dB are becoming available toward theoretical capacity and lifespan.

As Table 38 points out, 1 Gbps requires that split to move up to about the 200 MHz range under DOCSIS upstream SNR constraints. 200 MHz is in fact well over 1 Gbps of theoretical capacity, but we assume DOCSIS remains in use for 5-85 MHz, and that the 85-200 MHz region is exploited more aggressively. With new modulation profiles enabled by new FEC, less than 200 MHz will be required, as has been previously discussed.

DOCSIS' maximum profile today (64-QAM@6.4 MHz) itself filling the band out to 200 MHz falls short of 1 Gbps. With 256-QAM, this would no longer be the case. In the case of using split technologies (5-85 MHz of DOCSIS and 85-200 MHz of something else), a shortcoming that could come into play is the inability of that architecture, or at least the added complexity, of supporting 1 Gbps of peak service rate across potentially different systems.

10.1.3.2 Supported by HFC Optics

An attractive advantage of a diplex-based return of 200 MHz or higher is the ability to use analog return optics. However, the additional bandwidth comes with a power loading SNR loss associated with driving a fixed total power into the laser over a wider bandwidth.

Figure 67 compares 200 MHz optical link performance, fully loaded, to 85 MHz and 42 MHz cases. As previously, the lines representing 64-QAM and 256-QAM are SNRs representing theoretical BER without the use of error correction. The power loading loss is easily predictable, as simply the dB relationship among total bandwidths. For the optical link at least, using typical

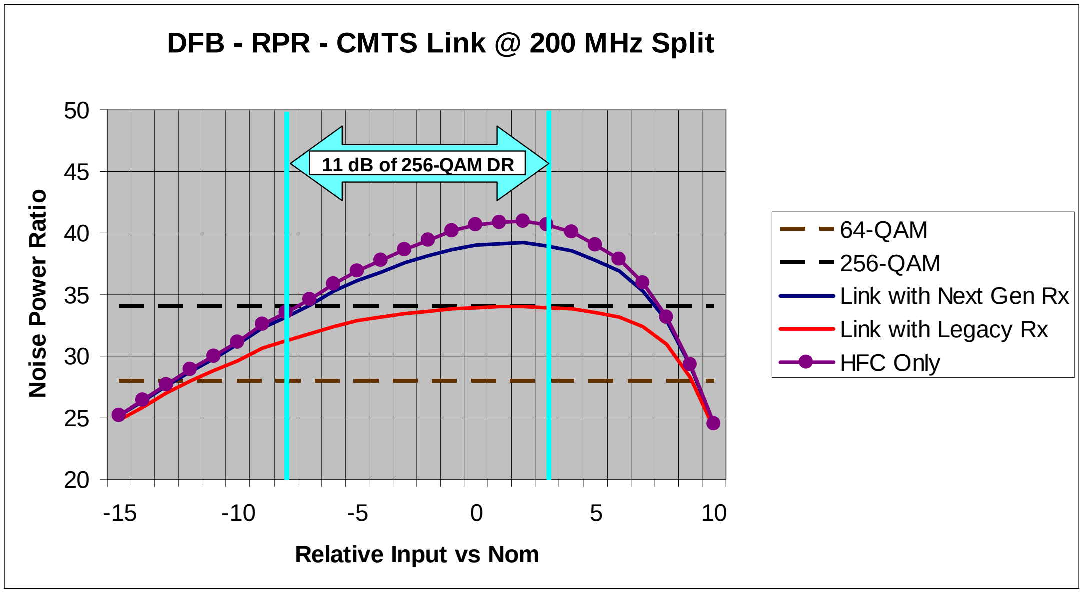


Figure 68 – Projected 256-QAM Dynamic Range Over 200 MHz Split

performance delivered by an analog DFB link, 10-11 dB of dynamic range exists across the HFC optics – a reasonable margin to accommodate alignment, drift, and plant behaviors, but borderline itself for robust, wide-scale roll-out, particularly given degradations that the link will inherent from the rest of the plant.

A comparison of the link using equivalent legacy CMTS receiver performance and modern, lower-noise receivers, is shown in Figure 68. Figure 68 helps to make the point noted in the beginning of this section. The minimum SNR limit assumed for DOCSIS is itself a very dated, and unfortunately conservative and constraining with respect to available capacity.

We now can observe in Figure 68 how the combined effect of the evolution of cost effective, high quality return optics coupled with low noise DOCSIS receivers is opening up new possibilities for extracting capacity from more capable upstream spectrum over wider band.

Based on Figure 68, the full low diplex migration approach has the flexibility of being supported over currently available linear optics. Note once again that we also observed DWDM lasers operating in Figure 20 over high split with NPR performance slightly better than the 1310 nm projection showed here under different link assumptions. This once again shows that today's HFC linear optics is at, or on the verge of, compliant performance for bandwidth efficient profiles over high-split, even without considering new FEC.

Furthermore, High-Splits that exceed current return path optical bandwidth, such as 300-400 MHz, could, in principle, be delivered over linear optics as well. The optics used would simply instead be forward path lasers, which would obviously be high performance.

The preferred, long-term, architectural direction for the long term is a solution based on digital transport over fiber to the node, such as Ethernet or EPON protocol based, to the node, and RF transport over coax. However, an approach based on a low diplex expansion does not require this architecture to operate, offering flexibility to the operator during the difficult transition phase of the network.

When such an architecture is available, the benefits of removing linear optical noise and distortion from the access link budget have very powerful capacity benefits to a low diplex, whose SNR performance is typically set by the optics.

10.1.3.3 Spectrum Evolution

If 85 MHz Mid-Split is a "natural" extension of the Sub-Split (42 MHz) for long-term growth, then a "natural" extension of Mid-Split for long-term peak rate support and FTTH competiveness is the 200-300 MHz High-Split. This concept is diagrammed in Figure 69, along with a summary of the pros and cons.

Unlike Mid-Split, a high split can achieve the 1 Gbps rate foreseen as possibly the next threshold in the upstream after 100 Mbps. And, in doing so, it does not suffer the very high RF attenuations that the alternatives that rely on frequencies above the forward band do. The exact upper band edge is a function of modulation profile, which again is tied to architecture and FEC.

This translates into more cost-effective CPE. As we have seen, implementation of today's HFC optics is possible, as modern HFC optics is based on 5-200 MHz and 5-300 MHz RF hybrids. And, to reiterate, this architecture, too, would benefit from any migration in the plant that relies on digital fiber delivery and RF carried only in native form on the coaxial leg of the plant.

By maintaining fundamentally a diplex architecture, there is still but one guard band in the architecture, preserving use efficiency. Lastly, at the low end of the HFC spectrum, there would not necessarily be a compelling reason to require an OFDM system, unlike other portions of the band.

The channel quality would not necessarily demand a multi-carrier waveform, and it would have modest advantages at best in a clean channel environment anticipated. Extensions that further empower DOCSIS become more reasonable to consider without a fundamental change in the waveform used, silicon architecture, specification, or new technology learning curves.

At the same time, because the linear optical return architecture anticipates a broadband, noise-like signal, the addition of OFDM channels, even wideband, can be carried within the linear optical architecture as well if the high split band evolves to

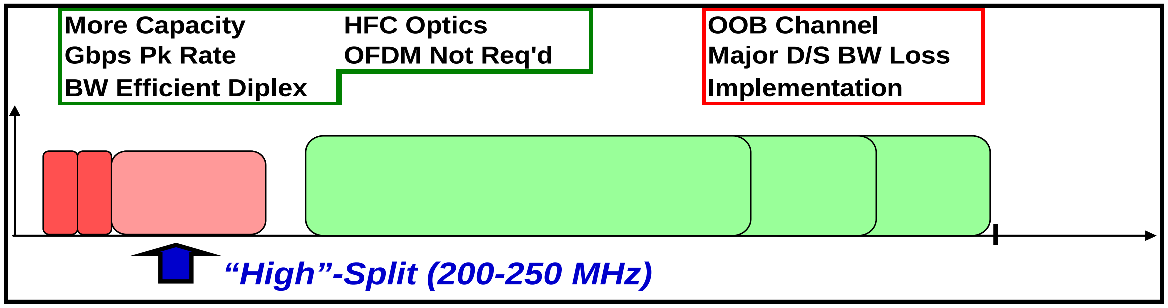


Figure 69 – High-Split Concept, Pros and Cons

include multi-carrier formats. Again, in comparison to other alternatives, this is an added degree of implementation flexibility.

The loss of the OOB downstream channel is an important consideration. However, the logic of this approach is that by the time it becomes necessary – again, likely at least 10 years down the road – the MSO has had ample opportunity to retire through natural attrition or actively manage down legacy STB relying on this OOB channel.

Again, knowing what steps are in place and coming over time, decisions can be made about handling legacy STB either through DSG or Home Gateways associated with an IPV transition.

10.1.3.4 Notable Obstacles

Unlike Mid-Split, High-Split is now a major imposition on downstream spectrum. However, it is expected that downstream spectrum will also undergo expansion over time as traffic in both directions continues to grow. There is already potential spectrum to be mined above the top end of the forward path in many cases, and it is anticipated that if the upstream is to continue to move "up" with high-split, there may be a need also to offset the loss of downstream spectrum by extending downstream as well beyond its current limitations.

By appending new spectrum to the end of the current downstream, this approach to exploiting new coaxial bandwidth is able to maintain a single diplex architecture. This concept is shown in Figure 70.

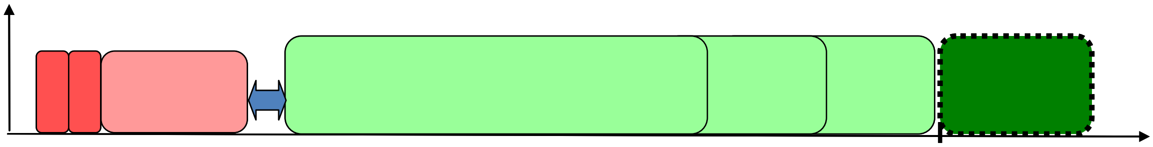


Figure 70 – Possible "Offset" Band Compensating for High Split

While this presents a potential solution from a capacity perspective, from a CPE perspective there are important limitations associated with legacy equipment. As the "Simulcast Bubble" winds down at the back end of this decade, models suggest that those savings will be able to compensate for the expansion of upstream into a high-split architecture.

However, under an assumption of persistent CAGR and a continued evolution of HD into even higher resolution formats, such savings will over time once again give way to spectrum management of a new phase of services growth. The window of savings, however, is an important component of a transition that includes the possibility of extending the forward spectrum. We will elaborate on the forward aspects in subsequent section.

10.1.3.5 High-Split Extension – Timing and Implications

The time frames required for a high-split migration are a key element of the strategy because of the intrusive nature of this magnitude of change, and the idea that we may wish to include as part of a transition plan the creation of new forward bandwidth. We touched on the expected timing of 1 Gbps solution in Section 2.6.

Even should the access network be evolved to enable a high-split in the 200-300 MHz band on-demand, such as putting the

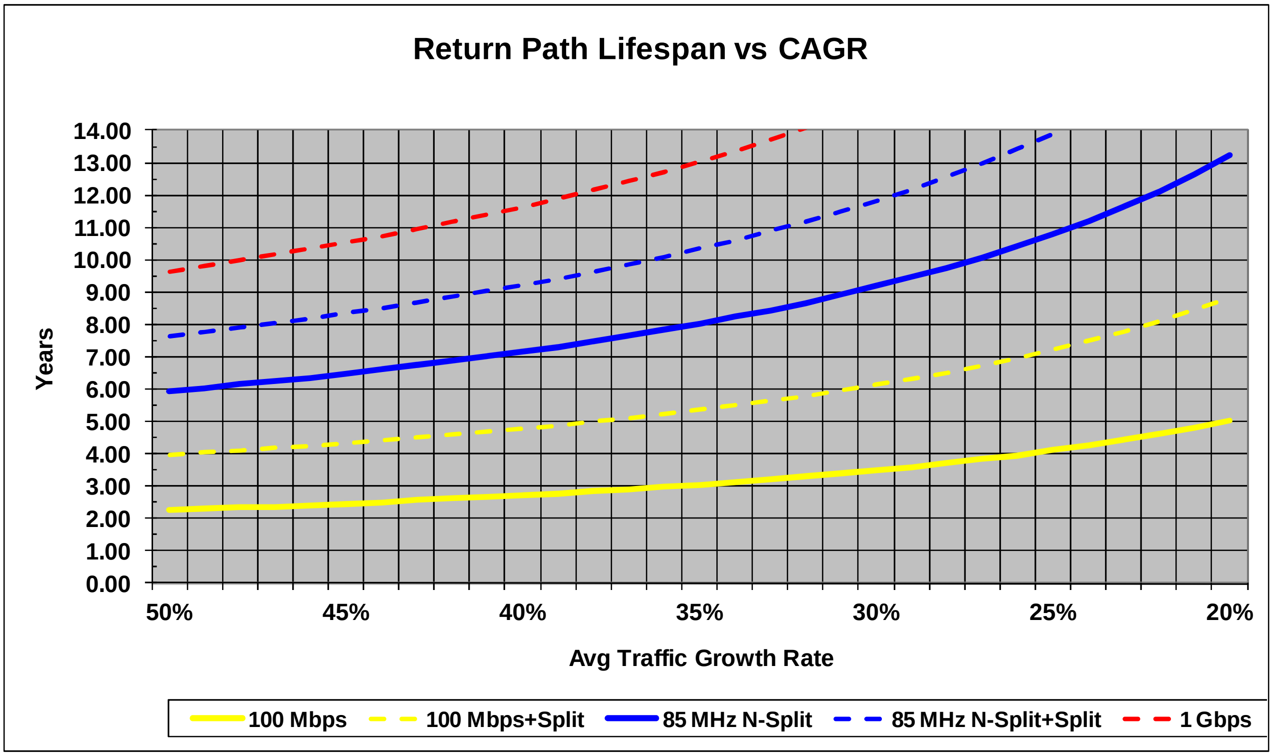


Figure 71 – Relative Lifespan and the Benefits of 1 Gbps

capability in when 85 MHz is deployed, the move to a high split has large impacts on the forward spectrum and return path transport that must be planned.

It is therefore important to get an idea of when we might need it. There are consumption and market pressure components of that, but let's view it in an apples-to-apples way with the prior analysis of the 85 MHz capability for extending return path lifespan. What does a Gbps of capacity imply for long-term traffic growth?

The answer to this question can be examined in Figure 71. It is an excellent illustration of how compounding works and the need to consider what it means if played out over the long term. It shows three threshold cases – 100 Mbps (A-TDMA only), 85 MHz Mid-Split and 1 Gbps (also with a split included).

Zeroing in on the gap between 85 MHz Mid-Split and 1 Gbps at 35% CAGR, we see that there exists about 2.5 years of additional growth after about 10.5 years of lifespan. When we think of "1 Gbps," this intuitively seems odd. Again, this is simply how compounding works. If we base analysis and decisions on the continuance of a compounding behavior paradigm, then the mathematical basis is quite straightforward.

With CAGR behavior, it takes many YOY (year-over-year) periods to grow from. For example, the 40 Mbps of upstream used by a service group today service today to the 440 Mbps that can be delivered by Mid-Split. That number, as Figure 68 shows, is 10.3 years of compounding at 35%. However, once there, the subsequent annual steps sizes are now quite large. That is the nature of compounding, resulting in what seem like small extra lifespan.

10.1.4 Summary

The spectrum migration shown and described above is repeated in Figure 72 and Figure 73. The role of the upstream migration phases in the larger picture of HFC spectrum evolution and the transition to an All-IP end-to-end system is shown in Figure 74 and Figure 75.

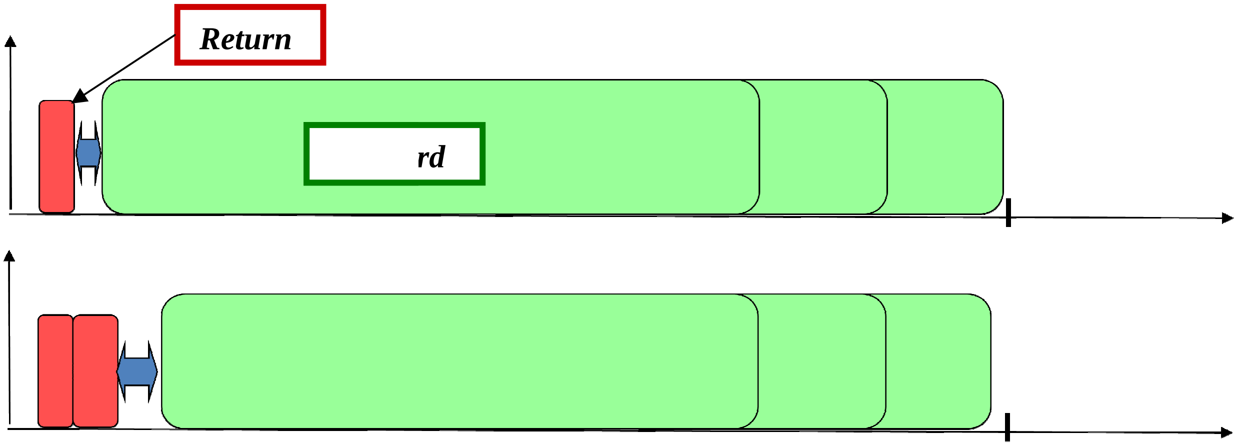


Figure 72 – Phase 1: 85 MHz Mid-Split

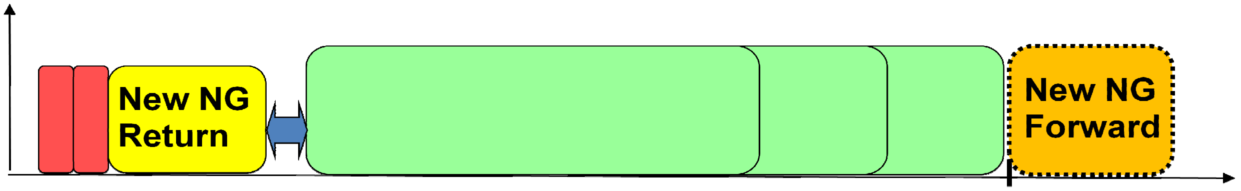


Figure 73 – Phase 2: 200+ MHz High-Split and Possible Relief Band Forward

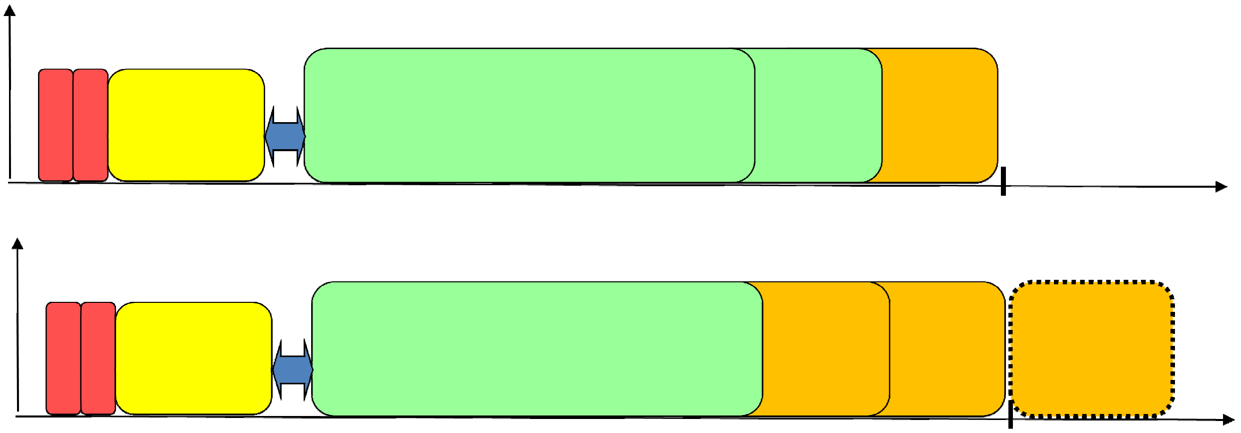


Figure 74 – IP Transition in Progress – Legacy Roll-Back

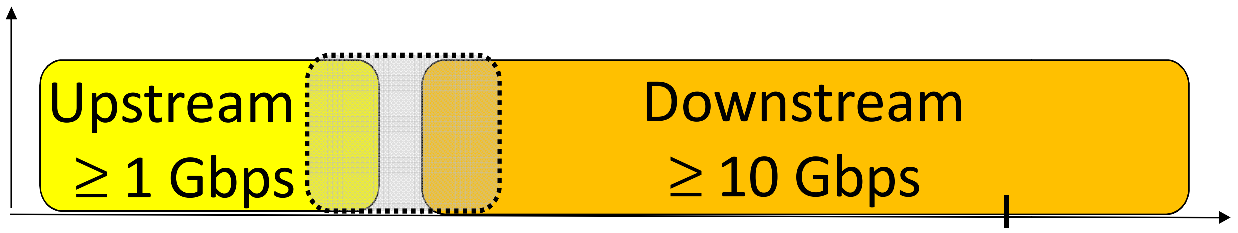


Figure 75 – Final State of All-IP Transition

Flexible/Selectable Diplex, Advanced PHY, Digital Transport-Based HFC Architecture, N+Small/N+0

10.2 Downstream Migration Strategy

10.2.1 Capacity and Lifespan Implications of IP Growth

Every individual HFC plant has evolved on an as-needed basis, and of course under CAPEX budget constraints that inherently come with a network of fixed assets expected to last a long time. As a result, HFC networks in North America have a range of top-end forward path bandwidths.

Typically, however, plant bandwidth is 750 MHz, 870 MHz, or 1 GHz – more so that former two. Absolute bandwidth is obviously important, but fortunately multiple additional tools are available to help manage downstream service growth, such as digital television (DTV), increasingly efficient DTV compression, more bandwidth efficient modulation formats, and switched digital video platforms (SDV). These are all complementary and are in addition to common network segmentation.

As cable advanced video services and data services have grown, however, it has become clear that powerful new dynamics are working against cable operators, and towards a capacity bottleneck in the downstream. The result has been a renewed interest in finding new spectrum, which to a first order directly translates to increased network capacity. Being aware that coaxial cable is not limited to any of the forward band limitations mentioned above, operators are exploring how to access what today is unexploited spectrum above these defined forward bands. There are no technology obstacles to its use, but significant legacy service, network, and equipment implications.

We have discussed in detail the capacity available in DOCSIS and DOCSIS NG as evolution phases take place. However, we have not discussed them in the context of the *available* HFC spectrum. While new DOCSIS capacity is powerful and important, most of the downstream spectrum today is locked down for video services. Finding new DOCSIS spectrum is a major challenge in the normal HFC band, and it is years away before we can exploit the extended bands. We can illustrate quite easily why finding new HFC capacity has become so important and difficult. Consider Figure 76.

Figure 76 projects two cases of IP traffic growth, modeled after the well-travelled Nielsen's Law approach to user bandwidth trends. In this case, it is taken in the aggregate, representing, for example, one service group or perhaps one node.

It assumes that eight DOCSIS downstream service this population today. This is represented on the y-axis, shown on a logarithmic scale because that is the nature of compounding growth. The axis is quite simple to translate in dB – 100 Mbps is 20 dB, 1 Gbps is 30 dB, and 10 Gbps is 40 dB. For eight DOCSIS channels (always using the transport rate in this example, since we are not quantifying service tiers), this works out to 25 dB as a starting point.

The trajectories proceed at 50% Compound Annual Growth Rate (CAGR), interrupted by service group segmentations (such as node splits). In this example, a simple, perfect split (in half) is performed mid-decade. A second, perhaps final, segmentation is done at the end of the decade that resembles an N+0 from a service group size perspective (40 hhp), although it is immaterial to the analysis whether there would physically need to be an amplifier in some particular plant geographies. We use N+0, as we subsequently discuss the implication this has for spectrum planning and capacity exploitation.

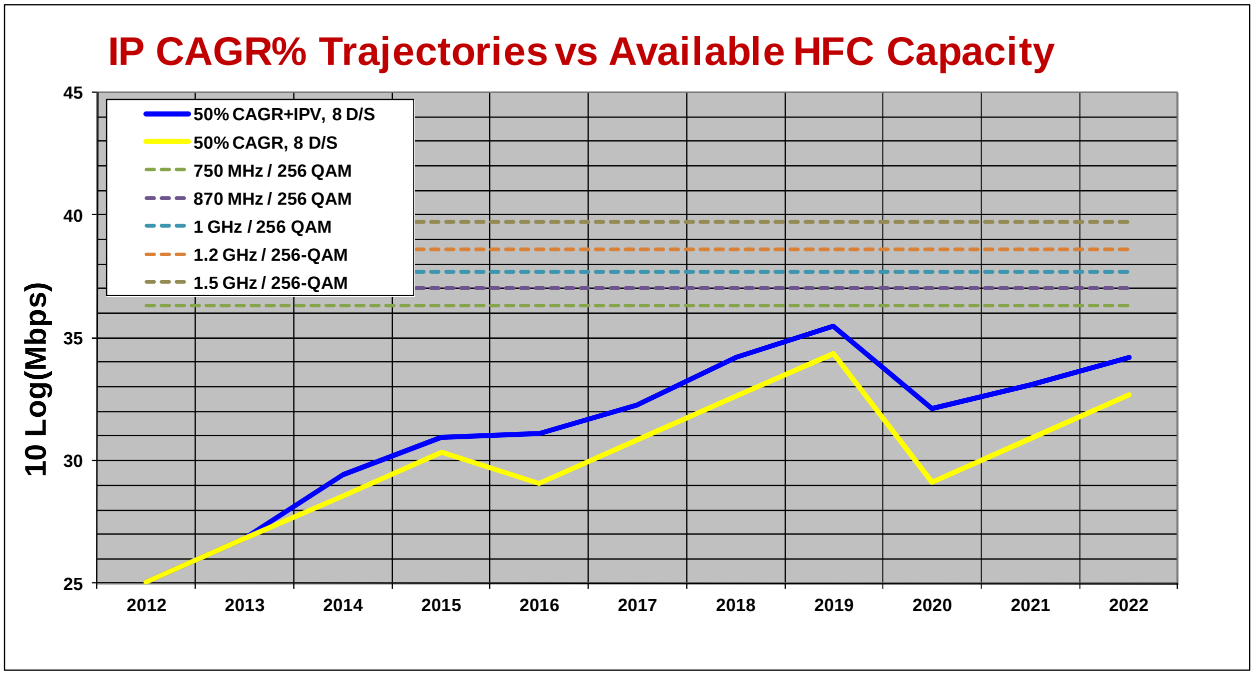


Figure 76 – HFC Downstream Capacity, IP Traffic Growth, and Segmentation

Finally, there are two trajectories because in one case we add dedicated IP Video channels to to IP traffic growth, in addition to the 50% CAGR itself. There is somewhat a philosophical discussion to be had about whether managed IP Video is the new engine of 50% growth (like OTT has been for years), or if CAGR plows ahead in addition to shifting the current video service onto the DOCSIS platform.

Here, the assumption is that blocks of DOCSIS carriers are added every other year beginning in 2014 – first four channels, then 8 channels, then 8 channels for a total of 20. It is a separate analysis how 20 DOCSIS slots represents an assumed video line-up that we will not go into here, but this has been analyzed and written about in many industry papers over the past 4 years.

Five thresholds are shown, consistent with five different assumptions of network bandwidth. In every case, it is assumed that the return bandwidth has been extended to 85 MHz, and the first forward channel is therefore in 109 MHz. It is also assumed, in the extended bandwidth cases of 1.2 GHz and 1.5 GHz, that 256-QAM can be supported.

This is a reasonable assumption – in fact minimally necessary to make turning that band on worth the effort – but obviously unproven at this point. Lastly, each of these thresholds can be incremented by about 1 dB (more) by making the assumption that 1024-QAM replaces 256-QAM (10 Log (10/8)). It was decided not to clutter this figure with those minor increments. But, as discussed, for DOCSIS NG, 1024-QAM downstream and up to 4096-QAM downstream are anticipated modulation profiles, with an objective for total downstream bandwidth of 10 Gbps (which is simply 40 dB in Figure 76 and Figure 77, however it is accomplished).

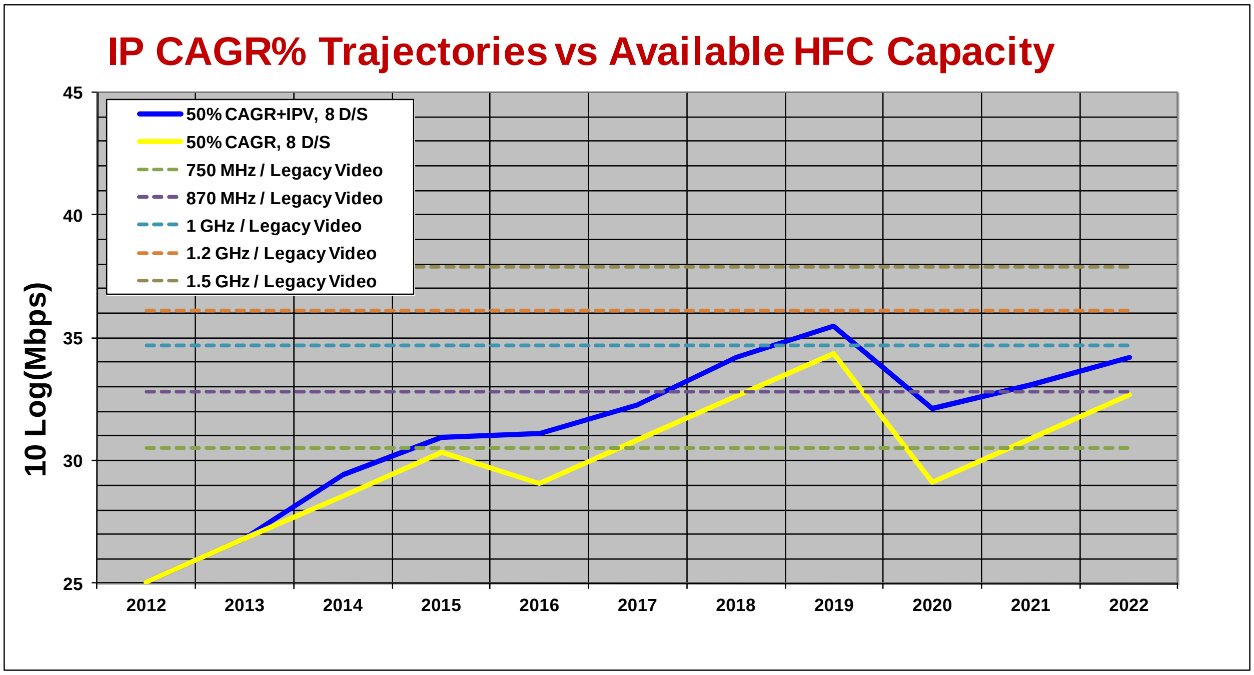


Figure 77 – Capacity, Traffic Growth, and Segmentation – Video Services Added

The thresholds are still based on the assumption of 6 MHz slots of 256-QAM, so represent "current" spectral usage efficiencies, and as such are conservative in that sense. The thresholds, thus, represent the integer number of 256-QAM slots, aggregated to a total based on 40 Mbps/per slot.

An obvious conclusion from Figure 76 would be that the HFC network is in fine shape to take on an extended period of aggressive growth. The network appears not threatened until (projecting to the right) the 2023-2024 time frame, worst case. Of course, there is something seriously missing from this analysis – current services.

Now consider Figure 77.

Figure 77 takes into account that most of the HFC spectrum is not available for new IP growth today. In fact, for most operators, have very little or no "free" spectrum to put new DOCSIS carriers in. When they need new ones, they shuffle other things around and use the tolls above to make it happen. This is much easier said than done as more spectrum, not less, is being consumed with the increasingly competitive environment around HD programming.

The programming line-up above assumes the following:

- Broadcast SD: 100 programs (10 slots)
- Broadcast HD: 40 programs (10 slots)
- SDV 24 slots: This increases the total programming to SD~300 and HD~150
- VOD 4 slots
- No Analog

Clearly, this is not particularly aggressive. First, it is assumed that there are no analog carriers – everyone's long term goal, but executed on by only a few. Also, not all operators are using SDV to this degree, the VOD count is modest, and objectives for HD are for 200-300 programs (not to be confused with “titles”). Finally, there is a real possibility that upstream congestion will require that this band be extended beyond 85 MHz, up to the 200 MHz range or beyond. This would significantly impose on available capacity.

And the result? A 750 MHz is in immediate danger without a service group split, and an 870 MHz network is not far behind. In all cases that do not go above 1 GHz, the “N+0” phase is required before the end of the decade to manage the growth.

The extra runway offered above 1 GHz is apparent – relatively modest for an extra 200 MHz (but this would offset a 200 MHz return at least), and substantial for a 1.5 GHz extension. In the context of the evolution of video services, then Figure 76 can be viewed as the capacities available when the full IP Video transition is complete, and no legacy analog or MPEG-2 TS based video services exist.

As such, they are not “phony” capacities – they merely represent the available capacity, under today’s limitations of technology, at the point in time when the legacy service set is fully retired. In this sense, then, they are very valuable thresholds for guiding plant migration and bandwidth management.

A final note on the Figure 76 thresholds is to note that 1 GHz of ideal 1024-QAM bandwidth, at 10 bits/s/Hz efficiency, adds up mathematically to 10 Gbps. We almost achieved this only considering 256-QAM @ 1.5 GHz, and clearly would have done so under a 1024-QAM assumption (one more dB on added to this threshold).

This order of magnitude is important relative to competitive PON deployments. With respect to subscribers served, the PON port is shared by 32 or 64 subscribers. With cable, the access leg is shared by one node port as a minimum, or more generally one complete node. Today, a typical single node average is about 500 homes passed, and this is headed downward. At N+0, it will reside likely in the 20-50 HHP range. For cable then, the subscriber base sharing a 10 Gbps-capable node will be similar to 10 Gbps PON networks in the downstream.

10.2.2 Making Room for Gbps Upstream with New Downstream

Moving to the 85 MHz Mid-Split adds 43 MHz of return bandwidth, doing so at the expense of modest imposition on forward bandwidth. When factoring in the new guard band, possibly nine or ten forward path slots in the traditional analog band are eliminated. Mathematically, converting these channels to digital allows them to all fit into one slot.

As such, as analog reclamation continues, this forward loss does not represent a major capacity concern. The primary operational concern is that the nature of the channels in this region. They are often a basic service tier, and therefore cannot simply be transitioned into the digital tier and off of the analog tier, practically or contractually in some cases, as perhaps some of the longer tail of the analog service could.

Instead, some channel re-mapping and/or more aggressive deployment of digital adaptors would be required. In any case, given the powerful set of tools available to provide downstream capacity, 85 MHz does not present significant imposition on the forward bandwidth in terms of capacity loss.

In the case of a 200 MHz extension, however, this is no longer the case. Cable operators generally use all of their spectrum, and a changed such as high-split, even if it phased in, will call for some significant impacts to the downstream services line-up.

The issue is magnified further when considering that while we are looking to extract downstream capacity and give it to the upstream, the downstream itself continues to see rapid CAGR – more rapid and consistent that the upstream. This amount of lost downstream capacity will have to be replaced, and, in fact, capacity above today's available forward capacity will have to grow over time. 1 GHz worth of 256-QAM slots today adds up to about 6.3 Gbps of total transport capacity, and 7.9 Gbps by enabling 1024-QAM. A 300 MHz starting frequency for the downstream removes about 1.6 Gbps – too big to ignore. That means we must find new downstream bandwidth. In Section 4.5 to 4.7, we identified performance of spectrum above 1 GHz for upstream use, and argued that the obstacles to effectively using the band for upstream make it much more suitable for extending the downstream. Here, we elaborate on this possibility and the potential new data capacity available.

So, where would new bandwidth come from above today's forward band? Virtually any new (actually new, not reseller) plant equipment purchased today will be of the 1 GHz variety. This is clearly at odds with trying to use bandwidth above 1 GHz. Industry discussion around enabling new bandwidth is along three fronts:

(1) What bandwidth do 1 GHz devices actually have? We observed "1 GHz" Taps for out-of-band performance in Section 4.5. Because there is always design margin, is there " free," but unguaranteed, spectrum to exploit? Some operators already place channels above the "official" downstream bandwidth, perhaps at a lower modulation order for robustness, which indicates that there is obviously exploitable capacity in some cases.

It can be shown that some of the friendliest taps in the field have about 20% of imperfect excess bandwidth to mine before difficult to manage roll-off kicks in. Field testing of this grade of tap has been extensively performed. In live plant conditions, a typical tap cascade of nominal coaxial spacing showed useable bandwidth to 1160 MHz with high efficiency for wideband (50 MHz) single carrier QAM [1]. Not all deployed taps will have this amount of useful bandwidth. Of course, the best way to mine bandwidth in such difficult conditions would entail a different modulation approach, and this is particularly the case where discussion of multi-carrier modulation (OFDM) is often introduced for cable networks. Aside from the flexible use of spectrum it allows in periods of transition, and through its use of narrow QAM subcarriers, OFDM would more effectively extract bandwidth, and make more bandwidth able to be exploited.

(2) Some suppliers have developed a 1.5 GHz tap product line. However, there is not very much new build activity, so the market for such products has not grown. Extended bandwidth is also available for some taps already in the field by "simply" swapping out faceplates. This is very intrusive and time-consuming, but of course it is also much *less* intrusive and much *less* time consuming than a full tap swap-out.

Some suppliers have developed this technology specifically for existing plant (versus new build which could, in principle, purchase 1.5 GHz taps). The "swap out" approach yields taps with a specified bandwidth to 1.7 GHz. There is more bandwidth than the 1.5 GHz taps, but it comes at the expense of minor degradation in other specifications. However, field testing has been encouraging that these taps extend bandwidth to at least 1.6 GHz [1].

(3) Full tap swap outs for models that increase bandwidth to up to 3 GHz (or use in new builds). This, of course, is a very intrusive plant modification.

It is important to note that suppliers have not yet developed node or amplifier platforms, at least not in volume scale, that extend beyond 1 GHz. There are no technology reasons this could not be done, although there are likely major redesigns involved in most cases right down to the housing, circuit boards, and connectors.

This is viewed as unlikely to take place for RF amplifier platforms, but perhaps not so for nodes. As N+0 is potentially a logical “end state” for an HFC architecture, the ROI picture is somewhat clearer to make for equipment manufacturers. In addition, nodes have undergone generally more R&D investment than RF platforms have, as they have kept up with the optical technology evolution.

Many fielded RF platforms have not changed very much since they were originally designed, and have been had their bandwidth limits continuously pushed. It is unclear how many new MHz are easily available, and the range of RF platforms is much larger.

This limitation on the bandwidth of the RF amplifier is important in the context of accessing new bandwidth and understanding the enabling architectures to do so. We will elaborate and quantify aspects of this in subsequent sections.

10.2.3 Excess Bandwidth Calculations on the Passive Plant

The first place to look for more downstream spectrum is simply in the band that continues directly above today’s forward path band edge. While this was shown to be a difficult band for an upstream service to efficiently and cost effectively support, it is much easier to consider as much for the downstream.

The downstream channel is already very linear, has a very high SNR, and these features of the access equipment are shared by the homes passed common to a piece of equipment in the plant. And, fortuitously, in many 1 GHz tap models there is that significant “free” bandwidth available.

Figure 78 shows the frequency response on the “through” port of the particular 1 GHz tap described in the field trials above that yielded an 1160 MHz net useful band edge. This port would be in series with other taps on the way to a connected home. The response on the tapped port also has essentially parasitic, low-loss properties over the first 200 MHz above 1 GHz.

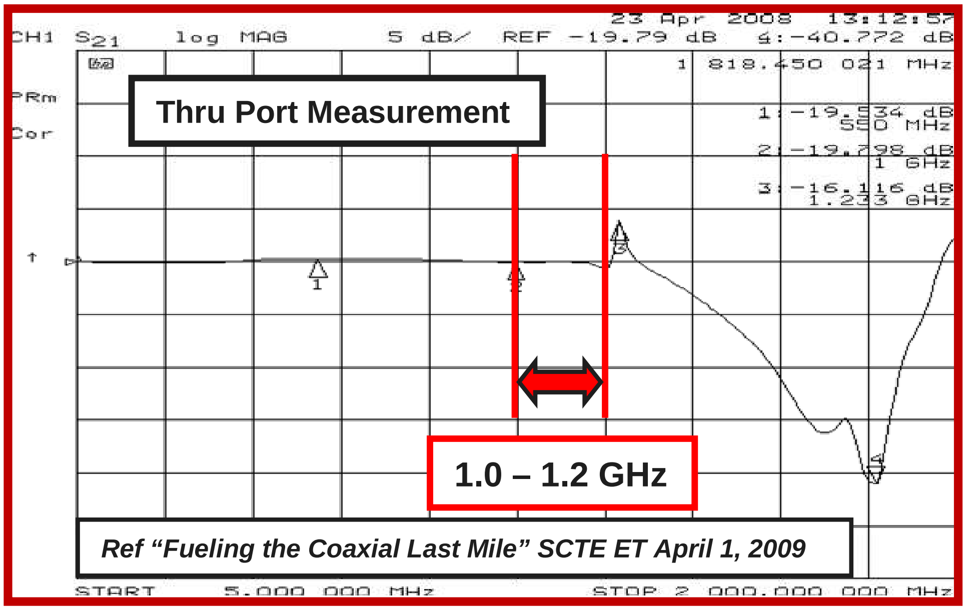


Figure 78 – 1 GHz Tap Frequency Response, "Thru"

Though not as perfectly flat, it creates no significant distortion burden to RF signals in the band, and in particular when considering that a new generation of OFDM technology will almost certainly be created to operate in that regions, and if so will run an adaptive bit loading algorithm.

The same is the case for some families of 750 MHz taps (available bandwidth exists above 750 MHz) and 870 MHz taps (available bandwidth exists above 870 MHz).

The amount of useful bandwidth and loss properties are vendor dependent, but cable operators already often use slots above these limits. Conveniently, as Figure 78 shows, the amount of available new bandwidth simply trickling over the top of the band is virtually the same the amount of bandwidth that would be removed from the forward by a 200 MHz high-split architecture.

With the support of the supplier community, CableLabs has undertaken an investigation to statistically quantify this excess bandwidth across Tap models and manufacturers so that operators can better understand in their specific plants what useful bandwidth is available, and how that changes with time with shorter cascades.

An important item to re-emphasize is that there is no guard band involved when this spectrum is operated as only a downstream extension, as there would necessarily be if upstream were to be deployed in this band. This "replacement" bandwidth amount provides adequate spectrum to facilitate new downstream capacity.

The ability to fully exploit this bandwidth in the passive plant obviously depends heavily on the band coverage of the actives themselves and the depth of the cascade. Clearly, this is where shortening cascades and "N+small" continue to payoff for HFC evolution.

The tapped port, of course, also contributes to the frequency response, and a sample of this port on the same 1 GHz tap model (2-port, 20 dB) is shown in Figure 79. The response on the tapped port also has essentially parasitic, low-loss properties over the first 200 MHz above 1 GHz.

Though not perfectly flat, it creates no significant burden to RF signals in the band, and in particular when considering a new generation of modem technology, such as multi-carrier. The same is the case for some families of 750 MHz taps (available bandwidth exists above 750 MHz) and 870 MHz taps (available bandwidth exists above 870 MHz).

It is clearly evident that the band between 1.0 GHz and 1.2 GHz is not flat, having about 2 dB of what can best be described as a broadband ripple in the response.

10.2.3.1 Excess Bandwidth SNR Model

In order to calculate the capacity associated with this "extra" bandwidth, we must numerically model this frequency response. This is easily accomplished for parasitic-type roll-offs, more so even that with classic RF filter responses such as diplexers.

We can, in fact, fit the attenuation response to some fundamental filter shapes and use those to calculate attenuation. And, by proxy, SNR for a fixed transmit power. In this case, the roll-off response can be fairly well represented by scaled versions of a 5th order Butterworth response, as shown in Figure 80.

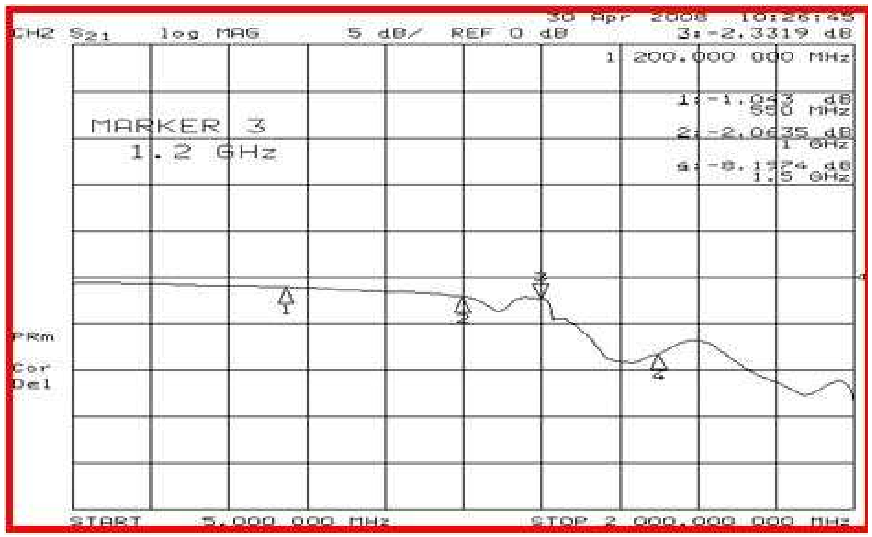


Figure 79 – 1 GHz Tap Frequency Response, Tapped Port to Home

Here, the thru attenuation (blue) of approximately 10 dB across the 1-2 GHz band, as well as the roughly 20 dB of attenuation over 600 MHz represented by the port (red), is represented. Note that increasing stop-band attenuation typically means correspondingly poor *return loss*, which is an RF reflection mechanism – a mechanism already part of DOCSIS, and that has become very sophisticated with DOCSIS 3.0. Of course, if a multi-carrier PHY is adopted in this band, it too is robust to this distortion, but through different means, such as use of a cyclic prefix.

Filter roll-off regions also typically correspond with regions of high group delay variation – another challenge taken on by the 24-Tap equalizer. For A-TDMA, however, there are limits to how successful the equalizer can be with combined micro-reflection, amplitude response, and group delay distortion.

Performance has been shown to be far, far beyond the conditions called out in DOCSIS specifications. Nonetheless, multi-carrier evolutions to the PHY minimize the potential concerns over operating in these regions as well. System parameters (subchannel widths, cyclic prefix guard times) can be used very effectively to overcome these obstacles where the channel performance degrades.

Consider the two narrowest bandwidth curves of Figure 80. These represent the composite frequency response of an N+0 cascade of five taps (N+5T, pink) or ten taps (N+10T, brown), and an accompanying length of coax governed by a typical attenuation model.

A subscriber at the end of a ten tap run will of course see nine thru responses and a tapped port (and quite possibly an active that would need to support this band or bypass it), and this response is represented by the brown curve. The pink curve represents a five tap scenario, which is a more typical run of taps between actives.

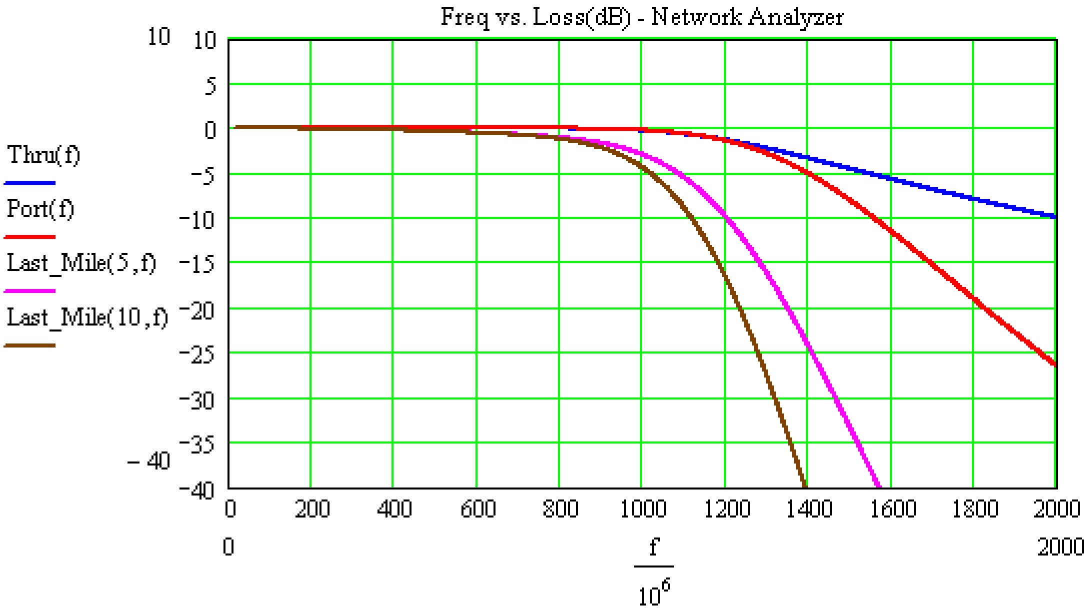


Figure 80 – Modeled Tap + Coax Performance

These attenuation curves for a cascade of taps, plus interconnecting coaxial runs, can be used to quantify the attenuation profile, and, given a transmit power profile (is it tilted or not), the SNR delivered from the network for a given power, and thus the capacity available as a function of new spectrum. We can thus see the efficiency with which this new part of the band delivers capacity.

10.2.3.2 Capacity Derived from Excess Spectrum

Figure 81 quantifies available capacity, assuming an HFC forward digital band starting SNR of 45 dB at 1 GHz in the HFC plant and using the frequency response of Figure 80. An HFC downstream link at the output of a node would be expected to deliver at least 51 dB of SNR as a common objective in the analog band, leaving the digital band 6 dB removed from that performance.

Thus, this represents an N+0 case ideally, but could also reasonably apply to a short cascade that includes RF amplifiers that pass this band with a flat response as long as there are not more than 5 taps in the series (the 10 Tap case is not shown in Figure 81). It also conservatively assumes a flat transmit response, and, while increasing in frequency, calculates the resulting capacity as this band edge moves to the right.

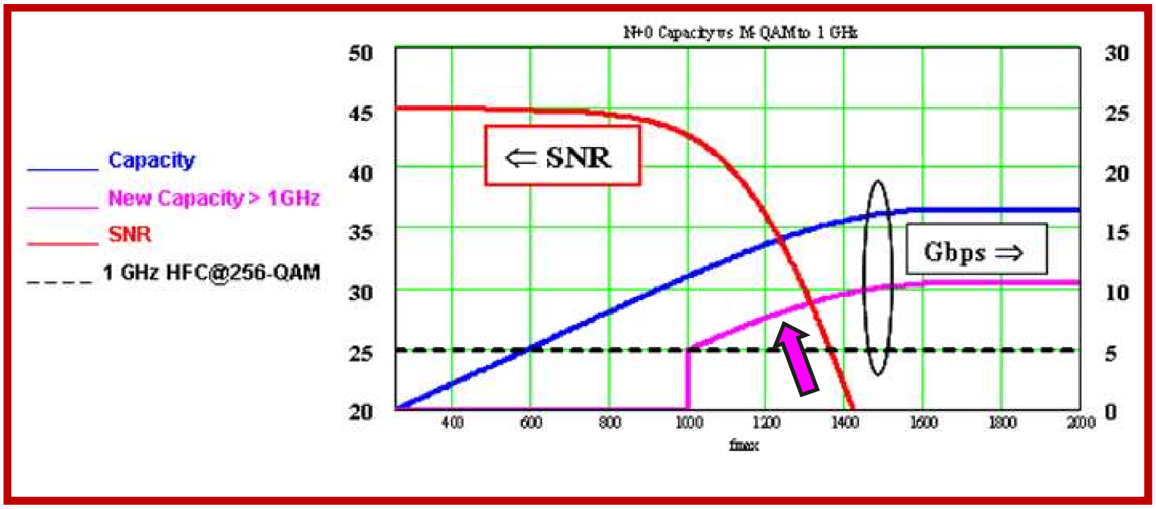


Figure 81 – N+0 Capacity vs. M-QAM to 1GHz

It is reasonable that an uptilt may be applied to compensate for the cable effect at least, but this would amount only to about 3 dB from one band edge to the other. Today's RF outputs are already tilted so as an extension of the payload this could be inherent.

The curves in Figure 81 show a full forward band throughput of 256-QAM , along with the theoretical capacity in Gbps (blue, right vertical axis), for a given maximum upper edge of the band shown on the x-axis. These capacities are shown along with the SNR vs. frequency delivered from a 5-tap cascade made up of taps such as that shown in Figure 78, and one coupled port from the same as shown in Figure 79.

The final trace (pink) recognizes the 256-QAM legacy spectrum as a given, already occupied bloc, and above that identifies new theoretical capacity potentially that can be exploited above 1 GHz in the passive segment as a function of the maximum upper frequency used.

Clearly, within the first 200 MHz above 1 GHz, more than a Gbps of capacity can be extracted. Also apparent is how much latent capacity still exists as the cascades shrink and open up new RF bandwidth potential, considering that 256-QAM is today's maximum modulation profile.

Of course, the expectation of 1024-QAM and perhaps even higher order modulations [1] are expected with the help of new FEC, allowing the "actual" to get closer to the capacity curve. Figure 81 also indicates that beyond 1.4 GHz there is diminishing return on new capacity as attenuation begins to take its toll on SNR.

For high SNR, such as those used in Figure 81, capacity is directly proportional to both bandwidth and SNR expressed in dB with very small error, a relationship observable in Figure 81.

10.2.3.3 Multicarrier Modulation Optimizes Channel Efficiency

Multicarrier techniques(OFDM)have made it possible to work through seriously impaired frequency response characteristics with high performance. As we observed in Section 7.3 “OFDMA, OFDM & LDPC”, the use of narrow subcarriers vastly simplifies the equalization function, and simultaneously provides the ability to consider each subcarrier independently in terms of the bandwidth efficiency of the modulation profile it can support on a dynamic basis.

Implementing multi-carrier technology for cable is a potentially attractive way to make use of the extended bandwidth of the coax, and because of this is a fundamental recommendation for the DOCSIS NG PHY. Much like xDSL before it, cable can leverage the powerful capabilities of OFDM techniques to most effectively use the current media, and this becomes more important as the use of the spectrum changes over time.

10.2.3.4 Excess Capacity Summary

In summary, here are plenty of available bits per second left to be exploited on the coax. It is expected that the DOCSIS NG PHY, using LDPC for most efficient use of SNR, and OFDM for most efficient use of unpredictable and changing bandwidth, will close the gap considerably on theoretical capacity over the HFC network. The most straightforward way to access this bandwidth is by continuing to migrate to fiber deeper, with a likely end state landing at an N+0 architecture of passive coax, and perhaps for practical purposes in some case N+1 or N+2.

Other useful elements of the migration include new RF technologies, such as GaN amplifiers that deliver more power at equivalent distortion performance can be used in multiple ways to enable this capacity to be accessed – allowing more economical deployment of N+0 long term (more hhp/node), using the additional RF drive capability to drive the new forward spectrum, or taking advantage of analog reclamation to deliver broadband performance based on QAM-only performance requirements.

Lastly, the same architectural option that delivers more capacity from the plant (N+0), bringing the last active and CPE closer together, works also from the receive end of the downstream link. Tied closely to optimal use of new spectrum is the ability to implement a point-of-entry (POE) home gateway architecture long-term.

This approach abstracts the HFC plant from inside the home, terminates downstream PHYs, delivers the bandwidth within the home on an IP network, and rids the access plant of having to overcome uncontrollable in-home losses and architectures.

10.2.4 Architectures for More Excess Bandwidth in The Passive Plant

As comforting as it might be that some plant segments already have some useable bandwidth above the specified top end of the equipment – used in some cases already for legacy extension – Figure 81 obviously behaves asymptotically because of the limitations of existing equipment. In the case evaluated above, it is due to the ultimate limitations of the 1 GHz Taps used in the analysis.

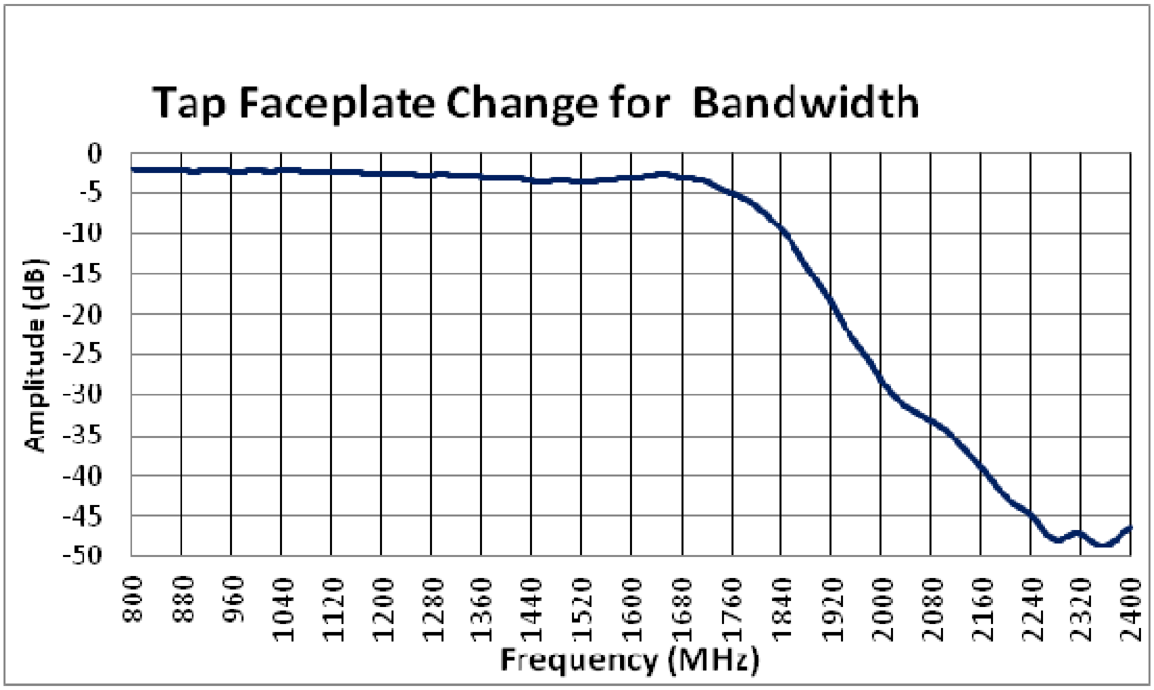


Figure 82 – Modifying Taps to Increase Bandwidth on the Passive Plant

If this limitation could be addressed, then the blue and pink curves shown in Figure 81 would continue to climb, providing access to more capacity, and with only the inherent coaxial attenuation contribution to shaping of the frequency response.

While there is little appetite for the intrusive nature and cost of exchanging all Taps in the plant, an elegant solution to freeing up more very useful spectrum is one that allows more spectrum without a wholesale cut-out of the existing Taps.

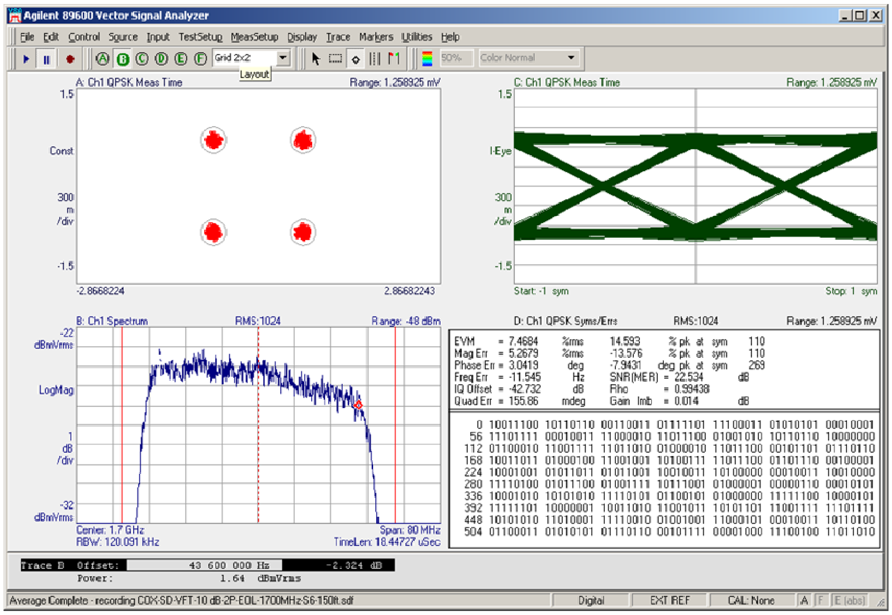
Tap models, such as those developed by Javelin, Inc., that allow for only a faceplate change of the existing Tap housing have been on the market to support this concept for some models of Taps in the field.

This is a much more simplified and time-efficient process for a field technician, and thus potentially a manageable option to operators looking for the sweet spot of "quick fix" versus bandwidth extraction. Wholesale change-outs can extend the Tap bandwidth to almost 3 GHz.

Figure 82 shows a frequency response of a sample Tap that has had its faceplate removed for the purpose of having the bandwidth extended.

Figure 82 shows a well-behaved passive response to 1.7 GHz. It is straightforward to estimate the additional capacity this provides using Figure 81. The first 200 MHz of spectrum added slightly less than 3 GHz of new capacity to the forward path. The additional 500 MHz shown in Figure 82 under the same assumption increases the total new capacity available to a little more than 10 Gbps theoretically.

This is a compelling number, as it immediately brings to mind the ability of the properly architected and engineered HFC plant to deliver GEPON-like speeds to its subscribers, without the need to build fiber-to-the-home. Indeed, as pointed out in [1], exploiting all of the available coaxial plant instead of just the legacy spectrum allows HFC to be directly competitive with FTTH rates and services.



```
EVM       =  7.4684    %rms     14.593     % pk  at  sym    110
Mag Err   =  5.2679    %rms    -13.576     % pk  at  sym    110
Phase Err =  3.0419     deg    -7.9431  deg pk  at  sym    269
Freq Err  =  -11.545     Hz    SNR(MER) =  22.534       dB
IQ Offset =  -42.732     dB    Rho      =  0.99438
Quad Err  =  155.86    mdeg    Gain Imb =  0.014        dB

  0 10011100 10110110 00110011 01111101 11100011 01010101 00010001
 56 11101111 00010011 11000010 11011100 01001010 10110110 10000000
112 01100010 11001111 11011010 01000010 11011100 00101101 01110110
168 10011011 01000100 11001001 10100111 11011100 01101110 00100001
224 10001001 01011011 01011001 10010011 10100000 00010011 10010000
280 11110100 01011100 01001111 10111001 01000001 00000110 00010101
336 10001010 10101010 11110101 01100101 01000000 11111100 10000101
392 11111101 10000001 10011010 11001011 10101101 11001111 11101111
448 10101010 11010001 11110010 01001001 11000101 00010011 10110100
504 01100011 01010101 01110110 00101111 00001000 11100100 11011010

Trace B   Offset:        43 600 000 Hz        -2.324 dB
          Power:              1.64   dBmVrms

Average Complete - recording COX-SD-VFT-10 dB-2P-EOL-1700MHz-S6-150ft.sdf
```

Figure 83 – Wideband (50 Msps) Characterization on Extended Tap BW

Even more simply, using just 1024-QAM, or one order of full modulation profile increase above 256-QAM (not full capacity), we need about 1.2 GHz of spectrum to aggregate to 10 Gbps of transport. Cable is not far from having the tools in place to achieve this already, and new LDPC FEC will make this actually quite simple to achieve.

Figure 83 shows a snapshot of the signal quality measured through an RF leg in the field made up of Taps of the type shown in Figure 82, transmitted *from* the end of a typical 150 ft drop cable (i.e. though passive, a measurement in the upstream direction).

There is some obvious droop at the band edge of this unequalized signal, with the drop cable contribution a primary culprit, but it is nonetheless easily corrected. The most important characteristic of Figure 83 has nothing to do with frequency response, but instead with the measured link loss from the end of the drop to the measurement station, sitting at the point where it would represent the first active in an N+0.

This is where "top split" architectures struggle to effective for return path applications. They must overcome in the 60 dB range – potentially worse when considering in-home variations – all tied simply to the relative attenuation characteristics of the low diplex band versus above 1 GHz.

The extended bandwidth taps relieve some of this through loss, but the impact on new CPE is significant in terms of generating broadband, linear, high RF outputs to overcome the loss and enable bandwidth efficient link budgets.

10.2.5 Summary

Many “1 GHz” Taps have significant, useable excess bandwidth above 1 GHz, although this is not guaranteed by specification. A practical cutoff point for family of Taps with the behavior shown in Figure 78 and Figure 79 for a 5-TAP cascade is between 1.16 GHz and 1.22 GHz.

It is expected that the same can be said above 750 MHz for “750 MHz” Taps and above 870 MHz for “870 MHz” Taps. However, because performance above 1 GHz is unspecified, different TAP models from different vendors are likely to vary in performance.

Faceplate replacement Taps represent a less-intrusive bandwidth extension option for the passive plant than 100% Tap replacement, and yield significant excess capacity.

The primary system issue is simply the RF loss entailed at these frequencies, and for this reason this capacity is most easily accessed for downstream use. The downstream channel already operates to 1 GHz, is highly linear across multiple octaves, delivers very high SNR for QAM, and is designed for broadband high power cost effectively to many users.

Each level of investment in bandwidth corresponds, as expected, to increased intrusiveness and operational expense. For some Tap models, there is virtually free bandwidth on the passive plant to at least 160 MHz above 1 GHz.

With the intrusiveness of a tap faceplate change, there is at least 700 MHz of new bandwidth made available. Finally, if all TAPs are completely replaced, bandwidth out to 2.75 GHz is freed up.

In all cases, standard 1 GHz HFC actives do *not* support the extended bands. And, in all cases, the rules governing RF loss versus frequency across the coaxial cable still exist and become the primary link budget obstacles to high order QAM transmission.

10.3 System Implications of HFC Evolution and Extended Bandwidth

There is already some flexibility in existing outdoor plant platforms. Modern nodes are very modular in nature and offer the flexibility to segment by port. Figure 84 shows the type of modularity most modern HFC nodes have today.

While amplifier platforms have seen less evolution than nodes in the past decade, there has been substantial investment in one area – fielded amplifiers today that can become nodes tomorrow through the swapping of internal plug-ins.

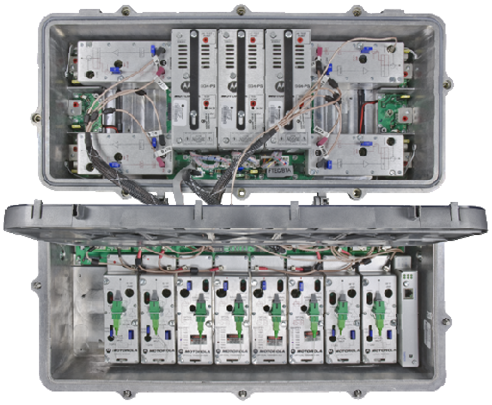


Figure 84 – Modern Node Platforms are Inherently Modular and Increasingly Flexible

This allows incremental bandwidth improvements as required within the context of the well-understood HFC infrastructure. Some suppliers have developed this capability for their entire RF amplifier portfolio, and it then becomes quite straightforward to envision at least a lower touch evolution to an N+0 deployment built around an existing plant.

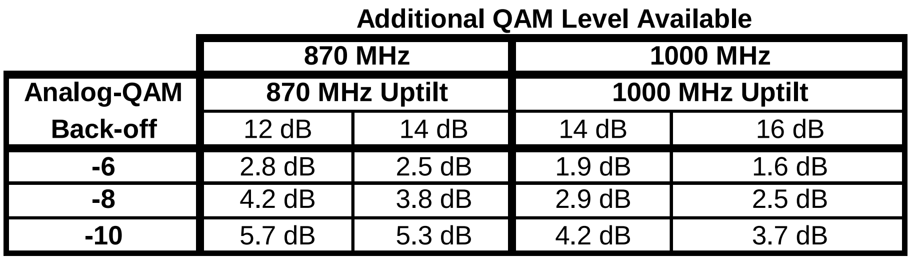
Taking the idea of node splitting to it logical conclusion, it ultimately leads to a natural N+0 end-state architecture. It is the final incarnation at which the coaxial cable last mile medium remains, leaving this passive part of the network and infrastructure investment in place.

Now, since these deeper nodes will correspond with adding bandwidth and average bandwidth is about serving group size, practical geography (subscribers don't always tend towards a uniform physical density) may dictate that an active element is still required. And, getting to an N+0 by successively splitting nodes repeatedly until there is nowhere else to go is probably not the most effective way to accomplishing the objective.

Plant geography and diminishing returns on average bandwidth per SG due to imbalance are likely to make this approach and less effective than a managed transition plan, and likely more costly as well.

Note that the march of nodes deeper into the network to N+0 leaves high similarity at the block diagram level to FTTC architectures used in the telco domain. Of course, there are significant differences in signal types on the fiber (at least for now), what is inside the node, and in the electrical medium – copper pair or coaxial. At some point, and possibly within the window of this fiber-deep evolution, the fiber delivery may become more common, leveraging 10 GbE or EPON technologies in both cases.

Table 39 – Total QAM Power with *All* Analog Removed



	Additional QAM Level Available			
	870 MHz		1000 MHz	
Analog-QAM	870 MHz Uptilt		1000 MHz Uptilt	
Back-off	12 dB	14 dB	14 dB	16 dB
-6	2.8 dB	2.5 dB	1.9 dB	1.6 dB
-8	4.2 dB	3.8 dB	2.9 dB	2.5 dB
-10	5.7 dB	5.3 dB	4.2 dB	3.7 dB

10.3.1 Bandwidth and Power Loading

The highest order deployed QAM modulation today is 256-QAM, which delivers a 1e 8 BER at a 34 dB SNR, ignoring coding gain improvements for simplicity. Meanwhile, a modest analog channel requirement is on the order of 45 dB – or 11 dB different.

Some of that large margin is eaten up in the relative signal level back-off, used on the QAM load. Use of 64-QAM levels 10 dB below analog and 256-QAM levels 6 dB below analog are common – and yet still leave significant SNR margin (7 dB and 5 dB in the examples given). These digital offsets can be used as tools in the RF power loading plan, to a degree.

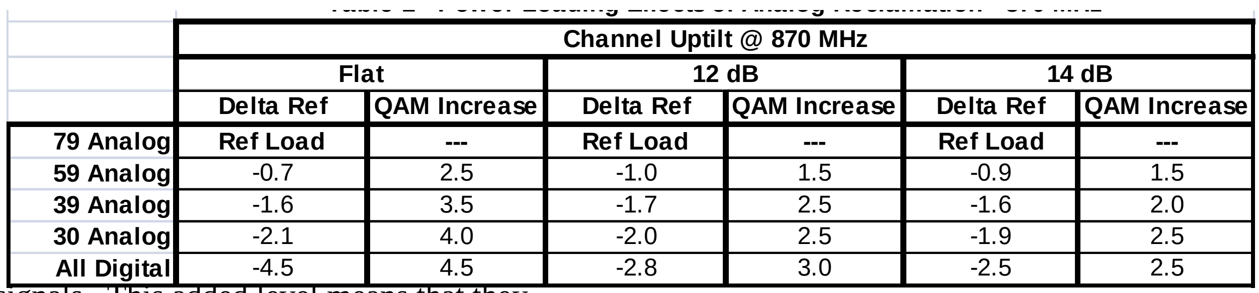
Because of the relationship between analog and digital power and their contribution to the total, when considering analog reclamation, additional power potentially becomes available for QAM signals. This added level means that they could absorb more attenuation from an SNR perspective.

Table 39 shows an example of the theoretically available increase in digital power on the multiplex, given that a fixed total RF output power is required for the mixed multiplex or for an all-digital load.

While this analysis is done for a full digital load, the analysis is easily adaptable to any number of analog carriers. For a small analog carrier count, the difference with "all-QAM" is relatively minor, because the limited set (such as 30) of analog channels are carried at the low end of the band, where their individual powers are smallest under commonly applied RF tilt. An example of stages of analog reclamation is shown in Table 40 for 870 MHz for comparison.

The case of "flat" would represent the change in the forward path multiplex sent across the optical link, while the uptilted cases represent the case out of the node or of

Table 40 – Power Loading Effects of Analog Reclamation - 870 MHz



	Channel Uptilt @ 870 MHz					
	Flat		12 dB		14 dB	
	Delta Ref	QAM Increase	Delta Ref	QAM Increase	Delta Ref	QAM Increase
79 Analog	Ref Load	---	Ref Load	---	Ref Load	---
59 Analog	-0.7	2.5	-1.0	1.5	-0.9	1.5
39 Analog	-1.6	3.5	-1.7	2.5	-1.6	2.0
30 Analog	-2.1	4.0	-2.0	2.5	-1.9	2.5
All Digital	-4.5	4.5	-2.8	3.0	-2.5	2.5

an amplifier where the RF level is tilted to compensate for cable attenuation versus frequency. Typically, it is the optical link which sets HFC SNR, and the RF amplifier cascade that is the dominant contributor to distortions.

What is clear from Table 39 and Table 40 are that the process of analog reclamation offers the potential; for SNR recovery. In the case of beginning with 79 analog slots and migrating to an all digital line-up, there is 4.5 dB of increased digital level available per carrier into the optical transmitter in theory, which can be converted to a better digital SNR.

10.3.2 Extended Bandwidth Loading

If the use of coax is to be extended to frequencies above 1 GHz, power loading will be affected accordingly for non-RF overlay approaches. For the sake of simplicity, we consider two cases:

1) Assume that the applied tilt will be required to extend this band according to the coaxial relationship previously discussed

2) Consider a flat signal band is delivered in the 1-1.5 GHz range, and new technology is burdened with overcoming the limitations of higher attenuation

We will use 1.5 GHz to be consistent with the above discussion on capacity and tap bandwidths. Example cases under these assumptions are shown in Table 41, which illustrates some key points. The starting point is the 1 GHz reference load of sufficient level and performance.

From a power loading standpoint, continuing the tilted response to 1.5 GHz adds a significant power load. However, variations to the tilt approach create a seemingly manageable situation (small dB's) from a power handling standpoint. Hybrids today are typically designed, through their external circuit implementations, to purposely roll-off.

Several 1-1.5 GHz RF loading implementations in Table 41 are relatively non-stressful. If the 1 1.5 GHz band is flat, the additional power load is between 0.4 dB to 3.9 dB. In the situation where the band is extended to 1.5 GHz in conjunction with analog reclamation leaving 30 channels in analog, the increase in total load is limited to 1.2 dB.

In order to maintain a tilted output to 1.5 GHz, an overall digital band de-rate of -10 dB instead of 6 dB keeps the power load hit to less than 4 dB. Given that this may be accompanied by perhaps anN+0 architecture,

Table 41 – Power Loading of Extended Bandwidth

	Analog BW	Digital BW	Digital Derate Relative to Analog (dB)		Digital BW Tilt (dB)		Relative Pwr
	MHz	MHz	550 MHz-1GHz	1-1.5 GHz	550 MHz-1GHz	1-1.5 GHz	dB
Reference	**550**	**450**	**-6**	**Unused**	**14**	**Unused**	**0.0**
Case 1	550	450	-6	-6	14	14	7.4
Case 2	550	450	-10	-10	14	14	3.9
Case 3	550	450	-6	-6	14	0	3.9
Case 4	550	450	-10	-10	14	0	0.9
Case 5	550	450	-6	-15	14	14	2.0
Case 6	265	735	-6	-6	14	14	7.2
Case 7	265	735	-10	-10	14	14	3.3
Case 8	265	735	-6	-6	14	0	1.2
Case 9	265	735	-10	-10	14	0	0.4
Case 10	265	735	-6	-13	14	14	2.2

the 4 dB of power may be available while maintaining sufficient performance because no noise and distortion margin needs to be left for an amplifier cascade. This approach may be more costly in terms of added power, but it is more straightforward to implement a uniform frequency response in a single circuit, than one that tilts part of the band but not another.

A final set of cases that show reasonable loading increase are the 79 channel and 30 channel cases with the tilt maintained, but new derate applied in the 1-1.5 GHz band. To maintain a load increase of <2 dB, an additional 9 dB and 7 dB derate should be applied for 79 and 30 channels, respectively. However, considering the link budgets associated with HFC networks today, dropping the levels this low likely creates a challenge to most efficiently using this band, as this would is then lost SNR and lower capacity.

Summarizing, it appears that various implementation scenarios are eligible for maintaining a reasonable power loading situation while extending the band of the output to 1.5 GHz. This does not account for possible changes in hybrid capability for an extended band. The hybrids themselves have bandwidth up to 1.5 GHz, but the circuits they are designed into are purposefully limiting and optimized for today noise and distortion requirements over legacy bandwidths.

Table 42 – Performance Effects of N+6 to N+0 Conversion

Performance of 1 GHz Multiplex with 79 Analog			
Parameter	**N+6**	**N+0 (nom)**	**N+0 (high)**
CCN	48	51	51
CSO	56	64	62
CTB	58	70	67

10.3.3 Reduced Cascade Benefits

It is well-understood cable math how shorter cascades result in higher SNR and lower distortion, as the link degradation of adding a relatively short length of fiber is a favorable trade-off with a run of active and passive coaxial plant.

Let's look at a typical example and evaluate this cascade shortening impact. In this case, the link is a 1310 nm link in an N+6 configuration in its original state, and the noise and distortion performance calculated for a 1 GHz multiplex of 79 analog channels.

The link is then modified to an N+0, and the analysis re-run at the same nominal output levels. It was also run for a 4 dB increased output level mode, as the extension to N+0 architectures today may entail a higher output requirement to accommodate the likelihood that the plant geography is not well suited to 100% N+0, and recognizing that the removal of the RF cascade gives distortion margin back that may allow higher output levels. The results of this analysis are shown in Table 42.

Note the emergence of 3-4 dB of additional SNR (CCN or Composite Carrier to Noise). This is independent of any SNR gain due to increasing digital levels that may be possible with analog reclamation per Table 39.

Increasing QAM levels while adding QAM in place of analog is not a fixed dB-per-dB SNR gain, as adding digital channels adds contributors to CCN (composite carrier-to-noise). However, this conversion to CCN also creates a significant drop in CSO and CTB distortions, which are significant impairments for higher order QAM

Table 43 – Performance Effects of N+6 to N+0 Conversion

Performance of 1 GHz Multiplex with 30 Analog			
Parameter	**N+6**	**N+0 (nom)**	**N+0 (high)**
CCN	48	52	52
CSO	67	70	70
CTB	68	74	73

performance [1].

Table 43 shows the same parameter set and HFC architecture as used in Table 42, but with an analog channel count of 30. Note the significant improvements in analog beat distortions, as well as the SNR (CCN) behavior. Clearly, the added digital distortion that contributes to CCN is mitigated by the improvements obtained by eliminating the cascade effects.

10.4 Importance of the CPE in the DOCSIS NG Migration Plan

We are proposing that DOCSIS NG have a minimum of two (2) PHYs and a common MAC across these independent PHYs. These PHYs will be at least one of the existing DOCSIS 3.0 upstream PHYs and the downstream PHY. In addition there will be a modern PHY. The placement of DOCSIS NG CPEs in the homes that have both DOCSIS 3.0 and DOCSIS NG PHY provides an evolutionary migration strategy.

This will allow the MSO to use the legacy DOCSIS 3.0 PHYs while the cable operator grows the installed base of DOCSIS NG CPEs in their subscriber homes. At such time there are sufficient numbers of DOCSIS NG CPE deployed, the MSO may allocate a few channels to the new DOCSIS NG PHY.

By supporting legacy and modern PHYs within the same CM, the MSOs can smoothly transition to the modern PHY as the legacy CPEs decrease in numbers.

11 RECOMMENDATIONS

This section summarizes the recommendation of the authors. A more extensive explanation of each decision can be found the in the rest of this white paper.

11.1 Areas of Consensus

Compatibility

The recommendation is to define a backwards compatibility goal that would allow the same spectrum to be used for current DOCSIS CMs and new DOCSIS NG CMs.

In this context, co-existence refers to the concept that DOCSIS NG would use separate spectrum but coexist on the same HFC plant. Backwards compatibility would refer to the sharing of spectrum between current DOCSIS and DOCSIS NG.

One example of this strategy would require a 5 to 42 MHz spectrum to be used for four carriers (or more) of DOCSIS 3.0. At the same time, a DOCSIS NG CM would be able to use the same four channels (or more) plus any additional bandwidth that a new PHY might be able to take advantage of.

Upstream Spectrum

The immediate goal with DOCSIS NG is to get as much throughput as possible in the existing upstream 5 to 42 MHz (5 to 65 MHz) spectrum.

This goal recognizes that it will take time, money, and effort to upgrade the HFC plant. The initial goal will to see how more advanced CMTS and CM technology can extend the life of the current HFC plant.

The short-term recommendation for upstream spectrum is mid-split.

Mid-split can be achieved with today's DOCSIS 3.0 technology. If an HFC plant upgrade strategy could be defined that would allow a cost effective two-stage upgrade, first to mid-split, and then later to high-split, then the advantage of higher data rates can be seen sooner.

Conversely, if downstream spectrum is available, an HFC plant could be upgraded to high-split sooner, but would start by deploying mid-split DOCSIS 3.0 equipment.

The long-term recommendation for upstream spectrum is high-split.

High-split offers the best technical solution that should lead to the highest performance product at the best price. The logistical challenges that high-split encounters are not to be underestimated but they are both solvable and manageable, and significantly less imposing than a “top-split” approach.

Downstream Spectrum

The short term goal is to make use of any and all available tools to manage downstream spectrum congestion, such as analog reclamation, SDV, H.264 and deploy 1 GHz plant equipment whenever possible.

This goal includes an expanded upstream spectrum within the current operating spectrum of the HFC plant.

The long-term goal is to utilize spectrum above 1 GHz, and push towards 1.7 GHz.

Field measurements have shown that the spectrum up to 1.2 GHz is available in the passive RF link. Measurements also show that up to 1.7 GHz is available with modest plant intrusiveness. Spectrum above 1 GHz is unspecified, and inherently more challenging than the standard HFC band and thus should take advantage of advanced modulation techniques such as OFDM.

New US PHY Layer

The recommendation for DOCSIS NG upstream is to add OFDMA with an LDPC FEC.

There is considerable new spectrum with DOCSIS NG that only requires a single modulation. Although ATDMA and SCDMA could be extended, now is a unique time to upgrade the DOCSIS PHY to include the best technology available, which the team feels is OFDMA and LDPC FEC.

New DS PHY Layer

The recommendation for DOCSIS NG downstream is to add OFDM with LDPC FEC.

Using the spectrum above 1 GHz will require an advanced PHY such as OFDM. To minimize the cost impact on CMs, a cap could be placed on the number of QAM channels required. OFDM will also be used below 1 GHz, and likely supplant legacy QAM bandwidth over time.

PAPR

We do not anticipate PAPR issues with multicarrier modulation for the upstream or the downstream when compared with single carrier channel bonded DOCSIS.

It is recognized that PAPR for multi-carrier technologies such as OFDM is worse than a single isolated QAM carrier. However, as the number of SC-QAMs in a given spectrum are increased, multiple SC-QAM and OFDM exhibit similar Gaussian characteristics.

Higher Orders of Modulation

The recommendation is to study the option to define up to 4K QAM for OFDM in both the upstream and downstream.

These new modulations may not be usable today. However, as fiber goes deeper coax runs become shorter, and other possible architectural changes are considered (POE home gateway, digital optics with remote PHY), there may be opportunities to use higher orders of modulation. The DOCSIS NG PHY will define these options.

SCDMA Support in a DOCSIS NG CM

The recommendation is to not require SCDMA in a DOCSIS NG CM that employs OFDMA

It is generally agreed that OFDMA with LDPC will be able to replace the role that SCDMA and ATDMA perform today. Thus, in a DOCSIS NG CM, SCDMA would be redundant.

US MAC Layer Baseline

The recommendation is to use the SCDMA MAC functionality as a basis for designing the OFDMA MAC layer.

The SCDMA MAC layer is very similar to the ATDMA MAC layer that has allowed upstream scheduling and QOS services to be near seamless between the two current modulations. This structure is to be extended over OFDM so that the new PHY has a less impact on the rest of the DOCSIS system.

11.2 Areas of Further Study

Some of these decisions require additional information. Some of these decisions have most of the required information and just lack consensus.

High-Split Cross-Over Frequencies

Further study is required to determine the upper frequency of the high-split upstream spectrum and the lower frequency of the downstream spectrum.

At this time, we are not sure the right choice of upstream band edge to achieve 1 Gbps throughput with satisfactory coverage and robustness. This will depend upon the base modulation chosen, FEC overhead, and if there are any areas of spectrum that cannot be used. There will likely be a reference configuration that will pass 1 Gbps and other configurations that will run slower or faster.

There may even be a set of frequencies that matches a 1.0 GHz HFC plant, and a different set of frequencies that matches up to a 1.7 GHz HFC Plant.

There may also be the ability to configure the cross-over frequency in the HFC plant so that it can be changed over time with shifts in traffic patterns. Similar flexibility in the CM could also be considered.

ATDMA in the Upstream

Further study is required to determine how may ATDMA channels a CM and a CMTS should support in the upstream.

Many cable operators are already deploying three full-width carriers or four carriers of mixed widths between 20 MHz and 42 MHz. In order to fully utilize a 5 to 42 MHz spectrum, a DOCSIS NG CM would need to support these channels, so four is the minimum. Newer DOCSIS 3.0 CMs promise 8 upstream channels. It depends upon the market penetration of these CMs as to the impact on backwards compatibility.

Some networks may have migrated to an 85 MHz mid-split before any DOCSIS NG CMs are available, and these would then be A-TDMA channels. Timing of such activity might define minimum channel requirements for the NG CM.

The CMTS may need more QAM channels than the CM. The CMTS needs to have a spare ATDMA channel to support DSG. It also needs to have an ATMDA channel running at a lower rate to support DOCSIS 1.1 CMs. These may be in addition to the 3-4 channels for DOCSIS 3.0.

SCDMA in the CMTS

Further study is required to determine if SCDMA should be retained.

It is generally agreed that SCDMA does offer better performance below 20 MHz (in North America, higher in other countries with worse plant) than ATDMA. For DOCSIS 3.0, SCDMA may be required to get that extra fourth full-size carrier, and is an important component for maximizing the throughput available in 5-42 MHz band.

Retaining SCDMA in addition to ATDMA and OFDMA potentially adds product cost, development cost, and testing cost. This has to be weighed against any significant market penetration of SCDMA prior to DOCSIS NG being available.

One possible approach is to specify a small number of channels of SCDMA as mandatory and more channels optional. However, an overall objective is to try and get to only one or two PHY technologies in the CMTS silicon that would imply the elimination of SCDMA.

Early deployment of mid-split would also help negate the need for SCDMA, as that would provide the extra spectrum to relieve the congestion in 5-42 MHz

Advanced FEC for Single Carrier Systems

Further study is required to determine if LDPC FEC functionality should be added to enhance the existing upstream and downstream PHY.

The argument for doing this is that the bulk of new capacity comes from advanced FEC, and existing SC QAM that co-exists on the silicon should benefit from this investment to optimize efficiency in systems that will be operating single carrier mode for many more years. The argument for not doing this is to cap the legacy design and only expand capability with OFDM.

Expansion of Upstream ATDMA Capabilities

Further study is required to determine if ATDMA functionality should be extended with wider channels, more channels, higher order modulation formats, and improved alpha.

The argument for doing this work is that they represent simple extensions of DOCSIS 3.0, and field experience and RF characterization of A-TDMA tools suggests a high probability of success. The argument for not doing this is to cap the legacy design and only expand capability with OFDM, and that an OFDM implementation would be less complex.

Expansion of Downstream QAM Capabilities

Further study is required to determine if downstream QAM functionality, currently defined by ITU-T J.83, should be extended with wider channels and higher order modulation formats.

The argument for doing this work is that they represent simple extensions of DOCSIS 3.0 and field experience and characterization of A-TDMA SC tools suggests a high probability of success. The argument for not doing this is to cap the legacy design and focus on expanding capability only with OFDM, and that an OFDM implementation would be less complex.

US MAC Improvements

Further study is required to determine if any changes not directly related to OFDM are worth pursuing.

Current suggestions include changing the request mechanism from request-based to queue-based, elimination of 16-bit minislots, and not including request slots on each upstream carrier.

Modifications need to be weighed against increases in performance, decrease in cost, and the need for backwards compatibility.

12 ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of others.

From Cisco, the authors would like to thanks Ron Hranac, Fred Hirt, John Ritchie, and Bill Wall for worrying us about the aeronautical band and GPS frequencies.

From Arris, the authors would like to thank Scott Shupe, Dr. Derald O. Cummings, Dr. Tom Cloonan, Dr. Frank O'Keeffe, Dr. Zoran Maricevic, Dr. Marcel Schemmann, Brent Arnold, Bill Dawson, Venk Mutalik, Jeff Howe, and Dr. Ayham Al-Banna for their contributions.

From Motorola, the authors would like to thank Jack Moran, Rob Thompson, John Ulm, Gerry White, Dr. Amarildo Vieira, Mark Schmidt, Mike Aviles, Steve Frederick, Phil Miguelez, Dean Stoneback, Vipul Rathod, Chuck Moore, Mike Cooper, John Holobinko, and Dr. Sebnem Ozer.

From Intel, the authors would like to thank Dr. Bernard Arambepola who provided the OFDM simulation results.

13 REFERENCES

[DOCSIS 2.0]
CM-SP-RFIv2.0-C01-081104, DOCSIS Radio Frequency Interface Specification, CableLabs, August 11, 2004

[DOCSIS DRFI]
CM-SP-DRFI-I11-110210, DOCSIS Downstream RF Interface Specification, CableLabs, issued Feb 10, 2011

[DOCSIS MACUP]
CM-SP-MULPIv3.0-I15-110210, DOCSIS 3.0 MAC and Upper Layer Protocols Interface Specification, CableLabs, Issued Feb 2, 2011

[1] Dr. Robert L., Michael Aviles, and Amarildo Vieira, "New Megabits, Same Megahertz: Plant Evolution Dividends," 2009 Cable Show, Washington, DC, March 30-April 1

[2] White, Gerry and Mark Schmidt, 64-, 256-, and 1024-QAM with LDPC Coding for Downstream, DOCSIS 3.0 Proposal, Motorola Broadband Communications Sector, Dec 10, 2004.

[3] Howald, R., Stoneback, D., Brophy, T. and Sniezko, O., Distortion Beat Characterization and the Impact on QAM BER Performance, NCTA Show, Chicago, Illinois, June 13-16, 1999.

[4] Prodan, Dr. Richard, "High Return Field Test: Broadcom Frequency Response Model vs. Motorola Data," Report to CableLabs AMP Committee, April 11, 2011.

[5] Moran, Jack, "CableLabs High Frequency Return Path Passive Coaxial Cable Characterization Results," AMP Report, Nov 18, 2011

[6] Howald, Dr. Robert L, Fueling the Coaxial Last Mile, SCTE Conference on Emerging Technologies, Washington DC, April 2, 2009.

[7] Finkelstein, Jeff, "Upstream Bandwidth Futures," 2010 SCTE Cable-Tec Expo, New Orleans, LA, Oct. 20-22.

[8] Howald, Dr. Robert L., Phillip Chang, Robert Thompson, Charles Moore, Dean Stoneback, and Vipul Rathod, "Characterizing and Aligning the HFC Return Path for Successful DOCSIS 3.0 Rollouts," 2009 SCTE Cable-Tec Expo, Denver, CO, Oct 28-30.

[9] Howald, Dr. Robert L., "Maximizing the Upstream: The Power of S-CDMA" Communication Technology Webcast, Sept. 9, 2009.

[10] Howald, Dr. Robert L. and Michael Aviles, "Noise Power Ratio the Analytical Way," 2000 NCTA Show, New Orleans, LA.

[11] Miguelez, Phil, and Dr. Robert Howald, "Digital Due Diligence for the Upstream Toolbox," 2011 Cable Show, Chicago, IL, June 14-16.

[12] Dr. Robert Howald and Phil Miguelez, "Upstream 3.0: Cable's Response to Web 2.0," The Cable Show Spring Technical Forum, June 14-16, 2011, Chicago, Il..

[13] Stoneback, Dean, Dr. Robert L. Howald, Joseph Glaab, Matt Waight, "Cable Modems in the Home Environment," 1998 NCTA Show, Atlanta, Ga.

[14] Ulm, John, Jack Moran, Daniel Howard, "Leveraging S-CDMA for Cost Efficient Upstream Capacity," SCTE Conference on Emerging Technologies, Washington DC, April 2, 2009.

[15] Thompson, Robert, Jack Moran, Marc Morrissette, Charles Moore, Mike Cooper, Dr. Robert L. Howald, and "64-QAM, 6.4MHz Upstream Deployment Challenges," 2011 SCTE Canadian Summit, Toronto, ON, Mar 8-9.

[16] Thompson, Rob, "256-QAM for Upstream HFC", NCTA Cable Show, Los Angeles, CA, May 2010

[17] Thompson, Rob, Jack Moran, Marc Morrissette, Chuck Moore, and Dr. Robert Howald, "256-QAM for Upstream HFC Part II", The Cable Show, Atlanta, Ga, NOV 2012

[18] Woundy, Richard, Yiu Lee, Anthony Veiga, Carl Williams, "Congestion Sensitivity of Real-Time Residential Internet Applications", 2010 SCTE Cable-Tec Expo, New Orleans, LA, Oct. 20-22.

[19] North American Cable Television Frequencies, Wikipedia, http://en.wikipedia.org/wiki/North_American_cable_television_frequencies

[20] CEA-542-C, Cable Television Channel Identification Plan, Consumer Electronics Association, Feb 2009

[21] Chapman, John T, "Taking the DOCSIS Upstream to a Gigabit per Second", NCTA Spring Technical Seminar, Los Angeles, May, 2010.

[22] Chapman, John T., "What the Future Holds for DOCSIS", Keynote speech, Light Reading Conference, Denver, May 18, 2012

[23] Department of Commerce, "United States Radio Spectrum Frequency Allocations Chart", 2003, United States Department of Commerce, http://www.ntia.doc.gov/osmhome/allochrt.pdf

[24] "Capture Effect", Wikipedia, http://en.wikipedia.org/wiki/Capture_effect

[25] SCTE 40 2011, "Digital Cable Network Interface Standard", Society of Cable Telecommunications Engineers, Inc, 2011

[26] "Instrument Landing System", Wikipedia, http://en.wikipedia.org/wiki/Instrument_landing_system

[27] David J.C. MacKay and Edward A. Ratzer, "Gallager Codes for High Rate Applications" published January 7, 2003,

[28] ETSI EN 302 769: "Digital Video Broadcasting; Frame Structure, Channel Coding and Modulation for a Second Generation Digital Transmission System for Cable Systems"

[29] United States Nuclear Detonation Detection System (USNDS), http://www.fas.org/spp/military/program/nssrm/initiatives/usnds.htm

[30] Committee on Radio Astronomy Frequencies, http://www.craf.eu/gps.htm

[31] Grace Xingxin Gao, "Modernization Milestone: Observing the First GPS Satellite with an L5 Payload", Inside GNSS Magazine, May/June 2009, www.insidegnss.com/auto/mayjune09-gao.pdf

[32] Dennis M. Akos, Alexandru Ene, Jonas Thor, "A Prototyping Platform for Multi-Frequency GNSS Receivers", Standford University

[33] Emmendorfer, Michael J, Shupe, Scott, Maricevic, Zoran, and Cloonan, Tom, "Next Generation – Gigabit Coaxial Access Network" 2010 NCTA Cable Show, Los Angeles, CA, May 2010

[34] Emmendorfer, Michael J, Shupe, Scott, Cummings Derald, and Cloonan, Tom, "Next Generation - Cable Access Network, An Examination of the Drivers, Network Options, and Migration Strategies for the All-IP Next Generation – Cable Access Network", 2011 Spring Technical Forum, Chicago, IL June 14-16

[35] Emmendorfer, Michael J, Shupe, Scott, Cummings Derald, Cloonan, Tom, and O'Keeffe, Frank "Next Generation - Cable Access Network (NG-CAN), Examination of the Business Drivers and Network Approaches to Enable a Multi-Gigabit Downstream and Gigabit Upstream DOCSIS Service over Coaxial Networks", 2012 SCTE Canadian Summit March 27-28, Toronto, Canada.

[36] Emmendorfer, Michael J, Shupe, Scott, Maricevic, Zoran, and Cloonan, Tom, "Examining HFC and DFC (Digital Fiber Coax) Access Architectures, An examination of the All-IP Next Generation Cable Access Network," 2011 SCTE Cable-Tec Expo, New Atlanta, GA, Nov. 14-17.

MPEG DASH: A Technical Deep Dive and Look at What's Next

Andy Salo
RGB Networks

Abstract

The MPEG DASH standard was ratified in December 2011 and published by the International Organization for Standards (ISO) in April 2012. This paper will review the technical aspects of the new MPEG DASH standard in detail, including: how DASH supports live, on-demand and time-shifted (NDVR) services; how the two primary video formats – ISO-base media file format (IBMFF) and MPEG-2 TS – compare and contrast; how the new standard supports digital rights management (DRM) methods; and how Media Presentation Description (MPD) XML files differ from current adaptive streaming manifests. In addition, the paper will discuss how MPEG DASH is likely to be adopted by the industry, what challenges must still be overcome, and what the implications could be for cable operators and other video service providers (VSPs).

INTRODUCTION

For much of the past decade, it was quite difficult to stream live video to a mobile device. Wide bandwidth variability, unfavorable firewall configurations and lack of network infrastructure support all created major roadblocks to live streaming. Early, more traditional streaming protocols, designed for small packet formats and managed delivery networks, were anything but firewall-friendly. Although HTTP progressive download was developed partially to get audio and video streams past firewalls, it still didn't offer true streaming capabilities.

Now, the advent of adaptive streaming over HTTP technology has changed everything, reshaping video delivery to PCs, laptops, game consoles, tablets, smartphones and other mobile devices, as well as such key home devices as Web-connected TVs and pure and hybrid IP set-top boxes (STBs). As a result, watching video online or on the go is no longer a great novelty, nor is streaming Internet-delivered content to TV screens in the home. Driven by the explosion in video-enabled devices, consumers have swiftly moved through the early-adopter phase of TV Everywhere service, reaching the point where a growing number expect any media to be available on any device over any network connection at any time. Increasingly, consumers also expect the content delivery to meet the same high quality levels they have come to know and love from traditional TV services.

Even though the emergence of the three main adaptive streaming protocols from Adobe, Apple and Microsoft over the past three and a half years has made multiscreen video a reality, significant problems still remain. Each of the three proprietary platforms is a closed system, with its own manifest format, content formats and streaming protocols. So, content creators and equipment vendors must craft several different versions of their products to serve the entire streaming video market, greatly driving up costs and restricting the market's overall development.

In an ambitious bid to solve these nagging problems, MPEG has recently adopted a new standard for multimedia streaming over the Internet. Known as MPEG Dynamic Adaptive Streaming over HTTP, or MPEG DASH, the new industry standard attempts to create a universal delivery format for streaming media by incorporating the best elements of the three main proprietary streaming solutions. In doing so, MPEG DASH aims to provide the long-sought interoperability between different network servers and different consumer electronics devices, thereby fostering a common ecosystem of content and services.

This paper will review the technical aspects of the new MPEG DASH standard in detail, including: how DASH supports live, on-demand and time-shifted (NDVR) services; how the two primary video formats (ISO-base media file format (IBMFF) and MPEG-2 TS) compare and contrast; how the standard supports DRM methods; and how Media Presentation Description (MPD) XML files differ from current adaptive streaming manifests. In addition, the paper will discuss how MPEG DASH is likely to be adopted by the industry, what challenges must still be overcome, and what the implications could be for cable operators and other video service providers (VSPs).

AN ADAPTIVE STREAMING PRIMER

As indicated previously, the delivery of streaming video and audio content to consumer electronics devices has come a long way over the past few years. Thanks to the introduction of adaptive streaming over HTTP, multimedia content can now be delivered more easily than ever before. In particular, adaptive streaming offers two critical features for video content that have made the technology the preferred choice for mobile delivery.

First, adaptive streaming over HTTP breaks down, or segments, video programs into small, easy-to-download chunks. For example, Apple's HTTP Live Streaming (HLS) protocol typically segments video content into 10-second chunks, while Microsoft's Smooth Streaming (MSS) protocol and Adobe's HTTP Dynamic Streaming (HDS) usually break video content into even smaller chunks of five seconds or less.

Second, adaptive streaming encodes the video content at multiple bitrates and resolutions, creating different chunks of different sizes. This is the truly 'adaptive' part of adaptive streaming, as the encoding enables the mobile client to choose between various bitrates and resolutions and then adapt to larger or smaller chunks automatically as network conditions keep changing.

In turn, these two key features of adaptive streaming lead to a number of benefits:

1. Video chunks can be cached by proxies and easily distributed to content delivery networks (CDNs) or HTTP servers, which are simpler and cheaper to operate than the special streaming servers required for 'older' video streaming technologies.

2. Bitrate switching allows clients to adapt dynamically to network conditions.

3. Content providers no longer have to guess which bitrates to encode for end devices.

4. The technology works well with firewalls because the streams are sent over HTTP.

5. Live and video-on-demand (VoD) workflows are almost identical. When a service provider creates a live stream, the chunks can easily be stored for later VoD delivery.

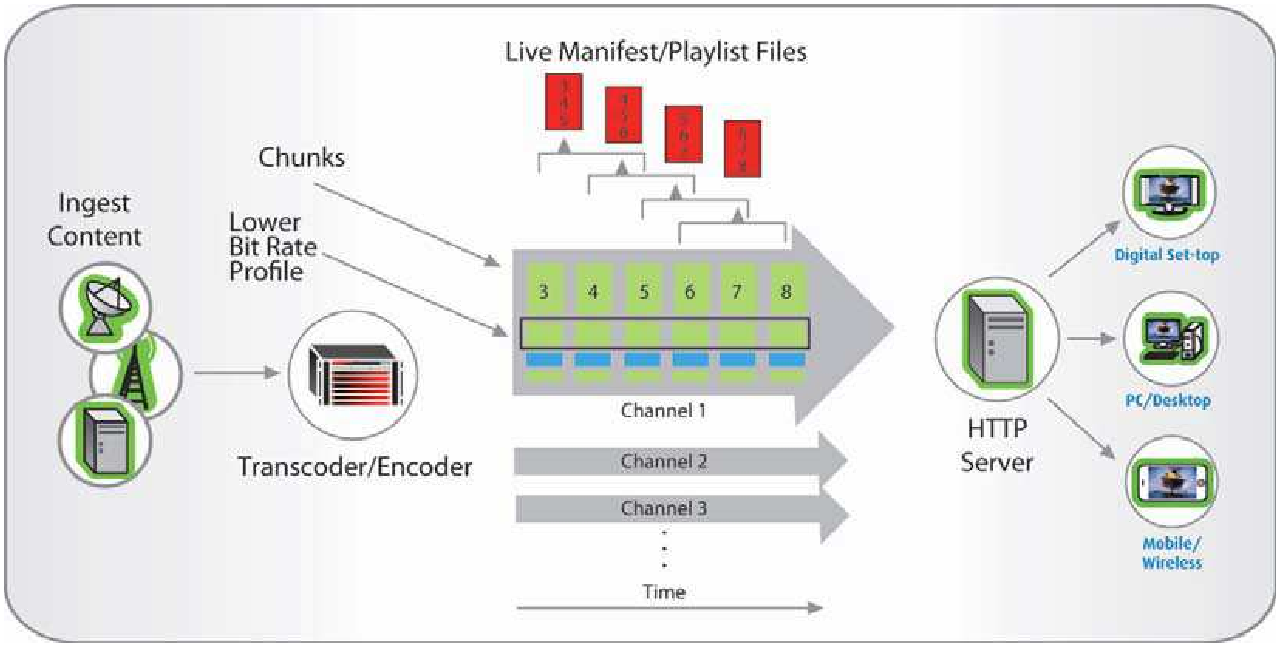


Figure 1: Content Delivery Chain for Live Adaptive Streaming

Sensing the promise of adaptive streaming technology, several major technology players have sought to carve out large shares of the rapidly growing market. Most notably, the list now includes such prominent tech companies as Adobe, Apple and Microsoft.

While the streaming of video using HTTP-delivered fragments goes back many years (and seems lost in the mists of time), Move Networks caught the attention of several media companies with its adaptive HTTP streaming technology in 2007. Move was quickly followed by Microsoft, which entered the market by releasing Smooth Streaming in October 2008 as part of its Silverlight architecture. Earlier that year, Microsoft demonstrated a prototype version of Smooth Streaming by delivering live and on-demand streaming content from such events as the Summer Olympic Games in Beijing and the Democratic National Convention in Denver.

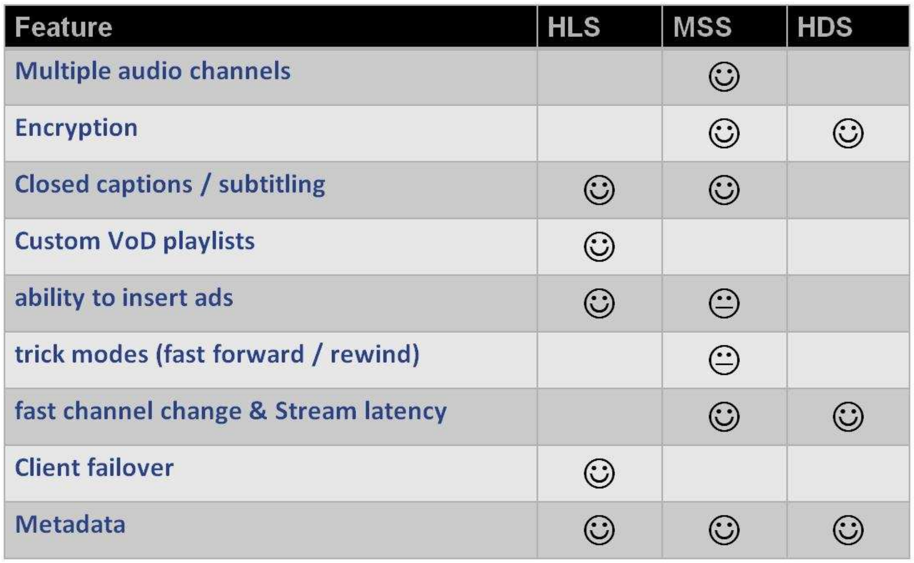
Smooth Streaming has all of the typical characteristics of adaptive streaming. The video content is segmented into small chunks and then delivered over HTTP. Usually, multiple bitrates are encoded so that the client can choose the best video bitrate to deliver an optimal viewing experience based on network conditions.

Apple came next with HLS, originally unveiling it with the introduction of the iPhone 3.0 in mid-2009. Prior to the iPhone 3, no streaming protocols were supported natively on the iPhone, leaving developers to wonder what Apple had in mind for native streaming support. In May 2009, Apple proposed HLS as a standard to the Internet Engineering Task Force (IETF), and the draft is now in its eighth iteration.

HLS works by segmenting video streams into 10-second chunks; the chunks are stored using a standard MPEG-2 transport stream file format. The chunks may be created using several bitrates and resolutions – so-called profiles – allowing a client to switch dynamically between different profiles, depending on network conditions.

Adobe, the last of the Big Three, entered the adaptive streaming market in late 2009 with the announcement of HTTP Dynamic Streaming (HDS). Originally known as “Project Zeri,” HDS was introduced in June 2010. Like MSS and HLS, HDS breaks up video content into small chunks and delivers them over HTTP. Multiple bitrates are encoded so that the client can choose the best video bitrate to deliver an optimal viewing experience based on network conditions.

HDS is closer to Microsoft Smooth Streaming than it is to Apple’s HLS protocol. Primarily, this is because HDS, like MSS, uses a single aggregate file from which MPEG-4 container fragments are extracted and delivered. In contrast, HLS uses individual media chunks rather than one large aggregate file.



Feature	HLS	MSS	HDS
Multiple audio channels		☺	
Encryption		☺	☺
Closed captions / subtitling	☺	☺	
Custom VoD playlists	☺		
ability to insert ads	☺	😐	
trick modes (fast forward / rewind)		😐	
fast channel change & Stream latency		☺	☺
Client failover	☺		
Metadata	☺	☺	☺

Figure 2: Feature Comparison of Three Major Adaptive Streaming Platforms

THE DUELING STREAMING PLATFORM PROBLEM

The three major adaptive streaming protocols – MSS, HLS and HDS – have much in common. Most importantly, all three streaming platforms use HTTP streaming for their underlying delivery method, relying on standard HTTP Web servers instead of special streaming servers. They all use a combination of encoded media files and manifest files that identify the main and alternative streams and their respective URLs for the player. And their respective players all monitor either buffer status or CPU utilization and switch streams as necessary, locating the alternative streams from the URLs specified in the manifest.

The overriding problem with MSS, HLS and HDS is that these three different streaming protocols, while quite similar to each other in many ways, are different enough that they are not technically compatible. Indeed, each of the three proprietary commercial platforms is a closed system with its own type of manifest format, content formats, encryption methods and streaming protocols, making it impossible for them to work together.

Take Microsoft Smooth Streaming and Apple's HLS. Here are three key differences between the two competing platforms:

1. HLS makes use of a regularly updated "moving window" metadata index file that tells the client which chunks are available for download. Smooth Streaming uses time codes in the chunk requests so that the client doesn't have to keep downloading an index file. This leads to a second difference:

2. HLS requires a download of an index file every time a new chunk is available. That makes it desirable to run HLS with longer duration chunks, thereby minimizing the number of index file downloads. So, the recommended chunk duration with HLS is 10 seconds, while it is just two seconds with Smooth Streaming.

3. The "wire format" of the chunks is different. Although both formats use H.264 video encoding and AAC audio encoding, HLS makes use of MPEG-2 Transport Stream files, while Smooth Streaming makes use of "fragmented" ISO MPEG-4 files. The "fragmented" MP4 file is a variant in which not all the data in a regular MP4 file is included in the file. Each of these formats has some advantages and disadvantages. MPEG-2 TS files have a large installed analysis toolset and have pre-defined signaling mechanisms for things like data signals (e.g. specification of ad insertion points). But fragmented MP4 files are very flexible and can easily accommodate all kinds of data, such as decryption information, that MPEG-2 TS files don't have defined slots to carry.

Or take Adobe HDS and Apple's HLS. These two platforms have a number of key differences as well:

1. HLS makes use of a regularly updated "moving window" metadata index (manifest) file that tells the client which chunks are available for download. Adobe HDS uses sequence numbers in the chunk requests so the client doesn't have to keep downloading a manifest file.

2. In addition to the manifest, there is a bootstrap file, which in the live case gives the updated sequence numbers and is equivalent to the repeatedly downloaded HLS playlist.

3. Because HLS requires a download of a manifest file as often as every time a new chunk is available, it is desirable to run HLS with longer duration chunks, thus minimizing the number of manifest file downloads. More recent Apple client versions appear to check how many segments are in the playlist and only re-fetch the manifest when the client runs out of segments. Nevertheless, the recommended chunk duration with HLS is still 10 seconds, while it is usually just two to five seconds with Adobe HDS.

4. The "wire format" of the chunks is different. Both formats use H.264 video encoding and AAC audio encoding. But HLS makes use of MPEG-2 TS files, while Adobe HDS (and Microsoft SS) make use of "fragmented" ISO MPEG-4 files.

Due to such differences, there is no such thing as a universal delivery standard for streaming media today. Likewise, there is no universal encryption standard or player standard. Nor is there any interoperability between the devices and servers of the various vendors. So, content cannot be re-used and creators and equipment makers must develop several different versions of their products to serve the entire streaming video market, greatly driving up costs and restricting the market's overall development.

INTRODUCING MPEG DASH: A STANDARDS-BASED APPROACH

Seeing the need for a universal standard for the delivery of adaptive streaming media, MPEG decided to step into the void three years ago. In April 2009, the organization issued a Request for Proposals for an HTTP streaming standard. By that July, MPEG had received 15 full proposals. In the following two years, MPEG developed the specification with the help of many experts and in collaboration with other standards groups, such as the Third Generation Partnership Project (3GPP) and the Open IPTV Forum (OIPF).

The resulting MPEG standardization of Dynamic Adaptive Streaming over HTTP is now simply known as MPEG DASH.

MPEG DASH is not a system, protocol, presentation, codec, middleware, or client specification. Rather, the new standard is more like a neutral enabler, aimed at providing several formats that foster the efficient and high-quality delivery of streaming media services over the Internet.

As described by document ISO/IEC 23009-1, MPEG DASH can be viewed as an amalgamation of the industry's three prominent adaptive streaming protocols – Adobe HDS, Apple HLS and Microsoft Smooth Streaming. Like those three proprietary platforms, DASH is a video streaming solution where small chunks of video streams/files are requested using HTTP and then spliced together by the client. The client entirely controls the delivery of services.

In other words, MPEG DASH offers a standards-based approach for enabling a host of media services that cable operators and telcos have traditionally offered in broadcast and IPTV environments and extending those capabilities to adaptive bitrate delivery, including live and on-demand content delivery, time-shifted services (NDVR, catch-up TV), and targeted ad insertion. DASH enables these features through a number of inherent capabilities, and perhaps most importantly, through a flexibility of design and implementation. Its capabilities and features include:

- Multiple segment formats (ISO BMFF and MPEG-2 TS)
- Codec independence
- Trick mode functionality
- Profiles: restriction of DASH and system features (claim & permission)
- Content descriptors for protection, accessibility, content rating, and more
- Common encryption (defined by ISO/IEC 23001-7)
- Clock drift control for live content
- Metrics for reporting the client session experience

A Tale of Two Containers – MPEG-2 TS and ISO BMFF

Under the MPEG DASH standard, the media segments can contain any type of media data. However, the standard provides specific guidance and formats for use with two types of segment container formats – MPEG-2 Transport Stream (MPEG-2 TS) and ISO base media file format (ISO BMFF).

MPEG-2 TS is the segment format that HLS currently uses, while ISO BMFF (which is basically the MPEG-4 format) is what Smooth Streaming and HDS currently use.

This mix of the two container formats employed by the three commercial platforms allows for a relatively easy migration of existing adaptive streaming content from the proprietary platforms to MPEG DASH. That's because the media segments can often stay the same; only the index files must be migrated to a different format, which is known as Media Presentation Description.

Media Presentation Description (MPD) – Definition and Overview

At a high level, MPEG DASH works nearly the same way as the three other major adaptive streaming protocols. DASH presents available stream content to the media player in a manifest (or index) file – called the Media Presentation Description (MPD) – and then supports HTTP download of media segments. The MPD is analogous to an HLS m3u8 file, a Smooth Streaming Manifest file or an HDS f4m file. After the MPD is delivered to the client, the content – whether it's video, audio, subtitles or other data – is downloaded to clients over HTTP as a sequence of files that is played back contiguously.

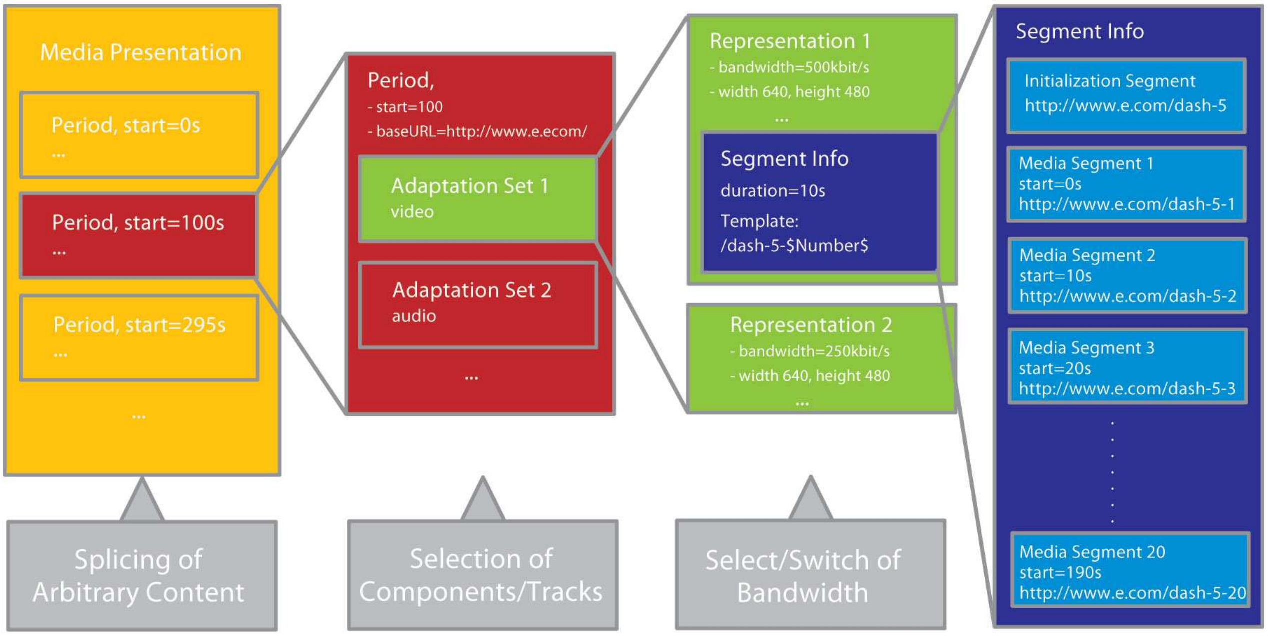


Figure 3: Media Presentation Data Model
(Diagram originally developed by Thomas Stockhammer, Qualcomm)

Like a manifest file in the three commercial platforms, the MPD in MPEG DASH describes the content that is available, including the URL addresses of stream chunks, byte-ranges, different bitrates, resolutions, and content encryption mechanisms. The tasks of choosing which adaptive stream bitrate and resolution to play and switching to different bitrate streams according to network conditions are performed by the client (again, similar to the other adaptive streaming protocols). In fact, DASH does not prescribe any client-specific playback functionality; rather, it just addresses the formatting of the content and associated MPDs.

To see what an MPEG DASH MPD file looks like compared to an HLS m3u8 file, consider the following example. The files contain much of the same information, but they are formatted and presented differently.

Figure 4: Comparison of MPEG DASH MPD and HLS m3u8 Files

Index.m3u8 (top level m3u8)

```
#EXTM3U
#EXT-X-STREAM-INF:PROGRAM- ID=1,BANDWIDTH=291500,RESOLUTION=320x180
stream1.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=610560,RESOLUTION=512x288
stream2.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=2061700,RESOLUTION=1024x576
stream3.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=4659760,RESOLUTION=1280x720
stream4.m3u8
```

Index.mpd

```
<?xml version="1.0" encoding="utf-8"?>
<MPD
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="urn:mpeg:DASH:schema:MPD:2011"
  xsi:schemaLocation="urn:mpeg:DASH:schema:MPD:2011"
  type="static"
  mediaPresentationDuration="PT12M34.041388S"
  minBufferTime="PT10S"
  profiles="urn:mpeg:dash:profile:isoff-live:2011">

  <Period>
    <AdaptationSet
      mimeType="audio/mp4"
      segmentAlignment="0"
      lang="eng">
      <SegmentTemplate
        timescale="10000000"
        media="audio_eng=$Bandwidth$-$Time$.dash"
        initialisation=" audio_eng=$Bandwidth$.dash">
        <SegmentTimeline>
          <S t="667333" d="39473889" />
          <S t="40141222" d="40170555" />

       ...

          <S t="7527647777" d="12766111" />
        </SegmentTimeline>
      </SegmentTemplate>
      <Representation id="audio_eng=96000" bandwidth="96000" codecs="mp4a.40.2"
audioSamplingRate="44100" />
    </AdaptationSet>
    <AdaptationSet
      mimeType="video/mp4"
      segmentAlignment="true"
      startWithSAP="1"
      lang="eng">
      <SegmentTemplate
        timescale="10000000"
        media="video=$Bandwidth$-$Time$.dash"
        initialisation="video=$Bandwidth$.dash">
        <SegmentTimeline>
          <S t="0" d="40040000" r="187" />
          <S t="7527520000" d="11678333" />
        </SegmentTimeline>
      </SegmentTemplate>

      <Representation id="video=299000" bandwidth="299000" codecs="avc1.42C00D"
width="320" height="180" />
      <Representation id="video=480000" bandwidth="480000" codecs="avc1.4D401F"
width="512" height="288" />
 codecs="avc1.4D401F" width="1024" height="576" />
      <Representation id="video=4300000" bandwidth="4300000"
codecs="avc1.640028" width="1280" height="720" />
    </AdaptationSet>
  </Period>
</MPD>
```

MPEG DASH’S PRIME CAPABILITIES – OVERVIEW

As mentioned earlier, MPEG DASH offers a great number of capabilities for adaptive streaming. This section goes into greater detail about many of the prime capabilities.

Codec Independence: Simply put, MPEG DASH is audio/video agnostic. As a result, the standard can work with media files of MPEG-2, MPEG-4, H.264, WebM and various other codecs and does not favor one codec over another. It also supports both multiplexed and unmultiplexed encoded content. More importantly, DASH will support emerging standards, such as HEVC (H.265).

Trick Mode Functionality: MPEG DASH supports VoD trick modes for pausing, seeking, fast forwarding and rewinding content. For instance, the client may pause or stop a Media Presentation.

In this case, the client simply stops requesting Media Segments or parts thereof. To resume, the client sends requests to Media Segments, starting with the next sub-segment after the last requested sub-segment.

DASH’s treatment of trick modes could prove to be a major improvement over the way that the three existing streaming protocols handle these on-demand functions now.

Profiles: Restriction of DASH and System Features (Claim & Permission): MPEG DASH defines and allows for the creation of various profiles. A profile is a set of restrictions of media formats, codecs, protection formats, bitrates, resolutions, and other aspects of the content. For example, the DASH spec defines a profile for ISO BMFF basic on-demand.

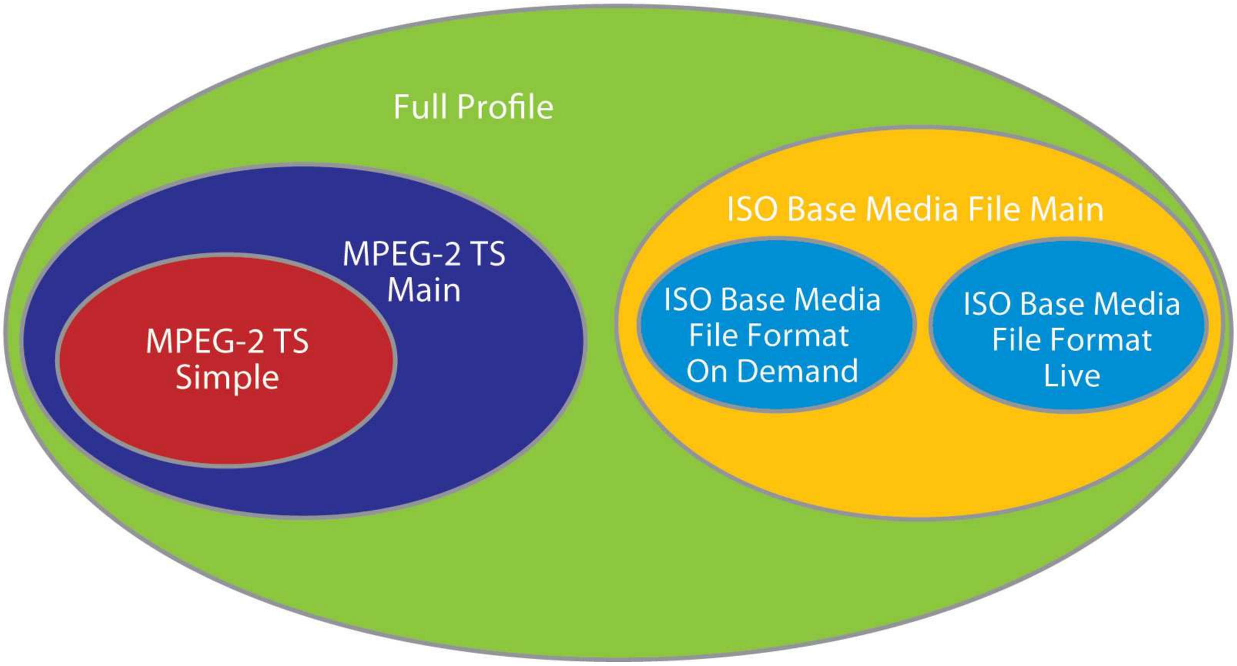

Full Profile
MPEG-2 TS Main
MPEG-2 TS Simple
ISO Base Media File Main
ISO Base Media File Format On Demand
ISO Base Media File Format Live

Figure 5: Describing MPEG DASH Profiles
(Diagram originally developed by Thomas Stockhammer, Qualcomm)

Content Descriptors for Protection, Accessibility, Content Rating: MPEG DASH offers a flexible set of descriptors for the media content that is being streamed. These descriptors spell out such elements as the rating of the content, the role of various components, accessibility features, DRM methods, camera views, frame packing, and the configuration of audio channels, among other things.

Common Encryption (defined by ISO/IEC 23001-7): One of the most important features of MPEG DASH is its use of Common Encryption, which standardizes signaling for what would otherwise be a number of non-interoperable, albeit widely used, encryption methods. Leveraging this standard, content owners or distributors can encrypt their content just once and then stream it to different clients with different DRM license systems. As a result, content owners can distribute their content freely and widely, while service providers can enjoy access to an open, interoperable ecosystem of vendors. In fact, Common Encryption is also used as the underlying standard for Ultraviolet, the Digital Entertainment Content Ecosystem's (DECE's) content authentication system. Common Encryption will be discussed in a bit more detail later in this paper.

Clock Drift Control for Live Content: In MPEG DASH, each media segment can include an associated Coordinated Universal Time (UTC) time, so that a client can control its clock drift and ensure that the encoder and decoder remain closely synchronized. Without this, a time difference between the encoder and decoder could cause the client play-back buffer to starve or overflow, due to different rates of video delivery and playback.

Metrics for Reporting the Client Session Experience: MPEG DASH has a set of well-defined quality metrics for tracking the user's session experience and sending the information back to the server.

MULTIPLE DRM METHODS & COMMON ENCRYPTION

As mentioned earlier, one of MPEG DASH's most important features is its use of Common Encryption, which standardizes signaling for a number of different, widely used encryption methods. Common Encryption (or "CENC") describes methods of standards-based encryption, along with key mapping of content to keys. CENC can be used by different DRM systems or Key Management Servers (KMS) to enable decryption of the same content, even with different vendors' equipment. It works by defining a common format for the encryption-related metadata required to decrypt the protected content. The details of key acquisition and storage, rights mapping, and compliance rules are not specified in the standard and are controlled by the DRM server. For example, DRM servers supporting Common Encryption will identify the decryption key with a key identifier (KID), but will not specify how the DRM server should locate or access the decryption key.

Using this standard, content owners or distributors can encrypt their content just once and then stream it to the various clients with their different DRM license systems. Each client receives the content decryption keys and other required data using its particular DRM system. This information is then transmitted in the MPD, enabling the client to stream the commonly encrypted content from the same server.

As a result, content owners can distribute their content freely and widely without the need for multiple encryptions. At the same time, cable operators and other video service providers can enjoy access to an open, interoperable ecosystem of content producers and equipment vendors.

USE CASES

The MPEG DASH spec supports both simple and advanced use cases of dynamic adaptive streaming. Moreover, the simple use cases can be gradually extended to more complex and advanced cases. In this section, we'll detail three such common use cases:

Live and On-Demand Content Delivery: MPEG DASH supports the delivery of both live and on-demand media content to subscribers through dynamic adaptive HTTP streaming. Like Adobe's HDS, Apple's HLS and Microsoft's Smooth Streaming platforms, DASH encodes the source video or audio content into file segments using a desired format. The segments are subsequently hosted on a regular HTTP server. Clients then play the stream by requesting the segments in a profile from a Web server, downloading them via HTTP.

MPEG DASH's great versatility in supporting both live and on-demand content has other benefits as well. For instance, these same capabilities also enable video service providers to deliver additional time-shifted services, such as network-based DVR (NDVR) and catch-up TV services, as explained below.

Time-Shifted Services (NDVR, catch-up TV, etc.): MPEG DASH supports the flexible delivery of time-shifted services, such as NDVRs and catch-up TV. For the enabling of time-shifted services, VoD assets, rather than live streams, are required. VoD assets formatted for MPEG DASH can be created using a transcoder. Additionally, a device commonly referred to as a Catcher can "catch" a live TV program and create a VoD asset, suitable for streaming after the live event. Because the VoD asset can be streamed in MPEG DASH in the same manner as the live content, the asset can be re-used and monetized by the operator.

Targeted Ad Insertion: Wherever there is video service, there is usually some kind of advertising content to monetize the service. 'Traditional" ad insertion methods rely on a set of technologies based on the widely used protocols for distributing UDP/IP video: ad servers, ad splicers, and an ecosystem based on zoned ad delivery. But as video delivery transport has evolved via the new set of adaptive HTTP-based delivery protocols from Apple, Microsoft and Adobe, the ad insertion ecosystem has had to evolve to employ new, targeted technologies for insertion and delivery of revenue-generating commercials. The difficulty of inserting ads with the three existing delivery methods is that the protocols don't support the same ad insertion methods, due to the inherent nature of how the protocols work.

MPEG DASH offers the dramatic potential to help enable adaptive bitrate advertising on many different types of client devices. DASH supports the dynamic insertion of advertising content into multimedia streams. In both live and on-demand use cases, commercials can be inserted either as a period between different multimedia periods or as a segment between different multimedia segments. As in the case with VoD trick modes, this would represent a significant improvement over the way that the three leading streaming protocols currently handle targeted ad insertion.

It is worth emphasizing that DASH supports a network-centric approach to ad insertion, as opposed to a client-centric approach in which the client pre-fetches ads and splices them locally based on interactions with external ad management systems. In DASH, the information about when ads play, which ads play, and how ads are delivered is transmitted through the MPD, which is created and distributed from the network.

PROSPECTS FOR INDUSTRY ADOPTION – CATALYSTS & CHALLENGES

With the development, ratification and introduction of the MPEG DASH platform, MPEG is attempting to rally the technology community behind a universal delivery standard for adaptive streaming media. Many tech companies have already enlisted in the effort, joining the new MPEG DASH Promoters Group to drive the broad adoption of the standard.

Not surprisingly, equipment vendors and content publishers are especially enthusiastic about the new standard. For instance, content publishers savor the opportunity to produce just a single set of media files that could run on all DASH-compatible electronics devices.

The key to MPEG DASH's success, though, will be the participation of the three major proprietary players – Adobe, Apple, and Microsoft – that now divvy up the adaptive streaming market. While all three companies have contributed to the standard, their levels of support for DASH vary greatly. In particular, Apple's backing is still in question because of the competitive advantages that its HLS platform stands to lose if DASH becomes the universal standard.

Besides such competitive issues, MPEG DASH faces potential intellectual property rights challenges as well. For example, it is still not clear if DASH will be saddled with royalty payments and, if so, where those royalties might be applied. This section will look at the intellectual property rights and other issues that may yet bedevil the new standard.

Unresolved Intellectual Property Rights Issues: In addition to the competitive issues, there are some unresolved intellectual property rights issues with MPEG DASH. For instance, when companies seek to contribute intellectual property to the MPEG standards effort, the contribution is usually accepted only if the property owner agrees to Reasonable and Non-Discriminatory (RAND) terms. In the case of DASH, though, it is not clear that all of the intellectual property rights (IPR) in the standard are covered by RAND terms.

Non-Interoperable DASH Profiles: Although MPEG DASH may have a single, unified name, it actually consists of a collection of different, non-interoperable profiles. So DASH doesn't solve the problem of different, non-interoperable implementations unless DASH clients support all profiles. This would basically be equivalent to having a client that supports HLS, HDS and Smooth Streaming (which, incidentally, would also address the interoperability problem). Thus, the adoption of DASH doesn't immediately imply a unified, interoperable ecosystem – a DASH world may suffer from the same interoperability issues that HLS, Smooth Streaming and HDS create today.

CONCLUSION

Now that MPEG DASH has been published by the ISO, it seems well on its way to becoming a solid, broadly accepted standard for the streaming media market. Three years in the making, DASH is poised to provide a universal platform for delivering streaming media content to multiple screens. Designed to be very flexible in nature, it promises to enable the re-use of existing technologies (containers, codecs, DRM, etc.), seamless switching between protocols, and perhaps most importantly, a high-quality experience for end users.

Furthermore, most of the tech industry's major players have already lined up firmly behind DASH. The list of prominent supporters includes Akamai, Dolby, Samsung, Thomson, Netflix and, most notably, such leading streaming media providers as Microsoft and Adobe. Apple stands out as one of the few major tech players that haven't fully enlisted in the effort yet. So there's a great deal of hope in the industry that MPEG DASH could actually bring in all of the major players and realize its full market potential.

Yet several critical hurdles remain in the way of DASH's dash to destiny. For one thing, Apple, Adobe and Microsoft must throw their full weight behind the standard and agree to make the switch from their proprietary HLS protocols in the future despite some clear competitive disadvantages of doing so. For another, all industry stakeholders must agree to make their intellectual property contributions to the standard royalty-free.

Neither of these developments will likely happen overnight. So it's not clear yet if MPEG DASH will end up superseding the existing adaptive streaming formats as a true universal industry standard or merely co-existing with one or more of them in a still-fragmented market. As usual, the outcome will depend on what the major vendors decide to do. It will also depend on whether cable operators and other video service providers shift their multiscreen deployments and content offerings to DASH or continue on their current streaming paths. Only time will tell.

OPTIMIZING FAIRNESS OF HTTP ADAPTIVE STREAMING IN CABLE NETWORKS

Michael Adams
Chris Phillips
Solution Area Media
Ericsson

Abstract

This paper describes a novel approach to traffic management for HTTP adaptive streaming that optimizes fairness across multiple clients and increases network throughput. Readers of the paper will gain an understanding of the network impacts of implementing HTTP adaptive streaming, and how network management techniques may be applied to optimize fair bandwidth allocation between competing streams.

Benefits for the network operator include:

- *enforcing fairness in the network (without resorting to techniques such as deep packet inspection),*
- *managing and ensuring consumers' overall quality of experience, and*
- *preventing network instability that can be caused by competing clients in a shared access network.*

Benefits for the consumer include:

- *a more consistent overall quality of viewing experience, and*
- *the ability to simultaneously use multiple devices within the home.*

The concepts described in this paper have been prototyped to show improvements in fairness and overall network throughput without placing special constraints on the client implementation (which is typically outside of the operator's control). The results are being published here for the first time.

BACKGROUND

There is a great deal of interest in HTTP adaptive streaming because it can greatly improve the user experience for video delivery over unmanaged networks. Adaptive streaming operates by dynamically adjusting the play-out rate to stay within the actual network throughput to a given endpoint, without the need for "rebuffering". So, if the network throughput suddenly drops, the picture may degrade but the end user still sees a picture.

Although adaptive streaming was originally developed for "over-the-top" video applications over unmanaged networks, it also brings significant advantages to managed networks applications. For example, during periods of network congestion, operators can set session management polices to permit a predefined level of network over-subscription rather than blocking all new sessions. This flexibility will become more and more important as subscribers start to demand higher quality feeds (1080p and 4K).

HTTP adaptive streaming is the generic term for various implementations:

- Apple HTTP Live Streaming (HLS) [1]
- Microsoft IIS Smooth Streaming [2]
- Adobe HTTP Dynamic Streaming (HDS) [3]

Although each of the above is different, they have a set of common properties (see Figure 1):

- Source content is transcoded in parallel at multiple bit rates (multi-rate transcoding). Each bit rate is called a profile or representation.
- Encoded content is divided into fixed duration segments (or chunks), which are typically between two and 10 seconds in duration. (Shorter segments reduce coding efficiency while larger segments impact speed to adapt to changes in network throughput).
- A manifest file is created, and updated as necessary, to describe the encoding rates and URL pointers to segments.
- The client uses HTTP to fetch segments from the network, buffer them and then decode and play them.
- The client algorithm is designed to select the optimum profile so as to maximize quality without risking buffer underflow and stalling (rebuffering) of the play-out. Each time the client fetches a segment, it chooses the profile based on the measured time to download the previous segment.

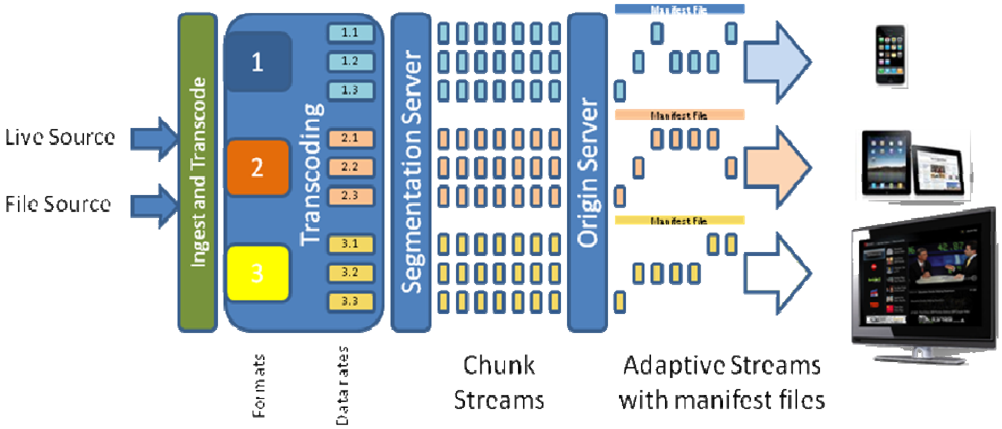


Figure 1: Ingest, transcoding, segmentation and adaptive streaming.

MPEG DASH

MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH) is certain to become a significant force in the marketplace [4]. While HLS uses the MPEG-2 transport stream format (which is widely deployed in most conventional digital TV services), Smooth Streaming and MPEG-DASH use an MPEG-4 Part 14 (ISO/IEC 14496-12) transport format known as fMP4 or ISO MP4FF.

Smooth Streaming and MPEG-DASH include full support for subtitling and trick modes, whereas HLS is limited in this regard. MPEG-DASH enables common encryption, which simplifies the secure delivery of content from multiple rights holders and to multiple devices.

Another key difference is the way in which MPEG-DASH and Smooth Streaming play-out is controlled when transmission path conditions change. HLS uses manifest files that are effectively a playlist identifying the different segments so that, for instance, when path impairment occurs, the selection of the URL from the manifest file adapts so that lower bit-rate segments are selected. In Smooth Streaming the client uses time stamps to identify the segments needed and thus certain efficiencies are gained. Both HLS and Smooth Streaming handle files in subtly different ways, each claiming some efficiency advantage over the other. Both use HTTP, which has the ability to traverse firewalls and network address translation, giving it a clear advantage over RTSP, RTMP and MMS.

Adaptive Streaming Standardization

There are a number of initiatives aimed at large parts of the overall solution for streaming video. A document called *MPEG Modern Transport over Networks* was approved at the 83rd MPEG meeting in January 2008, which proposed a client that was media aware with optimization between the transport and content layers to enable video to traverse networks in an adaptive manner. However, at that time, its focus was on the widespread adoption of a variant of AVC/MVC called SVC (Scalable Video Coding) that would allow the client to generate acceptable video from a subset of the total aggregated transport stream.

Subsequently, at the 93rd meeting, the focus was changed to HTTP streaming of MPEG Media called *Dynamic Adaptive Streaming over HTTP* (DASH) using 3GPP's Adaptation HTTP Streaming (AHS) as the starting point. MPEG has standardized a Manifest File (MF), a Delivery Format (DF), and means for easy conversion from/to existing File Formats (FF) and Transport Streams (TS) as illustrated in Figure 2.

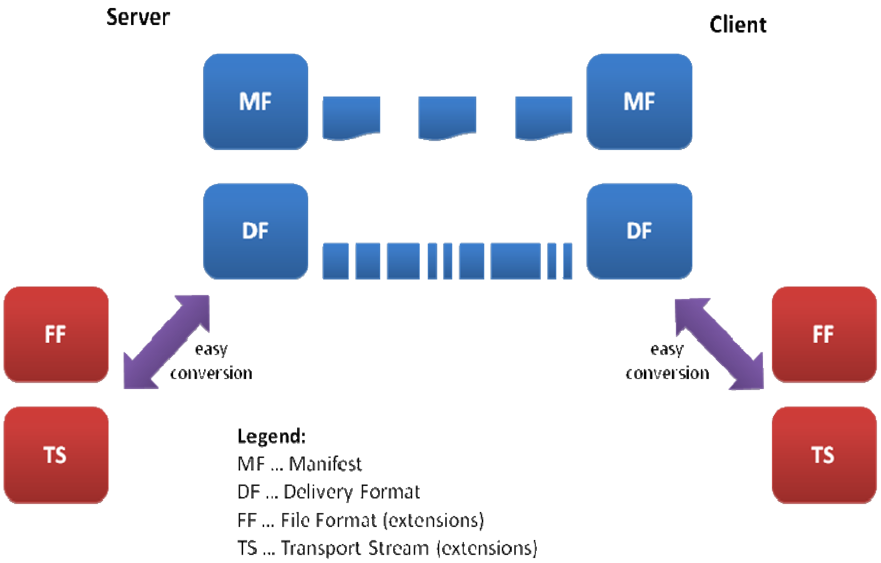


Figure 2: Dynamic Adaptive Streaming over HTTP (DASH).
Courtesy: Christian Timmerer, Assistant Professor at Klagenfurt University (UNIKLU)

The MPEG-DASH standard got Final Draft International Standard status in December 2011. MPEG-DASH has the potential to simplify and converge the delivery of IP video, provide a rich and enjoyable user experience, help drive down costs and ultimately enable a better content catalog to be offered to consumers, because more revenues can be re-invested in content, rather than paying for operating overheads. It will help streamline and simplify workflows and enable operators and content providers to build sustainable business models to continue to deliver the services that consumers demand.

FAIRNESS IN ADAPTIVE STREAMING

HTTP adaptive streaming clients implement a “greedy” algorithm, in which they will always seek to achieve the maximum bit rate available. This can lead to instability, oscillation and unfairness, where some clients will win and others will lose in times of network congestion [5], [6].

Reference Architecture

Figure 3 illustrates a typical arrangement where HTTP adaptive streaming is used to deliver video and audio programming to a device in a subscriber's home. Note that a CDN is typically used to replicate segments within the core network and this is assumed to have infinite bandwidth. At the edge of the network, the bandwidth is constrained by:

1. The downstream path over DOCSIS.
2. The wireless network path to a Wi-Fi connected device.

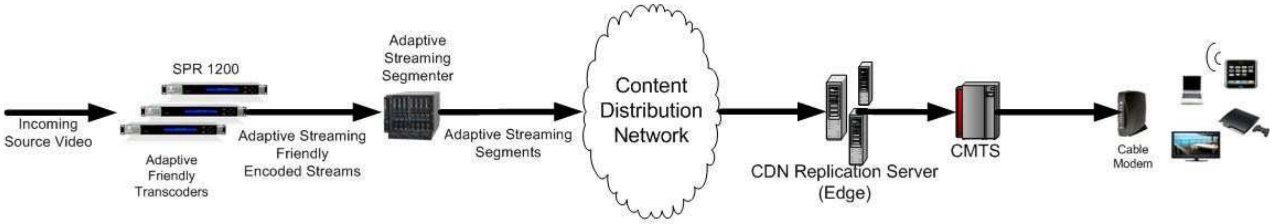

Incoming Source Video
SPR 1200
Adaptive Friendly Transcoders
Adaptive Streaming Friendly Encoded Streams
Adaptive Streaming Segmenter
Adaptive Streaming Segments
Content Distribution Network
CDN Replication Server (Edge)
CMTS
Cable Modem

Figure 3: Reference architecture

Prototype System Description

Figure 4 shows the prototype system that was developed to analyze the behavior of standard HLS adaptive streaming clients on a shared access network.

Packet scheduling is determined by the bandwidth monitor/allocator. It also creates a virtual pipe and constrains all packet delivery within it. The virtual pipe can be dynamically resized while the system is running. Two scheduling algorithms can been implemented within the virtual pipe; best-effort and weighted fair queuing. Best-effort implements first-come, first-served packet delivery. The weighted fair queuing algorithm schedules transmission according to the virtual pipe size and compares the amount of data transmitted through various classes to ensure that each class is allocated its fair share.

The bandwidth monitor/allocator also logs bandwidth utilization data for use by the real-time statistics monitor and the graphing module. This data is used to visualize the behavior of the system and to understand the behavior of the adaptive streaming clients.

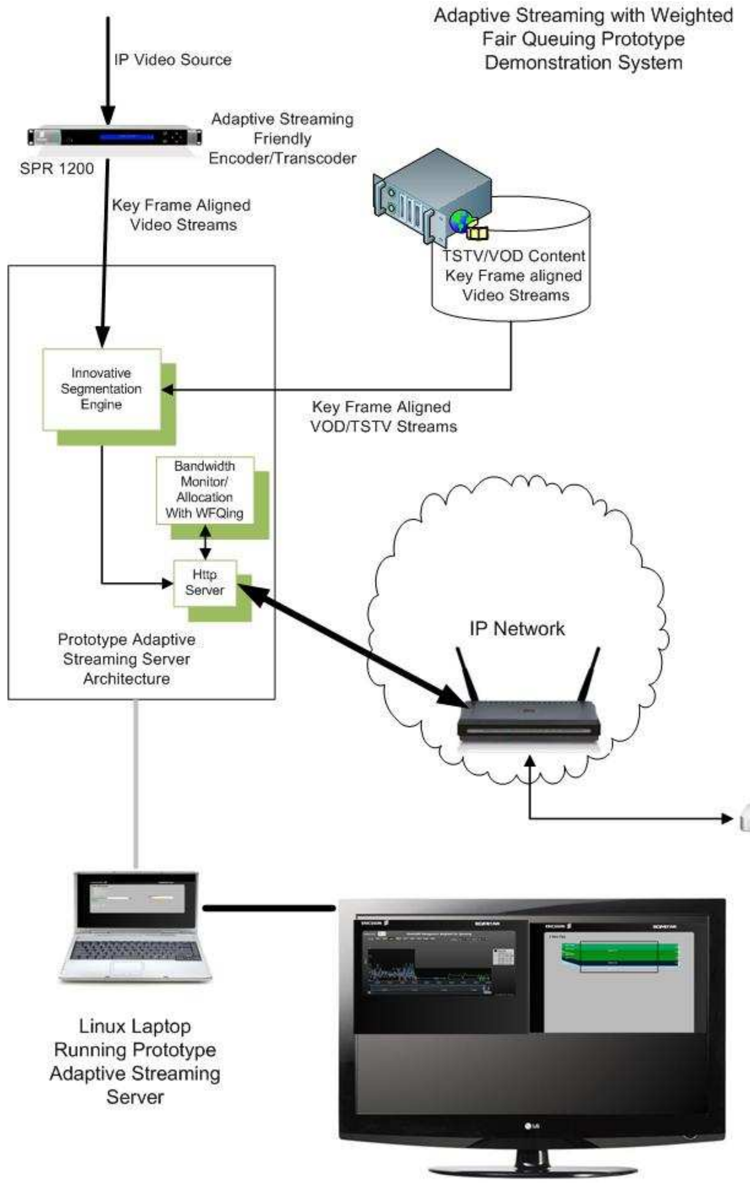

Adaptive Streaming with Weighted Fair Queuing Prototype Demonstration System
IP Video Source
SPR 1200
Adaptive Streaming Friendly Encoder/Transcoder
Key Frame Aligned Video Streams
TSTV/VOD Content Key Frame aligned Video Streams
Innovative Segmentation Engine
Key Frame Aligned VOD/TSTV Streams
Bandwidth Monitor/ Allocation With WFQing
Http Server
Prototype Adaptive Streaming Server Architecture
IP Network
Adaptive Streaming Clients
Linux Laptop Running Prototype Adaptive Streaming Server

Figure 4: Prototype system

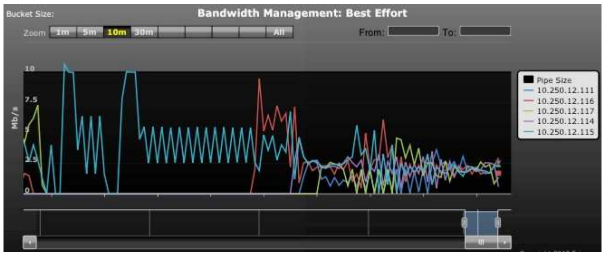
EXPERIMENTAL RESULTS

The experimental approach followed was to compare results of the best-effort (no traffic management) behavior with that of weighted fair queuing. The first best-effort run was repeated with identical parameters except for enabling the weighted fair queuing algorithm in the second and third runs. A fourth and fifth run were done with a different set of encoding profiles to investigate the effect that this would have on the behavior of the adaptive streaming clients.

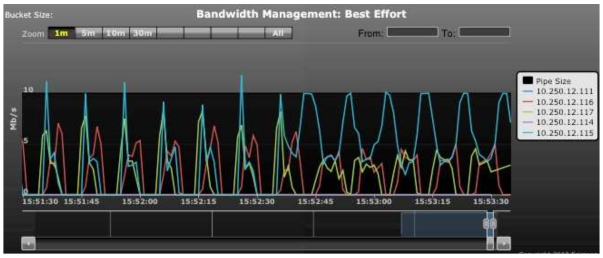
Run 1: Best Effort

Start Time	15:30:00 GMT
Pipe	10 Mbps
Content	How to Train your Dragon
Profiles	560, 660, 760, 860, 1000, 2000, 4000 Kbps
Clients	Mac Mini (115), iPad (112), iPad (114), iPad (111), iPhone (117)
Server	Best Effort

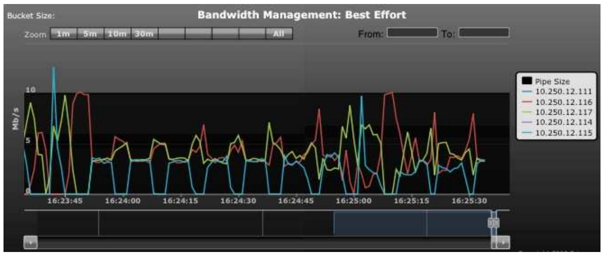
1. Each client started in sequence; all clients settled to 2 Mbps profile (Graph 1).
2. After 10 minutes iPad (114) goes to 1 Mbps profile; 9 Mbps pipe utilization.
3. Stopped iPad (114); 8 Mbps or 100% utilization (Graph 2).
4. Stopped iPad (111); 6 Mbps or 60% utilization (Graph 2).
5. After approximately 10 minutes, Mac Mini (115) jumped to 4 Mbps profile; still at only 80% pipe utilization.
6. Eventually, after approximately 30 minutes, system achieved 10 Mbps or 100% utilization (Graph 3).



Graph 1: Five clients started in sequence.



Graph 2: From four to three clients, 60% utilization.



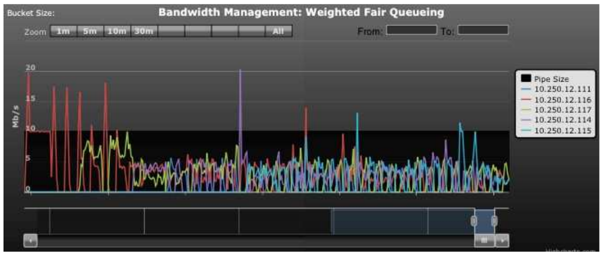
Graph 3: Final state achieved: 100% utilization.

This result was unexpected. It appears that the three clients (Graph 2) became locked in a synchronous pattern and as a result measured a lower segment-download rate than expected. As a result, none of the clients moved to a higher-rate profile, even though adequate bandwidth (4 Mbps) remained unused. Eventually (Graph 3) this pattern corrected itself, but not until approximately 30 minutes had elapsed.

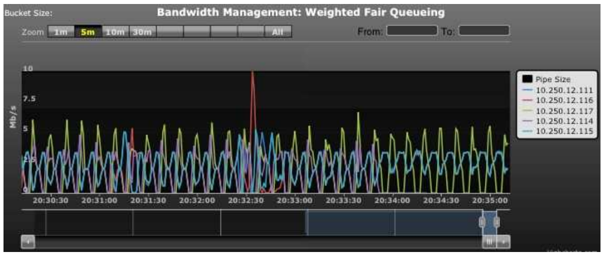
Run 2: Weighted Fair Queuing

Start Time	15:27:00 GMT
Pipe	10 Mbps
Content	How to Train your Dragon
Profiles	560, 660, 760, 860, 1000, 2000, 4000 Kbps
Clients	Mac Mini (115), iPad (112), iPad (114), iPad (111), iPhone (117)
Server	Weighted Fair Queuing
Weighting Factor	All clients set to 1

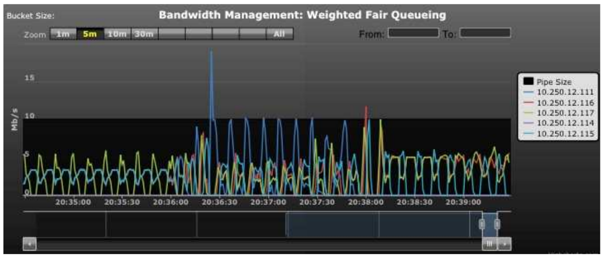
1. Each client was started in sequence (as before); same result as best-effort case (Graph 4).
2. iPad (111), iPad (116), and Mac Mini (115) all occasionally reached the 1 Mbps profile. Pipe stays at very close to 100% utilization.
3. Stopped iPad (114); immediately remaining clients achieved 100% pipe utilization. After 2 min iPad (111) reached the 4 Mbps profile (Graph 5).
4. Stopped iPad (111); pipe at 90% and then increased to 100% utilization (Graph 6).



Graph 4: Five clients, 100% utilization.



Graph 5: Four clients, 100% utilization.



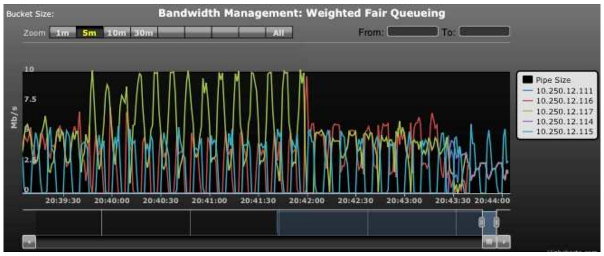
Graph 6: Three clients, 100% utilization.

It is apparent from Graphs 5 and 6 that the throughput of the system is much higher than in the best-effort case. In all cases, the pipe was utilized at, or close to, 100%. This may be understood by considering the scheduling algorithm at the HTTP server sequences through each partial segment download. Hence the clients take turns to achieve a more efficient segment download and therefore request a higher profile than in the best-effort case. The pipe throughput is maximized by the scheduling algorithm.

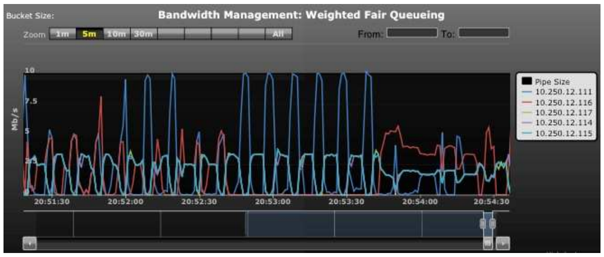
Run 3: Weighted Fair Queuing

Start Time	20:39:00 GMT
Pipe	10 Mbps
Content	How to Train your Dragon
Profiles	560, 660, 760, 860, 1000, 2000, 4000 Kbps
Clients	Mac Mini (115), iPad (112), iPad (114), iPad (111), iPhone (117)
Server	Weighted Fair Queuing
Weighting Factor	iPad(116) = 2, 3, 4

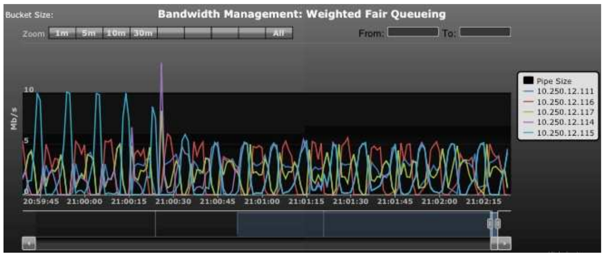
1. All clients started - no premium factor applied (Graph 7).
2. iPad (116) stopped.
3. Premium factor 2 applied to iPad (116) and started - quickly ramped to 2 Mbps (Graph 8).
4. Premium increased to 3. 115 went to 4 Mbps (momentarily).
5. Premium increased to 4. 114 went to 4 Mbps (Graph 9).



Graph 7: Fair network queuing.



Graph 8: Weighted fair queuing: iPad (116) weighting factor = 2.



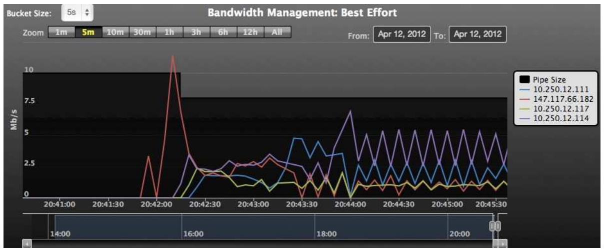
Graph 9: Weighted fair queuing: iPad (116) weighting factor = 4.

The weighting factor allows the premium client to achieve a higher profile than in Run 2, but it did not achieve the highest profile, probably because it was such a large jump in bit rate from 2 Mbps to 4 Mbps. Therefore, it was decided to re-run the test with a closer set of profile rates.

Run 4: Best Effort

Start Time	20:42:00 GMT
Pipe	8 Mbps
Content	Promo Reel
Profiles	400, 600, 910, 1200, 1600, 2000 Kbps
Clients	iPad (111), Mac Mini (182), iPhone (117), iPad (114), iPad (116)
Server	Best Effort

1. iPad (111), Mac Mini (182), iPhone (117), and iPad (114) started (Graph 10 and Figure 5).
2. 5th client iPad (116) started at 20:46:25 (Graph 11 and Figure 5).



Graph 10: Best effort, four clients.

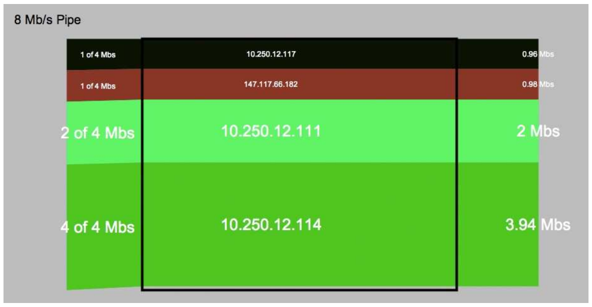
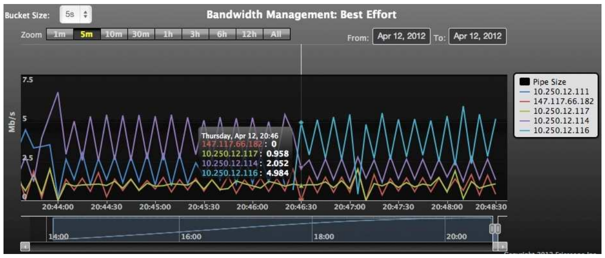


Figure 5: Best effort, four clients



Graph 11: Best effort, fifth client started at t = 20:46:25.

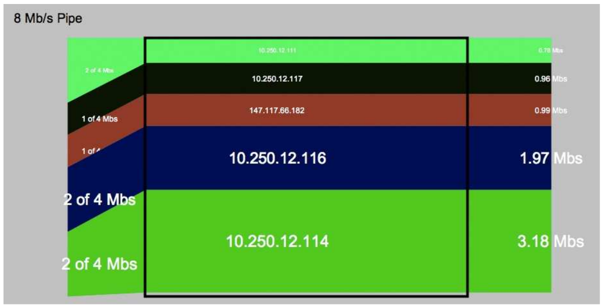


Figure 6: Best effort bit-rate allocation.

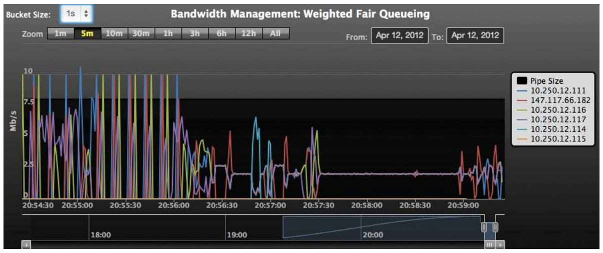
In this case, the bandwidth was shared unfairly. We hypothesize that the clients that first achieved the highest profile (2 Mbps) were able to maintain an unfair share of the pipe because of a positive feedback effect.

Run 5: Weighted Fair Queuing

Start Time	20:54:30 GMT
Pipe	8 Mbps
Content	Promo Reel
Profiles	400, 600, 910, 1200, 1600, 2000 Kbps
Clients	iPad (111), Mac Mini (182), iPhone (117), iPad (116), iPad (114)
Server	Weighted Fair Queuing
Weighting Factor	iPad(114) = 3

1 iPad (111), Mac Mini (182), iPhone (117), and iPad (114) started (Graph 12 and Figure 7).

2 iPad(114) with weighting factor of 3 started, and quickly ramped to 2 Mbps (Graph 13 and Figure 8)



Graph 12: Four clients, fair network queuing.

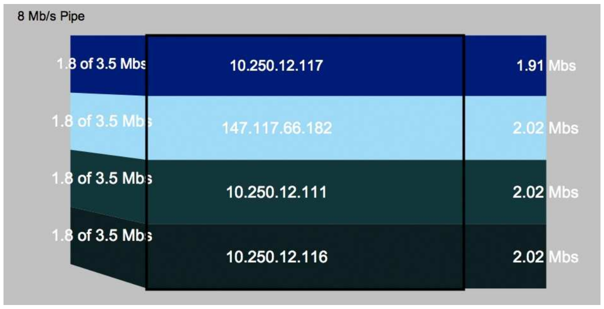
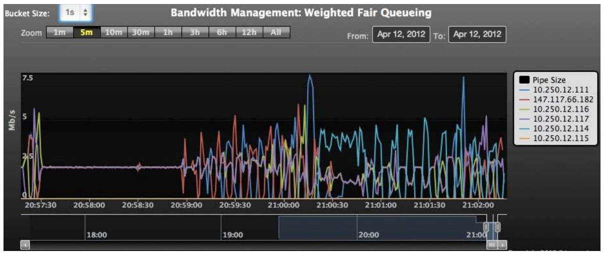


Figure 7: Four clients, fair network queuing.



Graph 13: Five clients, WFQ iPad (114) weighting factor = 3

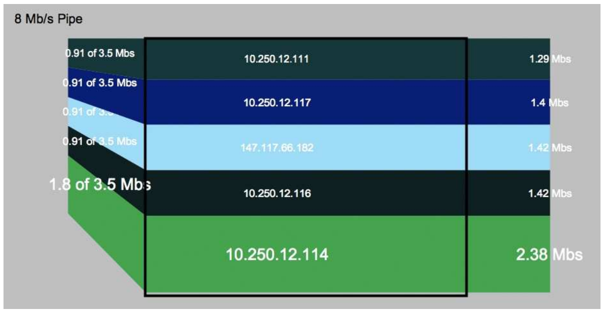


Figure 8: Five clients, WFQ iPad (114) weighting factor = 3

In this case, bandwidth was allocated equally (Figure 7) with four equally weighted clients. Subsequently, a greater share of bandwidth was allocated to a premium client (Figure 8) with a weighting factor of 3.

CONCLUSIONS

Implementation of a bit-wise, round-robin scheduler at the HTTP server can be used to effectively enforce fairness amongst HTTP adaptive streaming clients. In addition, a weighting factor may be established for a "premium" client to ensure that it experiences greater throughput during periods of access network congestion.

It appears that the weighted fair queuing mechanism [7] implemented in the prototype system is effective because it operates at the HTTP server layer, which is at a higher layer in the network stack than TCP congestion control alone [8], [9], and because it operates over a significantly longer time frame. If a client tries to "cheat" the system by requesting a higher profile than can be sustained by the network, it will only impact its own performance and not that of other clients.

IMPLEMENTATION OPTIONS

In order for the weighted fair queuing algorithm to be effective it must be implemented at the point in which traffic converges on a shared link in the access network. In the case of a DOCSIS access network, each downstream service group must be treated separately. The algorithm is implemented at the HTTP server (at the edge cache) and a virtual pipe must be created to each downstream service group (as illustrated in Figure 9).

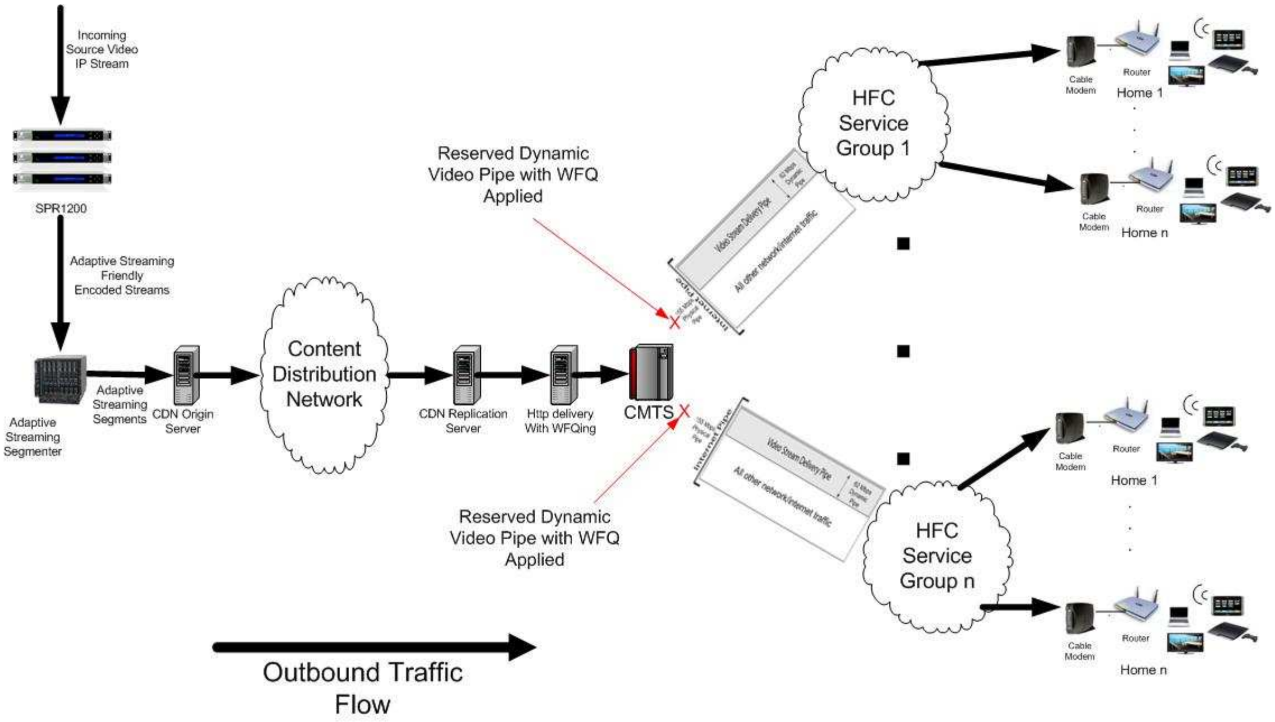

Incoming Source Video IP Stream
SPR1200
Adaptive Streaming Friendly Encoded Streams
Adaptive Streaming Segmenter
Adaptive Streaming Segments
CDN Origin Server
Content Distribution Network
CDN Replication Server
Http delivery With WFQing
CMTS
Reserved Dynamic Video Pipe with WFQ Applied
Reserved Dynamic Video Pipe with WFQ Applied
Internet Pipe
100 Mbps Physical Pipe
Video Stream Delivery Pipe
40 Mbps Dynamic Pipe
All other network/internet traffic
Internet Pipe
100 Mbps Physical Pipe
Video Stream Delivery Pipe
40 Mbps Dynamic Pipe
All other network/internet traffic
HFC Service Group 1
HFC Service Group n
Cable Modem
Router
Home 1
Cable Modem
Router
Home n
Cable Modem
Router
Home 1
Cable Modem
Router
Home n
Outbound Traffic Flow

Figure 9: Implementation in DOCSIS Network

References

1. R. Pantos and W. May. HTTP Live Streaming. IETF Draft, June 2010.
2. A. Zambelli. IIS smooth streaming technical overview. Microsoft Corporation, 2009.
3. Adobe HTTP Dynamic Streaming. http://www.adobe.com/products/hds-dynamic-streaming.html.
4. ISO/IEC 23009-1:2012 – Information Technology – Dynamic Adaptive Streaming over HTTP.
5. Saamer Akhshabi, Ali C. Begen, and Constantine Dovrolis. An Experimental Evaluation of Rate-Adaptation Algorithms in Adaptive Streaming over HTTP. Mac MiniSys'11, February 23–25, 2011, San Jose, California, USA.
6. Bing Wang, Jim Kurose, Prashant Shenoy, and Don Towsley. Multimedia streaming via TCP: An Analytic Performance Study. Department of Computer Science, University of Massachusetts.
7. Martin J. Fischer, Denise M. Bevilacqua Masi, and John F. Shortle. Simulating The Performance of a Class-based Weighted Fair Queuing System. Proceedings of the 2008 Winter Simulation Conference.
8. Robert Kuschnig, Ingo Kofler, and Hermann Hellwagner. An Evaluation of TCP-based Rate-Control Algorithms for adaptive Internet streaming of H.264/SVC. Institute of Information Technology (ITEC) Klagenfurt University, Austria.
9. Luca De Cicco and Saverio Mascolo. An Experimental Investigation of the Akamai Adaptive Video Streaming. Dipatimento di Elettrotecnica ed Elettronica, Politecnico di Bari.

Optimizing Wireless Networking of Wi-Fi and LTE

Marty Glapa, Amit Mukhopadhyay
Bell Laboratories, Alcatel-Lucent

Abstract

The proliferation of smart devices such as iPhones, DROIDs, tablets and others has resulted in huge increases in data traffic across cellular networks. These devices support multiple wireless technologies including 3G, 4G LTE and Wi-Fi. The massive adaptation of these devices can enable Wi-Fi to play a significant role in addressing the 3G/4G mobile data networks' increasing capacity requirements. Wireline and cable operators can both provide Wi-Fi offload for wireless operators. In this paper, we show how to optimize the performance and cost of heterogeneous networks comprised of cellular and Wi-Fi technologies.

INTRODUCTION

Most smartphones, tablets and PCs in today's world support Wi-Fi technology. While historically there have been hesitations on the part of wireless operators to embrace Wi-Fi as a complimentary technology to cellular, developments over the last several years in 3GPP interoperability have been breaking down the barriers (see, for example 1, 2). It's no longer an "either/or" discussion or debate but rather a complimentary use of both cellular and Wi-Fi technologies by operators to provide their end-users with optimized access to a rich set of services. Note that some Wi-Fi network deployments may include applications that drive Wi-Fi only traffic and not cellular traffic.

In this paper, we deal with the following key questions, which have not been commonly addressed, to the best of our knowledge:

- How much traffic can potentially be offloaded to Wi-Fi networks? This helps in sizing the Wi-Fi as well as the cellular network, since the latter now needs to carry only the remaining load.
- How do we overcome the well-known problems of interoperability between Wi-Fi and cellular networks, e.g., user authentication and admission control, mobility between the two networks, interference issues, guaranteed Quality of Service (QoS), etc.
- What are the best locations to deploy Wi-Fi hotspots? Access Point footprints are quite small compared to macro cellular footprints and deploying and clustering Access Points at the right locations, especially in high-traffic areas, is critical to the service provider for getting the most out of their investment and maintaining a consistent coverage footprint for nomadic users.
- How does the economics of combining Wi-Fi and cellular networks compare with the cellular network alone? A smart combination of Wi-Fi and cellular networks can keep the costs down and yet satisfy traffic demands.

Cable operators are in an excellent position to leverage their networks not only for using Wi-Fi as an extension of fixed access broadband services, but also in partnering with wireless service providers to use Wi-Fi and offload cellular network traffic.

We present a model to assess Wi-Fi offload potential in a network, based on applications, user behavior, etc. We show techniques of creating traffic density maps and identifying high traffic areas. Finally, we present a techno-economic model to compare the network options.

CONSUMERS, DEVICES AND TRAFFIC DRIVE "INTERWORKING"

A Wireless World

Nearly every mobile device in the foreseeable future will support multiple wireless technologies including 2G and 3G[1], 4G LTE[2] and Wi-Fi[3]. Wi-Fi plays a significant role in addressing the 3G/4G mobile data networks' increasing capacity requirements. Wi-Fi, in addition to providing a wireless extension for fixed wireline broadband, has emerged as a way to gain alternative connection to the 3G/4G cellular network services while off-loading data traffic from its radio access network (RAN). The complimentary use of LTE and Wi-Fi in providing wireless services enables the network operator to balance network and transport costs, while providing the consumer with services to meet their bandwidth needs.

Forces Driving Traffic Explosion

While early days of mobile data traffic primarily consisted of applications such as occasional web browsing, running search engines or instant messaging, today's mobile data traffic is dominated by richer applications such as video streaming, social networking and large file transfers. In many markets, voice and SMS traffic is being replaced by various web-based applications. Looking into the future, the five main applications for mobile data are considered to be cloud computing, different types of streaming, back-up and storage, full motion gaming and video communications.

There are several factors that are combining to trigger the mobile data explosion. On one end of the spectrum are technology factors like advancements in wireless technologies as well as end user devices. At the other end, cloud-based applications are encouraging social networking behaviors that were unthinkable of only a couple of years ago.

While early cell phone devices were not ideally suited for data communications, the introduction of QWERTY keyboards was the first game changer. Touch-screen phones brought on another round of evolution along with dramatic improvements in human-machine interfaces and software applications, all triggered by advancements in computing power and storage that can be packed in a small form factor. While PC data consumption on mobile networks remains high, the data usage by hand-held devices/tablets has been increasing sharply.

Figure1 below provides Bell Labs' projection of data traffic over the next several years.

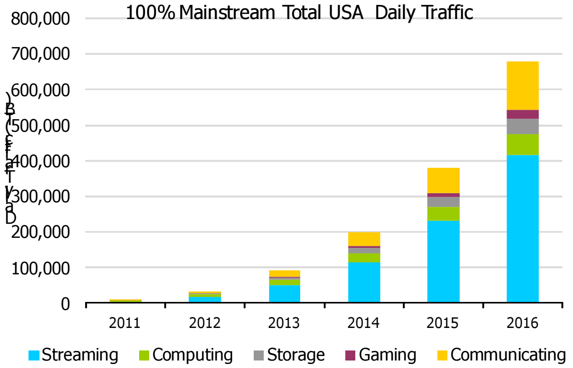


Figure1: Mobile Data Projections

Wi-Fi and LTE Applications

At the highest level, cellular networks can be used both as a mobile broadband solution, such as making a video call riding a train or bus, and as a non-mobile broadband solution, such as sitting in the backyard watching a video clip. Wi-Fi, on the other hand, is primarily used today as an extension to fixed broadband solutions. Table 1 below

[1] 2nd and 3rd Generation wireless standards that use licensed spectrum for Wide Area Networks.

[2] Long Term Evolution, a 4th Generation (4G) wireless standard that uses licensed spectrum for Wide Area Networks.

[3] A wireless technology that uses unlicensed spectrum for Local Area Networks.

illustrates the common types of applications vis-a-vis technologies.

Applications	Cellular	Wi-Fi
Fixed	Yes	Yes
Nomadic	Yes	Yes
Mobility	Yes	Very limited

Table 1: Applications & Technologies

Additional discussions describing key characteristics of cellular and Wi-Fi technologies are provided below.

Cellular

- 3G/LTE enables a high speed data connection to services when a user is mobile, in a fixed location, or when Wi-Fi is not available in a wireline broadband extension (fixed location) scenario.
- Cell site serving areas of several Km, depending on antenna height, location and geography; coupled with complex robust mobility algorithms; these help facilitate effective mobility hand-offs at vehicular speeds.
- A comprehensive security framework maintains secure connections and enables fast handoffs.

Wi-Fi

- An extension of wireline broadband via radio for "the last 100m". This includes the use of Wi-Fi hotspots in public locations, homes and enterprises.
- Data offload of licensed spectrum RAN networks using radio for "the last 100m" and offloading to broadband wireline connections.
- Nearly every mobile device and broadband modem today has built-in Wi-Fi capabilities. Many devices today can automatically search for available hotspots or can even themselves serve as Wi-Fi hotspots for other Wi-Fi devices.

Using Wi-Fi in Real-Time Mobile Applications

There are major challenges of using Wi-Fi in a real-time mobile solution in an uncontrolled public environment. These are:

- Interference – Wi-Fi uses unlicensed spectrum; a limited number of overlapping channels and uncoordinated neighboring Access Point deployments and spectrum used by competitive providers or even residential or enterprise users. This can result in interference, which in turn can limit capacity, mobility and service continuity.
- Mobility – Wi-Fi is intended as a short range wireless solution. Mobility is limited to slow pedestrian speeds. Wi-Fi mobility is defined in IEEE standards 802.11r and is generally supported within major vendor products. Not all vendors have implemented 802.11r and it is not clear whether 802.11r will be required for Wi-Fi Alliance certification. IETF is also involved in defining Wi-Fi mobility with RFC3990. Mobility at vehicular speeds is impractical due to small wireless coverage areas of Access Points, the challenges associated with hand-offs and admission control, and a lack of algorithms needed for service continuity. There is also CAPWAP, which is an IETF standard defined in RFC3990, which addresses mobility.
- Admission control (in the form of resource management) – at the time of a session handover from one Access Point to another Access Point. In the worst case, it would be similar to starting a new session for best effort traffic, though there is separate signaling that is used in 802.11 for resource reservation. Major vendors are addressing this and some may be providing seamless handoffs.
- Re-association with the target Access Point – requiring a large number of roundtrips for authentication. Security throughout mobility events, like handoffs, is not maintained, and has to be fully re-

established. 802.11r may help in reducing the number of roundtrips for the delay.

- Radio resource management granularity limits the ability to share channels between many users. This limitation is generally not noticeable in fixed and nomadic applications; it is an impediment for dense use mobile applications.
- Propagation characteristics at 2.4GHz ISM band are subject to significant interference; at 5.1GHz, signal strength fades away rather quickly resulting in smaller cell ranges and the device eco-system is still developing.

Addressing Key Wi-Fi Challenges

3GPP, working with other industry bodies, has developed two fundamental approaches for integrating Wi-Fi with cellular technologies. Figure 2 shows the architecture where the cellular operator has no control over the Wi-Fi Access Point, and Figure 3 shows the architecture when the operator has full control over the Access Point.

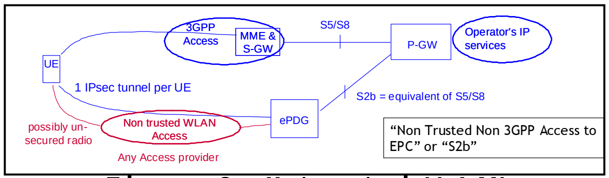


Figure 2: Untrusted W-LAN

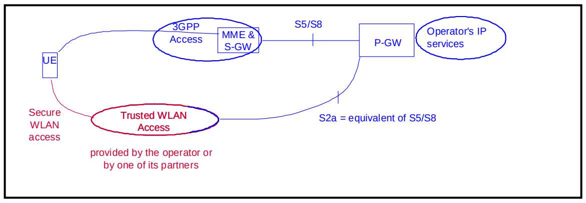


Figure 3: Trusted W-LAN

Additional developments in 3GPP continue in the form of initiatives like Access Network Delivery Selection Function (ANDSF) where the cellular network assesses the quality of experience in the Wi-Fi and Cellular networks for given applications and based on policies, may switch the user from one technology to the other. 802.11u also defines another way to achieve this. HotSpot 2.0 and Wi-Fi Alliance activities not only enable Wi-Fi roaming among operators but also open the doors for further integration between Wi-Fi and cellular networks.

OPTIMIZING WI-FI AND LTE NETWORKS

Traffic Offload to Wi-Fi

A significant part of mobile data traffic is considered nomadic and not necessarily mobile, thus making many cellular users amenable to Wi-Fi offload. It is likely that Wi-Fi offload may grow today from roughly 22% of traffic in North America to over 30% within the next four years. The amount of offload will depend upon various factors like residential broadband penetration; ubiquity of public Wi-Fi hotspots, mobile data tariffs, and technology evolution for seamless Wi-Fi-cellular integration etc., the potential for offload could be greater than 70% as seen from various studies in certain international markets.

Optimizing Wi-Fi Hot Spot Locations

Wi-Fi offers good user throughput in an interference-free environment. Figure 4 provides a comparison between different technologies, based on 3GPP simulations. The Wi-Fi value is based on typical environment for today's 802.11b/g deployment with 20 MHz channels. Pure 802.11n environment is expected to achieve 50 Mbps+ average user throughput with a 40 MHz channel.

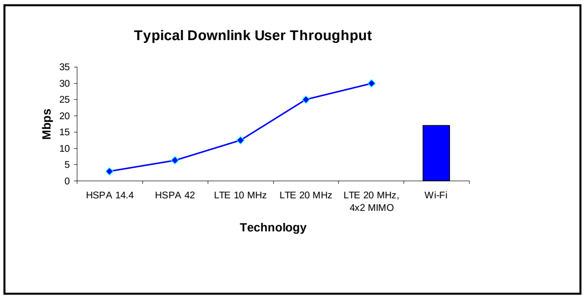


Figure 4: Throughput Comparison

However, Wi-Fi Access Points have small coverage areas, compared to macro cells. A comparison between technologies of capacity per unit area of coverage for typical dense urban environment is shown in `Figure 5` below. While the cell range for a typical macro cell is 1.2 – 1.5 km in an urban environment, typical Wi-Fi Access Point range is around 30m and generally not too much more than 100m. Typical downlink sector throughput for 3G HSPA is around 6.7 Mbps whereas for 20 MHz LTE, it can be around 30 Mbps. For 802.11g, typical downlink user throughput is around 17 Mbps.

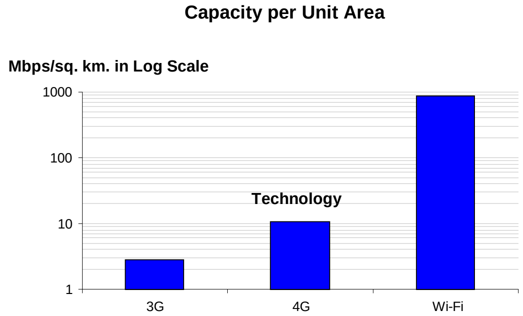


Figure 5: Unit Area Capacity

The challenge, thus, becomes how to enable maximum traffic offload Wi-Fi hotspots. Bell Labs analysis from various real networks has shown that a relatively large volume of mobile data traffic (50% - 60% or more) is often contained in a relatively small geographical area (10% - 15%) under a macro cell coverage area. `Figure 6` shows the relationship between geographical area and amount of traffic during busy hour in a macro cell in a large North American city – in this particular example, only about 8% of the geographical area contains 60% of mobile data traffic.

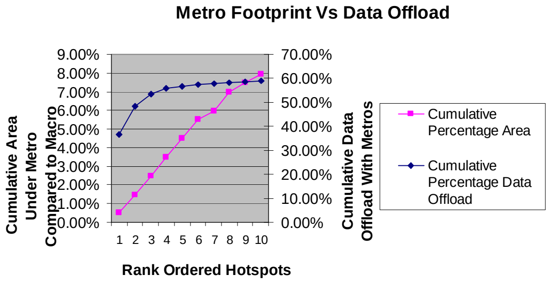


Figure 6: Traffic Density

To help address this challenge, industry techniques have been developed to create traffic density maps. This helps make a well-informed decision on placement of the Wi-Fi Access Points.

Techno-Economic Analysis

While the cost of a consumer Wi-Fi Access Point is almost negligible compared to the cost of cell site equipment, the cost of carrier grade Access Points is significantly higher than the cost of consumer Access Points. First, environmental hardening and security costs add significantly to capital expenses. Additionally, ongoing costs of backhaul and site rental significantly impact the Total Cost of Ownership (TCO). But overall, carrier grade Access Point costs are lower than macro cell site equipment.

Whether Wi-Fi deployment is economical or not depends upon a wide range of factors, including technical as well as commercial factors. In `Figure 7` below, we provide a simple normalized cost comparison, using a subscriber's monthly usage as a reference.

The cost points are used from a typical large wireless operator in Europe. The reference coverage area is the footprint of a macrocell in a large European city. It may be noted that the cost points for the Wi-Fi Access Points are for environmentally hardened network elements as required for outdoor deployment, which are significantly higher than indoor Access Points.

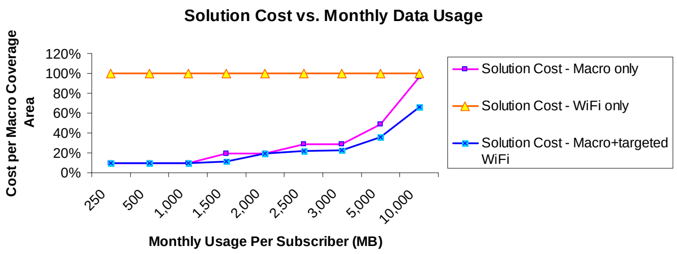


Figure 7: Technology Cost Comparison

The figure clearly shows that covering the entire macro footprint with Wi-Fi Access Points is an impractical solution. A macro-only solution is suitable for low data usage per subscriber but as the traffic per subscriber increases, macro complemented with targeted Wi-Fi deployment becomes the cost-optimal solution.

SUMMARY

Wi-Fi and LTE each have their own set of applications, but are most importantly complimentary:

- Wide area coverage with full mobility using licensed spectrum base stations with higher power and operating via higher towers (e.g., LTE) as a compliment to lower power unlicensed spectrum street level or campus environment deployments (Wi-Fi).
- Coverage and capacity limited network design that is independent of local deployments of other WLANs where the design can be impacted negatively if another WLAN is deployed nearby.
- Effective offloading of data traffic from congested cellular networks can be achieved by transporting this traffic over wireline and Wi-Fi facilities while enabling the user to enjoy the rich applications provided by these networks.
- Roaming capabilities and common authentication methods using the Wi-Fi network are adopted and certified by the Wi-Fi Alliance4.
- Careful identification of dense traffic areas and locating Wi-Fi Access Points at those locations is a key to efficient cellular-Wi-Fi integration.

Cable operators can leverage their networks not only for using Wi-Fi as an extension of fixed access broadband services, but also in partnering with wireless service providers to use Wi-Fi for offloading cellular network traffic. Optimizing the performance and cost of heterogeneous networks comprised of cellular and Wi-Fi technologies is critical for performance, customer satisfaction and cost management.

ACKNOWLEDGMENTS

The authors would like to thanks our Bell Labs colleagues Perminder Dhawan, Mohamed El-Sayed, Bob Novo and Carlos Urrutia-Valdés for providing valuable references and critique of this document.

REFERENCES

1. 3GPP TS 22.234, Requirements on 3GPP system to Wireless Local Area Network (WLAN) interworking
2. 3GPP TS 23.234, 3GPP system to Wireless Local Area Network (WLAN) interworking; System description
3. 3GPP TS 23.402, Architecture enhancements for non-3GPP accesses
4. 3GPP TS 24.312, Access Network Discovery and Selection Function (ANDSF) Management Object (MO)

[4] HotSpot 2.0 standardization effort and certification is expected in mid 2012.

PUSHING IP CLOSER TO THE EDGE

Rei Brockett, Oleh Sniezko, Michael Field, Dave Baran
Aurora Networks

Abstract

The ongoing evolution of cable services from broadcast video to narrowcast digital content (both data and video) has fuelled corresponding technical innovations to solve and support operators' operational and capital requirements. One area of particular interest is the QAM modulator. Accelerating subscriber demand for data and narrowcast video services will require a surge of new QAM deployments over the next several years, giving rise to a host of operational difficulties.

In this paper, we present the case for distributed headend architecture for HFC networks and discuss architectural and operational benefits of the Node QAM form factor, where the conversion of digital payload into QAM-RF signals is pushed from the headend to the cable TV optical node. In addition, we analyze the Node QAM in the context of the CableLabs® Converged Cable Access Platform (CCAP) architecture.

BACKGROUND

Distributed Architecture Drivers

A key topic when discussing next-generation cable infrastructure is the balance between analog optical transmission, including the transmission of multicarrier QAM-RF signals, and baseband digital transmission of signals such as native Internet Protocol (IP) signals. Cable operators have gone through several transitions already, with the introduction of digital television; the growth of high-speed data; the use of IP-based distribution in the headend; and the use of native baseband IP-based communication between headends and hubs. The driving force has always been efficiency and cost.

The imperative to meet subscriber demands results in certain bottlenecks: physical space and power within the headend, bandwidth capacity in the deployed HFC, distance between headend and subscriber, limitations of hard-wired infrastructure.

For each of these areas, there are solutions, but a distributed headend architecture that extends the boundary point where content enters the RF domain addresses all of these:

- Headend space and power consumption can be mitigated by consolidating functionality and increasing port densities in next-generation CMTSs and Edge QAMs. Alternatively, functionality can be distributed to the hubs and nodes, leaving only the IP network and MPEG2-TS processing in the headend. Direct generation of RF output at the edge of the network eliminates the need for an RF combining network at the headend. This reduces headend space and power requirements and simplifies network operations by avoiding the need to mix signals in the RF domain.

- Distance limitations can be relaxed by pushing deeper the conversion of digital signals to RF. Analog optical transmitters and amplifiers are at the limits of their capabilities, and add expense and design complexity. However, by extending the headend IP domain to the node, not only is optical transmission distance extended, but RF signal loss budgets are mitigated and higher loss budget at higher frequencies can be accommodated, thus increasing bandwidth capacity of the subsequent coaxial section of the HFC network. For example, baseband optical links to the node would eliminate analog link contributors to signal degradation, thus allowing for higher modulation levels and hence better spectral efficiency in the available coaxial bandwidth. This can be especially effective and fruitful in passive coaxial networks (PCN), also known as Fiber Deep, Fiber to the Curb (FTTC), or Node-plus-zero (N+0) HFC networks.

- In addition to the effect of explicit signal impairments due to analog optical transmission, bandwidth capacity in the HFC network is further constrained by the complexity of carrying analog (RF) signals over distance. In the optical links to the nodes, the use of multiwavelength systems, while justified by fiber scarcity and revenue opportunities, introduces severe constraints on the usable number of wavelengths and their link performance. Impairments from analog (RF) modulated optical transmitters and erbium-doped fiber amplifiers (EDFAs) further limit the capacity of individual wavelengths. Converting from RF modulated transmitters to baseband digital optics would eliminate these impairments and increase the number of cost-effective wavelengths to 88 (yielding 880 Gbps of capacity to each node) using current technology, with room for growth in the number of wavelengths and the wavelength capacity of next-generation optics.

- The challenge of managing bandwidth allocation between unicast, multicast, broadcast, and data QAM signals is eliminated by mixing content dynamically in the headend IP network. This allows bandwidth to be allocated as-needed in response to market requirements without requiring “hands on” labor.

Accelerating Demand for Narrowcast Services

Rapidly evolving subscriber behavior surrounding the consumption of multimedia is driving cable operators to confront two challenges. The first is the need to significantly accelerate the deployment of narrowcast services while also accommodating bandwidth-intensive services such as HDTV and 3DTV. These narrowcast services typically include high-speed data and packet voice, video on demand (VoD), and switched digital video (SDV), but also encompass other unicast and multicast services such as cable IPTV, network-based digital video recording (nDVR), and other services that leverage the IP cloud at the headend. The second challenge is the difficulty of planning a graceful and cost-effective migration from inefficient and obsolete service silos to new, dynamic methods of flexibly allocating capacity to different services in the face of constantly shifting customer demands.

The need to deploy an unprecedented volume of new QAM modulators is common to both challenges, and this raises concerns over issues including headend environmental constraints, flexibility of service allocation, RF combining issues, HFC transmission considerations, and the need to accommodate legacy equipment.

In these circumstances, one viable solution that achieves the benefits listed above is to relocate the QAM modulators to the HFC node, pushing the native baseband IP domain even further to the edge (closer to the user — the ultimate edge of the HFC network).

DESIGNING A NODE QAM

A Confluence of Technology and Need

Quadrature Amplitude Modulation (QAM) is a spectrally efficient way of using both amplitude and phase modulation to transmit a digital payload on an analog carrier. Cable QAM modulators[1] operate on packets in the MPEG2-TS format, and modern QAM modulators include integrated upconverters as well.

In the decades since the first baseband QAM modulators were assembled out of discrete components, silicon technology has increased a thousand-fold in processing price performance, and decreased a hundred-fold in size, giving rise to a surprisingly rich selection of special-purpose, general-purpose, and programmable chips, based on which we can re-design our modulators.

These advances can finally be used to their full advantage now that demand for modulators has swelled from tens and twenties per headend to hundreds and even thousands. Part of the advantage is in the availability of brute-force processing power, but a companion advantage is in algorithmic efficiencies derived from being able to perform certain steps in bulk. One result is that existing headend Edge QAMs can be made much denser, with thousands of QAM channels in a chassis. Another result is that it is now operationally feasible to put a full gigahertz' worth of QAM channels (or more) in the node.

Node QAM Requirements

The node is a hostile environment for advanced electronics. Power budget and space are limited; cooling is passive; operating temperatures can be extreme; and accessibility is limited. In order for a Node QAM to be operationally neutral when compared to a headend Edge QAM, it must meet the following criteria:

- Low power. In order to avoid the need for non-standard node powering, a full-spectrum Node QAM must be able to generate at least 158 (6 MHz) QAM channels using the same amount of power as a traditional optical receiver. This eliminates the need for active cooling.

- Compact. The Node QAM should be designed to fit within the existing, field-proven node housings.

- Industrial grade operating temperature range (–40°C to +85°C). Unlike climate-controlled headends, or even cabinet-based hubs, components in the node must be able to withstand large fluctuations in temperature.

- Reliable. Servicing a node is logistically cumbersome and operationally expensive. A Node QAM must be robust and uncomplicated. Additionally, remote monitoring is critical. Ideally, cost, space, and power consumption profiles can be kept low enough to enable the deployment of spare modules, which would allow operators high levels of redundancy, even at the node level.

- Simple to install — "Set it and forget it". Installing a Node QAM must be as simple as plugging in a module and verifying the output with a field meter. Complex procedures such as configuration and management should be done centrally, to simplify operations.

- Low cost. Per-channel equipment costs need to keep pace with the cost of headend Edge QAMs.

- Future-proofed. Given the logistical difficulties of servicing nodes, the distributed Node QAM modules should have a margin for upgradability so future technological changes and additions can be accommodated by re-programming the existing modules. This not only simplifies architectural evolution, but also extends the operational lifetime of each module.

This is also applicable to the interfaces between the node modules and the headend/hub infrastructure; new modules can be introduced in a very scalable manner if they leverage the standard data networking interfaces used in the IP network in the headend.

These requirements, while difficult to achieve, are attainable given modern silicon capabilities and careful design, opening up the option to move to a more distributed architecture, with many of the benefits.

ARCHITECTURAL BENEFITS

Generating some or all QAM signals at the node results in a number of advantages.

Exploiting Digital Optics

A major advantage of moving the QAM modulator to the node is the ability to shift to digital optics between the headend and the node. In traditional usage, electrical RF signals are amplitude-modulated onto an optical signal. These signals are extremely sensitive to various fiber nonlinear distortions like cross phase modulation (XPM), stimulated Raman scattering (SRS) and optical beat interference (OBI) caused by the four-wave mixing (4WM) products that come into play depending on power, distance, wavelength count, and other factors. Together with other nonlinear and linear fiber impairments, they limit the capacity of the links and significantly impair transported signals. Designing and “balancing” optical links to the nodes in an HFC system is a delicate art. Furthermore, the lasers modulated with analog (RF) multicarrier signals have limited Optical Modulation Index (OMI) capacity due to the fact of high sensitivity of these signals to clipping. The limits reach up to 30% for directly modulated lasers and approximately 20% for externally modulated lasers. These limits, with the operational back-off of 2-3 dB, severely limit the capacity of every single wavelength in any multiwavelength system of practical distance.

Using baseband digital transmission is much simpler. Because data is not as sensitive to nonlinearities and other impairments, not only can distances be extended, but more wavelengths within a single fiber can be employed, resulting in higher bandwidth capacity to the node. Simpler and more economical optics and amplifiers can be used, as well. With their OMI approaching 100%, digital optics enable significant increases in the capacity and distance of each fiber optic link. Using existing technologies, they can support cost-effective transmission of 88 wavelengths with 10 Gbps/wavelength over distances in excess of 100 km from the IP headend/hub infrastructure. This opens significant opportunity to provide unparalleled bandwidth to the nodes for residential services as well as significant opportunity for additional revenue.

More cost optimizations for capacity and distance can be achieved by leveraging lower-cost optical amplifiers, simplified optical filters, and symmetric and asymmetric SFP, SFP+ and XFP transceivers. Furthermore, deploying distributed architecture and transmitting native baseband IP signals to the node finally enables HFC to take advantage of the high-volume economies of scale in modern digital (data) networking infrastructure, which outperforms the economies of scale for analog cable TV optics a thousand-fold.

Another benefit of using baseband digital optics between the headend and the node is the elimination of an HFC weakness: the analog link contribution to end-of-line noise budgets. The analog (RF) optical links to the nodes with analog (NTSC or PAL) video signals are designed for 47 to 50 dB carrier to noise ratio (CNR) for occupied bandwidths ranging from 700 to 950 MHz. For QAM signals placed on the same link, it translates to modulation error ratio (MER) between 39 and 42 dB. For links with QAM-only load, this limit is usually lowered by designers to 37 dB MER to take advantage of cost tradeoffs and increase fiber utilization efficiency and reach. This is sufficient to support a modulation order of 256-QAM, but it limits the capacity of the HFC link to between 5 and 6.4 Gbps. Improved noise budgets by using the Node QAM would allow the support of 1024-QAM modulation, over a bandwidth range up to 1800 MHz, resulting in throughput capacity of 15 Gbps, nearly triple the current capacity.

Digital baseband transmission would unlock practically unlimited capacity in the fiber links to the node. With the proximity of the node to the furthest service user, especially in PCN networks, distributed fiber to the home (FTTH) solutions like Next-Gen RFoG and xPON can be extended from the nodes selectively, based on the demand and opportunities.

Simplification of RF Combining Network

Generating QAM signals in the node allows those QAM signals to bypass the RF combining network. Node QAM output signals can be combined *at the node* with traditionally carried HFC signals in a single stage. New narrowcast QAM signals can be added at the node as needed, with no impact on either the existing RF combining network or the HFC plant alignment.

Besides removing the complexity of recalculating the headend combining plant each time new RF ports are added, it avoids both the signal and power losses associated with combining, splitting, and directional coupling, as well as the power, cooling burden, and significant space inefficiencies. Many of these advantages are delivered by the CableLabs Converged Cable Access Platform (CCAP)[2] architecture, as described later. A distributed architecture goes a step further by allowing legacy signals to be combined in a single passive combining stage at the node.

RF Signal Quality and Node Alignment

When QAM signals are generated in the node (Figure 1), with given output levels and the same or better output signal quality as headend-generated QAM signals, the resulting RF signal in the node is much cleaner. This is because it bypasses the signal losses, noise, attenuation, and distortions that are typically introduced in the RF combining network and the amplitude-modulated optical links to the node.

Operationally, it reduces the amount of RF aligning needed at the node; output power and tilt are generated exactly according to configured specifications, defined by the operator. The signal is not subject to any of the traditional distortions. The impairment contribution of combining network and analog (RF) optical links to nodes is eliminated, with the benefit of unlocking coaxial plant capacity as described above. This allows a 43+ dB MER (see Figures 1 and 2) at the node and gives the operator more options in the coaxial portion in terms of loss budget/coverage and, most importantly, bandwidth. In certain conditions, it makes higher-order modulation rates possible as well, resulting in better spectral efficiency.

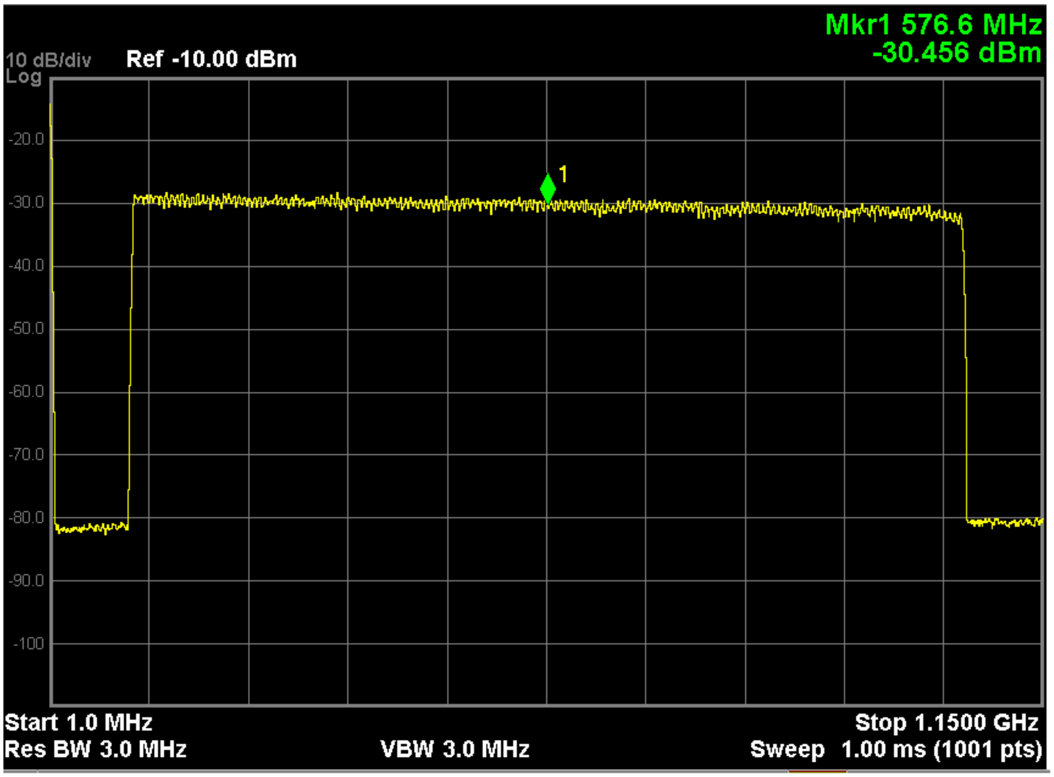
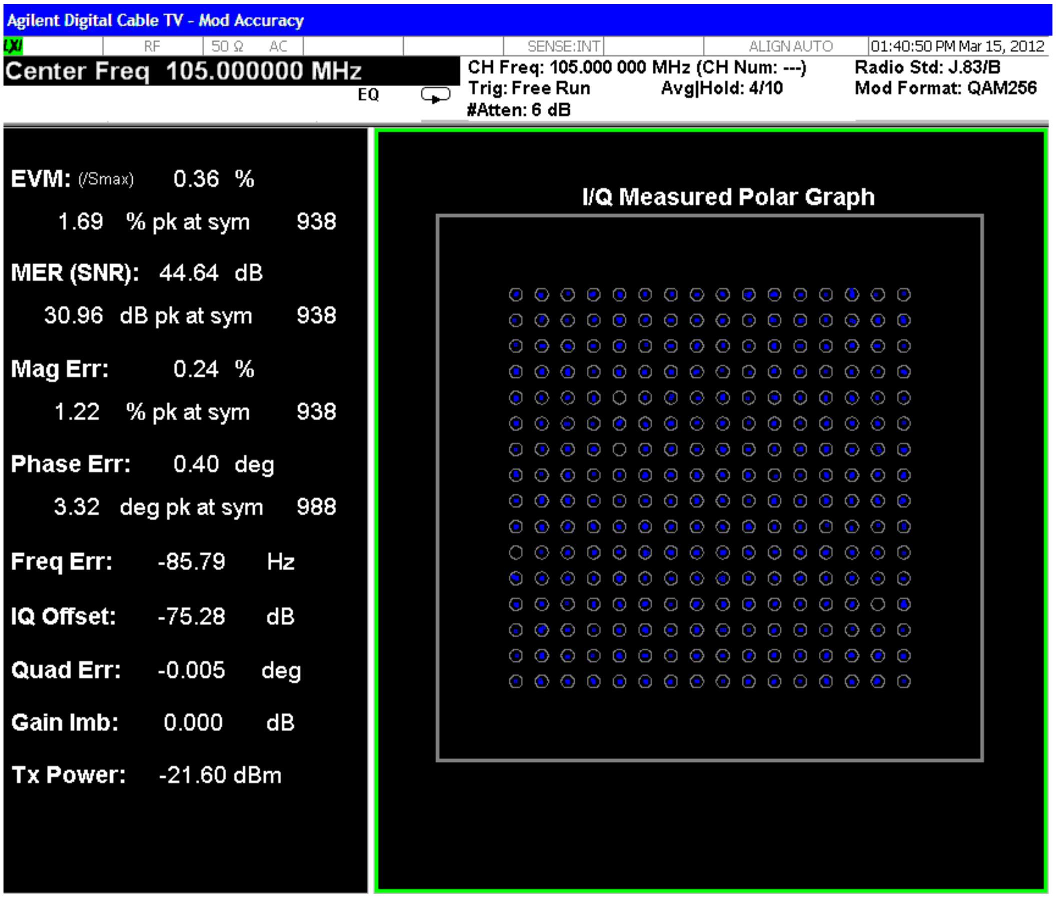


Figure 1: 158 Node QAM channels



01:40:50 PM Mar 15, 2012

Center Freq 105.000000 MHz

CH Freq: 105.000 000 MHz (CH Num: ---)
Trig: Free Run
Avg|Hold: 4/10
#Atten: 6 dB

Radio Std: J.83/B
Mod Format: QAM256

```
EVM: (/Smax)    0.36  %
    1.69   % pk at sym     938
MER (SNR):  44.64  dB
   30.96  dB pk at sym     938
Mag Err:        0.24  %
    1.22   % pk at sym     938
Phase Err:      0.40  deg
    3.32   deg pk at sym   988
Freq Err:     -85.79     Hz
IQ Offset:    -75.28     dB
Quad Err:     -0.005     deg
Gain Imb:      0.000     dB
Tx Power:     -21.60 dBm
```

Figure 2: Node QAM increases RF loss budget or bandwidth capacity.

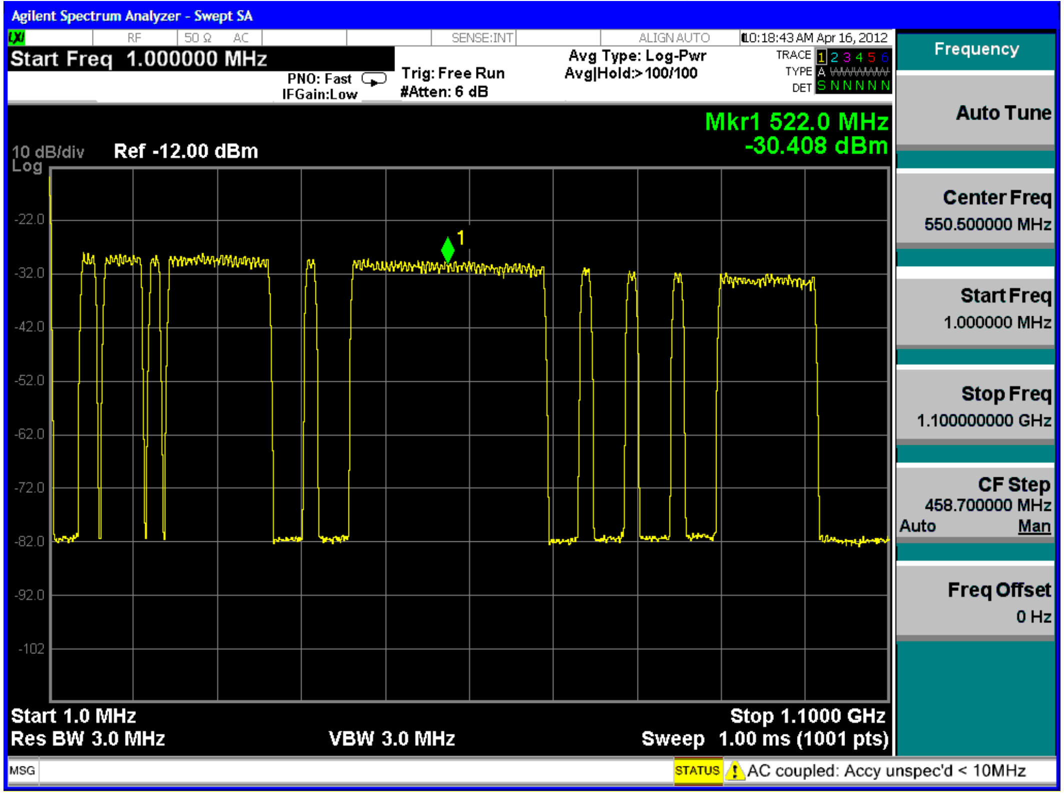
Service Flexibility

An important side effect of the Node QAM is that the optical network feeding it is a de-facto extension of the headend IP network, with access to all of the system's digital content — broadcast, narrowcast, unicast, and data.

The Node QAM itself is agnostic to the digital payload; it simply modulates the MPEG2 formatted transport streams that are delivered over the optical interface. The payload carried within the transport stream could be a groomed and re-quantized statistical multiplex; it could be an encrypted variable bit rate broadcast multiplex; it could be a simple multiplex of fixed-rate VoD streams; or it could be a DOCSIS M-CMTS-compliant data stream.

The contents of the transport streams are dependent only on the capabilities and sophistication of the headend service manager(s) and resource manager(s), and switched IP connectivity. Artificial service group constraints imposed by the hard-wired RF combining network are removed, leaving only a general-purpose pool of QAM signals to feed the population of subscribers attached to each node or node segment.

An enhancement enabled by the generation of QAM signals in the node from native IP input is the ability to selectively reserve local bands of frequency for other modulation and encoding schemes as well. See Figure 3.



Agilent Spectrum Analyzer - Swept SA

RF 50 Ω AC SENSE:INT ALIGN AUTO 10:18:43 AM Apr 16, 2012

Start Freq 1.000000 MHz

PNO: Fast
IFGain:Low

Trig: Free Run
#Atten: 6 dB

Avg Type: Log-Pwr
Avg|Hold:> 100/100

TRACE 1 2 3 4 5 6
TYPE A WWWWW
DET S N N N N N

Mkr1 522.0 MHz
-30.408 dBm

10 dB/div Ref -12.00 dBm
Log

Start 1.0 MHz Stop 1.1000 GHz
Res BW 3.0 MHz VBW 3.0 MHz Sweep 1.00 ms (1001 pts)

MSG STATUS AC coupled: Accy unspec'd < 10MHz

Figure 3: Spectrum Allocation Agility. Individual QAM signals can be turned on or off.

Some examples of practical applications include:

- Customized broadcast lineups. Certain niche customers, such as hotels, apartment complexes, hospitals, and campuses can receive their own broadcast lineups, created on the fly, without affecting the existing RF combining network.

- Uneven service usage. Usage of individual types of narrowcast and unicast services may vary unpredictably from node to node. Node QAMs with headend service switching allows each node to have a different service mix, without having to pre-allocate resources.

- Dynamic service allocation. Service usage may also vary within a single node, based on time of day or season. For example, a suburban node might experience heavy VoD usage during the day due to toddler addictions to children's programming, but switch to heavy internet usage late at night when parents use Netflix. With the Node QAM, a single pool of QAM signals can feed all services, without having to provision under-utilized service silos.

- Mixed services within a single channel. With sufficient sophistication from the headend multiplexers and resource managers, the Node QAM can deliver any mix of QAM services — broadcast, narrowcast, CMTS, VBR, CBR in a single channel, giving the operator complete flexibility.

Environmental

While modern headend QAM modulators are an order of magnitude more energy-efficient than earlier incarnations, and two orders of magnitude more compact, the addition of large quantities of new QAM channels via traditional methods creates a significant impact on the headend, in two ways. Headend Edge QAMs create a direct impact by their intrinsic consumption of power, rack space, and cooling mechanisms. They also have an indirect impact, due to the rack space occupied by the combining network; the power loss due to combining, splitting, and directional coupling of service groups, as well as the power consumption of intermediate amplification stages; and the power burden of heating, ventilation, and air conditioning (HVAC).

By moving QAM modulation to the node, not only are power and rack space requirements distributed, but overall per-QAM power and space consumption are reduced due to the fact that lower output levels are needed to drive the existing node RF amplification modules. This helps the Node QAM to live within the design constraints imposed by the node housing, including the use of passive cooling instead of fans. Node QAMs also eliminate the Edge QAMs' impact on the headend HVAC system.

In addition, by bypassing the RF combiner network at the headend, Node QAMs avoid wasting the signal power maintained by the RF combiner network's amplification stages, which end up being discarded when the signal is carried in its baseband digital format. Furthermore, power and space requirements are reduced when optical analog (RF) transmitters are replaced by low-power optical digital baseband transceivers.

These Node QAM benefits mesh well with the fundamental goals of the CCAP architecture, with the added advantages that Node QAM leverages digital optics, and that these benefits accrue on a node-by-node basis, allowing both small and large operators to migrate gracefully to CCAP.

CCAP

CableLabs' CCAP architecture is a bold step in addressing many of the challenges related to the growth of narrowcast services. It leverages heavily the existing body of Data-Over-Cable Service Interface Specifications (DOCSIS) with the goals of increasing the flexibility of QAM usage and configuration; simplifying the RF combiner network; possibly adding content scrambling; creating a transport-agnostic management paradigm to accommodate native support of Ethernet Passive Optical Network (EPON) and other access technologies; improving environmental and operational efficiencies; and unifying headend configuration and management capabilities. CCAP includes a new Operations Support System Interface (OSSI)[3] specification and also takes particular care to ensure compatibility with existing DOCSIS resource management and service management and configuration specifications, in order to facilitate the migration from current CMTS/Edge QAM infrastructure.

CCAP Reference Architectures

CCAP unifies digital video and high-speed internet delivery infrastructures under a common functional umbrella, allowing a CCAP device to be operated as a digital video solution, a data delivery solution (both CMTS and M-CMTS), a Universal Edge QAM, or any combination. Each of the CCAP reference architectures (Video, Data, and Modular Headend) describe physical and functional interfaces to content on the "network" side, operational and support systems within the headend, and the HFC/PON delivery network terminating in various devices at the subscriber premises. Ancillary service and resource managers are allowed to exist both within and externally to a CCAP device.

CCAP OSSI

The lynchpin of the CCAP architecture is the CCAP OSSI, which defines a converged object model for dynamic configuration, management, and monitoring of both video and data/CMTS functions, but also makes provision for vendors to innovate within the framework. By creating a unified standards-based operational front-end to the video and data delivery infrastructure, CCAP OSSI provides a solid foundation for the headend's metamorphosis from a collection of separately managed service silos into an efficient service delivery "cloud".

CCAP and Node QAM

In the CCAP video and data reference architectures, the CCAP interface on the subscriber side is the HFC network. Traditionally, that interface exists within the headend. However, there is nothing inherent about the provisioning and management of QAM signals that *requires* the QAM modulators to be in the headend. Extending the logical boundary of the headend out to the node and minimizing the analog portion of the HFC remains consistent with the goals and specifications of CCAP.

NODE QAM EVOLUTION

Initial Architecture

The initial configuration of the Node QAM topology can be envisioned as one presented in Figure 4. In this configuration, analog and operator selected QAM broadcast channels (*e.g.*, from a different location than the remaining QAM channels) are transported to the node in a traditional fashion but without the burden of combining with the remaining QAM channels in the headend/hub. The number of QAM channels originating in the Node QAM can be adjusted dynamically by the operator.

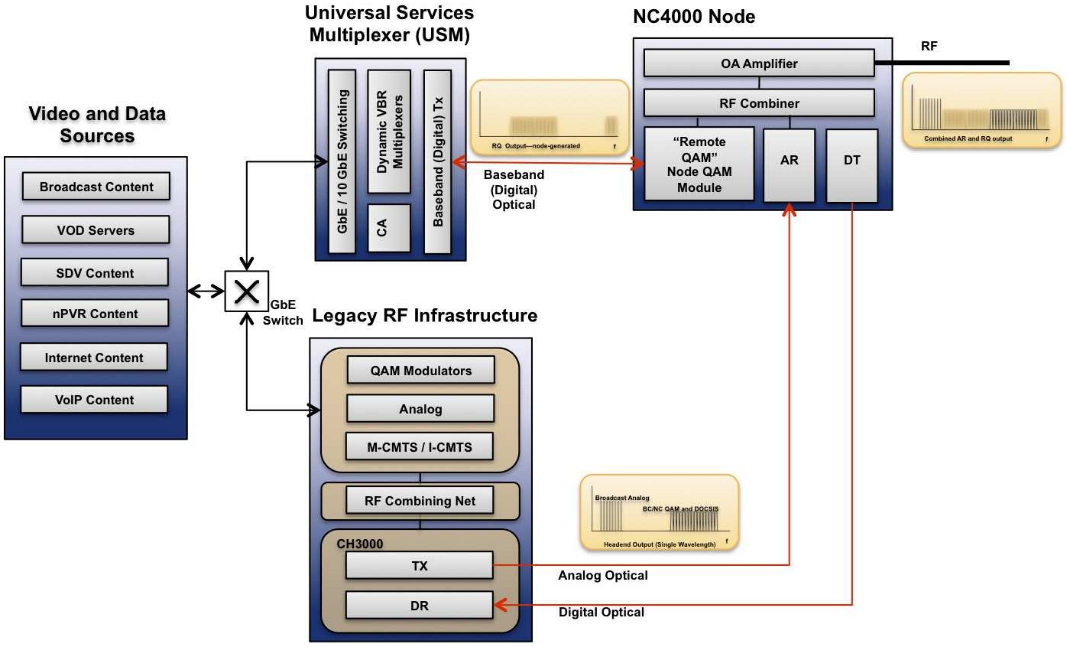


Figure 4 Node QAM Initial Implementation

Conversion to Complete Digital Baseband Node Transport

The next incarnation of the distributed architecture is presented in Figure 5. All analog channels and maintenance carriers are digitized in the headend and transported over the same transport (capacity allowing) to the node where they are frequency-processed and converted back to analog channels at their respective frequencies on coaxial plant. Some additional carriers (*e.g.*, ALC pilot signals) are synthesized in the Node QAM module.

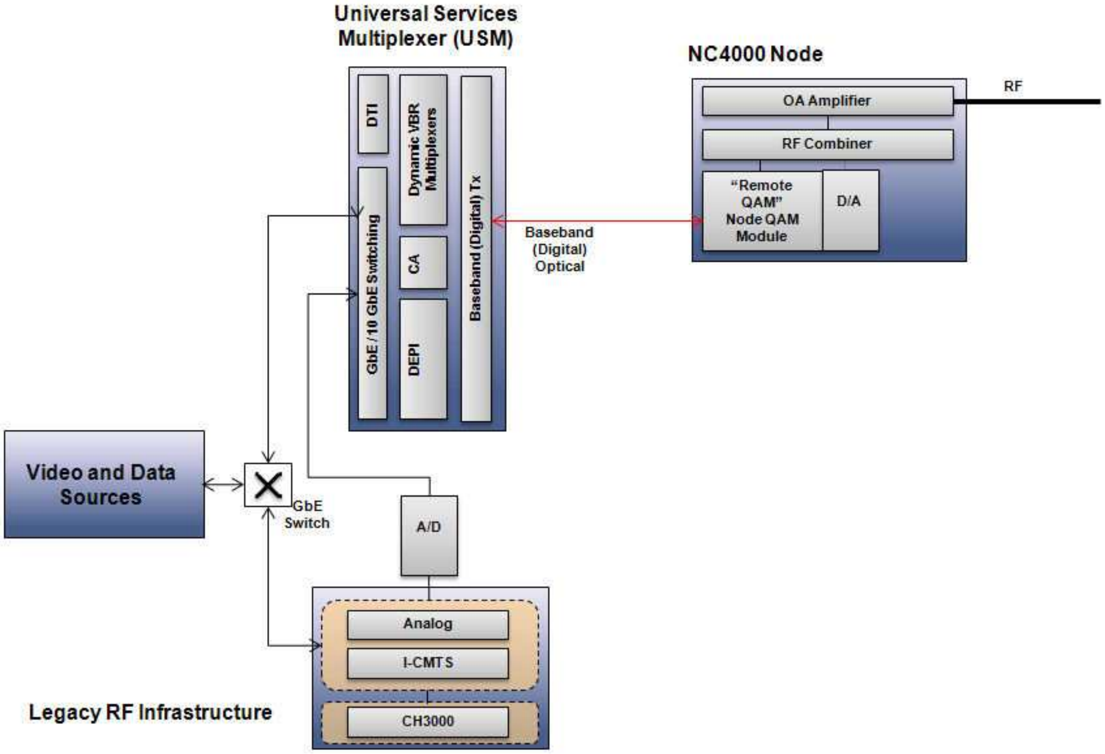


Figure 5: Node QAM Next-Generation

The reverse channel(s) from the node to the headend can also be converted to baseband digital optics, resulting in similar benefits. Options include traditional digital return (digitization of the return spectrum at the node), developing a node-based CMTS (or node-based DOCSIS burst receivers), or even next-generation native IP-over-coax technologies.

A related enhancement arising from the Node QAM's dynamic frequency agility is the ability to support flexible, remotely configurable frequency splits or capacity allocation between downstream and upstream communication, either using frequency division duplex (FDD) or time division duplex (TDD) transmission. This would enable full flexibility and adaptability to downstream and upstream traffic patterns and capacity/service demands.

Future Enhancements

The Node QAM is an ideal platform to be modified to support other modulation schemes for next-generation transport mechanisms, such as EPON Protocol over Coax (EPoC). Implementing EPoC in the node allows significant reach expansion, preserving and facilitating headend and hub consolidation without deploying additional signal conditioners or RF-baseband-RF repeaters with their additional cost, power consumption, added operational complexity of provisioning and additional space/housing requirements in the field or hubs.

OTHER ELEMENTS OF DISTRIBUTED ARCHITECTURE

Node PON

A distributed node-based EPON architecture shares the Node QAM architectural advantages. Node PON modules allow for selective fiber placement from the node for commercial services in node areas where construction costs and effort are limited to fiber extension from the node. In PCN architecture, this is usually below 1 km, and mostly below 300 m if the node is placed strategically. In conjunction with DOCSIS Provisioning of EPON (DPoE) and CCAP, Node PON can address the needs of fast deployment of dedicated fiber links to selected high capacity demand users.

Next-Generation RFoG[4]

In situations where fiber exists all the way to the subscriber, RF over Glass (RFoG) in a distributed architecture has the potential, with minor changes, to exceed the throughput of 10G PON/EPON, without the complexity of adding a PON overlay. This allows for seamless expansion of fiber from RF optical nodes to residences without replacing the distributed architecture node modules. Taking fiber from the node all the way to the subscriber with a FTTH network would allow for additional capacity enhancement beyond 15 Gbps downstream and 1 Gbps upstream facilitated by distributed coaxial architecture, especially with PCN and residential gateways deployed. With RFoG in a distributed architecture, 20+ Gbps downstream and 3 Gbps upstream is achievable today without PON overlay.

SUMMARY

No-one knows precisely what the future will bring but it is clear that subscriber-side demand for IP-delivered multimedia continues to grow as “smart” home and mobile electronic devices proliferate. The cable industry is blessed with the most extensive and highest bandwidth conduit to that last-mile “IP cloud”. At the same time, cable headends have largely already made the transition to IP-based distribution. Moving the native baseband IP-to-RF transition point from the headend to the node brings the convergence of IP headend and IP home one step closer.

As discussed in this paper, there are many advantages to extending the digital headend domain as far into the network as possible, in terms of performance, resource utilization, operational simplicity, and service flexibility. There are many paths for the evolution to digital HFC: the Institute of Electrical and Electronics Engineers (IEEE) is proposing a new physical layer standard called EPON-Protocol-over-Coax (EPoC) to deliver IP traffic natively at 10 Gbps over last-mile HFC; fiber vendors continue to innovate on bringing fiber to the home; new silicon may enable conversion of large bands of RF spectrum at the headend into digital bitstreams that can be converted back to analog at the node. By bringing IP closer to the edge, the Node QAM helps pave the way to a distributed headend and digital HFC.

ABBREVIATIONS AND ACRONYMS

10G-EPON	IEEE 802.3 Ethernet PON standard with 10 Gbps throughput
3DTV	3D Television
4WM	Four Wave Mixing
ALC	Automatic Level Control
BER	Bit Error Rate
CBR	Constant Bit Rate
CCAP	CableLabs® Converged Cable Access Platform
CMTS	Cable Modem Termination System
CNR	Carrier-to-Noise Ratio
DOCSIS®	Data over Cable Service Interface Specification
DPoE™	DOCSIS Provisioning of EPON
EDFA	Erbium-doped Fiber Amplifier
FDD	Frequency Division Duplex
FTTC	Fiber to the Curb
FTTH	Fiber to the Home
EPoC	EPON Protocol over Coax
EPON	IEEE 802.3 Ethernet PON standard with 1 Gbps throughput, a.k.a. 1G-EPON, G-EPON or GEPON
Gbps	Gigabits per second
HDTV	High Definition Television
HFC	Hybrid Fiber Coaxial
HVAC	Heating, Ventilation and Air Conditioning
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPTV	IP Television
M-CMTS	Modular Cable Modem Termination System
Mbps	Megabits per second
MER	Modulation Error Ratio
MPEG2	Motion Picture Experts Group 2 standard
MPEG2-TS	MPEG2-Transport Stream
nDVR	Network-based Digital Video Recording
NTSC	National Television System Committee
OBI	Optical Beat Interference
OMI	Optical Modulation Index
OSSI	Operations Support System Interface
PAL	Phase Alternating Line
PCN	Passive Coaxial Networks
PON	Passive Optical Network
QAM	Quadrature Amplitude Modulation
RF	Radio Frequency
RFoG	Radio Frequency over Glass
SDV	Switched Digital Video
SFP	Small Form-factor Pluggable
SRS	Stimulated Raman Scattering
TDD	Time Division Duplex
VBR	Variable Bit Rate
VoD	Video on Demand
XFP	10 Gigabit Small Form-factor Pluggable
XG-PON	ITU-T's broadband transmission standard with 10 Gbps throughput
XPM	Cross Phase Modulation
*x*PON	any of a family of passive optical network standards (*e.g.*, GPON, GEPON, 10G PON (BPON, GEPON or GPON)

[1] ITU-T J.83 Digital multi-programme systems for television, sound and data services for cable distribution. April 1997.

[2] TR-CCAP-V02-110614. CCAP Architecture Technical Report. June 2011.

[3] CM-SP-CCAP-OSSI-I02-120329 . Converged Cable Access Platform Operations Support System Interface Specification. March 2012.

[4] O. Sniezko. *RFoG: Overcoming the Forward and Reverse Capacity Constraints.* NCTA Spring Technical Forum 2011.

RECLAIMING CONTROL OF THE NETWORK FROM ADAPTIVE BIT RATE VIDEO CLIENTS

John Ulm & Gerry White
Motorola Mobility

Abstract

This paper provides a brief introduction to adaptive bit rate (ABR) video and discusses why handling this class of traffic well is very important to the cable operator. It then examines the major differences between ABR and the current IP and MPEG video delivery mechanisms and looks at the impact these differences have on the network. Some interesting experimental results observed with real world ABR clients are presented. A number of problems which may develop in the network as ABR clients are deployed are discussed and possible solutions for these proposed. Finally, the paper looks at the cable modem termination system (CMTS) as a potential control point that could be used to mitigate the impact of the ABR clients and regain control of the access network for the operator.

INTRODUCTION

Adaptive bit rate is a delivery method for streaming video over IP. It is based on a series of short HTTP progressive downloads which is applicable to the delivery of both live and on demand content. It relies on HTTP as the transport protocol and performs the media download as a series of very small files. The content is cut into many small segments (chunks) and encoded into the desired formats. A chunk is a small file containing a short video segment (typically 2 to 10 seconds) along with associated audio and other data. Adaptive streaming uses HTTP as the transport for these video chunks. This enables the content to easily traverse firewalls, and the system scales exceptionally well as it leverages traditional HTTP caching mechanisms.

Adaptive streaming was developed for video distribution over the Internet. In order to deal with the unpredictable performance characteristics typical of this environment, ABR includes the ability to switch between different encodings of the same content. This is illustrated in Figure 1. Depending upon available bandwidth, an ABR client can choose the optimum encoding to maximize the user experience.

Each chunk or fragment is its own stand-alone video segment. Inside each chunk is what MPEG refers to as a group of pictures (GOP) or several GOPs. The beginning of each chunk meets the requirements of a random access point, including starting with an I-frame. This allows the player to easily switch between bit rates at each chunk boundary.

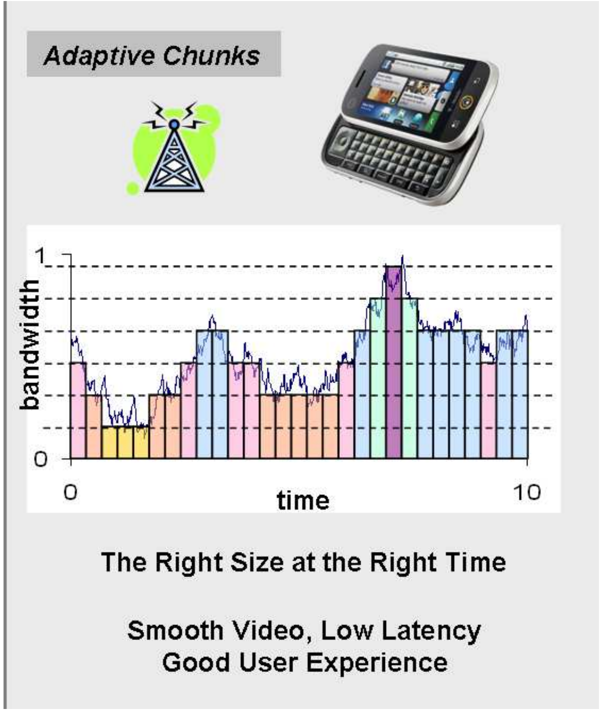


Figure 1 Adaptive Streaming Basics

Central to adaptive streaming is the mechanism for playing back multiple chunks to create a video asset. This is accomplished by creating a playlist that consists of a series of uniform resource identifiers (URIs). Each URI requests a single HTTP chunk. The server stores several chunk sizes for each segment in time. The client predicts the available bandwidth and requests the best chunk size using the appropriate URI. Since the client is controlling when the content is requested, this is seen as a client-pull mechanism, compared to traditional streaming where the server pushes the content. Using URIs to create the playlist enables very simple client devices using web browser-type interfaces. A more in-depth discussion of ABR video delivery can be found in [ADAPT]

IMPORTANCE OF ABR

Second and Third Screens

ABR based video streaming has become the de-facto standard for video delivery to IP devices such as PCs, tablets and smart-phones. ABR clients are typically shipped with (or are available for download to) these devices as soon as they are released. Given the short lifetime of this class of device this is a key enabler, especially compared to the time required to deploy software to traditional cable devices. As mentioned previously, ABR delivery simply requires an HTTP connection with sufficient bandwidth so that it is available both on net and off net. With these advantages, both over-the-top (OTT) and facilities based service providers are leveraging ABR so that essentially all video delivery to second and third screen devices uses this mechanism.

Primary Screen

ABR is also used to deliver a significant quantity of video to television screens in both standard and high definition formats. Over-the-top providers of video service leverage ABR clients installed in platforms such as gaming consoles, Blu-ray players, set-top box-like devices and smart TVs to provide video services to the primary screen. This content rides over the service providers' high speed data (HSD) service and, in many cases, constitutes the bulk of the HSD traffic.

ABR Traffic Load

Studies of Internet traffic patterns [SAND], [VNI] show that video has become the dominant traffic element in the Internet, consuming fifty to sixty percent of downstream bandwidth. Netflix alone constitutes almost thirty-three percent of peak hour downstream traffic in North America. Thus, how well the network supports ABR based IP video is obviously crucial to providing a satisfactory customer experience. In addition, delivery of Internet video to televisions is predicted to grow seventeen-fold by 2015 to represent over sixteen percent of consumer Internet video traffic (up from 7 percent in 2010) [VNI]. Thus, many of the customers will not only be viewing IP video, but will be doing so on a large screen device with expectations of high quality.

In addition to the Internet video explosion, significant amounts of managed service provider video will also migrate to an ABR mechanism, further increasing the percentage of ABR traffic on the network.

Having this much ABR traffic on the network means that it will be a key driver of network costs and with ABR delivering prime entertainment services, how well it is supported will be a key metric for customer satisfaction going forward. Therefore, understanding the issues around delivery of ABR over the DOCSIS network will be crucial for MSO's video service delivery, and for their ongoing profitability.

ABR vs. CURRENT VIDEO DELIVERY

ABR video delivery has a number of very significant differences to both MPEG video delivery and streamed IP video delivered over Real-time Transport Protocol/User Datagram Protocol (RTP/UDP) as used in a Telco TV system such as Microsoft Media Room [MMR]. A number of these differences are discussed below.

Client Control

ABR has been developed to operate over an unmanaged generic IP network in which bandwidth decisions (i.e. choosing the video bit rate to request) are made by the client device based on its interpretation of network conditions. This is fundamentally different from the approaches used for existing MPEG or conventional streamed UDP video delivery, where devices under the direct control of the network operator make the important decisions relating to bandwidth. Thus, in MPEG delivery, the encoding, statistical multiplexing and streaming devices determine the bit rate for a given video stream. These devices are under control of the service provider. Similarly for a UDP streaming solution, the video is encoded and streamed at a selected rate from devices owned by the service provider. In contrast, the behavior of ABR clients is specified by the developer which, in general, will be a third party outside the service provider's control.

Variable Bit Rate

As described previously, an ABR client will select a file chunk with a bit rate which it believes to be most appropriate according to a number of factors including network congestion (as perceived by the client) and the depth of its playout buffer. Thus the load presented to the network can fluctuate dramatically. This is in stark contrast to both MPEG and UDP video streams which are either constant bit rate (CBR) or are clamped variable bit rate (VBR) (i.e., bandwidth can vary up to a maximum bit rate but not beyond it).

A more detailed discussion on the impact on network loading of a number of factors is found in a later section of this paper.

Admission Control

ABR clients join and leave the network as users start and stop applications. From a network perspective, there is no concept of a session with reserved resources or admission control. Again this is the antithesis of MPEG or UDP video in which a control plane operates to request and reserve network resources and determines whether to admit a user. In a controlled network, adding a new user session can be guaranteed not to impact existing users. Once resources are exhausted, any additional session requests will be denied, introducing a probability of blocking into the system. In an ABR model under network congestion, each new session will reduce the bandwidth available to all existing sessions rather than be denied. Thus, users may see a variation in video quality as other ABR clients start and stop. This reduction in quality during peak times is analogous to statistical multiplexing in legacy MPEG video. During peak times, the statmux reduces bit rates across the various video streams to fit within its channel. The ABR system has an advantage in that it will be over a larger channel using DOCSIS bonding.

Congestion Control

With MPEG or UDP streaming video delivery, congestion control is not relevant as the control plane provides admission control to ensure it does not occur. When ABR is used for video delivery, congestion control is a potential issue. The situation is complex in that three levels of congestion control mechanisms are involved operating at different layers in the protocol stack. At the media access control (MAC) level, the CMTS is responsible for scheduling downstream DOCSIS traffic [MULPI]. Operating at the transport level is standard Transmission Control Protocol (TCP) flow control based on window sizes and ACKs, [TCP] and, finally, at the application level the client can select the video bit rate to request. The latter two levels of control (TCP and application) are the responsibility of the ABR clients and as such are outside the control of the network operator. Interaction between these three flow control mechanisms is not well understood at this time and may have unforeseen impacts.

Prisoners Dilemma

As noted above, ABR clients have the responsibility to select the quality (bit rate) of the video they request to download. The algorithms and parameters used by each client to make this decision are outside the control of the network operator. Each client is faced with a decision not unlike the classic "prisoner's dilemma" [PDIL] in that they can elect to optimize for their own benefit or they can optimize for the common good of all clients on the network (including their own). For example, a very selfish client may never request a lower quality file even during network congestion based on the assumption that other clients will do so, and thus resolve the congestion for them. Commercial pressures to create "better" clients may drive in this direction, but if all clients move to this mode the network will fail. This is not an issue with MPEG or UDP streaming delivery as the network operator has the incentive and necessary controls to offer a quality service to all customers.

Imperfect Knowledge

Clients base their decisions on what to request based on their local knowledge rather than on an overall view of the network conditions. This is in contrast to MPEG or UDP streaming where the network operator provisions the video bit rates based on knowledge of the end-to-end network and expected loads.

The following section on potential problems will address these issues in more depth and attempt to develop some potential solutions.

ABR CLIENT CHARACTERIZATION

As discussed previously, the ABR client plays a critical role in the operation of adaptive protocols. For an operator trying to provide a differentiated quality of experience, it is important to understand how different ABR clients behave under various circumstances.

Motorola research teams took multiple different types of clients into the lab to analyze their behavior. Previous work [Cloonan] discussed results from a simulator. Our goal was to capture live client interaction. Operation during steady state was relatively stable. The interesting observations occurred during startup and when video bit rates were forced to change.

At startup time, clients try to buffer multiple segments as fast as they can. This was particularly obvious for video on demand (VOD) assets where the entire content stream is accessible. Live content tends to have a limited playlist available to the client, preventing large buffer build up. During this startup period, the clients are also calculating the available bandwidth and may decide to switch bit rate. This action may cause some segments to be re-fetched with the new resolution. Overall, the differences between clients seemed fairly subtle for startup.

In our lab environment, the amount of bandwidth available to the ABR client was adjusted. In this manner, the client was induced to switch video bit rates. After reducing available bandwidth, the clients in general made a smooth transition to a lower bit rate. Some clients reacted more quickly than others in down shifting. When the available bandwidth is opened up again, clients started searching for new higher bit rates with the associated buffering of segments, similar to startup. It was in this phase where we saw the most differences between clients. In fact, we saw differences from the same device running different revisions of their protocol.

POTENTIAL PROBLEMS

Based on the above characterization, operators must be aware of some potential problems. As was discussed, there is a burst of additional traffic during startup and when switching to higher bit rates. The system must be capable of handling this additional traffic burst.

Actively managing ABR video traffic may be challenging given that every ABR client may be operating its own disjoint algorithm. This is also compounded since client behavior may change with the download of an updated revision. Bandwidth stability may become a concern if multiple clients become synchronized. For example, the network becomes congested causing a group of clients to lower bit rates. If these clients then sense that bandwidth is available (i.e. it is released due to downshifting by other clients), there may be a surge in traffic that causes congestion, and the cycle repeats.

In general, ABR clients are designed for general Internet usage, so they tend to back off quickly and may be slow to ratchet their bit rates back up. This will create some stability and should prevent the above oscillation, but this may make it challenging to fully utilize the network bandwidth.

There are several fairness concerns that must be taken into consideration. If the current bandwidth utilization is high, then new clients just starting their video may select a lower rate than other clients are currently using. Other forms of unfairness may be introduced when network congestion causes video bit rate changes. Some clients may decide to change while others remain at current bit rates, resulting in disparity between clients.

Another concern, especially for a managed video service, is maintaining a good Quality of Experience (QoE). The more that clients change bit rates, the more potential impact there is to QoE. The system should be designed to minimize unneeded bit rate changes.

For future research, Motorola will expand its investigation to system-level behavior for a large number of disparate ABR clients. It is important that the industry grasps the system dynamics for adaptive protocols.

POTENTIAL SOLUTIONS

In a discussion of potential solutions to problems with ABR video delivery under network congestion, two types of ABR traffic must be considered: managed and best effort. Best effort video traffic is OTT types of service which, in general, would be indistinguishable from general Internet traffic.

Managed traffic would typically be video sourced by the service provider, or by a third party with whom the service provider has negotiated a carriage agreement. How well managed traffic is supported is a significant problem for a service provider as it is, in effect, a branded service for which customers will have a higher expectation.

In general, the following potential solutions apply to a managed IP video service. We will highlight where it also applies to OTT traffic.

Controlled Client

Managed ABR services may be made available only from a specific service provider application downloaded by the user. This removes the issues relating to client misbehavior and enables the operator to predict how the client will handle network congestion events.

It has the disadvantage that the operator must keep the application up to date both in terms of feature parity with other clients and with new devices and operating systems as they are released. It also makes it likely that the user must have multiple applications to access different video sources.

This is not applicable to OTT video from third parties, which will be typically be delivered to either a native client on the device or a client provided by the OTT service.

Session Control

One option to control ABR traffic is to implement a session mechanism similar to those used for more traditional video streaming. In this case a user (or possibly a proxy for the user such as a Fulfillment Manager) requesting a video asset would invoke resource checking and reservation mechanisms in the network control plane. The control plane would reserve access network bandwidth for the video session. Mechanisms such as PacketCable™ Multi-Media (PCMM) [PCMM] are in place today to enable quality of service (QoS) bandwidth reservation over DOCSIS. This is detailed later in the paper.

A problem with this approach is knowing when to start and terminate a session and specifically when to acquire and release the resources. For managed video this could be achieved by using a service provider application as described above. The application would invoke the session setup and teardown as part of the video selection and playing process. Even a controlled application implementation would need a back up mechanism to release resources as the user may simply power off a device or lose connectivity. At the minimum, a “no traffic timeout” would be needed (refer to CMTS section below for more details).

Network Override

In conventional ABR video distribution, the ABR client determines the bit rate of the next file to download from the options in the playlist and retrieves this directly from the content delivery network (CDN). This decision could potentially be overridden from the network in a number of ways.

The playlist file provides the bit rate options specified by the service provider. Normally this selection would be statically provisioned and implemented by the encoding and packaging processes as the video asset was processed. For example, each asset could have files created for 1, 2, 4 and 6 megabits per second (Mbps) and the client allowed to select between these. Modifying the selection options in the playlist file provides a potential mechanism for the network to influence the client operation. Thus in times of congestion, the high bandwidth option could be removed by providing a playlist with only 1 and 2Mbps options. This of course requires run time manipulation of the playlists. A potential problem is the lag from playlist manipulation to actual changes in bit rate selection. Even a short playlist file would probably need to represent video content lasting for a significant time so that this mechanism would have a very slow reaction time to network events. Thus, it would not respond to short term congestion events. However, if the network had well known congestion periods (e.g. 8:00 pm through 10:00 pm) it could be used to reduce congestion during these times. Alternatively, the Session Manager might provide notification when the system is congested. This mechanism would not be applicable to OTT traffic as detecting the playlist files would be problematic, and modifying the third party data is unlikely to be permitted.

CMTS AS CONTROL POINT

For users on an HFC network, IP traffic will always flow through the same CMTS port to reach a user at home. As the shared CMTS to CM link is normally the "narrow pipe" in the video distribution network, this is where congestion would be expected. Therefore the CMTS can potentially provide a useful control point to manage the ABR traffic.

Downstream Scheduling and Queuing

The DOCSIS standard provides very complete QoS functionality which may be useful for managing ABR traffic. DOCSIS QoS is based on the IntServ model of filter and flow specifications [INTS]. If a packet matches an installed filter (i.e. classification) it will be mapped to a specific service flow and then forwarded based on the parameters associated with that flow. Classification is based on matching fields in the packet header such as IP address and Differentiated Services Code Point (DSCP) fields. Thus it could be possible to recognize a managed ABR video packet from a well known source address (e.g. video server) or IP subnet. Alternatively all managed video traffic could use a DSCP marking indicating a preferential forwarding class [DSCP]. Inbound traffic to the network from non-trusted sources such as over-the-top (OTT) video would be subject to DSCP overwrite and set to a base priority such as best effort. The CMTS could then provide preferential treatment for the operator's managed video flows.

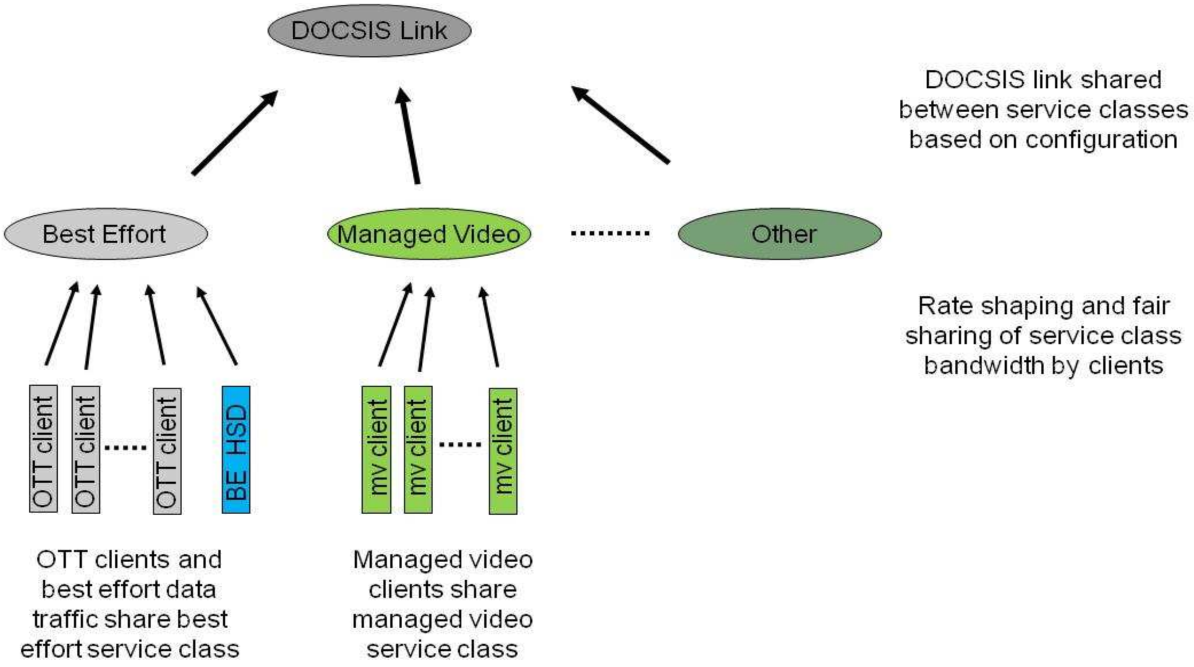

DOCSIS Link
DOCSIS link shared between service classes based on configuration
Best Effort
Managed Video
Other
Rate shaping and fair sharing of service class bandwidth by clients
OTT client
OTT client
OTT client
BE HSD
mv client
mv client
mv client
OTT clients and best effort data traffic share best effort service class
Managed video clients share managed video service class

Figure 2 DOCSIS Link Sharing

If the CMTS supports multi-level scheduling and per-flow queuing as shown in Figure 2, then it can provide fairness between video flows. In this case, each video packet would be mapped to an individual queue (based on the header fields in the packet) within a particular scheduling class such as managed video or best effort traffic. All queues within the same scheduling class share the bandwidth assigned to the class equally so that a single user receives only their fair share and cannot disrupt other video sessions. This mechanism applies to both managed video and OTT ABR video. OTT traffic will be put into the best effort class but will still receive a fair share of the assigned bandwidth for this class. It will, of course, share this with all general Internet traffic. Each scheduling class would be assigned a percentage of the available bandwidth proportionate to its expected load.

Session Control at CMTS

The DOCSIS infrastructure has a mechanism to reserve bandwidth for a flow based on the PacketCable™ Multimedia specification [PCMM]. This provides a potential mechanism to implement resource reservation at the session level. It requires a session establishment and teardown mechanism. In the PCMM model, client applications communicate with an application server (AS) that initiates the QoS requests to the policy server and CMTS. The ABR client application server might be co-located with a session/fulfillment manager, edge server, or user interface (UI) server depending on an operator's control plane infrastructure. Therefore, it would be suited to a managed video service but not OTT. The PCMM / CMTS mechanisms are well understood and include error recovery functions such as the timeout of orphaned sessions.

A potential problem arises in that a video asset may be delivered from one of multiple sources within the CDN. Thus, the filter specification used to identify the packets associated with the session would need to be capable of handling this. This may be as simple as using a known sub-network for the video sources. A more complicated problem is that within the single session, multiple file chunks at different bit rates may be requested due to local events in the client device. The resource reservation for the session could be selected to provide the maximum data rate expected from the client. However, if the client downshifted, this reserved bandwidth would not be used for the managed video but released for use by best effort traffic.

The lab investigations showed that the ABR clients tend to require additional bandwidth during startup and following bit rate increases. The PCMM mechanism can be used to provide a "turbo" mode in which additional bandwidth bursts are allowed for these periods.

CONCLUSION

The impact of ABR traffic on the network is already considerable and is likely to grow significantly as more video is distributed using this mechanism. ABR traffic operates very differently from existing video delivery mechanisms, and in the conventional use case, control over access network bandwidth is essentially abrogated to the device clients. Motorola experiments indicate that these clients vary from device to device and are not necessarily well behaved. Given that they have an incentive to be greedy rather than cooperate for the common good, it seems imperative that the operator finds other mechanisms to control ABR traffic impacts.

A number of options are discussed and the CMTS appears to be a promising location to implement this control. For OTT ABR traffic, the CMTS can provide rate limiting and fair sharing of bandwidth between both ABR clients and other best effort users. This is implemented using existing DOCSIS QoS and CMTS downstream scheduling. For managed ABR traffic, these QoS and scheduling mechanisms may also be used and can also provide segregation of the managed traffic from best effort traffic. With the addition of a session management function in the network, additional control is possible. This enables PCMM control mechanisms to be used to establish service flows for the video streams with defined QoS and reserved bandwidth.

The existing functions provided by the CMTS appear to provide the operator with an excellent control point to impose order on the access network despite the potential for aberrant client behavior.

REFERENCES

[ADAPT]	Adaptive Streaming – New Approaches for Cable IP Video Delivery J. Ulm, T. du Breuil, G. Hughes, S. McCarthy, The Cable Show NCTA/SCTE Technical Sessions spring 2010
[SAND]	Global Internet Phenomena Report Fall 2011; Sandvine
[VNI]	Cisco® Visual Networking Index (VNI) 2011
[MMR]	Microsoft Media Room -www.microsoft.com/mediaroom/
[MULPI]	DOCSIS 3.0 MAC and Upper Layer Protocols Interface Specification www.cablelabs.com
[TCP]	RFC 2581 TCP Congestion Control M. Allman, V. Paxson, W. Stevens
[PDIL]	Kuhn, Steven, "Prisoner's Dilemma", The Stanford Encyclopedia of Philosophy (Spring 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2009/entries/prisoner-dilemma/>.
[INTS]	RFC 1633 Integrated Services in the Internet Architecture: an Overview R. Braden, D. Clark, S. Shenker
[PCMM]	PacketCable™ Multimedia Specification www.cablelabs.com

ABBREVIATIONS AND ACRONYMS

CCAP	Converged Cable Access Platform
CDN	Content Delivery Network
CMTS	DOCSIS Cable Modem Termination System
COTS	Commercial Off The Shelf
CPE	Customer Premise Equipment
DOCSIS	Data over Cable Service Interface Specification
DRM	Digital Rights Management
DVR	Digital Video Recorder
DWDM	Dense Wave Division Multiplexing
EAS	Emergency Alert System
EQAM	Edge QAM device
Gbps	Gigabit per second
HFC	Hybrid Fiber Coaxial system
HSD	High Speed Data; broadband data service
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
MAC	Media Access Control (layer)
Mbps	Megabit per second
MPEG	Moving Picture Experts Group
MPEG-TS	MPEG Transport Stream
nDVR	network (based) Digital Video Recorder
OTT	Over The Top (video)
PHY	Physical (layer)
PMD	Physical Medium Dependent (layer)
PON	Passive Optical Network
RF	Radio Frequency
STB	Set Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VOD	Video On-Demand
WDM	Wave Division Multiplexing

STRATEGIC CAPITAL - A FORMALISM FOR INVESTING IN TECHNOLOGY

Marty Davidson
Society of Cable Telecommunications Engineers

Abstract

Spending decisions in cable today are complex. Long gone are the days of prioritizing OPEX over capital or purchasing via simple volume related discounts. Capital is now under an intense microscope. This paper presents a way to strategically and logically determine the optimal purchase price that will minimize the total cost of ownership, identify ways to drive efficiency into a workforce by identifying the proper division of labor and it will make way for the possibility of technological innovation through a 'creative destruction' process that will enable long-term growth.

INTRODUCTION

Currently, telecommunications service providers face stiff competition with new entrants every day and must search for solutions to the challenges and difficulties of growing revenue as well as margins. They must do this while dealing with the continued high fixed cost of doing business and the multitude of seemingly simultaneous priorities. Additional pressures exist due to operators being evaluated on a free cash flow basis. Under the existing economic climate, more often than not, this pressure is mis-prioritized and translates to demands for lower priced Customer Premises Equipment, or CPE. When this happens, a caustic force is unleashed that actually increases total costs and negates the scientific possibility of technological innovation. Due to the many ramifications of such a decision, the development of an evaluation schema is required.

This paper provides a formalism for a new way to think about how features in equipment that have the potential to translate into lower costs over time can be objectively and agnostically assessed. After this valuation is completed, decisions that optimize performance and lower OPEX can be made at the time of purchase. A specific example used is the consideration of strategic technical investment in CPE diagnostic elements that optimize operational costs by identifying applicable processes and the possibilities for the proper division of labor. It is shown via this formalism that through this type of upfront investment, service providers will reliably identify and improve not only their fiscal position, but also the quality of customer experience and will be well armed for the ever-evolving subscriber/revenue battle. Lower operational costs via these types of strategic technical investments in CPE will be shown to have additional advantages that can be evaluated using the formalism to determine how they would aid in improving capital efficiency and the ability of a cable operator to react even more quickly to new service needs and market forces. Finally, the formalism will provide a mechanism for operators to determine which new features are critical enough in long term cost-benefits to warrant standardization so that all equipment supports the features. A key goal of this formalism is to implement the type of industrial efficiency and quality envisioned by the likes of Frederick W. Taylor and W. Edwards Deming by specifically coupling equipment procurement decisions into a longer-term process of continued technology improvement to enhance the competitive position of cable operators. But another goal is to provide a mechanism for the type of disruptive process of transformation or ‘creative destruction’ via new equipment and service capabilities that accompanies the kind of radical and rapid innovation that is the force that sustains long-term economic growth.

FINANCIAL PRIORITIZATION

Opinions vary as to where, when and how our current economic climate started, be it deflation, deleveraging, debt accumulation, etc. associated with the housing & financial bubbles. Initially, the economic downturn actually benefited service providers as consumers limited their expenses for activities like going to the movies. The desire for entertainment was still strong so subscribers turned more and more often to home entertainment services. While the concern over the potential for a significant age of deflation was being ignored by the masses, some companies began to feel the real impact to their top and bottom lines. Telecommunication service providers seemed to initially weather the storm, however, as the economy kept declining and lagging, its impact to these providers began. To the credit of the industry, bold changes began happening, but not all the changes were for the betterment of the business in the long term. One example of this is when operators reduced expenses but cut not only the fat, but also the muscle and sometimes into the bone. In the short term, when these changes were looked at in a silo they appeared to be very reasonable; however, when you couple such decisions with being evaluated on a free cash flow basis, some very dangerous things happen. Purchasing organizations are incented to drive prices lower and lower, which in itself is the right intent. The danger is when decisions on capital purchases are based purely on purchase price. When this happens without taking into consideration the ‘hidden’ costs in operations, the total cost of ownership can far outweigh any purchase price savings. Additionally, technological innovation is stymied and the possibility of the ‘creative destruction’ process for sustained fiscal growth vanishes. Joseph Schumpeter popularized the idea of ‘creative destruction’ based on the economic theories of Karl Marx and he believed innovation shifted the powers in a market place by the introduction of new competitors and that ‘creative destruction’ described the dynamics of industrial change.

In order not to limit a new age of industry pioneers, a methodology is needed to holistically evaluate purchasing decisions that will lead to the most strategic investments in capital possible. A formalism is presented here that identifies a new parameter called Optimal Purchase Price, which takes into account a wide array of considerations one could use when negotiating equipment purchases, whether that be with a vendor or with the purchasing department within their own company. This prescription for strategic capital purchases leverages a Total Cost of Ownership, or TCO, approach and is not a Cost Benefit Analysis. Performance differences between pieces of equipment should be evaluated relative to the importance to the purchaser. This formalism looks at capital investments from concept to test to deployment to operational integration to trouble resolution to future proofing.

OPTIMAL PURCHASE PRICE

To begin to define the Optimal Purchase Price or OPP, a base upon which can be built is required. That base is the traditional, actual purchase price that an operator would pay for a given piece of equipment. While this paper does consider equipment throughout the network, from the national distribution centers through the backbone, headends, hubs and HFC plant, the predominate evaluation comes from Customer Premises Equipment, or CPE.

Traditionally, the purchase price is evaluated on a Return on Investment, or ROI, basis. ROI is a function of base purchase price, BPP, average revenue per unit, ARPU, and average expense per unit, AEPU. Essentially it is the time period that operational cash flow takes to recover the capital purchase, usually expressed in a number of months.

$$\mathrm{ROI} = \frac{\mathrm{BPP}}{\mathrm{ARPU}-\mathrm{AEPU}} \qquad (1)$$

For the purpose of this formalism, a normalized payback can be considered. One characterization of this is seen in Figure 3.1.

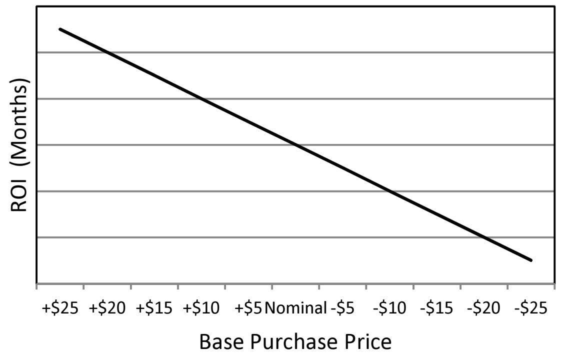


Figure 1

This curve is linear but it certainly has multiple Purchase Price Factors, or PPFs, which can influence it non-linearly such as PPF_1 which accounts for equipment volume discounting or other price influencing factors. Another adjustment that can be made on the base optimal purchase price is differential pricing for multiple organizations or PPF_2. One example of this is sometimes referred to as most favored nation pricing. For a given operator PPF_2 is ignored, but is a valuable tool for a comprehensive analysis across multiple perspectives.

Once the base OPP is established, the incremental components of total cost factors must be defined and evaluated.

Pre-Deployment Test Cost

Before the operational cost impacts of a deployed device can be considered, an evaluation of Pre-deployment Test Costs or PTC must be made. These apriori considerations include:

- Software, firmware and hardware related costs that come from issues that are identified in lab or field trial evaluations and require new versions prior to deployment. Each of these costs has a related scale factor based on the likelihood of needing multiple revisions. Software typically requires 10-20 times more revisions than hardware or firmware.
- Lab testing costs which encompass lab setup, test, evaluation, post analysis, tear down and personnel costs, whether performed internally or externally to a given operator.
- Field trial expenses including training, planning, trial management, field and customer care resources, increases in calls and truck rolls as well as tangential components to account for costs due to customer dissatisfaction and poor press.

$$\mathrm{PTC} = \sum_{i,j,k}^{n}(\mathrm{SF_{SW}} * \mathrm{Sw_i} + \mathrm{SF_{FW}} * \mathrm{Fw_j} + \mathrm{SF_{HW}} * \mathrm{Hw_k}) \ (2)$$

Each component has built into it the number of resources in the lab, field and management of the project, the associated costs for these resources and the time it takes to resolve the issues that have been identified.

Cost of Deployment

Once a piece of equipment has made it through the lab and field trial hurdles, deployment begins. Operators use multiple strategies for deploying new hardware, firmware and/or software. Deployments could start from a few friendly users to a small market with limited deployment, all the way up to a national or company-wide roll out. There have been numerous situations where small deployments did not identify operational issues until an appropriate level of scale was met. As such a fiscal evaluation of deploying new technology must be used. The Cost of Deployment, or COD, is proposed and is a major component of OPP.

The most influential factor in COD is the increase in trouble rate. This increase has been shown to add a significant cost to doing business. When a piece of equipment from a new supplier is introduced into the field, the customer-reported trouble rate can increase, CRT_i, as much as 30 percentage3032733840 points. There are numerous cost drivers when this happens such as: increased calls into customer care, CC_c, increased truck rolls, TR_c, both valid and in error (traditionally 10-15% of all trouble calls into customer care translate into a truck roll in error) and resources on the team that manages the tickets being worked, TM_c.

$$\mathrm{COD}=f(\mathrm{CRT}_{\mathrm{i},},\mathrm{CRT}_{\mathrm{t}}) * (\mathrm{TR}_{\mathrm{c}} + \mathrm{CC}_{\mathrm{c}} + \mathrm{TM}_{\mathrm{c}}) \quad (3)$$

Pick a dollar figure for a call into care, a truck roll and a hourly labor rate and you will see how significant this parameter can be. But that is just the beginning as this is a problem that just keeps on giving. There is a major influence on all of these expense increases, namely, the time it takes to get the customer-reported trouble rate back down to normal levels.

As seen in Figure 2 below, getting back to the normal trouble rate can take 18 months and with new technology or product introduction this can be even longer.

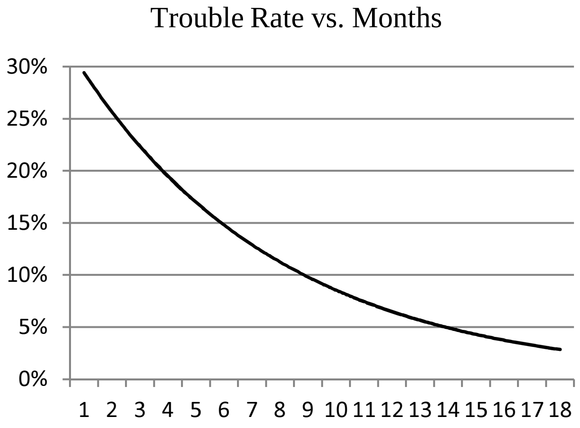


Figure 2

Equipment Combo Factors & Locale Weight

As mentioned in the introduction of OPP, the major focus of this paper is on CPE even though there are other network equipment influences (NEF) built into the formulation. Equipment Combination Factors, or ECF, take into consideration what services can be enabled on a given piece of CPE and what actual services a customer is paying for on that CPE, this is referred to as CPE_F. For example the lowest ECF components are stand-alone set top boxes and cable modems. Just above that are home gateways, WiFi enabled modems and eMTAs. Additional weighting is applied to devices that carry critical services like lifeline voice and home security, CPE_w. This is reflected in the matrix operation to determine ECF.

$$\mathrm{ECF} = [\mathrm{CPE}_{\mathrm{F}}] * [\mathrm{CPE}_{\mathrm{W}}] + f(\mathrm{NEF}) \quad (4)$$

The location of the equipment being deployed also has an impact on the overall total cost that needs to be considered and is reflected in this analysis as ELF. Factors considered in ELF include every locale where equipment could be deployed (EDL) from the home through the HFC network to the backbone and into national data centers. The degree of influence that errors associated with new deployment have on the customer population is weighted appropriately (EWF). This weighting function is proportional to the number of subs potentially impacted by it and a characteristic function of the device itself.

$$\mathrm{EWF} = f(\mathrm{device}) * \sum_{i=1}^{n} (\mathrm{subs})_i \quad (5)$$

The functional combination of these two elements provides the overall equipment locale weight, which can be seen in Figure 3.

$$\mathrm{ELW} = f\,(\mathrm{EDL},\ \mathrm{EWF}) \quad (6)$$

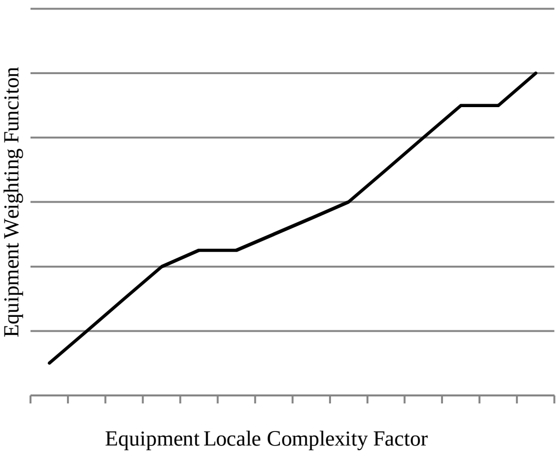


Figure 3

Optimal Ease of Use

Whenever a new piece of equipment or software is introduced into service, there are some differentiators between products that can have an impact on real operational costs. Training is the first element in the calculation of Optimal Ease of Use, or OEU. Training material must first be developed. These could be as small as talking points posted to a call center knowledge base or as involved as a multi-day, hands on session with a live instructor. Once training is developed, the degree of complexity, which can be correlated to time off the job, varies as described. But it is not only the length of training that is of concern, it is the complexity associated with it and the probability that repeat training would be needed. Representing the training aspects of this analysis, Training Development & Deployment, or TDD is used.

OEU is also influenced by the degree of difficulty or ease with which a user can debug and solve a problem on a given device. This is reflected in the Total Time Usage Factor, or TTU.

Standards are so well embedded into our daily life that the average worker rarely, if ever, considers the impact of standards. The Standards Product Factor, or SPF, is a factor that lends itself to the ease of integration, training, etc. when compared with non-standards based products. Standards based products allow for efficiencies to be realized and this can lead to a division of labor which can re-purpose resources to more important and complex challenges.

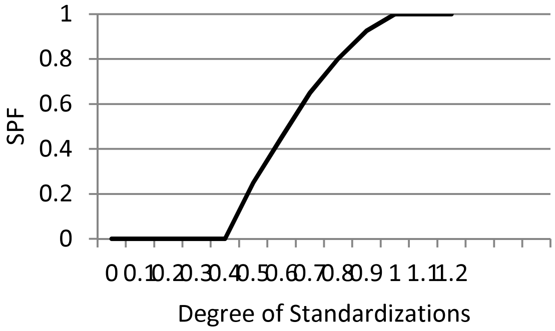


Figure 4

SPF can be looked at as an inverse function so that if a product is standards based, it will help lower the total cost of ownership. This leads to the formulation of OEU.

$$\mathrm{OEU} = f(\mathrm{TTF}) * \ f(\mathrm{TTU}) * f(\mathrm{SPF}) \quad (7)$$

There are multiple other considerations that could be included in the OEU calculation such as: how much a technician likes a particular product and thus an internally created efficiency of how it helps improve his or her daily work duties; the support provided by a particular vendor; or the creativity and innovation instilled in an employee inspired by the technology and associated ease of use.

Customer Type Factor

The customer must not be forgotten in this analysis, so the introduction of an OPP parameter for the customer is necessary. CTF, or the Customer Type Factor is a complex, non-uniform variable that is heterogeneous in nature. If only one service was provided to a customer and each customer had the same propensity for calling when things didn't work correctly, assessing the CTF would be a much simpler effort, as opposed to the ever growing number and complexity of products a customer may have, as well as the level of, or lack thereof integration that exists. CTF is a function of the products or services a customer has, their likelihood to call into customer care based on a characteristic distribution, the number of different revisions of software, firmware and/or hardware and the types of equipment and level of integration of such devices.

$$\mathrm{CTF} = f(\mathrm{products}) * \mathcal{L}(\alpha|\mathrm{x}) * f(\mathrm{revisions}) * f(\mathrm{integration}) \quad (8)$$

An additional component that could be considered in CTF, but is not reflected here, is if an operator were to prioritize service for their most valued subscribers.

Technical Advancement Advantage

Every piece of equipment has its merits and its opportunities for improvement. As rapidly as technology evolves, as well as the associated operations and customer expectations, a relationship between a given piece of equipment and the technological advantages that it provides is proposed as the Technical Advancement Advantage, or TAA. TAA is the calculated as:

$$\mathrm{TAA} = (\mathrm{FPF} + \mathrm{HPF}) * f(\mathrm{CPD}) \quad (9)$$

Both FPF, the Future Proofing Factor and HPF, the Historical Performance Factor functions are characterized similarly as described by following which is then normalized.

$$\begin{array}{ll} \sigma_i \geq 1 & \mathrm{i}^{\mathrm{th}} \text{ device } 100\% \\ 0.3 < \sigma_i < 1 & \mathrm{i}^{\mathrm{th}} \text{ device } \sqrt{\sigma_i - 0.3} \\ \sigma_i \leq 0.3 & \mathrm{i}^{\mathrm{th}} \text{ device } = 0 \end{array}$$

FPF is essentially the ability of a given piece of equipment to extend its operational usefulness. An alternative, inverse way to think about this would be the less changes required over the life of products from a technological operations perspective. HPF is a confidence value in a vendor who is trusted and has demonstrated past performance of delivering what has been requested. The higher value in both of these factors correlates to a positive impact on TAA and overall OPP.

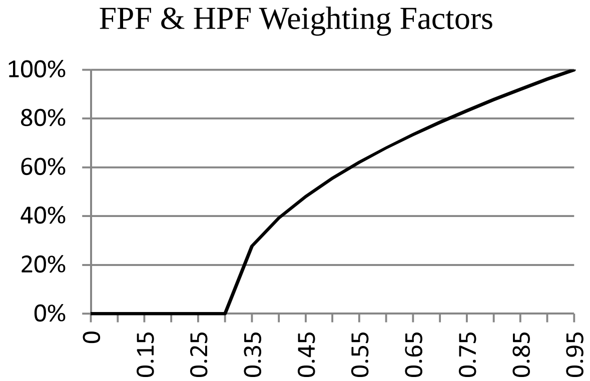


Figure 5

CPD, or Customer Platform Diagnostics, is another proposed functional scaling variable that highlights a piece of equipment's overall diagnostic ability to cross platforms and reduce overall time to repair. One representation of this function is that it gives an increasing, exponential positive benefit to the overall calculation because significant improvements in this area are challenging to come by to say the least.

Each of the attributes of this advanced technology element can be very individualistic and multiple other relationships could be used.

Smart Energy Adjustment

Economic times have made it more difficult for operators find costs savings in their business, but one of the more recent areas of focus is on energy use. Power bills are still a major component of cable expenses and both space and existing power are becoming rare resources. In order to factor energy into the equation of capital purchases, a Smart Energy Adjustment, or SEA, is needed. Presented here are four areas for consideration.

The first component of SEA is the Energy Efficiency Factor, or ε_F, which is a calculation of how efficient a given piece of equipment is. Proposed here is a ratio measure of average throughput and total power used, e.g. bits/watt.

$$\varepsilon_F = \frac{\bar{\mathrm{X}}_\mathrm{n}}{\mathrm{P}_\mathrm{T}} \quad (11)$$

Other elements that need to be included in the smart energy calculation are: density (a function of throughput and physical area), size (a function of how much space a given device occupies, particularly critical in centralized equipment locales) and diagnostic ability. For the purpose of this formalism they are reflected as:

$$\varepsilon_D = f(\tau, A) \quad (12)$$
$$\varepsilon_S = f(S) \quad (13)$$
$$\varepsilon_{PD} = f(t, I) \quad (14)$$

The power diagnostic factor, ε_{PD}, is an intriguing area that could have significant impacts on energy consumption and power availability. The ability here is for a device to understand the historical current (or voltage or power) use and correlate it to potential failure modes, essentially looking at energy as a proactive indicator of overall service availability and reliability. This is major focus related to the fiscal health of the industry and an opportune area for further research.

Diagnostic Capability Determinant

One of the most crucial areas of focus in this formalism is the value of investing in technology, particularly in CPE, that can include diagnostic elements that lead to the optimization of operational costs by identifying customer impacting issues throughout the lifecycle of the customer. This identification, as detailed below, can ultimately lead to process efficiencies and thus even greater savings and enable the possibility of even more technical innovation. Characterization of this capability is done through the definition of the Diagnostic Capability Determinant, or DCD. There are four drivers of DCD, the first of which is Pre-Customer Realization, or PCR. PCR outlines the ability of diagnostics to identify a service related issue before a customer would notice it. Ideally, the best scenario would be if an event could be identified before it happens, however, there are events that will always be impossible to prevent. The time variable in the PCR equation accounts for this situation and is reflected in the overall calculation as the duration of an event multiplied by a function of the percent of time that identified instance occurs times the frequency of occurrence. The calculation is shown with a summation because there may be multiple devices with identical alarms that are worked independently by the work force.

Additionally, unique issues can occur simultaneously. The summation is across the total of all of these events.

$$\mathrm{PCR_i} = \sum_{\mathrm{i}}^{\mathrm{n}} \Delta \mathrm{t_i} * f(\frac{\mathrm{I_i}}{\mathrm{I_T}} * \frac{\mathrm{F}_i}{\mathrm{F_T}}) \quad (15)$$

As mentioned, customer-impacting events are impossible to prevent, but PCR identifies how well the diagnostic capability works in a pro-active fashion. When an event actually impacts a customer it is critical that we identify a DCD component that measures how well the embedded software can identify and distinguish an issue and provide information to the service provider to remedy the situation, which is suggested here as the Diagnostic Activity Factor, or DAF. DAF is a nonlinear function that heavily weights quicker resolution of troubles, as is seen in Figure 6.

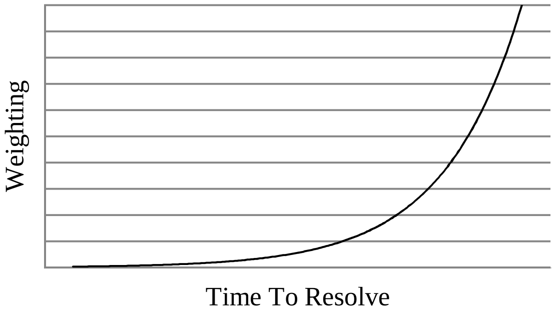


Figure 6

Even with a high DAF, the cause of the problem may not be known. For example, the problem or occurrence may be identified and the problem resolved quickly, which is the initial priority in operations, but the underlying cause was not determined. The Post Issue ID, or PID, addresses the intrinsic value in knowing what caused the problem. PID is a measure of how specific diagnostics are in their ability to identify the actual cause of the problem. Many times using posteriori data can help put new alarm parameters or thresholds in place or identify new process steps or errors in an existing processes.

Figure 7 articulates a scalar multiple that can be used in the DCD calculation. Notice that if no or minimal post problem identification exists, the value for PID is zero. Above that, a three tier value is proposed. These values should be evaluated based on the particular type of equipment in use and the services it supports. Another consideration is how many actual devices are or would be deployed.

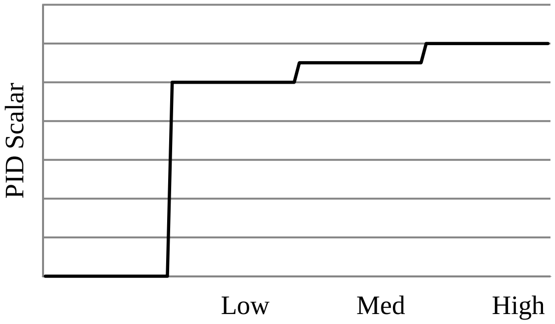


Figure 7

The last consideration in DCD is a proposed parameter that reflects the accuracy of diagnostic recommendations. In operating a network there are many times when data being presented point to a particular issue but when further due diligence is performed, the identified issue is inaccurate. The value of such a parameter is individualized on how important that is to a given user. Here we simply call it Error ID Avoidance, or EID and is a function of user ranked importance, ϕ.

$$\mathrm{EID} = f(\phi) \quad (16)$$

Combining the fore mentioned components of DCD leads to the following calculation. PID and EID are important factors but their influence is adjusted appropriately when compared to DAF or PCR.

$$\mathrm{DCD} = \mathrm{DAF} * (\mathrm{PCR} + \frac{\mathrm{PID+EID}}{\mathrm{PCR}}) \quad (18)$$

Workforce Effectiveness Principle

The last element of OPP is a parameter called the Workforce Effectiveness Principle, or WEP, which is composed of four parts. The first three parts are directly correlated to the technician using a given device and the fourth is a new concept addressing the possibility of quantifying the ability to distribute labor in the most efficient way.

Two of the WEP components measure a technician's ability to work with a given piece of equipment. Both are straightforward in the sense that their intent is to assess the technician's interactions during installation and troubleshooting. They are called Installation Ease, or IE, and Troubleshooting Ease, or TE. An ideal approach for these factors would be to perform a time and motion study, using the techniques to identify business efficiency through Frederick Winslow Taylor's Time Study work combined with work of Frank and Lillian Gilbreth on Motion Study. This will provide a historical baseline and then a static, multi-tier variable based on a suggested difficulty factor here called, δ, can be determined.

$$\text{IE} = f(\delta_{IE}) \quad (19)$$
$$\text{TE} = f(\delta_{TE}) \quad (20)$$

A less scientific measure, but perhaps even more valuable consideration in WEP is the technician's confidence in working with a given piece of equipment, which here is called the Technician Confidence of Use, or TCU. Anyone who has managed a workforce of technicians can readily articulate the benefits of a confident and enthused team. TCU is a proposed measure to capture just that.

Distribution of Labor, or DOL, is the main driver in WEP and on a macro scale can have the most significant impact on operational expenses. The reason that DOL is so significant is that it looks at the current workforce operations and processes through a 'Scientific Management' lens that Frederick Winslow Taylor proposed and used in the Efficiency Movement. DOL attempts to evaluate the most efficient ways to accomplish the tasks at hand by using advanced diagnostics that will enable the problems to be worked in a more efficient manner. As such WEP is defined as:

$$\text{WEP} = f(\text{DOL}) + \frac{\text{IE+TE}}{||DOL||} + f(\text{TCU}) \quad (21)$$

Due to the inherent complexity for this key component of OPP and because of its many interrelated degrees of freedom, computational algorithmic analysis is required.

Summarizing Optimal Purchase Price

There are a dozen main contributors that have been used to describe OPP. Each of these components has its own level of complexity and interrelatedness to the others. A structured model is required using computational algorithms with bounded, varying randomized inputs for the many individualized computations proposed in this formalism. A Monte Carlo type analysis is suggested that is specifically targeted at reducing the overall total cost of ownership, including resource reallocation efficiencies. The largest challenge of such a model will be integrating those components that are more "soft", less deterministic and highly dependent on the individual or company prioritization of such elements. One such example is how worker satisfaction is valued via parameters such as TCU, which was described earlier as part of WEP.

Once this is done, a new ROI can be evaluated taking OPP into account and the overall value of a product can be effectively evaluated, both from the operator and vendor perspectives. Next the varying lifecycles of a product should be considered. This may be done by creating different ROI analyses, e.g. via the Monte Carlo analysis mentioned above. If, within a given set of parameters, the ROI exceeds the expected lifecycle, one would need to iterate the formalism to identify areas of cost that could be removed from the business and thus creating a lower OPP with a potential a higher TCO. These realized savings can be thought of as 'insurance' against unforeseen costs. This approach is particularly applicable in the current business climate given how short life cycles can be.

INDUSTRIAL QUALITY & EFFICIENCY

There are many possible ways to divide where the work should be done vs. where it is being done. There are three primary tasks investigated here: issue identification, fix implementation and resolution confirmation. The most fundamental view of efficiency in this discussion is purely how much more efficient a worker can be doing the same tasks as were done previously. This worthwhile endeavor is the same approach outlined by W. Edwards Deming in his approach to Total Quality Management. Workers and management alike should continually assess their duties to look for opportunities for improvement. Detailed chronicling of this work is imperative to drive consistency into the services that are provided to end customers, i.e. standardization. Once the work is documented, methods and procedures can be built into training materials and subsequently the training being given, bringing efficiencies to the training resources as well. Once standardized, many tasks can then be modeled and implemented in software tools to remove menial labor tasks. When this happens, the proverbial flood gates open and one can investigate how to divide the workforce and reallocate the work in the most efficient manner. One example of this would be how the three primary tasks mentioned above could be divided. Taken in order, issue identification could be implemented in software and the verification of issues could be done in a centralized work group. This work group would have fewer total resources than a distributed model as they would be able to fill in the otherwise distributed, individual lows of work with the volume that comes from the total distributed load. Additionally, the speed of identification would also increase. The fix implementation process would also improve. Not all work could be centralized, but any fix that could be done remotely could move into a centralized group and similar efficiencies could be realized.

Finally, much like the issue identification scenario, issue resolution confirmation could be moved to a centralized work group, once again creating a way for the most efficient manner possible.

TECHNOLOGICAL INNOVATION

Some believe that the source of the Western idea of 'creative destruction' is the Hindu god Shiva, who was thought to be the destroyer and creator simultaneously. However, as mentioned earlier, Joseph Schumpeter is credited with introducing the term 'creative destruction' in his famous book, *Capitalism, Socialism and Democracy*. It was in this book where he described how innovation can cause the disruptive process of transformation.

So how does 'creative destruction' apply to this formalism? First, an example is necessary to baseline the concept. In the retail market, previously, many small, older, local companies historically offered retail consumer products. The distributed nature of this model left little opportunity for expense reductions. Then came the technological innovations that Wal-Mart introduced, including new ideas such as personnel, marketing and especially inventory management. While these innovations destroyed businesses like Montgomery Ward and Woolworths, it created a whole new set of technology that spawned other businesses and innovations. Another example is the destruction/creation cycle of 8-track to cassette to compact disc to MP3.

If the proper division of labor outlined in this paper is considered, work is moved from a distributed workforce to a centralized one requiring fewer resources overall. This destroys the structure of the legacy labor pool, but creates an increased level of customer service, reduces operational expenses and opens the door to a whole new set of technological innovations that could never have been imagined before this change, i.e. a disruptive innovation that helps create new business opportunities for existing and new vendors. Along with these new business opportunities comes the scientific possibility for further innovation, aka, 'creative destruction' which leads to rapid and sometimes radical change that yields economic growth over the long-term.

There are other tangible benefits that are realized from the suggestions in this paper, such as increased customer service and loyalty, i.e. reduced churn, marketing advantages as in brand strength, reduced advertising costs and thus lower costs of acquisition. Additionally, an interesting benefit is how capital may be used more effectively and spent in places where the greatest benefits are. Reduced cycles for new product introduction may also be realized as would softer advantages like internal and external public relations.

CONCLUSION

Solely choosing purchase price for equipment based on traditional volume discounts or on a 'lowest cost wins' basis is not sufficient. Today's high fixed cost of doing business, simultaneous priorities, ever increasing level of competition and return on investment expectations, demand a new approach.

Determining the optimum purchase price for equipment can be identified through the combination of the many factors described earlier. These factors require operators to take a look at the capital costs with a new total, comprehensive perspective. This approach requires that personnel in operations, engineering, marketing, finance and purchasing work together in evaluating the total cost vs. assessing it in silos with competing priorities.

Evaluating products in a manner described in this formalism provides: the possibility for operational performance optimization; product and operational standardization; lower short term costs; distinct competitive advantages; increased customer service levels; reduced product deployment times; the proper division of labor; and the radical and rapid innovation required to sustain long-term economic growth.

It is imperative that operators take into consideration the kind of upfront investment and implement a capital strategy that takes into account the vast and complex array of variables that contribute to the overall fiscal health of their business and set themselves up for the next generations of success.

References

- Schumpeter, Joseph A. (1994) [1942]. “Capitalism, Socialism and Democracy”. London: Routledge.
- Reinert, Hugo; Reinert, Erik S. (2006). "Creative Destruction in Economics: Nietzsche, Sombart, Schumpeter”
- Mitcham, Carl and Adam, Briggle “Management” in Mitcham (2005)
- Zandin, K. (2001), Industrial Engineering Handbook, 5th edition, McGraw-Hill, New York, NY.
- Deming, W. Edwards (1986). Out of the Crisis. MIT Press.
- In-Home Support Services, SCTE Home Networking Primer Series
- Innovator's Guide to Growth - Putting Disruptive Innovation to Work. Anthony, Scott D.; Johnson, Mark W.; Sinfield, Joseph V.; Altman, Elizabeth J. (2008). Harvard Business School Press
- Trends Over Time in Server Energy Use, Performance, and Costs: Implications for Cloud Computing and In-house Data Centers. SCTE SEMI Primer Series
- Network Management Fundamentals, Alexander Clemm, 2006, Cisco Press
- DOCSIS Checklist for PacketCable™ Reliability in the Outside Plant - Downloadable DOCSIS Checklist for the PacketCable™ Reliability. From SCTE Outside Plant Seminar
- How to Identify and Build Disruptive New Businesses, MIT Sloan Management Review Spring 2002
- Monitoring and Managing a DOCSIS™ 3.0 Network. SCTE DOCSIS Primer Series
- The W. Edwards Deming Institute, Fostering Understanding of The Deming System of Profound Knowledge

Strategies for Deploying High Resolution and High Framerate Cable Content Leveraging Visual Systems Optimizations

Yasser F. Syed PhD, Dist. Eng/Applied Research
& Dan Holden, Fellow, Comcast Labs

Abstract

This paper examines how and why to deliver higher resolution and framerate content in an HFC system, especially focusing on 4K Video delivery with an advanced audio experience. It examines how to deploy this content in a bandwidth constrained environment and concentrates on improvements to the viewer's quality of experience through video compression technologies and leveraging potential video compression gains through sensitivities in the human visual system.

INTRODUCTION

The launch of higher resolution video with greater frame rates will allow MSOs to develop new business opportunities, and provide a competitive advantage against new entrants in the video marketplace. In this paper we will examine the road to better delivered video quality, especially how to leverage the existing HFC infrastructure to deliver 4k video with an advanced audio experience. The paper will concentrate on video compression technologies and the potential for leveraging the human visual system model to provide 4K video in a bandwidth constrained environment. For deployment, we will look at required upgrades to the HFC infrastructure, and what engineering requirements are needed for 4K delivery. New technologies and approaches to reduce costs will also be examined, as well as how the complexity of high-resolution video changes delivery methodology.

4k television technology was introduced at the Consumer Electronics Show in 2012. It is based on a display that has approximately 4000 pixels in the horizontal resolution. 4k differs from previous television standards (480i, 480p, 720p, and 1080P/I) in which the vertical pixel count was annotated. In a 4k display the horizontal resolution is maintained around 4000 pixels, and the vertical resolution is allowed to vary as a function of source content. This technique was adopted to allow support for various aspect ratios and letterboxing. Figure 1 shows the scale of 4K content compared to the resolutions that are supported today.

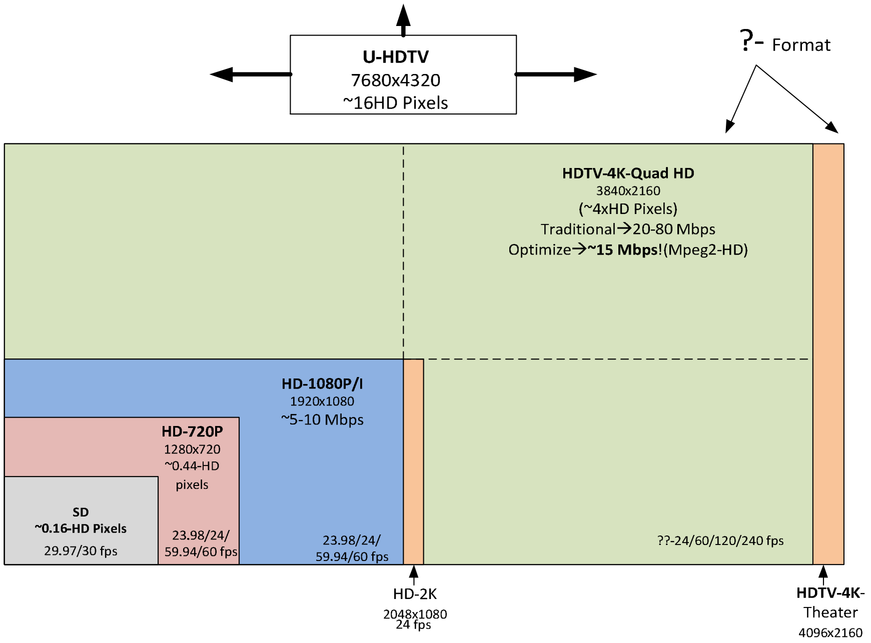


Figure 1 Comparison of 4K to Different Video Resolutions

BUSINESS OPPORTUNITY

One of the most compelling cases for higher quality video is to gain a competitive advantage in the video marketplace. 4k will require “big pipes” at a time when there is clear movement on the part of industry competitors to adopt a mobile strategy utilizing technology that will be limited by available spectrum. Newer video compression techniques will certainly reduce the size requirement for the pipes, while the demands of newer display technologies, (8k and 256 fps) will tax any future video distribution system.

Cable has a reputation for being the leader in delivering an exceptional video experience. First generation 4k delivery platforms will need a vast amount of bandwidth, which is most likely to require a full QAM in order to deliver a quality experience. If we look at historical data, early generation H.264/AVC video was around 9 Mbps for High Definition (HD) 1080i video. A few short years later, we have been able to reduce this bandwidth to 4.3 Mbps.

Until display technology retail prices drop to a reasonable level, it is expected early adopters for 4K televisions will be bars, restaurants, and high-end home theaters. Here are the key assumptions:

- Mass deployment of 4k televisions will not take place until the cost per unit is less than $3,000 per unit
- The introduction of 4k will follow the same general path as the

introduction of High Definition video, which has currently penetrated more than 70% of US households

- Adoption of 4k video will be slower than HD video
- Volume of 4k encoded VOD assets will grow exponentially over the next three years
- Studio post-production already supports a 4K workflow which can be extended to downstream VOD content delivery
- Additional revenue will be generated when customers select to watch assets in a 4k format
- 8k video will not be introduced until at least 2016
- MSOs will not simulcast 1080p60, but may select to offer this format in VOD
- Bars, restaurants, and elite home theaters offer a significant up-sale opportunity

MSOs should take the lead on the introduction of high resolution video delivery. Rather than focus solely on video, it is suggested by the paper authors that the entire sensory experience be enhanced, which includes the addition of 3D audio channels. Background noise in a bar can be very distracting, and providing a high quality audio experience will set our video offering apart from the competition. The adoption of 4k video with 3D audio will most likely not progress at the same pace as HD. HD had the added benefit of changing the format to 16:9 from 4:3, and the elimination of large cathode ray tubes, which drastically reduced the size of the television footprint in the living room. The adoption of HD televisions has been relatively quick historically, whereas, the migration to the distribution of higher resolution video has yet to be established.

ADOPTION WILL BE DIFFERENT FROM 3DTV

There have been many attempts to categorize 4k video to the 3D television experience. This type of comparison is probably not the correct model, as 4k will not suffer from the infamous 3D glasses gaffe. Additionally, massive libraries currently exist at studios that can be easily scanned or transcoded into higher resolution video for distribution. We believe comparing 4k adoption to 3D would be a mistake, since 4K will most likely follow the adoption and general operational patterns developed for HD.

The first linear 4k channel will most likely be an occasional feed that is brought up when a live 4k event is aired. Under this model, 4 HD channels would need to be taken down in order to broadcast a single 4k event. With a few enhancements to the backoffice systems, it should be possible to sell access to a 4k stream on a pay-per-view fashion. The broadcast of huge events, like the Olympics or Super Bowl, could lead to enormous up-sale opportunities.

4k VOD will most likely be the first place where we see significant inroads of high resolution video. Encoders have already been developed that can process 4k video, and it is believed VOD pumps will not have issues with the larger file sizes or MPEG-2 transport stream wrappers. Adaptive streaming technologies should also be suitable for 4k VOD distribution. The video encoding process for QAM and adaptive streaming can be identical. Fragmentors should not require modifications, unless they are "just in time," which may suffer from data transfer rates and latency. The largest gap in the distribution system will be the ability to handle 3D audio, and finding a suitable video player.

ALTERNATIVES

There are many alternatives to 4k video, including Quad HD and 8k. While Quad HD has slightly less resolution than 4k, 8k has twice the resolution and twice the bandwidth requirements. Should Quad HD TVs be introduced into the marketplace, it would be preferred if they have the capability to ingest true 4k content, as MSOs would not want to simulcast both Quad HD and 4k streams. For VOD delivery, it would be possible to support both formats, but as the VOD library grows the added storage expense would prove challenging.

Rather than rolling out 4k, another possibility is to move forward with 1080p60. Encoders and STBs were released in 2012 to support this format. Formal analysis of 1080p60 video quality is beyond the scope of this paper.

Current compression technology will most likely prevent the delivery of 8k content over a QAM, but 8k delivery could conceivably be done utilizing the CMTS and IP delivery methodologies. Both products deliver the same benefits as 4k, higher video quality.

BENEFITS OF 4K

The first implication of moving to 4k video is the size of streams and files will be massive. A single mezzanine, linear stream from a live event may reach up to 500 MBps and a stream sent to a set top box could be on the order of 38 MBps. This implies four high definition channels would need to be taken down in order to place one 4k signal on the plant (Figure 2).

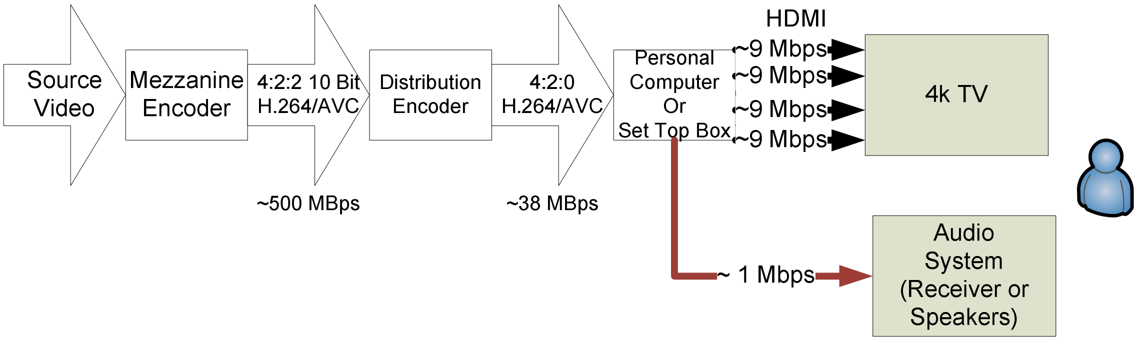


Figure 2 Delivery of 4K content from Ingestion to Consumer

Higher resolution video will allow MSOs to compete with both BluRay and local movie theaters. Many movie theaters currently delivery digital projects in 2k resolutions with a maximum audio experience of 11.1. It is theatrically possible to delivery 4k video with a 22.2 audio experience across an existing QAM to a personal computer (PC) which will replace the current functionality provided by a set top box (Table 1).

Thus, a completely optimized and compressed 4k/HEVC asset should be around 19 Mbps. When we compress this asset utilizing HEVC, we expect to gain around a fifty percent reduction in bandwidth, putting our 4k asset at approximately 10 Mbps. Next, consider that 50% of that potential gain is taken back 50% due to inefficiencies in first generation encoding technologies, frame-rate allocations, and make allowances for content types, then our 4k/HEVC asset can be distributed in the same band width as an HD asset compressed with MPEG-2 (~15Mbps).

Phase	Video Type	CODEC	Bandwidth (MBPS)	Notes
Initial	1080i	MPEG-2	19.3	19.3 was part of a specification. The first generation HD at some MSOs was set to 18 MBPS.
Today	1080i	MPEG-2	9.7	With 4:1 statistical multiplexing, it is possible to send 4 HD streams down a single QAM
Today	1080i	H.264	4.3	Average bit rate for H.264/AVC streams
Initial	4k	H.264	38	Target bit rate for lab trials
Production 4k[1]	4k/60 fps	HEVC	15	Target bit rate for 4k/60 with 22.2 audio

Table 1 Projected and Historic Bandwidth Consumption

[1] Note the projected bandwidth for a production 4k asset. The basis for the projection is calculated as follows:4k video is slightly larger than four HD signals: 4 * 4.3 ~ 18 Mbps in H.264/AVC Add additional audio bandwidth of approximately 1 Mbps for a total of 19 Mbps in HEVC

AUDIO

In addition to an enhanced video experience, the opportunity exists to upgrade the viewer's audio experience. Cable MSOs understand that audio can enhance or detract from the video quality of experience.

Many new audio technologies are under development that will put additional audio channels into the home. Old content can be remixed to support new formats, and additional microphones can be utilized to capture a true "3D" audio experience.

In the short term, consumers will need to add additional speakers to gain the improved audio benefit; and in the near future we will see sound bars that will reduce the complexity and cost of delivering this technology into the home theater and entertainment based businesses.

A typical 22.2 audio experience would require almost 1.5 Mbps when utilizing 24 channels at 48 kbps with constant bit rate (CBR) encoding. By switching this to capped Variable bit rate (cVBR) encoding, a substantial reduction in audio bandwidth utilization will be realized. Additionally, new sound bar technologies will reduce the cost, complexity, and number of speakers required to bring a true 3D audio experience to the customer.

As part of the distribution process, care must be taken to monitor every channel and to ensure multichannel audio is down-converted to basic stereo for playback though the television speakers. While it is assumed 4k content will be viewed with enhanced audio, consumers may select to view the content while utilizing the stereo audio capabilities of the display.

COSTS

The costs to enhance the end-to-end solution for 4k can be broken into their representative components. Here is a partial list of items that may require upgrades.

Encoders – Existing VOD encoders have the ability to deliver 4k video with few modifications, while linear encoders will need to be developed that can handle massive amounts of data in very short periods. Additional modifications will be needed to handle advanced audio technologies such as Dolby Adaptive Audio, SRS Multi-Dimensional Audio, and 22.2 specifications. There will need to be a clear roadmap to get from initial 4k video with H.264/AVC encoding to HEVC. For the initial launch, a single linear 4k encoder should suffice. It will allow a MSO the ability to replace four HD streams with a single 4k stream. For VOD, it is possible to scale the number of encoders to match the size and refresh rate of the library to be converted.
SRM – A next generation SRM will need to be deployed in order to allow the VOD pump to select a 4k asset.
Metadata – New fields will need to be included to indicate the asset is 4k.
Content Encoding Profile – New profiles for 4k encoding will need to be defined.
Storage – 4 times the storage per asset, as compared to HD.
Video Player – Support for new Video and Audio formats.

Adaptive Dynamic Streaming – Support for additional audio CODECs or video CODECs.
Backoffice – Enhancements for billing.
Set Top Box – Faster single or multi-core CPUs and bigger pipes.
Direct Fiber – Larger pipes for mezzanine sources.
Mixing new audio – New mixing technologies for audio.
Trucks, Cameras, Post – Enhancements to editing systems, graphics, and source acquisition equipment for live capture content.

SERVICE AND INFRA-STRUCTURE VIEWS

It has already been demonstrated in the laboratory that 4K video encoding for VOD can be accomplished on existing encoders. A single 4k transport stream is generated and sent across the plant for decoding on a Personal Computer (PC). This stream is then split into four separate streams for delivery to the display across four separate HDMI cables. Once the HDMI interface is upgraded, it is expected a single stream and HDMI cable will be attached to the television.

Linear encoding could be done by handling the encode as 4 separate HD processes that need to be synchronized (hence QuadHD) and distributed as a single stream on the wire. It is important to note this implies that within the video encoding process, the input stream could be split for processing and then combined into a single stream for transport.

While this approach may be viable, newer, multi-core CPUs will most likely be able to handle the entire encode as a single transport stream. A single stream approach across the entire plant will increase operational efficiencies and simplify the operational model. In the case of adaptive streaming, fragmenting a single transport stream would require the identification of a single boundary point in the source video.

The same intuitive logic applies to network DVR. As previously stated, utilizing the same encoding techniques for linear and VOD is optimal due to simplicity and overall operational models for distribution of 4k video (figure 3).

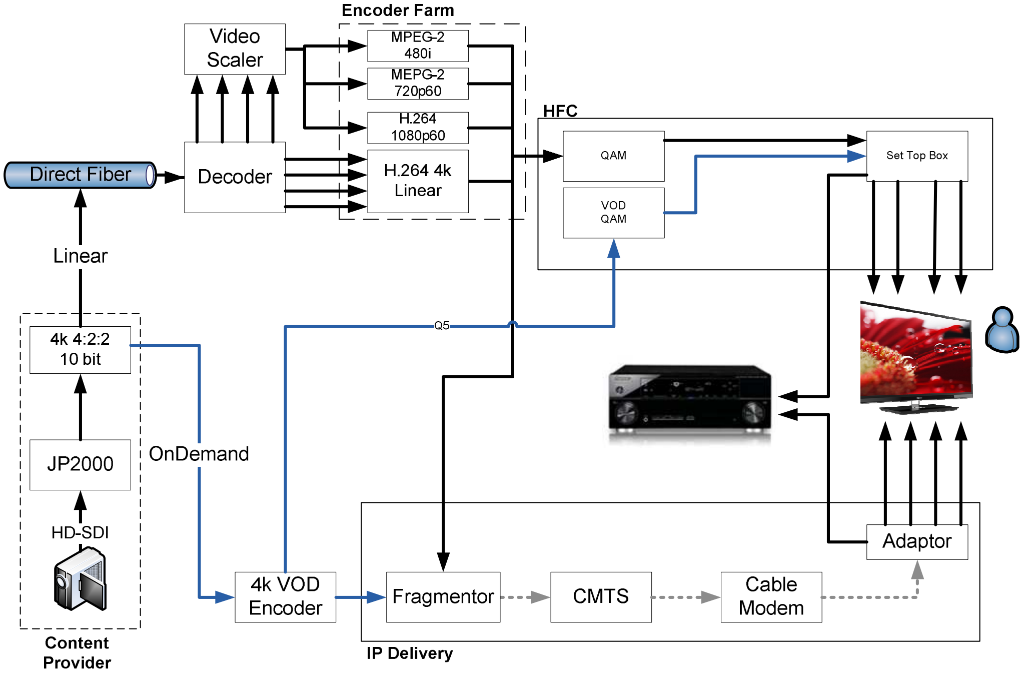


Figure 3 Operational Model for 4K Distribution Video

HD encoding of 1080I30/1080P24 using newer encoding techniques could range from 5-10Mbps when compressed with AVC/H.264. Offline VOD compression will most likely be superior due to multi-pass encoding. If 4K is supported at the same frame-rate, this could imply an encode bit rate from 20-40 Mbps in a cumulative data sense. This does not assume further compression efficiencies due to increased pixel density.

Can the infrastructure support a 40 Mbps 4K stream? A single 40 Mbps 4K channel would:

- Require the same bandwidth as 4-8 HD channels,
- Not fit into a 38.8 Mbps QAM
- And would likely not be carried by an ISP over the public internet

The bandwidth infrastructure modifications for this approach would be cost prohibitive. One bound stream could possibly be fit into a single QAM with bandwidth of 38.8 Mbps, which would replace about 2^+ MPEG-2 HD channels (or 4 HD streams on a 4:1 Mux). To meet a 4K service for HD, each QAM would need two 4K channels. This would mean each 4K channel would need to be bounded under 19 Mbps which would be about 1 HD channels and 1 SD channel.

Is it possible to move from a 1:4 upper bound bit processing ratio to a 1:1.3 ratio? With new coding tools from MPEG such as HEVC, a 50% improvement in compression can be expected. Additionally, having greater pixel density should create some further compression efficiencies to decrease the 1:4 ratio. Even more efficiencies can be gained by the way of improvements to perceptual modeling of our visual system and applying this to coding.

There is room to create more compression efficiencies, especially since encoder design is evolving and new compression tools are becoming granular. And even if a greater frame-rate is needed, pixel processing burden would be less than expected due to increased efficiency in motion vector accuracy and longer GOP length for the same amount of time.

As we examine all of the factors of better compression, filters and modulations, it does become possible to create a 4K stream that should ultimately approach 15 Mbps in the near future.

The next part of this paper will look at potential places to leverage the human visual system model to increase compression efficiencies through perceptual coding.

PERCEPTUAL CODING AND THE HVS MODEL

HVS (Human Visual Systems) attempts to describe how we actually see [from the photoreceptors in our eyes into the visual cortex and other parts of the brain]. Perceptual video coding is used in “lossy” compression at a target bitrate to mask, transform/quantize, or conceal information that is not seen by our visual systems (psycho-visual redundancies) or is optimized to improve what we can see. This is not coding efficiencies due to manipulation of the bit-stream to improve bit/symbol rate of the stream. It attempts to narrow the total information rate to what is just needed for our visual systems.

Our eyes are made up of 127 million photoreceptors in the retina (120 million rods and 7 million cones) that feed a million neurons in the optic nerve that is connected to the brain [Figure 4]. That already represents about 127:1 convergence of information. The rest of the eye is there to focus, shape, and control the amount of light going into the retina. The rods are used for vision at very low light levels (scoptic) and do not contribute very much to color perception. However, the cones deal with vision at higher light levels (photopic) and with resolving fine spatial details and color. These cones are divided into three types (S-short, M-medium, and L-long) that are sensitive to different wavelengths of light and they are the basis for our ability to match any color through a combination of three primary colors (trichromacy) [Figure 5]. The cones are concentrated in a central part of the retina called the fovea which provides the majority of information traveling along the optic nerve. The fovea matches to what we perceive as “the center of focus” for our vision.

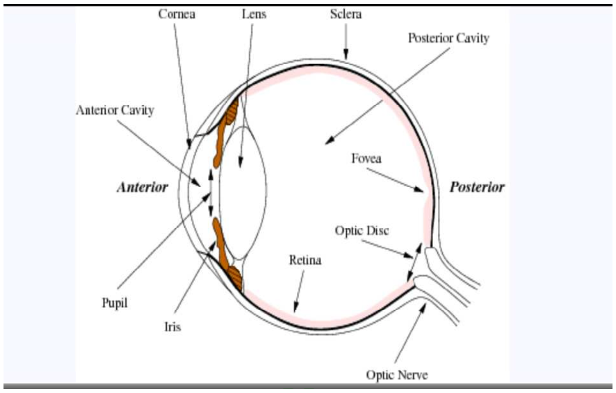


Figure 4 Eye

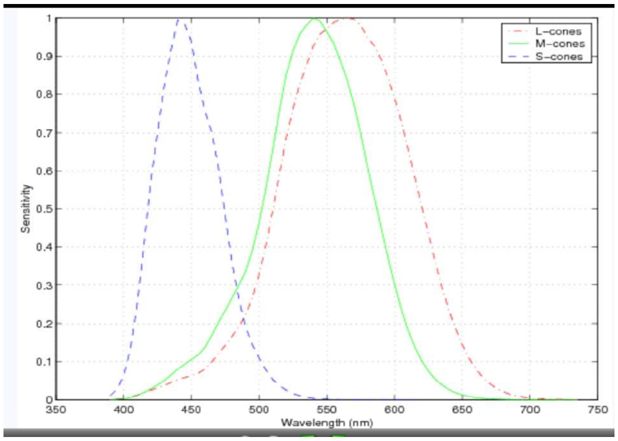


Figure 5 Different Cone Type Wavelength Sensitivity

This information electrically stimulates the optic nerve which feeds into the visual cortex of the brain for semantic and feature processing based upon differences to a windowed-steady state visual model. Eye movement, both right tied with left (saccades), is based on spatiotemporal sensitivities to capture these differences to the brain. From what we see in the human visual system, the visual cortex in the brain does not try to process all information but just what is needed to provide a semantic visual model. Perceptual coding attempts to move past the photoreceptor stage to keeping just the information that will make it into the visual cortex.

So, in trying to model HVS, it can be split up into three areas: 1) a visual attention model, 2) spatiotemporal visual sensitivity model, and 3) a visual masking model. This is basically what is interesting to see, what we can make out of it, and what we could never see at all. Our visual system is sensitive in a number of ways:

- **Contrast**- we aren't sensitive to a level of brightness, we are sensitive to differences in brightness between areas in our vision. This equates to sensitivity to edges in an image and can be affected by the brightness in the background.

- **Spatial Frequency**- as spatial frequency increases, we become less sensitive to variances in spatial details (when does edges become texture?). This can equate to tolerance in coding artifacts in high texture areas as opposed to more constant areas. In color we are even less sensitive to variances in spatial frequencies. Hence one of the reason we can sample color difference less frequently than luminosity (4:2:2 or 4:2:0).

- **Visual Acuity**- This is the ability for the eye to resolve details. One can have reduced visual acuity in fast moving objects (though eye tracking can reduce perceived motion of the object --- reduced retinal velocity). One can also reduce visual acuity by moving further from the object or screen. For ideal viewing, Viewer should be far enough away to not be able to discern pixels on the screen. Increased resolutions can allow for the observer to sit closer to the screen without being able to discern pixels [Figure 6].

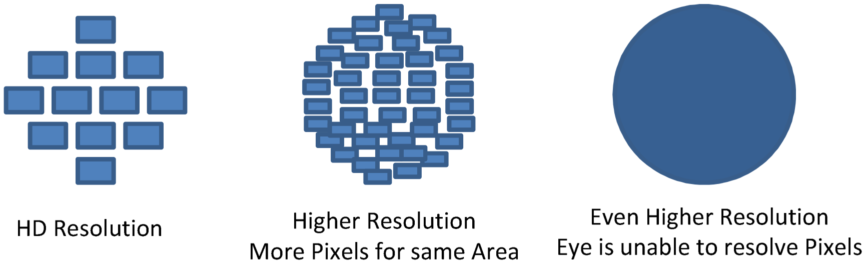


Figure 6 Visual Acuity and Denser Pixels

- **Noise**- These are unnatural changes in contrast due to the image capturing process. This could be due to the scatter on photo sensors in the CCD/ CMOS, heat on electronics carrying the pixel values, or celluloid processing leaving film grain artifacts [Figure 7]. The eye is sensitive to noise at different spatial frequencies which is why low-pass/

band-pass filtering is used as a preprocessing technique to remove these unnatural artifacts.

- **Temporal Frequency**- we are more sensitive to temporal cues rather than lack of spatial details. This is one of the reasons why interlacing can happen because it is a tradeoff of spatial frequency for temporal frequency to address bandwidth issues. It is believed below 50-60 Hz (fps), flicker can be perceived in a series of played out still frames. For this reason, 24fps material sometimes is flashed twice in frame playout on display devices and now material traditionally being shot at 24fps is being shot at 60 fps or even 120 fps for this reason. Additionally, movement that follows natural movement speed and direction is less surprising than erratic movement and speed.

- **Perceptual Uniformity**- This basically means keeping a consistent quality across a video sequence. We are sensitive to quality changes in spatial details of a moving object when viewed in the fovea area of the eye.

To mimic HVS, the attention model needs to identify areas of the image that are tracked by eye movement (saccade) to keep interesting areas in the fovea. Things outside of the fovea do not have to retain as much detail due to change blindness. Object size, and movement (predictable and unpredictable) can be used as cues to identify areas in the video sequence that need more spatial detail. Artifacts can cause a miscue in the eye to areas in the video sequence that are not natural areas of interest and need to be minimized where possible. The spatial temporal model can affect how to maintain a natural sequence with consistent quality over a content scene. Visual masking is a preprocessing function that can hide information in areas that don't need as much spatial detail such that it is coded in a fewer number of bits.

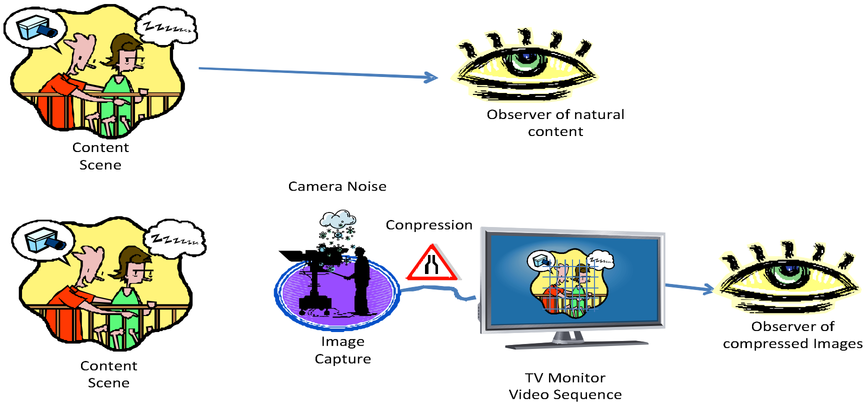

Content Scene
Observer of natural content
Camera Noise
Conpression
Content Scene
Image Capture
TV Monitor Video Sequence
Observer of compressed Images

Figure 7 Capturing Natural Content on Screen

EARLY PERCEPTUAL CODING TECHNIQUES IN COMPRESSION

When we directly see a natural scene, our eyes have a filter (mentioned in the sections above) that reduces the amount of information that reaches the visual cortex. We use our eye muscles, focus and movements to change what the cones in the fovea are seeing such that attention is there for important information in the scene.

To capture the image such that we can recreate what we see (Film/ TV without compression), we represent the scene through a series of still pictures being played at a specific temporal frequency (24 fps (2x)/ 30 fps (60 fields)/60 fps). Consistent quality is maintained between each frame, and interlacing techniques are used for further reducing bandwidth using a tradeoff of spatial resolution and temporal frequency.

However, in the capturing of the image, noise is introduced into the content scene through CCD/CMOS camera devices. To avoid seeing the pixels instead of the content scene, we sit back far enough (2H-4H) such that our visual acuity cannot discern a pixel and blends them together.

With the evolution of an analog medium (6MHz analog program) to a digital medium (10 Channels in 6MHz), we now have the ability to manipulate each pixel value and only send difference information between each frame (i.e. compression). In terms of pre-processing, the noise is being removed through low-pass, band-pass, and temporal filters like MCTF. The encoder then uses block-based transforms to change the coefficient values to be measures of spatial frequency energy.

At this point, the coefficients of higher energy frequencies can be quantized with less precision and use less bits because we have less sensitivity at high spatial frequencies. Additionally, this helps with reducing data redundancies in the bit streams since many of these coefficients are quantized to zero.

In terms of motion, movement of natural objects can only move at certain speeds and are predictable which factors in to some of the coding algorithms that reduce computational complexity. This allows for a reduction of motion search space, and a reduction of number of motion vectors based on size of the object. The “errored” differences between frames can also be quantized in the same manner since errors are mostly in high spatial frequency details. In post processing, the blocking artifacts along transform boundaries can then be removed from the image.

To avoid seeing artifacts from the medium (pixels) rather than the content scene, it is important to be able to view the display screen at the proper viewing distance. If one moves in closer, visual acuity increases to the point where pixels can be discerned (visual acuity is inversely proportional to distance). In terms of monitors, we are getting larger monitors going from 40” to 50” to now 60-70” sets, and the viewing distance from these monitors is remaining mostly constant.

Additionally, we are also getting display devices like tablets and PCs that are being viewed at much closer distances than the 2H-4H recommendations.

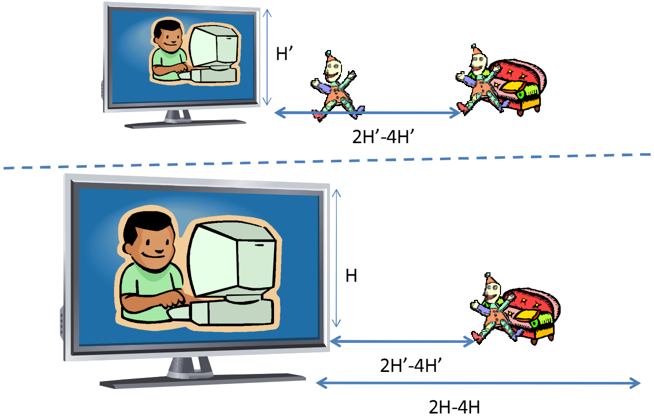


Figure 8 Screen Sizes, Distance, and Visual Acuity in Monitors and PC/Tablets

AFFECTS OF 4K AND HIGHER FRAME-RATES

Going to 4K can create more natural content scenes. Increasing pixel density does not have to create a larger picture; it creates a more densely sampled picture. Each pixel now represents a smaller area which allows for:

- Sharper Edges
  - ✓ Fonts on letters are sharper. The viewer can read documents. [It’s “Resolution-ary”].
  - ✓ Less aliasing artifacts and “jaggies” around edges
  - ✓ Textures are more detailed
- Increased pixel density
  - ✓ Approaches visual acuity limits. See less pixel definition and more of the picture at closer viewing distances and angles.
  - ✓ The Viewing distance becomes more flexible. We can get closer to pictures in both large and small displays (This aligns better with the attention model)
- Better contrast
  - ✓ Pictures look brighter/ more natural due to contrast differences and more gradient increases and decreases (This was always an issue for compression)
  - ✓ Neighboring pixels are more correlated since they represent a smaller area

Going to higher frame-rates can create more natural content scenes cues, by sampling motion in content scene to make it more linearly predictive. This is becoming more helpful as CGI (computer-generated imagery) effects in film and video content introduce faster moving objects in sequences. It is also very helpful in sports content where motion is quick and erratic. If the frame rate is too slow for the motion in the content scene, we can get "juddering" artifacts especially if the picture is flashed multiple times to simulate higher frame-rates:

- Smoother Motion
    - ✓ Movement between frames is shorter and can be predicted better
    - ✓ "juddering" can be reduced due to more sampling of motion and less repeated flashing of the picture
- Less Noise from Image Capturing devices
    - ✓ Noise is not temporally correlated and can be filtered through comparisons of sequential frames.

LOOKING AT CODING WITH RESPECT TO HIGHER RESOLUTION AND COMPRESSION

With increases in resolutions, there are going to be more pixels to process. The encoder picks a target bitrate and then tries to make decisions in coding based upon that. Generally, the encoder attempts to conduct:

1) ***Pre-filtering*****:** remove noise and apply a low-pass filter to remove information and details that would never be resolved at that bit rate anyways. Basically, to remove the information that makes the encoder work harder than it needs to be working.
2) ***Transform/Quantization*****:** change the information order of the data stream to make it more compact and quantize high spatial frequency information. Apply entropy coding to the output of this stream
3) ***Predict Subsequent Frames*****:** Use a reference frame(s) to produce a set of motion vectors and "errored" difference frames (P& B Frames). Calculation of motion vectors need to go through a motion vector search which can be a complicated encoding process.
4) ***Post processing*****:** Conceal artifacts created by the encoding process such a blocking and boundary artifacts through post filtering approaches

Places where we can improve this process, due to having higher resolutions and frame-rates, include:

1) ***Pre-filtering:*** Removing noise may be easier because it is approaching the granularity of our visual acuity while natural content scenes would not have this level of granularity. Using the stronger correlation between neighboring pixels, there can be improved techniques for filtering and dithering to handle noise. Additionally with the improvements in CMOS, we may be able to do this earlier at the point of image capture.
2) **Transform/Quantization**: The transform represents a smaller area and more correlation between the pixels which can help in energy

compaction. Some savings can be achieved as well because quantization levels can be changed for a smaller area. However, there are more transform blocks to deal with at higher resolutions.

3) **Predict Subsequent Frames**: With higher resolutions, movement can go beyond the motion search space, which would mean more bits to encode. With higher frame rates, movement is shortened between frames and is much more predictable, which could reduce the amount of bits that are expended. Objects are also bigger (have more pixel density), which would require less motion vectors to support this process. With ½ pel (pixel) motion accuracy across a smaller portion of the picture, the effect of this approach could be fewer errors in the "errored" difference frame. With more accuracy this can save on bits as well. Lastly another effect is longer GOPs over the same time period (just more frames in the same period) which can reduce the expected increase in data through temporal compression.
4) **Post-processing**: There would still be blocking artifacts that would need post processing it would just be smaller in the picture and may only need simpler post-processing techniques.

ENCODERS AND NEW CODING TOOL ABILITIES

With new demands for multiple bitrate encoders and addressing multiple devices, encoders have been evolving to output streams at multiple target bit rates. In many encoders, there is already a calculated quality metric used to make encoding decisions used for the purpose of meeting multiple target bitrates. Additionally encoders are also deploying "look ahead" to analyze the source content to optimize encoding decisions. Both these mechanisms help out in maintaining perceptual uniformity and enabling better visual masking throughout the video sequence through the use of dynamic adaptable filters.

The newer coding standards (i.e. AVC/HEVC) have also been evolving that are developing advancements in coding tools to handle each sub-area of the image and sequence in a different manner. The objective is to use as few bits to convey parts of the image or sequence that don't need as much detail such that more bits can be spent elsewhere. For instance, the background may not need as many bits as a moving object in the foreground. Also, a moving object may not need as much motion vectors since the object travels at the same relative speed against the background. Some tools being developed or refined are:

- Spatial Intra-frame compression Techniques
- Better motion pixel motion search down to ¼ or 1/8
- More granularity in quantization across coding units or transforms
- Changing the transform block size- 8x8, 4x4, 8x4, 4x8

- Changing the size of the macro-block (16x16 to 64x64)
- Changing the number of motion vectors needed for a macro-block
- Reducing the number of motion vectors needed for coding

These different tools contribute to being able to identify and handle separate areas of the image, treat specific bands of spatial frequencies with alternate options, and to code objects as separate temporal frequencies. Combine this with the ability to analyze content and a calculated quality metric in the encoder, and you have the basic tools for creating an attention model along with further refinements in the spatiotemporal sensitivity model and visual masking. From this, the HVS model used in encoding can rapidly improve encoding and reduce the amount of bits needed that can be processed by our HVS system beyond the 50 % reduction already claimed by the latest codecs.

CONCLUSION

The first phase of 4k video delivery should focus on a quality experience for the customer. It is expected that 4k will start with a single, linear occasional channel and a small library of 4k VOD assets encoded with H.264 compression techniques. Should 4k prove to be a success, it will be easy to expand the VOD library by transcoding studio content into higher resolution video.

In order to support new audio formats, assets would need to be remixed. MSOs could have a very basic 4k solution in place in the very near future; and HEVC encoding will allow a production 4k solution using substantially less bandwidth. Based on our calculations, and leveraging coding algorithms sensitive to the human visual system (HVS), 4k assets may in the near future consume the same bandwidth on the local loop as an existing HD asset encoded with MPEG-2.

References

[1] Tang, Chih-Wei, *"Spatiotemporal Visual Considerations for Video Coding"*, IEEE Trans. On Multimedia, Vol. 9, No. 2, Feb. 2007, pp. 231- 238.

[2] Naccari, Matteo and Pereria, Fernando, *"Advanced H.264/AVC-Based Perceptual Video Coding: Architecture, Tools, and Assessment"*, IEEE Trans. On CSVT, Vol. 21, No. 6, June 2011, pp.766-782

[3] Wu, H.R. and Rao, K.R. eds., Digital Video Image Quality and Perceptual Coding, CRC Taylor and Francais Group, New York, 2006.

[4] JCTVC- G1113 WD5: Working Draft 5 of High-Efficiency Video Coding, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 7th Meeting: Geneva, CH, 21–30 November, 2011

[5] ITU-T Rec. H.264 | ISO/IEC 14496-10, (2005), *"Information Technology – Coding of audio visual objects –Part 10: Advanced Video Coding"*

[6] *"Understanding CCD Read Noise"*, www.qsiimaging.com/ccd

[7] Additional Conversations and some eye diagrams Dr. Damian Tan and Dr Henry Wu, School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Melbourne, Victoria Australia.

SURFACES: A NEW WAY OF LOOKING AT TV

Simon Parnall, Kevin Murray and James Walker
NDS

Abstract

The rapid evolution of home display technology offers the potential for an ever-more realistic and immersive experience of media and, within a few years, we will see large and yet also unobtrusive 'lifestyle' surfaces that could cover a whole wall. In the face of such capability the obvious question is "How might the television experience evolve?" and our vision is of a better, more integrated system that provides viewers with both a collective and personal experience and which adapts to a range of sources, including metadata, from both outside and inside the home.

INTRODUCTION

The choice of type and size of television screen for the home is so often a compromise between the extremes of an exciting viewing experience when the device is switched on and the wall or corner space occupied by a dark and dull object when the device is switched off. And, when the screen is on, the size of the picture may well be inappropriate for the type of content and engagement of the occupants of the room.

Science Fiction overcomes such concerns by assuming an invisible and scalable screen – often taking the place of the wall itself, or a window or indeed in mid air. Science Fiction has also assumed an intelligent management of presented material, following the individual and assimilating and prioritizing a range of sources.

Today's mobile phones make the Star Trek communicator look somewhat bulky as advances over the years have successively removed the novelty of such a concept. In the same way today's screen, projection and graphics technologies are slowly and steadily bringing us closer to a reality of the vision of Science Fiction. In fact, we are now very nearly at the point where key aspects of this vision can be realized and could be adopted by consumers in the not-so-distant future.

Walk into a consumer electronics exhibition today and you will find many example components of this vision. There are thin-bezel screens that can be treated as tiles to create larger and larger surfaces, or glass screens that transparently reveal the wall behind when off. We already have sophisticated companion devices offering touch control and each year we are seeing ever more sophisticated gesture and voice recognition.

Our role in this opportunity space will be to create the technologies that integrate such components to produce a sophisticated and intuitive user experience that matches content and mood, and which produces pictures of an appropriate size and position for each circumstance. Furthermore the presented audiovisual content will be supplemented with additional content and so-called domotic feeds (that is material concerning the home).

In this paper, believing in the inevitability of this trend in display technologies and the opportunities this creates, we set out our vision for how the television experience will evolve, some lessons learned from our first prototype implementation of this vision, and touch on our plans for the second-phase prototype which we are currently constructing.

VISION

Our vision of the future is of a viewing environment with one or more large display surfaces. Surfaces that are a) frameless, b) unobtrusive, c) ultra high-definition and d) ambient. These surfaces can be adapted to fill or partially fill one or more walls of a room, and will co-operate to provide an integrated experience. The opportunity is to open up possibilities way beyond the limits of today's devices though:

- content comprising multiple visual elements that can be adapted spatially and temporally, freeing the user from choosing a single element, or the system from having to impose overlays;
- shared, co-operative usage of the surfaces, with connected companion devices becoming personal extensions;
- supporting connected applications and services operating in a more streamlined, integrated manner, reflecting and effecting changes in viewer engagement in TV content;
- dynamic adaption to, and control over, the environment the surface finds itself in – such as physical size, resolution and the room in which it sits (e.g. adapting to the wallpaper or controlling the lighting); and
- introducing domotic content into the TV display in a sympathetic manner.

Based on this vision, a prototype was constructed and demonstrated at both IBC 2011 and CES 2012. This prototype has a single surface occupying most of one wall and a photograph of this is shown in figure 1. This shows a single surface constructed from six screens and one of several companion devices that may be used simultaneously to control and interact with the system.

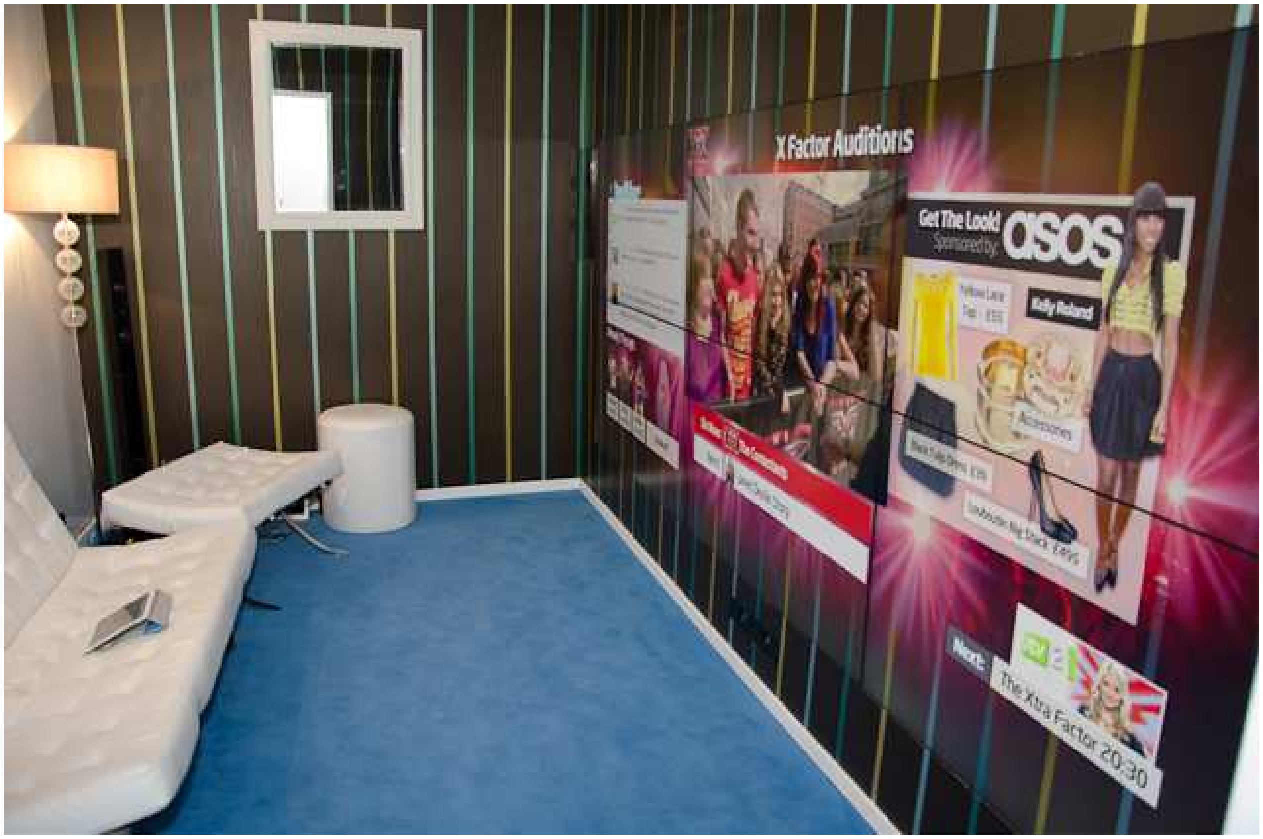


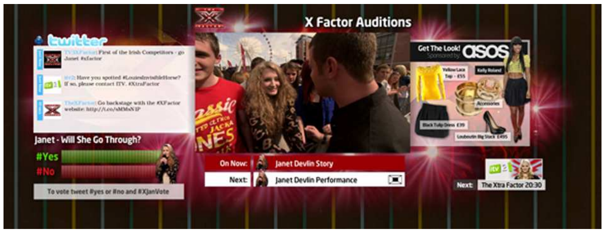
Figure 1: Prototype System

IMMERSION

Many programs have a natural flow and pace – points at which the viewer or viewers are extremely immersed and engaged in the content. Examples of this may be a critical part of play in a sports game, a news story of direct relevance or a very dramatic scene in a soap. Likewise there may be times of lesser immersion or engagement. Examples of this may be waiting for players to take their positions, an uninteresting news item or a section of the soap that is recapping past happenings. In these areas of lesser immersion, the viewer's interest may naturally be taken by other related items, such as the current scores in related games, the next news story or what is being said about the soap by their social contacts.

In our system we have introduced the concept of 'immersion'. Immersion is key to the way that the surfaces are used and the way that the content is presented on them. Put simply, the more immersed in the content the viewer is, the greater emphasis that is placed on the core video, and the less immersed they are the more emphasis comes to be placed on related content which may then be introduced. This related material could be social media, advertising, program graphics, additional material, or virtually anything.

Examples of high and low immersion are shown in figures 2 and 3 respectively, which are screen captures taken from our prototype system. In figure 2, we see how the video roughly shares the surface with other social, voting and advertising graphics and content sources during the scene setting and build-up to the main performance. By comparison, figure 3 shows the high immersion example where the program in figure 2 has moved on to the main performance, and the related items have been removed, and the video increased in size and prominence.



TV3XFactor: First of the Irish Competitors - go Janet #xfactor

itv2: Have you spotted #LouiesInvisibleHorse? If so, please contact ITV. #XtraFactor

TheXFactor: Go backstage with the #XFactor website: http://t.co/sMMsN1P

Figure 2: a low immersion example

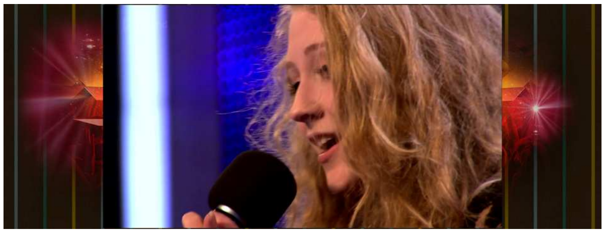


Figure 3: a high immersion example

In our prototype system, immersion is controlled in two ways – via "broadcast metadata" (as was used for the examples above) which indicated the broadcasters expected level of immersion, and also via a slider in the companion device which allows the user to modify the immersion (both up and down) as they wish. Clearly other mechanisms could also be employed, such as audio or video analysis of the room and the viewers, but the prototype shows that these two simple mechanisms work very effectively.

TECHNOLOGICAL MOTIVATORS

Displays

Display technology is continually improving. We have seen that relentlessly the average screen size is increasing year by year, as evidenced by (3). But there are two key technological changes which directly relate to our vision.

Firstly, screen bezel sizes are getting smaller. Our prototype system uses professional monitors with 5mm bezels, but LED backlit consumer displays are approaching similar, or better, bezel sizes and

OLED offers the prospect of a bezel width of near zero. Even with today's widths there is the real option of creating large ultra high definition surfaces out of tiled arrays of inexpensive displays.

Secondly, whilst still in the research laboratories, transparent displays which naturally allow the underlying environment to show through are starting to be developed. These would trivially allow the blending of displays into the room environment.

Video Content

We are also starting to see the first indications of the next jump in resolution beyond HD with the advent of 4K – both in displays and in content. At the same time as this higher resolution content is arriving, the importance of lower resolution content is not diminishing, whether from archives, citizen journalists or from challenging remote locations. Thus it is becoming hard to just assume that any content will look acceptable on any display size.

Non video Content

Outside the display arena, we are seeing ever more related data sources, from social media through games to dedicated websites. In the interconnected world, these are a crucial part of the entertainment experience, but today we are faced with the dilemma of either destroying the television experience by placing graphics over the video, or taking the viewer away from the lean back world of television into the very different and highly-interactive world of the internet.

BREAKING THE SCREEN BOUNDARIES

Today's television makes the basic assumption that "the display is always filled". Thus, video will fill the display, regardless of the size of display, quality of the video, or the resulting impact of an oversized face or object; and it also effectively does only one (main) thing at a time.

With larger, higher resolution display surfaces this implicit behavior and more can be challenged. Content need no longer necessarily fill the display surface, and the display surface can simultaneously be used for many different components.

In turn, these new capabilities mean that the traditional means of laying out video and graphics can be challenged. For instance we might:

- share the display between the content of more than one viewer, helping to make the TV a shared focal point rather than a point of contention;
- 'unpack' the constituent elements that are composited by a broadcaster in post-production, presenting these alongside the 'clean' audio-visual (AV) content, leaving it un-obscured. Obvious examples include digital on-screen graphics such as tickers, banners and sports statistics. To enable this, the composited elements would need to be delivered separately alongside the clean AV and then rendered in the client;
- 'unpack' all of the contextual assets that are composited in the Set-Top Box (STB), such as interactive applications and multi-screen content (e.g. multi-camera sports events);
- present contextually relevant online content alongside the video, for example, relevant web content, social comments (such as twitter hash-tags for the show), relevant online video etc;
- enable navigation and discovery user interfaces to be presented alongside video, going beyond today's 'picture-in-guide' presentation;
- present personal content, which whilst not directly related to the main television content, may still be desirable to end users to be seen on screen. Examples would

include personal social feeds, news feeds, images, discussion forums etc;

- present domotic content, such as user interfaces for in-home devices and systems, which can include video feeds from devices such as security cameras, door entry systems and baby monitors; and
- integrate visual communications, such as personal video calls, noting these may sometimes be used in a contextual way e.g. virtual shared viewing experiences between homes.

Thus, the way the TV experience takes advantage of the surface is by continuously managing a wide range of content sources and types that are combined appropriately for presentation.

Real Object Size

The tradition of a television picture scaling up to fill the display means that an object is effectively displayed at an unknown size. With this assumption broken, it now seems realistic to allow an object to be displayed at its real size, regardless of the display (as displays report their size though the standard connectors). For instance, in advertising it could be interesting to show just how thin the latest phone really is, just as is possible in print media today.

Content Opportunities

In the same way that the composition has always assumed a need to fill the rectangle, so has the creation of video content – which has followed the model of filling the proscenium arch of classical theatre. The proposed systems can offer new opportunities to the content creator.

One simple example of this is shown in figure 4. Here, the movie trailer is blended into the background to give the appearance that it tears its way through the wallpaper, dramatically conveying the unsettling nature of the promoted movie.

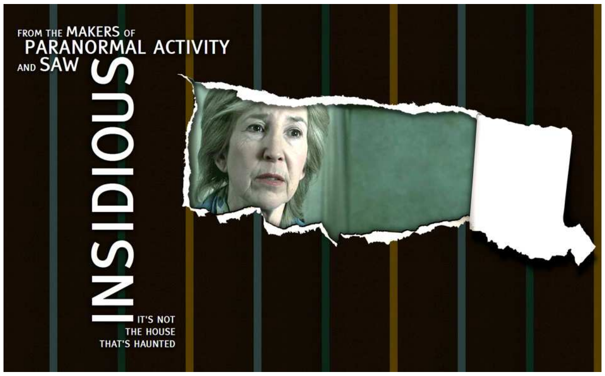


Figure 4: Non-rectangular content

There are numerous other areas where this technique opens up new opportunities. For example:

- editing could become more subtle with gentle fades, and several scenes can co-exist for longer and with less interference;
- content need no longer be fixed into a given size – if portrait content is provided from citizen journalists, then it can be displayed naturally in that form; and
- multiple synchronized videos could be used, in a fashion made popular in TV series such as 24, but without any requirement for their relative placement.

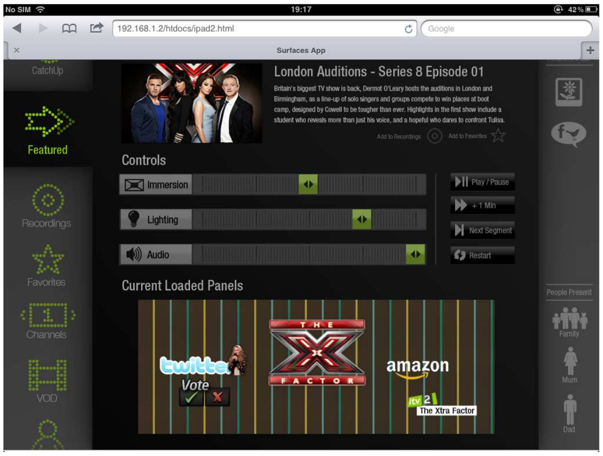
Implicit in this capability is the requirement to support an “alpha plane” style functionality that can be used both to describe arbitrary shapes and to allow for blending of the content into the background. This is, of course, not new and techniques such as luma and chroma keying are well known both in the professional head-end market place as well as supporting functionality in DVD and BluRay media. However, bringing this functionality into a traditional broadcast chain would represent a new usage.

A COMPANIONABLE EXPERIENCE

The growing importance of companion devices (tablets, phones, laptops etc) to the modern TV experience cannot be understated. Such devices permit us to construct an experience which is, at the same time, both collective (involving everyone in the room) and yet personal (allowing each person to interact with the various elements as they wish).

The companion device is key and integral to our prototype experience – and interactions with the companion device are directly connected with what is seen on the main surface(s). This is achieved through several means:

- The companion devices are able, within constraints, to adapt the content on display, including adding or removing components or re-arranging the layout. An example of this is interface is shown in the iPad screen capture of the web-browser in figure 5, where, for instance, the display can be re-arranged by dragging around the icons representing the parts of the content displayed on the surface.



192.168.1.2/htdocs/ipad2.html

London Auditions - Series 8 Episode 01

Britain's biggest TV show is back, Dermot O'Leary hosts the auditions in London and Birmingham, as a line-up of solo singers and groups compete to win places at boot camp, designed by Cowell to be tougher than ever. Highlights in the first show include a student who reveals more than just his voice, and a hopeful who dares to confront Tulisa.

Figure 5: A Companion Application Interface

- Interactions, such as voting or feedback is done on the companion device, but this directly feeds back into the graphics displayed (in addition to the normal feedback one would expect).
- Control over the level of immersion. Although, as discussed earlier, a change in the level of immersion can be triggered through broadcast data and sensors in the room, the companion device is fundamentally able to control the final immersion experienced. In the prototype, as shown in figure 5, this is managed through a slider.

This approach results in interactions with the companion device that end back at the main display surface(s), rather than just with the companion device itself. For example scores from a game played by the whole family during a show could be displayed on the main surface.

A SURFACES SYSTEM ARCHITECTURE

The prototype system constructed to explore our vision was built using a single, six-output computer (an AMD Eyefinity graphics card in a powerful PC) with software that was itself built on standard HTML5 technologies (e.g. javascript and CSS transitions) in functionality largely contained within a standard browser. This approach enabled a fast and flexible development and exploration of the principles. Whilst the HTML-5 toolset proved to be an excellent platform, the use of a single six output graphics card places fundamental limits on scalability, the number of display surfaces that can be supported and, of course, on cost.

In our current work we are moving from the architecture of our first prototype. We are doing so because we will be using multiple display surfaces within a single room, and exploring how these can be combined for the presentation of a single entertainment experience, and co-operate to support multiple simultaneous entertainment experiences (e.g. the big game and the soap).

To achieve the required flexibility in the number of surfaces, scalability, cost and content presentation dynamism, we are developing a more advanced architecture, based around several concepts, including:

- rendering the graphics and video on more than one independent device;

- utilizing synchronization between the rendering devices, such as used in SAGE (1), but tailored for the specific use cases we are tackling;
- a separation of layout policy issues and rendering issues; and
- a single layout with a "world view" of the entire set of surfaces in use.

A high-level overview of the current architecture is shown in figure 6. This shows two separate surfaces, each driven by its own client. These clients then interact with the layout and synchronisation server(s) to ensure a consistent experience across the surfaces. In addition, the diagram shows that the audio is driven from only one surface, a deliberate choice to simplify the architecture.

Synchronization Architecture

It is important to be able to synchronize content spread between different clients. In a more traditional broadcast architecture, this would theoretically be possible using mechanisms such as the PCR values contained within a transport stream, but our approach does not assume either a direct transport stream feed to each client, or even that the content is made available in transport streams (e.g. it could be streamed over HTTP using any one of a number of mechanisms such as HLS or Smooth Streaming).

Instead, we have chosen to synchronize to a master audio playback clock on the main audio output. Where broadcast content is being consumed, there are many techniques that can be used to match this clock to that of the live broadcast content. This master audio clock is then replicated and synchronized via the synchronization server to other clients that are involved in playing back synchronized media.

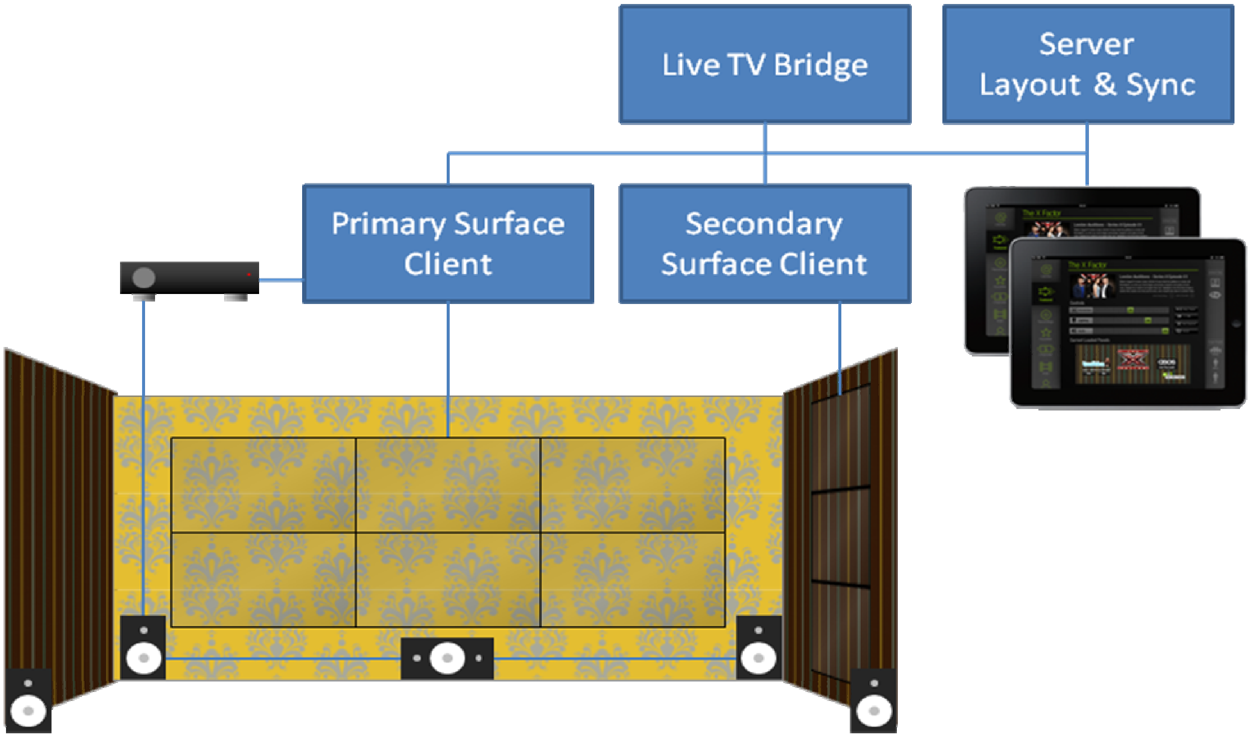


Figure 6: A New Surfaces Architecture

Our initial experiments with this architecture have shown that it appears to provide a reliable synchronization between different clients to a level that is acceptable for lip synchronization. Further experiments are underway to characterize and measure the accuracy that can be achieved.

Audio Architecture

Normally, audio is decoded and presented with simply a level control. However, in our proposed system the audio architecture becomes more complex than in a traditional approach, with various audio processing operations becoming an essential part of the overall architecture.

The most obvious audio processing requirement is positioning. From the proposed layout of surfaces in figure 6, it is clear that the secondary surface is not between the main speakers, and so any video that is presented on this surface with synchronized audio needs to have this audio repositioned. This repositioning needs to be dynamic, for instance as a video is moved from the primary to the secondary surface, the audio should be moved in synchronization. And, given the potential size of a surface, repositioning of the audio is desirable even when the content is moved within a surface. For example a video that occupies only the left third of the surface should have its sound stage correctly placed.

Earlier we discussed the concept of immersion, and how the video element of the experience can be balanced against other components to reflect the levels of interest both through a program's length and of a given viewer or viewers. This has a direct mapping to processing of the audio. Whilst the volume levels are one key part of this, this is best when combined with controlled compression – a reduction of the dynamic range of the content so that quieter parts become louder and the louder parts become quieter. Such processing allows the volume to be reduced in a fashion that retains access to the quiet sections of the content.

Much of the required functionality described above appears to be relatively easy to implement in the proposed Web Audio APIs that have recently become available on various platforms (2). This should make implementing the required audio architecture within an HTML5 environment relatively straightforward, and this work is currently underway.

Layout

One final component of the architecture deals with the layout of the media items to be displayed. Earlier in this paper we discussed how content typically packed together can be transmitted in an unpacked form, with the chosen and relevant components then laid out by the Surfaces system when the content is finally presented to the viewer. This process is not the highly constrained process we are used to where precise locations can be given for each item and, as the surfaces to be used might well be substantially different in each viewing environment, the process must be very flexible, and it is this flexibility that is an interesting challenge.

One aspect of the required flexibility comes from the number of inputs to the layout process to control what is displayed. These come from the local environment such as the range, sizes, locations and properties of the surfaces available and the immersion level of the viewer, and from the broadcaster, such as the list of potential components, their relevant priorities and a potential preferred immersion level. It is the layout engine that balances these inputs and selects a suitable set of components to display and locations for them.

In addition to the "what" of the layout is the "how', the appearance. More specifically, certain components may need to be adapted to the environment into which they are to be placed. For instance, if the room has white walls and the content item is white text, some means of making the text legible must be provided automatically. More generically, the design of an item should be able to adapt to the predominant background colors of the environment.

This introduces challenges at several levels that go beyond that of most current content presentation designs, such as may be found in many websites. Firstly we need an adaptive description of the requirements a broadcaster desires beyond those commonly in use today, and beyond even those of responsive web design (4). Next, we need a mechanism that can quickly and efficiently resolve these requirements in the face of a collection of local inputs. Finally, and perhaps most challengingly, we need the content producers and designers to understand that their content can and will be presented in many different ways, and a complete control over this presentation is potentially very counter-productive to the viewer's engagement.

CLOSING THOUGHTS

Our thinking started when considering the possibilities that the display industry will be offering in just a few years' time when the black boxes in the corners of out rooms disappear and unobtrusive, frameless, ultra high definition ambient surfaces take their place. In exploring the opportunities this technology will offer we have come to consider how content is presented, and the way in which its various components (current and future) will be assembled for the viewer. We have come to an appreciation of the way in which control and interaction with such an experience can work both in a personal and collective manner. And, in contrast to the 'lean forward' experience of today's connected TV we have seen how the 'lean back' experience of Surfaces requires a sophisticated automatic layout control engine.

As we have explored function, so we have explored form, and the PC based solution for a first stage demonstration now begins to give way to a believably scalable and cost effective hardware and software architecture.

It is often commented that the role of television in our lives has changed dramatically as other devices have fought for our time and won our attention. And yet, families and groups still wish to spend time together, sharing space and switching between personal and collective experiences. A developed television experience which embraces this truth, and which invites immersion and interaction at appropriate levels, must surely be for our industry a goal worth aiming for. Surfaces is, for us, a vehicle to explore this space and we are excited by the future we see before us, and the reaction we have received. The future is not one where the medium is marginalized, but a future in which people will truly find a new way of looking at TV.

REFERENCES

(1) High-Performance Dynamic Graphics Streaming for Scalable Adaptive Graphics Environment, *SuperComputing 2006, November 11-17 2006.* http://www.evl.uic.edu/files/pdf/SAGE-SC06-final.pdf

(2) Web Audio API, Chris Rogers, W3C, https://dvcs.w3.org/hg/audio/raw-file/tip/webaudio/specification.html

(3) Of Large LCDs, Unused Fabs, and Projector Killers, Pete Putman, Display Daily, April 9th, 2012. http://displaydaily.com/2012/04/09/of-large-lcds-unused-fabs-and-projector-killers/

(4) Responsive Web Design, Ethan Marcotte, 2011, ISBN 978-0-9844425-7-7, http://www.abookapart.com/products/responsive-web-design

TELEVISION 3:0 - THE MERITS AND TECHNICAL IMPLICATIONS OF CONTROLLED NETWORK AND CLIENT CACHING

Edmond Shapiro, VP Project Delivery Americas
NDS, Ltd.

Abstract

The cable industry has long debated the merits of using general purpose devices to access cached information in the network (commonly referred to as "cloud storage") as opposed to using cached information stored locally (on a device or within a home network), and in what combinations. In the past these trade-offs have involved the location of video on demand (VOD) and digital video recorder (DVR) storage. Today technical design decisions have become even more complex as engineers grapple with the growing number of caching permutations that will facilitate the deployment of Television 3.0, the next generation of IP based advanced digital cable services.

This paper analyzes network design considerations that cable engineers should consider when architecting Television 3.0, the next generation of IPTV applications using an array of cacheable information that includes: Application logic (cached JavaScript), Presentation logic (Remote and Local User Interface], Content (Adaptive Bit Rate (ABR) and Progressive Download (PDL) files as well as Metadata and Network Interfaces.

By highlighting the importance of an abstraction layer herein referred to as the Television 3.0 Common Service Framework, this paper explores hybrid architectures that permit network operators to dynamically cache information at multiple locations within a network – to dynamically adapt deployed services from one type of device to the next and from one region to the next – constantly evolving as new devices and network resources are made available in a rapidly changing technological environment. Smart software design builds an agile, future-proof foundation to increase deployment velocity of advanced services and enhance the operator's brand through improved system performance and better user experiences. Smart software design also avoids the many pitfalls of the past that have afflicted cable operators – from outdated devices and vendor lock-in, to degrading performance and feature bloat, to network-wide equipment upgrades in support of new services.

Specific applications and services highlighted in this paper include:

- *Content protection solutions – Conditional Access System (CAS), Digital Rights management (DRM)*
- *Content Delivery Networks – global, regional and federated*
- *Content recording and playback – DVR scheduling, resource allocation*
- *Ad insertion –graphic and video*
- *Multi-screen service delivery*

BACKGROUND

Television 1.0

The television application is a communication technology for the sharing of moving images with a group of people: the "mass media". The television transport network is more efficient when it can deliver the same information to more than one person at a time, as was originally the design of the analog terrestrial, cable and satellite networks.

The *Moving Pictures Expert Group* (MPEG) standards enabled the broadcasting of digital television services for the first time, within this paper referred to as the *Television 1.0 service*. As with analog, the Television 1.0 ecosystem was designed to efficiently transport broadcast digital information to a mass of people.

Digital television rapidly increased the amount of content or number of channels that a consumer could view at any moment in time and therefore the Electronic Program Guide (EPG) application was invented to improve content discovery.

With few exceptions (parental controls for one), the EPG user experience remains the same for every viewer. The underlying content changes (Linear and VOD events), but the EPG screens remain constant.

Web 1.0

Though the Internet is defined as the inter-connect of many different private and public IP networks for the purposes of sharing information, in the minds of most consumers, the Internet has become synonymous with the World-Wide Web (WWW) or just Web.

When the Web was invented, the Web user experience was as impersonal as the Television 1.0 experience. The Web was made up of pages written in a HyperText Markup Language (HTML), transferred from one computer to the next using the HyperText Transfer Protocol (HTTP). Each web browser eventually saw the exact same screens (same experience as Television 1.0).

Differences in user experience from one user to the next arose from the distance travelled by the HTML file. If a file crossed too many networks, it might be slowed down or even stopped altogether. The Web doesn't care whether a file is located on a computer in the same city or on the other side of the world. The response time for every Web request is unpredictable because there are no guarantees of reliable transport between the web server and the clients. User experiences on the web are considered "best effort" because of this unpredictability.

The first implementation of these standards came to be known historically as "Web 1.0". As the Hyper*Text* name implies, the first HTML files were simply text files carrying textual information. As long as text was the only content being broadcast on the web, the size of the text file made little difference to the user experience. Best effort was simply good enough, even when accessing files across very slow networks.

However, user experience designers quickly grew frustrated at the impersonal Web 1.0 experience and began to deploy "richer" graphical user interfaces including non-text based files, so called "binary files", such as music, photo and video files, which led to an explosion in the size and number of files managed by web designers (see Figure 1).

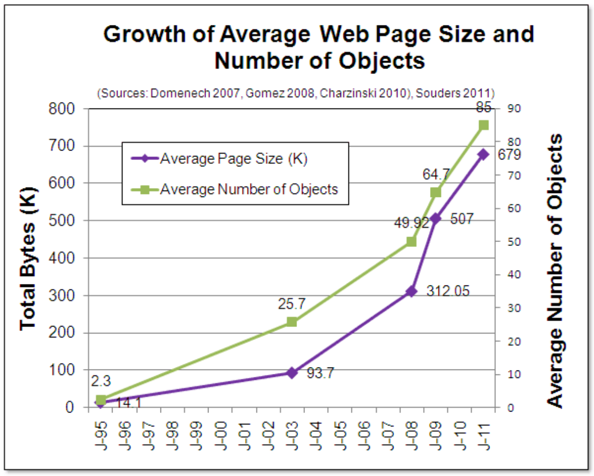


Figure 1: Growth of the Average Web Page

The best effort mode of the Internet was unable to adapt to the demands of the Television 1.0 application developers. Broadcasting (or multicasting) IPTV over the Internet has been notoriously difficult to achieve. Every network between the content distributor and the consumer has to agree to pre-allocate enough bandwidth to carry the fixed bandwidth required by the IPTV service. Only a private network, managed by a single network operator, has ever effectively scaled network capacity to implement this application (e.g. AT&T UVerse).

Web 2.0

User experience designers soon adapted and began to dynamically generate HTML files exposing a richer more personalized user experience for each user and device type. This personalization of web services became known as “Web 2.0”, or the social web, and has been exemplified by the success of web service providers such as Facebook and Twitter.

The successful Web 2.0 service providers learned to work around the “best-effort” design of the Internet by overlaying a virtual network on top of it, called a Content Delivery Network (CDN). The CDN was used to rapidly distribute and store (or cache) the many media files to as many edge networks as possible, as close as possible to the mass of the consumers, thereby avoiding network overload and reducing the number of network hops between the consumer and the media files, and thus reducing the unpredictability of the Web user experience.

CDNs take advantage of specialized algorithms to redirect HTML hyperlink requests to the nearest cache location of the requested media file. These algorithms are constantly and dynamically evaluating network boundaries to avoid bottlenecks and to determine optimal routes.

Over time this CDN virtual network adapted to many different uses as the various types of Internet connected devices exploded. Where once a web application could assume that all web browsers were located on a Personal Computer with a single screen size and a local cache, now mobile devices with smaller screens and no cache had to be accommodated. By delivering different size graphic files and deciding upon remote or local caches, the CDN was able to optimize content delivery to each type of device.

An equally important Web 2.0 development was the widespread adoption of a standardized programming language called JavaScript and the Extensible Markup Language (XML) that enabled web application developers to selectively retrieve parts of the HTML file based on local context or inputs (e.g. user actions, cookies, etc.), as opposed to retrieving the entire HTML file.

Television 2.0

With the growing capacity of CDNs to deliver rich media to Web 2.0 users, demand grew for the delivery of streaming media services such as Radio and Television. As traditional IPTV could not be predictably delivered over the Internet, new CDN friendly techniques were required.

Adaptive Bit Rate (ABR) technology arose from this challenge. Where traditional IPTV content was pushed at a fixed rate, ABR content is pulled at a number of different bitrates that can be influenced by the CDN as well as the local application context.

Instead of broadcasting a common monolithic media file to every consumption device, ABR technology delivers an index file (or manifest file) instead. This abstraction allows for different versions of the media file (size, resolution) to be cached and consumed at any time or place. Where the CDN is able to locate a faster network or when a client connection improves, then the user's viewing experience improves accordingly.

ABR technology has become the foundation for a generation of *Television 2.0* services. These Television 2.0 services have enabled service providers to deliver television to any type of device, and no longer just to televisions. ABR is particularly well suited for unmanaged or minimally managed networks, especially home WiFi networks.

Different ABR standards have competed in the marketplace, being driven by the commercial interests of major CE device manufacturers (e.g. Apple, Microsoft). MPEG recently standardized the *Dynamic Adaptive Streaming over HTTP* or DASH specification which incorporates a number of these approaches.

Cable's Challenge

Cable network operators have watched as popular Over-The-Top (OTT) service providers have taken advantage of ABR technologies to deliver high-quality (even HD quality) services over their fixed and mobile IP networks.

ABR technology is so efficient that it will squeeze nearly every bit of available bandwidth from the access network, limiting alternative services that the cable operator might wish to provide over that same network.

In some countries "net neutrality" regulations restrict cable operators from differentiating or prioritizing any type of broadband service. Cable operators are forced to implement service neutral bandwidth and data caps to control the growth of these OTT services.

The impersonal EPG of legacy Television 1.0 services cannot compare to the personalized and social media experience of many Television 2.0 applications. This is ironic for a cable operator given that the legacy broadcasting Television 1.0 network is generally more efficient at distributing television services.

To compete, cable operators have deployed their own multi-screen Television 2.0 services. However, these same net-neutrality regulations are being tested as subscriber usage limits may be ignored when using the cable operator's own Television 2.0 delivery networks.

From a technical perspective, there is little difference between a private IP sub network and the legacy Quadrature Amplitude Modulation (QAM) as both are based on the same MPEG networks that cable operators used to deliver legacy broadcast and narrowcast television services. All of these services must coexist on the same access network or "last mile".

As ABR encoding (or transcoding) can occur at any control point in the content delivery network, regulators will find it increasingly difficult to make these net neutrality distinctions.

A consumer who chooses to stream ABR content on their own has the option to purchase a CE device, such as Sling Media, to transcode and transmit set-top box content to any other device. A cable operator who delivers the same ABR service from the network will not only save the consumer the purchase price of such a device, but will reduce energy costs for all consumers by centralizing this functionality in the "cloud".

Television 3.0

OTT services delivered by global CDNs to any network on Earth are ideal for content creators and broadcasters who wish to expand viewing audiences, but only if the access networks are capable of delivering their service.

Broadband service providers who control the access networks will compete on the capabilities of their network infrastructure and will be judged by consumers on the user experience of these OTT services.

Though these various service providers may compete, there is also an incentive for them to cooperate. New *Television 3.0 services* will utilize these same CDN and ABR technologies but in a more network aware fashion.

Where Television 2.0 services utilized CDN overlays and local device context to optimize the user experience, Television 3.0 services will go further, benefiting from a deep and intimate knowledge of the underlying IP networks, and leveraging a *Common Service Framework* to abstract the user experience.

Caching and abstraction techniques led to the advances in Web 2.0 and Television 2.0 content delivery. Extending these techniques with the addition of new communication and collaboration tools will be the foundation for the common service framework of Television 3.0.

For example, publish and subscribe techniques will make it possible for an access network provider to expose to a Television 3.0 service provider the caching resources of the Edge CDN closest to the subscriber. Television 3.0 service providers who design their applications to account for these network variations will inevitably produce a better user experience.

Today most home networks are minimally managed by professionals. This makes the home network the last frontier for a managed IPTV service. Television 3.0 service providers will expand their management of the subscriber's home networks as well, such that the Edge CDN may very well be within the subscriber's home.

Television 3.0 services can be extended into the subscriber's home network only if a home gateway is capable of supporting these services. For instance, a managed gateway supplied by the cable operator could assure the bandwidth needed to supply a 3D service to every 3D capable device within the home network.

Such a technique will make possible other advanced services as well. For instance, managing the home network for the subscriber would enable plug and play video-conferencing, home security and energy management services to co-exist with television on most devices.

A Common Service Framework that is network aware will incorporate all of these advanced services into the service provider's branded user experience across any device and on any network. Television 3.0 features that are available only in the home network will be disabled on the road. Other features that are only available with a higher subscription tier will be managed in a common fashion across all devices.

The Television 3.0 service benefits from network awareness but should not be dependent upon it. All of the components of the Common Service Framework must be optimized for use within and outside of a managed IP network. For instance, this means that a DRM client application should provide robust enough security to validate a user and their consumption device anywhere in the world, and not just inside of their home or within their cable network.

Finally, as the Internet itself adapts to these new paradigms (e.g. Software Defined Networking), the Television 3.0 service will fully enable consumers everywhere to experience the true "TV Anywhere" service that they desire.

LEVERAGING ADVANCED CACHING TECHNOLOGY

Usage Context

The term "application framework" describes a software structure for developing software applications within a specific operating system or environment. The responsibility of the application framework is to provide "context" to the users of the framework, which are a set of software components called clients or "applications".

In a Television 1.0 service, the application framework that provides a common set of services for accessing the television transport stream is commonly referred to as "Middleware".

Middleware enables a common set of applications, such as the service provider's Electronic Program Guide (EPG), to share set-top box hardware resources, without being tied to a specific set-top box manufacturer. *Digital Video Broadcasting Multimedia Home Platform (*DVB MHP) and CableLabs *OpenCable Application Platform* (OCAP) are two standardized sets of middleware functions or APIs.

The Television 1.0 service has a relatively fixed or static usage context. The users of a middleware framework are expected to reside in a fixed location on a single set-top box attached to a single television. Dynamic events are limited to remote control or front panel user inputs or network control messages that are typically managed by a conditional access function.

The Television 2.0 service has an equally fixed or static usage context as well. OTT services are typically manipulated at the source (in the network) to conform with the requirements of a specific CE (Consumer Electronic) device manufacturer's chosen application framework, typically Microsoft (Windows), Apple (iOS) or Google (Android).

The Television 3.0 service has a much richer usage context, as it is designed to flexibly adapt to a more complex set of environmental variables. A Television 3.0 service may be executing on a fixed device such as a set-top box, or on a mobile device such as a Tablet.

On a fixed device, the Television 3.0 service must adapt to the same usage context as a traditional set-top box, for instance, by supporting a front-panel interface. However, the Television 3.0 service might also support the geospatial feedback that it acquires from the mobile hand-held device.

All Television 3.0 applications whether fixed or otherwise will respond to the same network control messages including subscriber entitlement or service feature changes. The service and content protection function of the Common Service Framework (historically referred to as Digital Rights Management or DRM) must constantly validate the usage context of the media consuming application, ensuring that the content distributor's commercial agreements are respected and that content will not be used for any purposes other than the intended ones.

Within such a complex operating environment, prioritization of usage context becomes critical. For example, in a telephony enabled device, such as a smartphone, the application framework may need to determine for each event whether the television application or the phone application takes precedence. The same Television 3.0 services running on a smart TV, smartphone and tablet will react differently to each type of event, and may in fact be programmed by the user to react differently.

Though service providers have the option of developing a different set of Television 3.0 applications, each optimized for the specific CE manufacturer's application framework and unique usage context, this will inevitably be seen as a costly and infinitely expanding endeavor, as all of these applications will need to be supported and maintained indefinitely.

Alternatively, Television 3.0 service providers will draw upon the experience of Web 2.0 service providers by abstracting their services through the use of the Common Service Framework.

The Television 3.0 service provider will balance the goal of a complete abstraction layer that minimizes device specific development, with the desire to leverage unique device specific capabilities (e.g. larger or higher resolution screens, better memory management, unique man-machine interfaces, hardware security hooks or greater portability).

The HTML5 standard, currently in development, is expected to facilitate the deployment of a common set of applications and services across compatible devices. HTML5 includes richer graphical capabilities and more complex JavaScript application logic. As device specific application frameworks standardize on the JavaScript standard interfaces, Television 3.0 service providers will deploy more features in a common fashion, and thereby reduce their dependency on device specific or downloadable application frameworks.

As an example, HTML5 utilizes JavaScript to abstract the usage context of the local device. Through the use of JavaScript to access the device's local storage, effectively extending the virtual CDN network into the device, Television 3.0 application developers will be able to cache service information that improves the predictability of the user experience, potentially approaching the reliability that consumers experienced in Television 1.0 applications (at least within managed networks).

Built-in DRM functions may already exist in many CE device specific application frameworks. However, in order to achieve a common set of security functions across every device type, the service provider will inevitably require a global set of content and service protection functions to be included within the Common Service Framework. These security functions may leverage device specific security capabilities or usage context, but must never be completely dependent upon them.

For example, to ensure that there exists a trust hierarchy, the Common Service Framework might leverage any hardware based personalization or security functions that are exposed by the device manufacturer (for example Unique Device IDs). Alternatively, DRM clients may be integrated with the device's application framework, such that a DRM application may be downloaded to provide dynamic security hooks that may be leveraged by the trust functions of the Common Service Framework.

Context Control

For a service provider who is designing a Television 3.0 service across fixed and mobile devices using a Common Service Framework, a key decision is whether that service should take advantage of network specific features or whether it should be agnostic to the underlying transport technologies.

If for example a cable operator deploys a network agnostic (Television 2.0) application on a tablet alongside a traditional set-top box application (Television 1.0) over the same access network, then the network resources required to deliver the same quality of service to both devices may end up being twice the resources that would have been required if both devices had implemented a common Television 3.0 application that utilized a network aware Common Service Framework.

Most broadband service providers have already adapted to the demands of Television 2.0 service providers by scaling their access networks to enable ABR to coexist alongside legacy analog and digital cable television services. To avoid the overhead cost of simulcasting ABR video over the same access network that already delivers similar content in a traditional digital video format, the network operator must allow the Television 3.0 Common Service Framework to interface with the legacy network controller systems.

To permit a Television 3.0 service to interoperate with the legacy digital cable plant requires a complex interoperability design that adapts existing Television 1.0 infrastructure to the Common Service Framework. This includes the ability to leverage legacy System Information (SI) and Conditional Access (CA) services already being transported alongside the broadcast digital video service.

The access network operators and service providers must agree on a context control interface to communicate the availability of hybrid or legacy services, and to enable the Television 3.0 service provider to control the access network transcode or transcrypt resources that would be required to convert content to the required consumption format. This includes transferring the service protection metadata of the legacy conditional access system to the Common Service Framework for use by the Television 3.0 applications.

The advantage of deploying a network aware Television 3.0 service is that it can be made to be more scalable by reducing demands on the access network resources. The disadvantage is that interoperability costs may be greater for the network aware Television 3.0 service (see Figure 2).

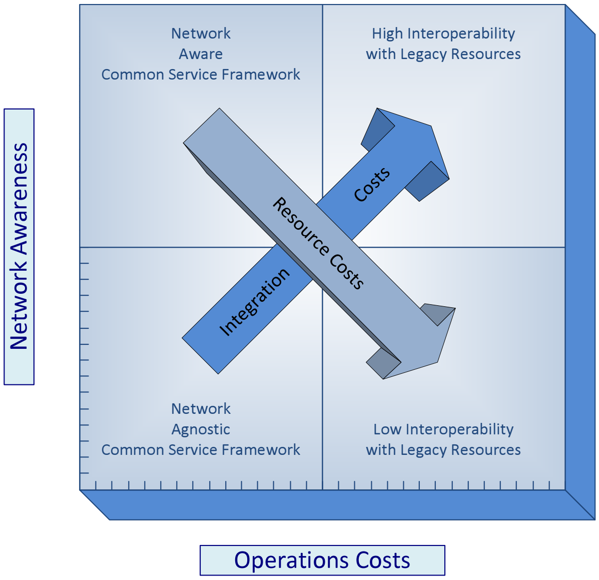


Figure 2: Cost of Network Awareness

Including legacy network awareness into the Television 3.0 Common Service Framework should be built-in from the start as it will become ever more difficult to retrofit fielded applications. Legacy network awareness need not be deployed into the field on day one. These features may exist in the

Common Service Framework, but may be turned off in the first deployments, thereby reducing and/or delaying the legacy interoperability costs.

Inevitably a hybrid gateway device will be required to support Television 3.0 service interoperability with the legacy television services. New Internet TV compatible transport standards such as MPEG DASH may never be feasibly deployed on all existing fielded hardware. Until a service provider forces every subscriber to replace their incompatible legacy equipment, a hybrid device will be useful (and cost-effective) in translating between the legacy transport and the new Internet standard transports.

A hybrid gateway device might be installed within the home or outside of it, but regardless of where it is situated it will serve the same function. The hybrid gateway translates between the legacy service and transport controls, for instance by leveraging MPEG2 transport streams (TS), system information (SI) and conditional access (CA) information to acquire and terminate the legacy broadcast service, and then transcoding, transcrypting and translating these services into the Television 3.0 service exposed to the newer multi-screen applications.

Note: A more detailed discussion of media gateway termination technology is available at: Architecting the Media Gateway for the Cable Home

Context Aware Services

As mentioned above, an application may require that the Television 3.0 service will distinguish between conflicting user priorities based on context.

As an example, when viewing television on a mobile phone, the user may choose to pause their viewing in order to answer the phone call. In a pure streaming model, the application would determine availability of pausing by validating whether a “catchup” version of the program is available for bookmarking and later streaming from the paused location.

But if the same mobile application happens to be situated within the user’s home network then a DVR capable application might already be recording a legacy broadcast version of the program to a local storage device, altogether eliminating the need for additional network “catchup” streaming resources. Further, if the Television 3.0 Common Service Framework were capable (e.g. DRM content controls permitted), then the application might be able to stage the content for offline as well as online viewing, enabling the DVR recorded content to be viewed in a park or on a plane.

If the subscriber exposes local storage in the home for the purposes of viewing television content, then the same Common Service Framework (in a managed network) might use Progressive Download techniques (PDL) to persist as many formats of the content in the home as are required by that home’s devices, avoiding the future need for in-home devices to go back to the network for viewing. The same PDL technique may be used to pre-position personalized advertising content.

Another example of a context-aware service framework is the ability to limit or expose service provider resources based on an application’s user privileges. For instance, limiting the quality of the television content might be dependent on a user’s data quota. The service provider might for instance permit the user to limit their household’s access to HD quality content when a certain threshold of usage is met every month, or for a specific household device or user.

Such authorizations may in fact be federated in the Television 3.0 model. For instance, additional personalized metadata about a specific television event might be available to a subscriber only if they subscribe to multiple service providers. As an example, the Common Service Framework may permit the CDN to be managed by one service provider, but the content discovery may be managed by different ones, only sharing a common content identifier. For instance, subscribers to Common Sense media or Rotten Tomatoes might have additional descriptive information about the current movie that the user is accessing from their cable subscription.

To deploy a Television 3.0 Common Service Framework capable of unlimited TV viewing anywhere, as has been described in this paper, the Common Service Framework must be capable of implementing a very robust contextual control interface over the content as well as the content delivery network, whether connected to a network or offline. The context control interface logic itself must be cacheable along with any associated context control metadata including related content and service information required for discovery, protection and transport of the content, so as to permit offline as well as online viewing. The ABR manifest in DASH may be used to implement such a context control interface. The XML manifest file standardized by DASH may be accessed from a stream or from a cache. The DASH manifest file may be adapted and/or extended on the fly by any intermediate control point or it may be kept in its original pristine form, untouched as any associated content is transported through the content distribution network.

For instance a service provider might choose to regionalize, localize and/or personalize the original broadcast manifest file as well as any associated sidecar files. Sidecar files may be used to extend the original manifest or index file, for instance by describing network specific abstractions (and might be required in cases where the original manifest file is write-protected). Examples of content personalization include frame accurate insertion of an overlay graphic, an alternate video replacement, or any other form of advanced advertisement, as well as the inclusion of a user or a group’s specific bookmarks.

At each point that the manifest file is transferred from one sub network to another over the entire content distribution network, the manifest or associated sidecar files will be subject to controlled manipulation as required by the context specific needs of the service provider or the end user.

As an example, in the case that a service provider is leveraging a local storage device to permit offline and online viewing of their controlled content, the service provider might pre-stage personalized content or metadata to be used in place of the broadcast content or at other pre-defined interstitial points. Such a Television 3.0 application would not only allow for the display of personalized advertisements, but would indeed allow for any type of personalized content – the same movie might be available in a specific subset of the five parental advisory formats for each user in the household.

Through these ABR synchronization techniques, the Television 3.0 application may be able to access new types of contextual metadata, for instance enabling users to skip through episodes in a series – or articles in a video news journal. The user might even personalize their application to prioritize content based on their location (for instance emphasizing movies shot in Paris over British comedies and then vice-versa as location changes).

Additional Television 3.0 contextual services will be made possible by the development of a robust context control interface. For instance, content discovery and recommendation searches may be persisted and prioritized by users to enable automatic organization of future programs (or versions of programs) in a much more personalized fashion.

The Television 3.0 content will adapt to the use of contextual metadata. Television series will include metadata to allow viewers to automatically catch-up to favorite plots. Movie directors will include metadata to allow viewers to experience their stories differently, depending on personal desires (e.g. family-friendly fare, racy endings, and mood-sensitive plot lines).

Consumers most likely will agree to reductions in privacy in return for more personalized Television 3.0 content – that will be delivered along with more personalized advertisements (which may actually be of interest to consumers). Contextual control of playback will assure advertisers that the consumers have actually viewed their information, and will enable consumers more instant gratification (e.g. immediate purchase of the actress's dress).

The concept of DVR scheduled and recorded content will evaporate over time as all content will be available at any time and in any place. Instead of recorded content, users instead will refer to "My Content Library" in order to distinguish between personally interesting content and everything else. DVR schedulers will evolve into personal recommendation and content discovery tools. All content the user ever viewed will be available to them, but only the recent content most likely to be viewed next will be displayed within personal recommendations list.

Network Aware Services

A few years from now, the content delivery networks of today will be considered as outdated as the Web 1.0 applications of a decade ago.

A service aware contextual gateway application might be deployed at every sub-network interface point on the content delivery network. The content delivery network control application itself might be virtualized and contextualized in the same fashion as the television applications described above. Today network switches and routers are fundamentally constrained by the Open Systems Interconnection (OSI) model to a very limited visibility of application needs within a specific OSI level.

Using the same kind of abstraction model described above for television content distribution, IP packet distribution can be equally freed from the constraints of the existing network control models.

New forms of "network aware services" will be enabled to adapt more easily to the physical constraints of the underlying sub-network. An example of this is a residential gateway that dynamically routes consecutive video packets across both a home wireless and home wire-line network (e.g. MoCA, 80211.AA) depending on temporal noise characteristics and error correction on each physical transport medium.

The poster art representing a movie to be displayed in a television application might be dynamically adjusted not only by the device screen size, but by the capacity of the underlying IP network on which it was transferred to the device.

As each network gateway application is empowered to make service aware optimization decisions, the network controller will coordinate and mediate conflicting needs of applications, service providers and access network operators.

CONCLUSION

In a world where every consumer desires to have their subscription services on any device at any time, service providers must learn to live with an endless variety of devices and an infinite number of services, delivered over any type of network, both offline and online. In the short-term, every service provider must be able to deliver their services in managed, unmanaged and hybrid environments equally well.

The rapid ascent of audio-video services delivered with ABR technologies and the rapid adoption of the MPEG DASH standards exemplify this trend. ABR manifest and associated content and metadata files may be cached and manipulated at every point in the content delivery path to assure consumers access to television services whether in the home or on the go.

Use of similar next generation caching techniques will be extended throughout service delivery platforms to assure that every operator service benefits from similar scalability paradigms, including user interfaces and collaborative communication features.

Just as Television 2.0 took advantage of techniques developed for the social web to optimize delivery of television over unmanaged networks, Television 3.0 applications will adapt those techniques to the needs of network operators who require consistent managed and branded television service to be delivered to any subscriber at any time and any place.

The Economics of IP video in a CCAP World

John Ulm & Gerry White
Motorola Mobility

Abstract

The paper outlines an IP video architecture and determines the relative cost contributions from the major components. For current equipment, the DOCSIS® downstream channel is shown to be the major contribution to infrastructure cost. As next generation Converged Cable Access Platform (CCAP) systems are deployed this will fall enabling a cost effective IP video platform to be realized. At this point other cost contributors become more significant. CDN and nDVR trade off options are discussed. Finally the paper looks at spectrum migration options to release the bandwidth needed to deliver IP video service.

INTRODUCTION

Delivery of IP video will be a major factor driving cable infrastructure during the next few years. Studies of Internet traffic patterns [SAND], [VNI] show that video has become the dominant traffic element in the Internet consuming 50 to 60% of downstream bandwidth. Cable's "Over The Top" (OTT) competitors account for much of this traffic, with Netflix alone constituting almost 33% of peak hour downstream traffic in North America.

To remain competitive cable operators need to deliver IP video to the rapidly expanding tablet, PC, smart-phone and gaming device market. They must leverage the same cost effective technologies as OTT competitors for this and for video delivery to the primary TV. Thus service providers must deliver two forms of IP video: unmanaged OTT off-net and managed video services on-net. This has caused the industry to become very focused on the implications of offering IP video over a DOCSIS® (Data Over Cable Service Interface Specification) channel.

Over the years, the relatively high cost of a DOCSIS channel has impacted potential solutions for IP video. In the past this spawned multiple alternate proposals. Bypass architectures such as DOCSIS IPTV Bypass Architecture [DIBA] were proposed as alternatives and bandwidth saving mechanisms such as multicast and variable bit rate (VBR) technologies investigated. These have become somewhat redundant with the recent surge of adaptive bit rate (ABR) protocols among consumer devices. This unicast delivery mechanism based on HTTP has become the defacto standard for IP video services to this class of devices. In fact, ABR may be used for all IP video traffic including primary screen [CS_2012].

Thus a critical question for operators is how to deliver unicast based IP video cost effectively. It is important to understand the cost implications for DOCSIS downstream channels in the future.

IP VIDEO ARCHITECTURE

To understand the economic impact of migrating to IP video, the system must be separated into key elements. Components of a Managed IP video Architecture are detailed in [CS_2012] and [Ulm_NCTA_2012].

Figure 1 shows a high level abstraction of an end to end functional architecture for delivering IP video from content providers to content consumers. The video service provider must ingest content from multiple content providers, process it appropriately and then transport it over multiple types of access networks to the destination consumer devices.

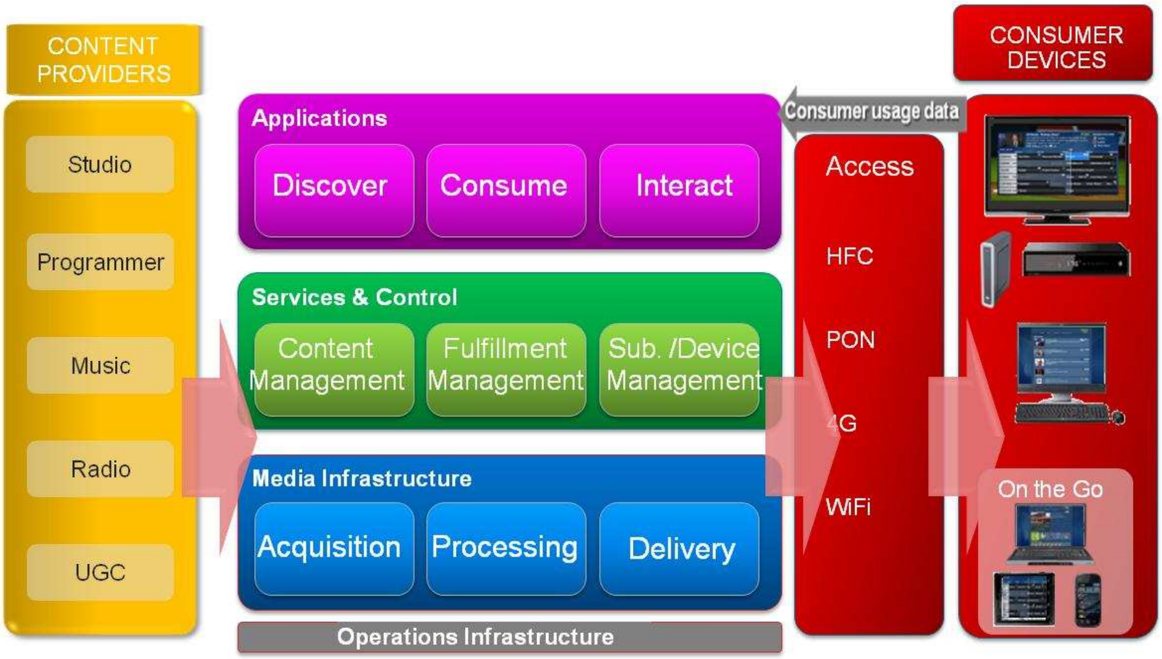


Figure 1 IP video Functional Model

This functional model is used to develop a high level breakdown of the costs for IP video delivery and to compare the relative contribution of each component. This will enable operators to understand the impact of the major cost drivers and make intelligent system trade-offs in their IP video architecture.

MAJOR COMPONENTS AND COST IMPLICATIONS

Content Providers

The number of content sources is increasing. Traditional streamed linear television broadcasts from studios and programmers may be received over satellite or terrestrial links in MPEG-2 and MPEG-4 formats. User-generated content and other Web based multimedia sources must also be supported, but will more typically be delivered as file-based assets.

Costs associated with ingesting content from content providers scale based on the number of program sources ingested and the cost of the material. Once purchased and ingested, these programs are shared across all subscribers. The cost per subscriber is not materially impacted by changes within the delivery infrastructure and thus these costs are not considered further in this paper.

Consumer Devices

One of the principal drivers towards a service provider IP video infrastructure is to be able to support generic IP-based consumer devices such as smart-phones, tablets and gaming devices. The range of consumer devices appears to be almost limitless in terms of screen sizes and resolution, network data rates, processing power, mobility, media format support and DRM support.

Most of these consumer devices are owned directly by the consumer. The one exception to consumer-owned devices might be the IP set-top box. For this analysis, it is assumed the operator will have some leasing revenue associated with the IP STB so it is not considered as part of the infrastructure costs. There are some cost tradeoffs in the use of home gateways and hybrid video gateways which will be considered later in the paper.

Access Networks

A primary reason to move to an IP video infrastructure is that it can be access network independent in contrast to existing MPEG/RF video infrastructure. For the purpose of this investigation only the hybrid fiber coaxial (HFC) access network will be considered.

Core IP Network

The components of the IP video architecture interconnect via the same generic IP core network used for all video and high speed data service delivery. The costs of the core network are amortized over multiple services and all subscribers. Thus the cost contribution to IP video service on a per subscriber basis is relatively low.

Application Layer

The Application layer provides interaction with the end user and is largely responsible for the user experience. It includes functions that: 1) discover content through multiple navigation options such as user interfaces (UI), channel guides, interactive search, recommendation engines and social networking links; 2) consume content by providing applications for video streaming, video on demand (VOD) and network DVR (nDVR) consumption; and 3) provide companion applications which enable user interaction in conjunction with media programs such as interactive chat sessions.

Applications are typically implemented in software running on servers in the data center with a thin client application on the consumer device. The applications may be provided by the service provider, the device provider or a third party. Costs associated with the applications layer are thus shared between these entities. On a per subscriber basis these are relatively small as they are amortized over a large number of subscribers.

Services & Control Layer

The Services & Control Layer is responsible for assigning resources within the network and for enforcing rules on content consumption that ensure compliance from a legal or contractual perspective. It includes functions that manage: content work flow from ingest through to delivery; the resources needed to ensure content is delivered to users when requested; and subscribers and devices to ensure that content is delivered to authorized consumers in the required format.

The Services & Control Layer is implemented as a set of software applications running on servers in the service provider network. These applications are typically licensed on a per subscriber or per session basis. Thus costs are a combination of hardware platform and software licensing. The basic control components required include: workflow and session management, DRM control, and resource management.

Media Infrastructure Layer

The Media Infrastructure Layer is responsible for video content delivery from the content provider to the consuming devices over the access network. This includes acquiring content from satellite or terrestrial sources (as either program streams or files), encoding it for ingest into the system and processing the content to prepare it for delivery. This is where the heavy video processing occurs and functions such as trans-coding, multiplexing, advertising insertion, encryption and publishing to a content delivery network (CDN) are found. This layer must also deliver the content to the target device through mechanisms such as web servers, CDNs, and streaming servers.

Costs for content reception and encoding scale on a per content stream basis. The content is shared across many subscribers so that the per subscriber cost is low. Packaging costs may scale based on content streams for a pre processing model or on subscribers if a just-in-time model is used. The choice between these is based upon a trade-off between packaging, storage and transport costs [PACK]. CDN costs do scale on a per subscriber basis.

RELATIVE END TO END INFRASTRUCTURE COST

The relative end to end cost of delivering IP video to a subscriber includes contributions from all of the components mentioned above and each component can have a wide range of variability. The Application Layer and Services & Control Layer products tend to be software on standard server platforms in a data center where costs are shared over a very large number of subscribers. The Media Infrastructure Layer is the component that contains the specialized hardware products and is where most of the operator investment occurs. Rather than attempt a detailed investigation of all of these components, the focus of the paper is on how changes in the network access pieces of the Media Infrastructure Layer impact the cost model.

Cable modem termination system (CMTS) ports to date have been deployed to provide high-speed data (HSD) and voice services. CMTS costs on a per subscriber basis have been relatively low. This cost point has been possible because HSD services could be heavily over-subscribed. IP video has a very different service model and cannot be over-subscribed to the same extent. A single CMTS channel can support anywhere from a half dozen high definition (HD) to a couple dozen smaller active video streams depending on the encoding rate used. The ratio of high definition to standard definition content now becomes very important. At historical CMTS pricing points, this translates to an order of magnitude of $100’s per IP video stream.

Each of the components above is highly configurable which can result in wide variations in the end to end cost analysis. To understand the relative costs of these components required a nominal use case based on data from actual products and bid responses. This is shown in Figure 2 on a cost per active video user basis. For this example, the CMTS cost is roughly ten times the costs ascribed to the other major components. It is clearly the most significant cost driver for IP video and will be the primary focus of the rest of the paper.

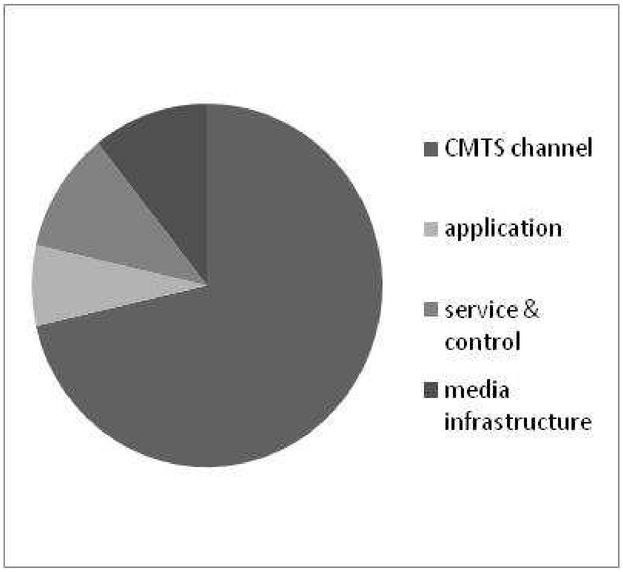


Figure 2 Per video user cost contributors

IP VIDEO & DOCSIS CHANNEL COSTS

CMTS Costs – Historical Perspective

DOCSIS is now 15 years old, having first been established in March 1997. Over that time, it has continued to evolve. In the early days, the cost per downstream channel was above $10,000. Early implementations had fixed downstream to upstream ratios (e.g. 2x8), so if more downstream bandwidth was needed, the system was burdened with the cost of more upstreams whether or not they were needed.

In addition to the fixed ratio, these early CMTS’s were focused on offering a robust voice service for the operators. This introduced significant costs as these CMTS became carrier grade incurring the associated redundancy overheads.

Thanks to Moore's Law, these costs were reduced over time. Two architectural changes accelerated this trend. First, the DOCSIS 3.0 specification (D3.0) was developed and released. This laid the groundwork to enable multiple bonded channels per downstream port. At the same time, CMTS architectures shifted to decoupled architectures where upstream and downstreams could scale independently of each other. Some vendors chose a modular CMTS (M-CMTS) path for this while others implemented decoupled architectures within their Integrated CMTS (I-CMTS). As D3.0 was deployed, this helped to accelerate the reduction in cost per downstream channel as multiple channels were now implemented per port and the upstream burden per downstream channel was reduced.

So where are we today? Based upon recent research from Infonetics (Q4 CY2011), the revenue per downstream (channel) will decline in calendar year (CY) 2012 to approximately $1,600. After several years of significant reductions following the introduction of D3.0, the industry is starting to see price declines level out. Infonetics has forecasted that CY12 will see a 10% drop over CY11 which is substantially less than the previous two years.

As we move forward with unicast based IP video, it is very important to understand the cost implications for DOCSIS downstream channels going forward.

CCAP Disrupts DOCSIS Density & Pricing

Recent industry and CableLabs® efforts have defined a new specification called CCAP that is a high density combination of CMTS and edge QAM (EQAM) in a single unit. Current CMTS products may only support 4 or 8 channels per downstream port. The initial version of CCAP is defined to support 64 narrowcast channels per port, with a flexible channel mix between DOCSIS and EQAM. Future CCAP products may support 128 or even 160 channels per port, enough to fill the entire 1GHz downstream spectrum. Clearly, CCAP causes a disruptive shift in downstream densities, increasing by a factor of sixteen! With these densities, there will be a corresponding decrease in the cost per downstream channel. For IP video deployment, it is very important to understand how CCAP will affect access network costs.

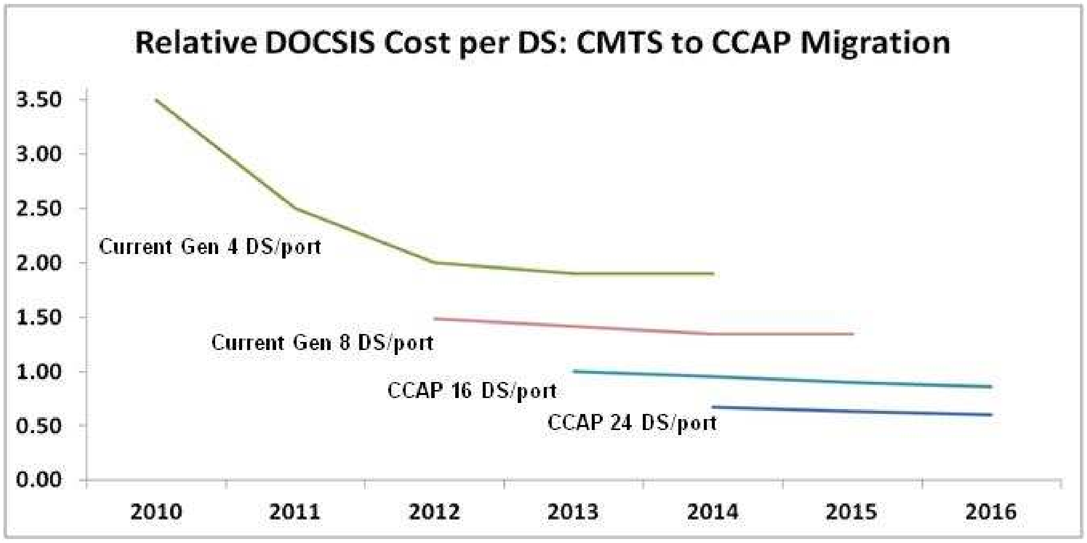


Figure 3 DOCSIS Downstream Cost

Initially, operators will only need a fraction of the CCAP capacity. Even if they wanted to deploy more channels, the spectrum required is a very scarce resource. For an operator to buy the full CCAP capabilities but only use a fraction of its capacity (e.g. 16 downstream channels) would cause a significant spike in the cost of downstream channels. CCAP would not be cost effective compared to current CMTS platforms. Therefore, vendors will need to license channels, similar to what is done today for high-density EQAM products. This allows CCAP products to be deployed while offering competitive downstream channel costs; vendors then defer revenue to a later time once additional channels are licensed and operators gain the benefit of deploying systems with longer lifetimes. Figure 3 above depicts the downstream channel cost trends over time for current CMTS with 4 and 8 downstream channels per port; then speculates where CCAP with 16 and 24 downstream channels per port might be positioned relative to current CMTS pricing.

To further explore this, Motorola developed an economic model for CCAP deployments around licensing algorithms. As discussed previously, a model where the full CCAP costs are paid up front will be difficult to justify on a cost per channel basis. On the other extreme, selling CCAP channels at the average price per channel based on a fully deployed product is also problematic. The system must be designed to support the full working load. If only a small number of channels are licensed to start, then vendors will lose money on initial deployments with no guarantees of future revenue. This would inhibit product development.

The ideal model required a licensing algorithm that would reflect the expected channel deployment. As referenced in [Howald], downstream capacity can be expected to continue at the 40-60% annual rate. Based on this along with an assumed starting point of 16 downstream channels per port, Table 1 shows how the downstream channel deployment is modeled.

Year	Total Downstreams	Incremental Downstreams
2013	16	-
2014	24	+8
2015	32	+8
2016	48	+16
2017	64	+16
2018	96	+32
2019	128	+32

Table 1 Downstream Growth

Note that this is reasonably close to the 50% growth per year that is often quoted.

Another factor that must be taken into consideration is that operators have a limited budget to spend in a given year. Infonetics forecasts show that CMTS revenue is only expected to grow 5% annually over the coming years while overall capacity above is growing at 50%. This implies that the CCAP downstream cost per channel must drop year over year (YOY) as larger number of channels are introduced in later years.

The results from our economic model are shown in Figure 4 and Figure 5 below. The baseline was 16 downstreams (DS) per port for the initial year and the average cost per downstream channel is shown for the sequence described in Table 1. Figure 4 shows the ratio with 16 DS being the 1.0 baseline. Figure 5 is interesting in that it plots the same data with a log scale. Even though Figure 4 shows each sequence getting progressively closer together, Figure 5 highlights that there is a roughly fixed percentage decrease YOY.

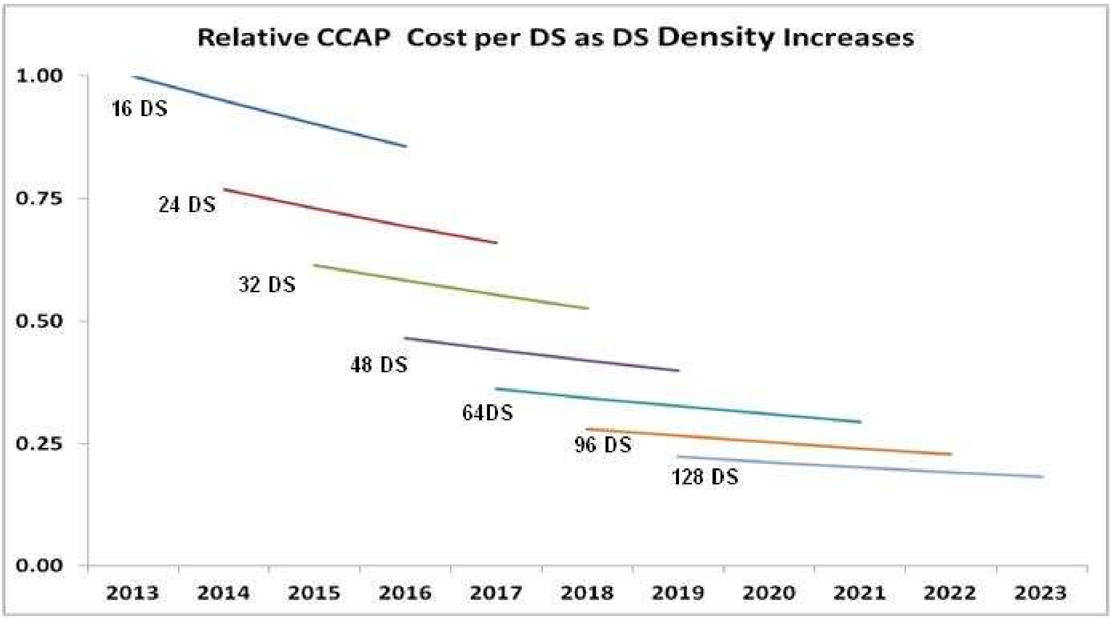


Figure 4 Cost Per DS at Higher Density

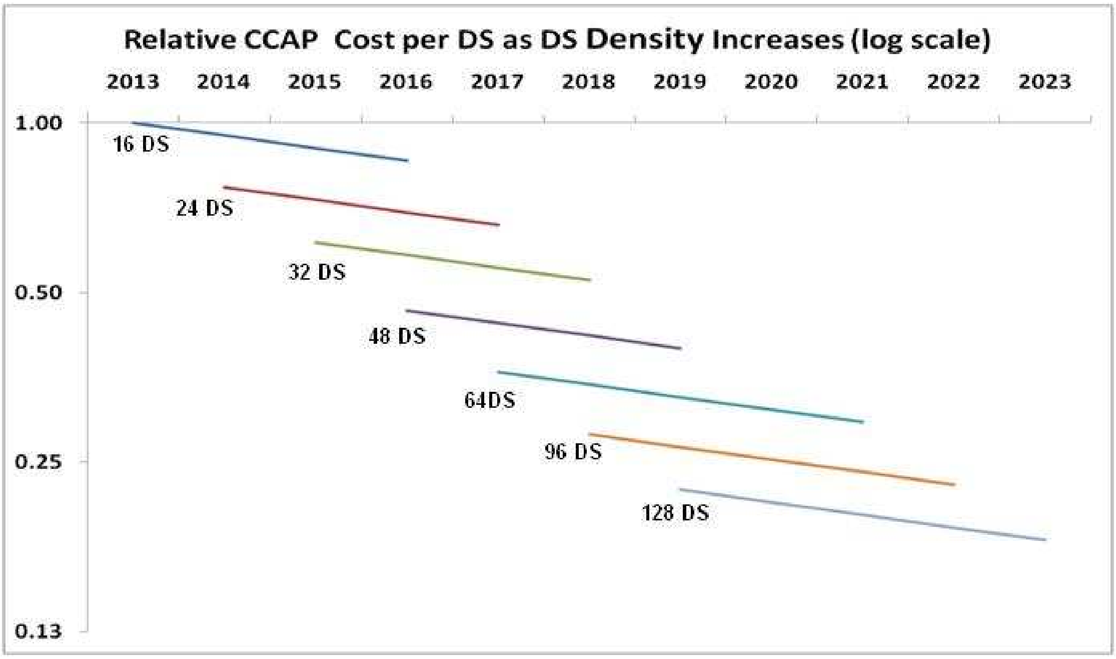


Figure 5 Cost Per DS at Higher density (Log Scale)

A licensing model like this is beneficial to both customers and vendors, assuming the initial starting point of 16 downstreams is sufficient to the vendor for initial installation and the YOY decrease in costs per downstream channel is sufficient to enable the operator to incrementally add channels in ever larger amounts within their budget.

Disclaimer: the above analysis is hypothetical and not based on any real products. It shows some possibilities for licensing algorithms that may be beneficial to vendors and customers. Every vendor may implement their own licensing algorithm and market conditions may cause these licensing algorithms to change over time.

As seen in Figure 4 and Figure 5, the economics around IP video deployment will vary over time. Costs will be higher initially but volumes will be lower. As IP video penetration ramps up, DOCSIS channel costs start to drop substantially.

Another important aspect is that IP video deployment is an incremental addition onto an existing DOCSIS HSD infrastructure. Therefore, it is critical to understand the incremental costs for downstream channels, not just the average costs which were previously discussed. This can be best explained by an example. Let’s start with 16 downstreams as a baseline cost. Now suppose once there are 32 downstreams, the average cost per channel is 75% of the baseline cost per channel. In reality, the first 16 channels cost 100% and the incrementally added 16 channels were just 50% of baseline, giving a weighted average of 75%. So the incremental cost of 50% is the number that should be used for IP video economic analysis.

Taking the analysis further, CCAP leverages high-density EQAM technology. In the extreme, the incremental addition of a downstream channel could approach that of a high density EQAM product. Infonetics research shows that the average QAM cost was $163 in CY11 and forecasts that it will drop to $86 by CY16. Note that these are the average cost per QAM.

From our previous analysis, the incremental cost per channel could be substantially less. So it would not be a reach to suggest that the incremental cost per QAM several years from now may reach $40 per channel. This is an interesting number as the industry will approach $1 per Mbps for downstream bandwidth.

Working with this number for IP video economics, an IP video HD stream @ 5Mbps would therefore cost $5 to transport. Note that a few years ago this may have been $200-$400 using older CMTS downstreams. This radically changes the IP video economics. An updated chart with relative infrastructure costs is shown in Figure 6 below and shows the DOCSIS component has fallen from being the major cost contributor to become comparable to the other elements in the total cost. At this point other components become just as significant to the overall cost model.

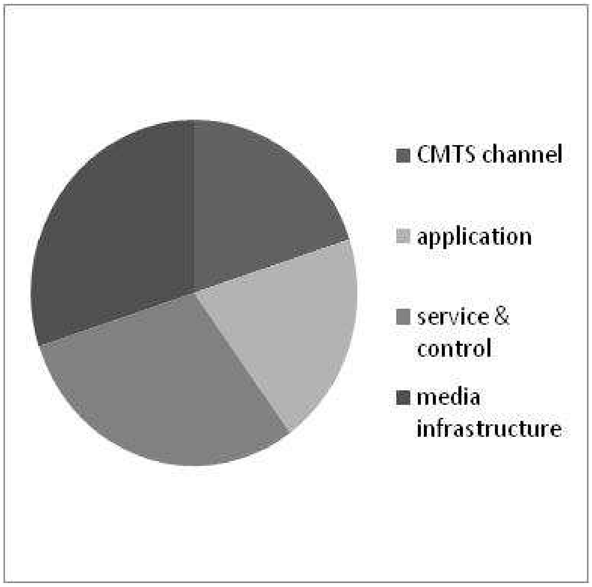


Figure 6 Post CCAP Cost Contributors

OTHER COST CONSIDERATIONS

CDN Options

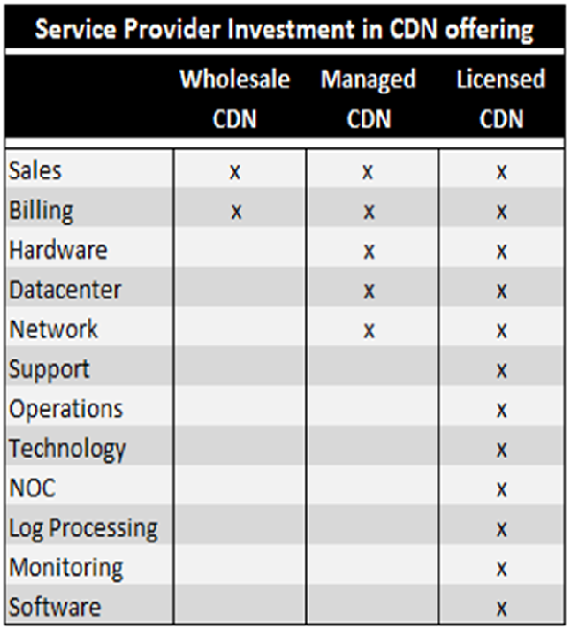
As operators migrate to IP video services using ABR, they will be able to leverage internet CDN technology for video delivery. There are a wide range of options to achieve this with a corresponding range of costs.

Initially many operators may purchase CDN services from one of the worldwide CDN providers. Eventually an operator may enter into a wholesale relationship with that CDN provider in order to resell CDN capacity directly to content providers and web site servers. This may allow the operator to extend their brand to the CDN services as well.

A possible next step in the CDN progression would be to install a managed CDN. In this step CDN nodes are added inside the service provider network but are still managed by the CDN provider. This allows the service provider to deliver content internally on their own nodes and network while still leveraging global access through the CDN partner company. The service provider minimizes operational expenses (OPEX) since the CDN partner still manages the internal CDN.

Finally, the service provider can install a licensed CDN. Equipment and software are deployed on the service provider's network and the provider assumes responsibility for operations and support. At this stage, the service provider can participate in a federated CDN exchange with other CDNs to deliver content outside their own CDN.

Table 2 shows the various functions associated with each of the three approaches. From a cost perspective, the wholesale approach requires the least amount of up-front investment but it is also the most expensive on a per-bit-delivered basis. Each step then requires more investment from both a capital expenditure (CAPEX) and OPEX perspective, but continues to result in lower costs for delivering each bit of content.



Service Provider Investment in CDN offering	Wholesale CDN	Managed CDN	Licensed CDN
Sales	x	x	x
Billing	x	x	x
Hardware		x	x
Datacenter		x	x
Network		x	x
Support			x
Operations			x
Technology			x
NOC			x
Log Processing			x
Monitoring			x
Software			x

Table 2 Service Provider CDN Options

Transcoder and Storage Trade-offs

For linear television service, there is traditionally no storage costs associated with it. The content is encoded/transcoded, prepared and delivered to the consumer. With the new world of IP devices arriving, operators will want to go beyond simple linear television service to these devices and offer the ability to time shift. Consumers have become accustomed to their DVR for the television screen and will demand the same service for their IP devices. This will create a need for network based or "cloud" DVR services (nDVR).

Some current legal rulings based on existing content contracts require that nDVR content have a unique copy for each subscriber that records it. Other services offered today with re-negotiated content agreements allow single copy storage provided the fast-forward feature is disabled. The relative cost impact of nDVR is affected dramatically by the ratio between these.

Multi-rate ABR also exasperates the problem since a unique copy of a piece of content must now be stored in multiple bit rate formats. An example of this cost impact is shown in Figure 7.

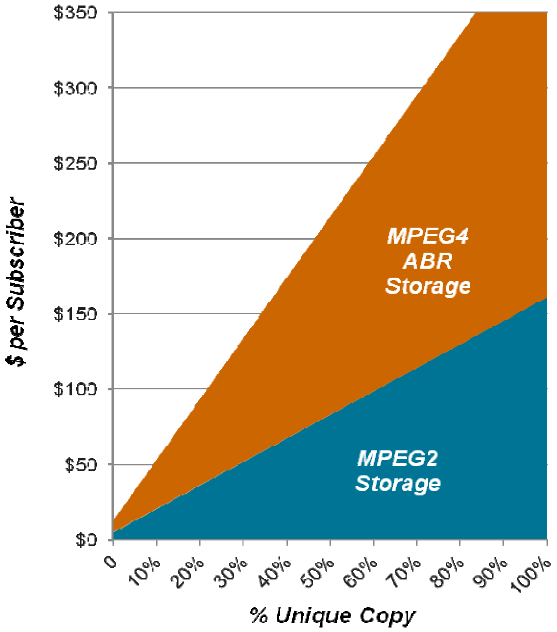


Figure 7 Storage Costs – nDVR

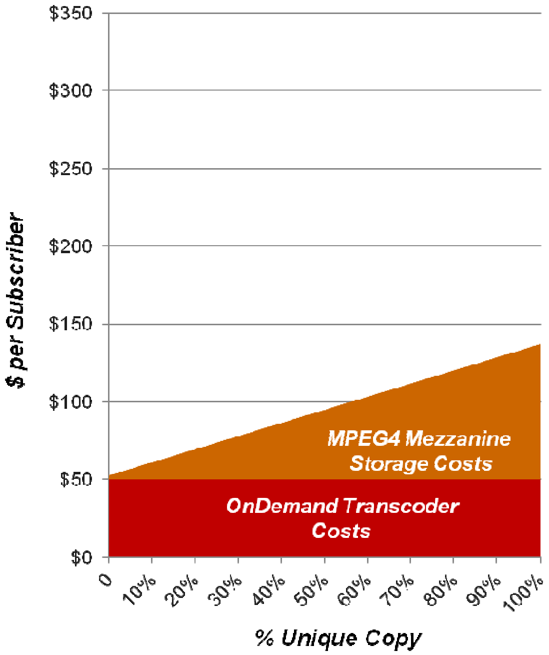


Figure 8 Transcoder vs. Storage

An alternative approach is to store a limited number of mezzanine formats in the nDVR storage and then transcode the content to the appropriate ABR bit rate on the fly when it is being viewed. Figure 8 shows an example of how costs may be impacted.

This creates a tradeoff between storage costs and transcoders costs that is constantly shifting. Many factors go into this analysis and the on-demand transcoder costs can vary significantly. This is an area where a disruptive change in transcoder costs could significantly change the landscape.

SPECTRUM MIGRATION STRATEGIES

Another very important aspect to IP video migration is finding sufficient spectrum. Some operators have already made more spectrum available by recovering analog TV channels using digital TV terminal adapters (DTA) while other operators have upgraded their HFC to 1GHz or used switched digital video (SDV). This available spectrum is being gobbled up today as more HD content is deployed, VOD requirements continue to increase and HSD services continue to grow at 50% annual rates. So there may still be a need for additional spectrum to ramp up IP video services with a corresponding economic impact.

Early Transition Plans

One way to significantly reduce spectrum requirements is to convert legacy MPEG-2 linear TV to IP video in a home gateway device. To support ABR devices in the home requires a transcoder in the home gateway device. Simple stand-alone devices are available today that accomplish this. This is an excellent approach for early deployments as it has almost no impact on infrastructure costs for rolling out linear IP video services. It also requires no new spectrum as this home gateway device appears as an STB to the system.

The next step in this migration is to introduce hybrid video gateways that also incorporate transcoding technology. These perform the same IP video conversion for linear TV described above for delivery to multi-screen IP devices. The video gateway also has the advantage that it is the single point of entry for video services and allows IP STBs to be deployed elsewhere in the home behind it. These devices can also operate as IP devices and are pivotal in the transition to an all IP system. As above, it can have a minimal impact on infrastructure costs to start and allows the operator to grow its IP video infrastructure at their own rate.

A detailed discussion of the home gateway migration is given in [CS_2012].

Complete Recovery of Legacy Bandwidth

The previous discussion on home gateway migration plans helps the operator begin the IP video transition. However, the end game is to eventually get to an all-IP system. Legacy MPEG digital TV services may continue to consume 50% to 80% of the available spectrum. Regardless of which path the operator took to free up spectrum, eventually they will need to install switched digital video (SDV) to reclaim all of the legacy bandwidth.

Adding SDV to the mix also increases the need for narrowcast QAM channels. This plays well into the previous CCAP analysis in this paper. Also, as the mix between legacy and IP subscribers change, an operator will need to re-assign SDV bandwidth to IP video bandwidth. This is also well suited for CCAP. A more detailed analysis of the SDV migration is in [Ulm_NCTA_2012].

CONCLUSION

Operators must deploy unicast ABR video to remain competitive. The infrastructure costs of providing this service are currently dominated by the cost of the downstream DOCSIS channels needed. With the development and deployment of high density CCAP platforms the cost per downstream is expected to fall dramatically, enabling the operator to deploy sufficient channels to meet demand while remaining within budget.

In the early days of CCAP deployment, not all channels will be used creating a potential disconnect between the capacity of the platform and the cost per channel deployed. The paper offers a framework for licensing which should be mutually acceptable to vendors and operators to circumvent this hurdle.

With the DOCSIS channel cost reduced significantly other cost components become more significant. ABR video is conveniently and cost effectively delivered via a standard internet CDN. A range of options to implement this are available from complete outsourcing to in house each offering different trade-offs in OPEX and CAPEX.

As nDVR is deployed into the ABR infrastructure another set of trade-offs will be required. For each recorded asset the multiple bit rate versions required can either be created at record time or created at play out from a recorded mezzanine format. In this case the trade off is between storage capacity and real time transcoding costs.

Operators will need to find the downstream bandwidth required for IP video delivery. Several options are available to do this. Home gateways may be used for early deployments in parallel with legacy MPEG-2 video. As the move to all IP video progresses the amount of MPEG-2 channels will decrease so that they can be economically delivered using SDV. CCAP is well suited for this.

The operator has multiple choices to make but will be able to deploy the technology required to remain competitive in an IP video environment.

REFERENCES

[SAND]	Global Internet Phenomena Report Fall 2011; Sandvine
[VNI]	Cisco® Visual Networking Index (VNI) 2011
[DIBA]	M. Patrick, J. Joyce, "*DIBA – DOCSIS IPTV Bypass Architecture*", SCTE Conference on Emerging Technology, 2007
[CS 2012]	J. Ulm, G. White, *"Architectures & Migration Strategies for Multi-Screen IP Video Delivery"*, SCTE Canadian Summit, March 2012.
[Ulm NCTA 2012]	J. Ulm, J. Holobinko, *"Managed IP Video Service: Making the Most of Adpative Streaming"*, NCTA Technical Sessions, May 2012.
[PACK]	Unified Content Packaging Architectures for Managed Video Content Delivery, Santosh Krishnan, Weidong Mao, SCTE Cable-Tec Expo 2011
Howald 2011]	Dr. Robert Howald, "*Looking to the Future: Service Growth, HFC Capacity, and Network Migration"*, 2011 SCTE Cable-Tec Expo Capacity Management Seminar,, Atlanta, Ga, November 14, 2011
Howald 2010]	Dr. Robert Howald, "*Boundaries of Consumption for the Infinite Content World"*, 2010 SCTE Cable-Tec Expo, sponsored by the , New Orleans, LA, October 20-22, 2010
[Howald 2012]	Howald, Ulm, *"Delivering Media Mania: HFC Evolution Planning"*, SCTE Canadian Summit, March 27-28, 2012, Toronto,

ABBREVIATIONS AND ACRONYMS

CCAP	Converged Cable Access Platform
CDN	Content Delivery Network
CMTS	DOCSIS Cable Modem Termination System
COTS	Commercial Off The Shelf
CPE	Customer Premise Equipment
DOCSIS	Data over Cable Service Interface Specification
DRM	Digital Rights Management
DVR	Digital Video Recorder
EAS	Emergency Alert System
EQAM	Edge QAM device
Gbps	Gigabit per second
HFC	Hybrid Fiber Coaxial system
HSD	High Speed Data; broadband data service
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
nDVR	network (based) Digital Video Recorder
OTT	Over The Top (video)
STB	Set Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VOD	Video On-Demand

THE GROWN-UP POTENTIAL OF A TEENAGE PHY

Dr. Robert Howald, Robert Thompson, Dr. Amarildo Vieira
Motorola Mobility

Abstract

Cable operators continue to see persistent annual increases in downstream traffic, most recently driven by aggressive growth in over-the-top video. Well-understood tools exist to manage the growth, including switched digital video (SDV), analog reclamation, service group splitting, improved encoding and transport efficiency, and RF bandwidth expansion. Beneath it all, however, the underlying RF transport approach has remained unchanged, relying on 1990's era ITU J.83 technology for PHY layer and FEC technology. Meanwhile, 15+ years of advancements in communications technology and processing power have since taken place. Many of these advances, which close the gap between Shannon theory and real-world implementation, are already being tapped in other industries. Some are now poised to enable cable to support a new generation of Gbps-class services and to mine completely the capacity of the coaxial last mile – a key element to guaranteeing an enduring HFC lifespan.

In this paper, we will present a comprehensive link analysis addressing the deployment possibilities of these communications technology advances over the HFC channel. We will focus in particular on the ability to support higher order QAM, such as 1024-QAM through 4096-QAM. We will discuss the role multi-carrier techniques (OFDM) could play and why. We will specify SNR implications to HFC, including considerations for fiber deep migration. We will describe the SNR repercussions of advanced QAM to CPE noise figure (NF), which is critical to understand as wideband, digitizing front ends replace analog STB tuners. In addition, we will dive deeper into the subtle link impairments that become potentially limiting factors as we push the boundaries of PHY technology on the cable plant. Previously less significant issues such as timing jitter and phase noise are magnified as constellations become increasingly dense. These items ultimately effect equipment requirements.

In summary, we will articulate and quantify the ability of the HFC network to support ever-increasing orders of bandwidth efficient modulation, and the impact these modern communications formats have on equipment and requirements.

INTRODUCTION

The industry is deeply engaged in long-term network planning, in recognition of the continuing growth of IP traffic and concern for the network's ability to support it. There are two key components to the problem. The first is simply determining if the infrastructure in place is physically capable of delivering on the growth, and, if so, for how long. There tends to be a consensus that the HFC architecture is capable enough, but that it has not been optimized as of today to ensure it [8, 11]. This brings us to the second part. If the answer to the first is yes, then how do initiate a transition plan from today's infrastructure to the architecture that does optimize what can be extracted from the network?

There are many spectrum and capacity management tools in the downstream. However, the operator has much less control over upstream congestion. Common to both downstream and upstream is the reliance on what is now aging, 1990's era, physical layer

(PHY) tools. In use on the downstream is the ITU J.83B Physical layer (PHY) and forward error correction (FEC) technology. The DOCSIS upstream is also QAM with a Reed-Solomon based FEC – powerful at the time but more powerful PHY techniques are available today. The result is less efficient use of cable spectrum that could be achieved with modern PHY tools.

Relying on 1990's era technology puts cable at a disadvantage. Perhaps the single most important long-term objective of the architecture transition is, in the end, to have created maximum bandwidth efficiency (and therefore maximum lifespan) cost-effectively. We will focus our attention on this one component of capacity management – more efficient use of spectrum – more bits-per-second-per-Hz (bps/Hz).

Capacity Levers

Note that theoretical capacity is a based on two variables – bandwidth (spectrum allocated in our case), and SNR. For high enough SNR, finding spectrum dominates the equation. The capacity equation can be simplified, and in so doing, it can be shown that capacity is essentially directly proportional to bandwidth, B and SNR expressed in decibels (dB):

$$C \approx [B]\,[SNR\,(dB)] / 3 \qquad (1)$$

This approximation is accurate asymptotically within 0.34% with increasing SNR. Because of the inescapable relationship of capacity to bandwidth, as cable looks to increase capacity, new spectrum is being sought after. Figure 1 is an example of a likely spectrum evolution [2], resulting in a final state of bandwidth allocations.

In the downstream, we are extending an excellent channel into an area where it will suffer more attenuation as a minimum. It will also likely have to deal with frequency response issues.

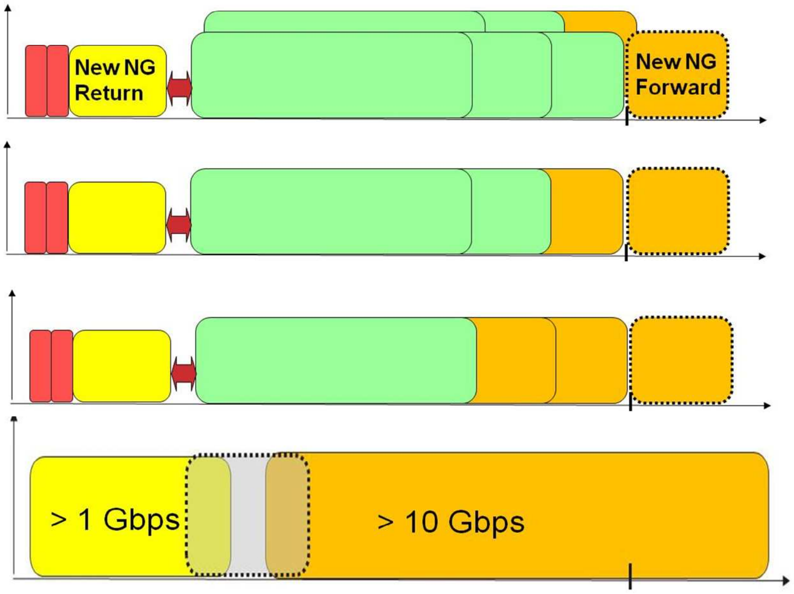


Figure 1 – Likely Cable Spectrum Evolution

In the upstream, we will be in some ways doing the opposite – extending a partially troubled channel into an area where we expect a much better behaved environment from which to extract new capacity.

We will be taking advantage of significant technology advances for enhancing the PHY. Much of what is being taken advantage of is continued advances in the real-time computing power of FPGAs and ASICs. The theoretical basis for modern PHY tools in some cases is, in fact, very old.

For example, Reed-Solomon coding itself was born in the 1959. Low Density Parity Check Codes (LDPC), the basis of today's most advanced forward error correction (FEC), is also quite old, first introduced in 1960. Information Theory is a linear algebraic discipline. However, the computing power to perform the algebraic operations required for efficient decoding of very large matrices, and using non-binary arithmetic (Reed-Solomon) came along much later.

Multi-carrier modulation (MCM, the generic name for OFDM and its variants – we will use both throughout) has a parallel history to advanced FEC in this sense. It was very difficult to implement, until the FFT version of the DFT came along, followed by computing power to calculate larger FFTs faster. IFFT/FFT algorithms form the core of the OFDM transmit and receive function. So, while we are talking about "new" technology, it is important to understand that these technologies are already very well grounded theoretically.

Of course, the single most important attribute of these advances is that they close the gap between theoretical Shannon capacity and real-world implementation – something the world has been trying to do since 1948. A simple crunching of today's HFC performance and throughput illustrates how far from this ideal we are today. Table 1 compares downstream and upstream as we use them today against the theoretical capabilities of the channel. We have accounted for code rate efficiency losses, but not framing, preamble, or other overhead unrelated to pure PHY channel transmission capacity.

Table 1 – Downstream and Upstream vs. Theory

	D/S	**U/S**
BW (MHz)	6	6.4
SNR (dB)	35	25
Capacity from (1)	70.00	53.33
Legacy QAM (no framing)		(t=10)
256-QAM	38.83	37.75
64-QAM	26.99	28.31
Delta Capacity		
256-QAM	**55%**	**71%**
64-QAM	**39%**	**53%**

As we can see in Table 1, today's commonly used modes – 256-QAM downstream and 64-QAM upstream – operate at 50-60% of capacity, and therefore are leaving a lot of bits on the table. With the help of new tools and supporting architecture evolution, some of the current limitations can be alleviated, putting cable in a position to deliver a new class of services and maximize the lifespan of its core architecture.

QAM LINK BUDGETS

Capacity Enhancements

Let's begin with the "SNR" part of (1). There are two elements of the SNR component of capacity. First, clearly, more capacity is available if higher SNR is available. Since it is related to SNR in dB, however, it is a compressing function, and its affect on capacity less effective in increasing C compared to spectrum. In practice 50% more spectrum yields 50% more SNR. This is also true for 50% more SNR in dB. However, in practice, 50% new SNR in dB, such as converting a 35 dB SNR into a 52.5 dB SNR, is not reasonable in most cases. Nonetheless, it is certainly the case that more SNR translates to more capacity, and architectures that create higher SNR are architectures that open up more potential capacity. This is why, for example, when fiber deep topologies are discussed, both average bandwidth per home (because of fewer homes per node) as well as a more robust, higher SNR channel for use are both important results. We will quantify architecture effects later in this section.

Part 2 of the SNR component of capacity is using the available SNR most efficiently. This is specifically where MCM and FEC advances come into play. A good way to understand the former is to use the “long” (but not longest!) form of (1).

$$C \approx (1/3)\sum_{\Delta f} [\Delta f]\ [P(\Delta f)\ H(\Delta f)\ /\ N(\Delta f)]_{dB} \quad (2)$$

This is the same information expressed in (1), just differently. Instead of bandwidth, we have used a summation over a set of small frequency increments, Δf. The sum of all Δf increments is the bandwidth available, B. Instead of SNR, we have identified the components of SNR – signal power (P), noise (N), and channel response (H) – each also over small frequency increments.

The capacity, then, is a summation of the individual capacities of chunks of spectrum. The purpose of (2) is to recognize that channels with changing SNR – such as any “new” bands to be exploited outside the normal cable bands – that may not have a flat response. In particular, above today’s forward band there will be roll-off with frequency. The capacity of this region can be calculated by looking at it in small chunks that approximate flat channels. More importantly, however, a technology that can actually *implement* small channel chunks can optimize each of those frequency increments to get the most capacity from them. This is the key advantage of MCM – very narrow channels, each of which can be loaded with the most bits possible. With a single, wide, transmission, it is difficult to achieve the same effect without very complex, and sometimes impractical equalization techniques and interference mitigation mechanisms. Thus, (2) effectively expresses why MCM is often better suited in channels with poor frequency response. MCM also accomplished this while overlapping these narrow channels. They are kept the independent through the orthogonality of frequency spacing – separated by the symbol rate of the sub-channels.

Figure 2 shows a capacity calculation of the forward band extension case described above. It plots the capacity including bandwidth above a 1 GHz network when the network is modeled as a lowpass roll-off, governed by the frequency response characteristics of 1 GHz taps [7]. The red curve shows the aggregate roll-off of five taps and a single coupled port, as well as interconnecting coaxial cable. An assumed 45 dB digital SNR at 1 GHz is used to calculate the capacity as signal power is attenuated above 1 GHz on a flat transmission profile.

The total capacity calculation if the entire forward band is taken into account (blue) is also shown, as well as the capacity over and above what is currently available in a 1 GHz network that is fully loaded with 256-QAM signals (pink). In both cases, the diminishing returns associated with the attenuation of current HFC passives – inherent implementation limitations, not barriers of technology – are obvious as the forward band goes above about 1.2 GHz. Analyses like Figure 2 point out why cable is bullish on the ability of the HFC network to support 10 Gbps data rates.

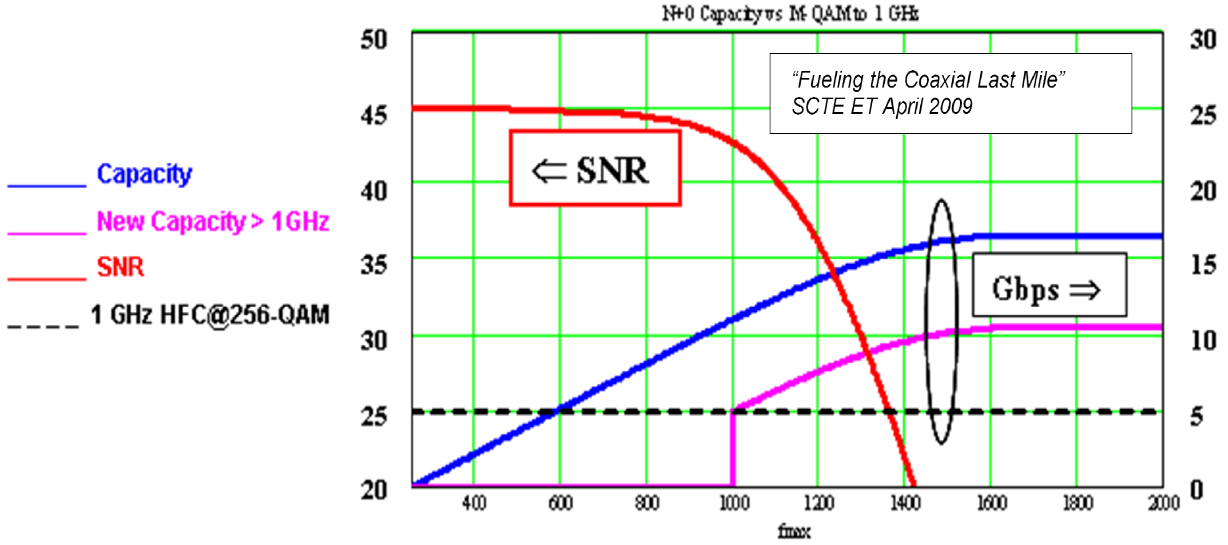


Figure 2 – Capacity Above 1 GHz on a Passive Coaxial Segment

Now let's consider the role of FEC. Using SNR most efficiently, from a coding perspective, is about finding the right codeword design. Major leaps in this capability have occurred in the past 20 years – Reed-Solomon (RS), Trellis Coded Modulation (TCM), Turbo Codes, and now LDPC. Again, the advances have mostly to do with the ability to process the complex decoding algorithms and the tools needed to design them specific to the application.

Coding theory has always been about trying to close the gap between the theory derived by Shannon and the reality on the wire or in the air. Quantifiably, this means simply getting more bps/Hz out of the same or lower SNR.

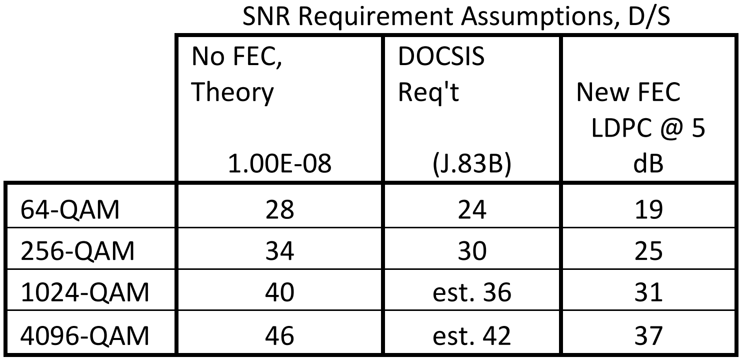
Defining SNR Thresholds

The impact on of advances in FEC is quite simple – it reduces the SNR required to achieve a particular modulation profile, increasing throughput. The SNR requirements for each QAM modulation profile without coding are theoretically well-founded. We will consider advanced FEC as we compare results of link analysis to determine what can be supported by a particular PHY and HFC architecture SNR, and to compare that to today's architecture and requirements. The thresholds that will govern the comparisons are shown in Table 2.

The three M-QAM BER columns are as follows:

1) 1e-8, No FEC
2) DOCSIS Specification (and extended estimates where QAM profile does not exist)
3) New LDPC-based FEC; assumption of 5 dB more gain

Table 2 - Downstream SNR Assumptions for M-QAM Profiles



SNR Requirement Assumptions, D/S

	No FEC, Theory 1.00E-08	DOCSIS Req't (J.83B)	New FEC LDPC @ 5 dB
64-QAM	28	24	19
256-QAM	34	30	25
1024-QAM	40	est. 36	31
4096-QAM	46	est. 42	37

Several important items must be noted with respect to Table 2.

- DOCSIS includes an allocation for implementation margin on top of an assumed coding gain impact. We are inherently carrying those implementation margins forward by using an LDPC gain factor and not an LDPC SNR versus QAM simulation.

- Coding gain may increase as M-QAM orders increase, but it is conversely more difficult to maintain a constant implementation loss across higher profiles. By using 5 dB, we are essentially calling these a wash. No effort has gone into infrastructure requirements to support, for example 4096-QAM, and hardware limitation can become exaggerated for these cases.

- There are no code designs selected using LDPC for North American cable. In Europe, DVB-C2, for example, has defined a range of code rates, and these are SNR reference points reported of the OFDM PHY + LDPC:

  256-QAM: 22-24 dB
  1024-QAM: 27-29 dB
  4096-QAM: 32-35 dB

  These numbers similarly require margin be applied in practice, but are nonetheless useful in understanding how efficient a network can be as other non-idealities of implementation are reduced.

- These are all AWGN-only SNR values, which is the fundamental construct of channel capacity.

- These are SNR thresholds for the downstream only. The first column, of course, is independent of downstream or upstream.

So, while the dB to use for a given profile can be debated, Table 2 gives us a ballpark starting point.

We similarly set thresholds to use for the upstream, which also includes margin for operations. Because of the variety of unknowns, the margin allotment is higher. Note that the upstream is not ITU J.83B, although it is Reed-Solomon-based with configurable error correction parameter. Thus, the RS code can be stronger or weaker, depending on configuration, although it is usually set at lower code rates (stronger).

As reference guidelines for upstream SNR, we choose what one particular operator uses as classification for a good upstream in terms of observable metrics. These metrics are based in part on upstream SNR (actually MER) reported. A good score for the upstream includes a consistent 30 dB reported SNR. We will assume this would represent a link capable of the highest order modulation profile at all times, which is 64-QAM today. This is what is reflected in Table 3.

Note that this threshold is actually above the no-FEC threshold for 64-QAM. This is simply the nature of the margin allotted to upstream as operated today. While the RS code offers a theoretical gain similar to the RS downstream, it is configurable from none up to strong. Also, the upstream channel has, in general, been difficult to fully exploit to date because of the range of impairments and field implementations. It is certainly reasonable to expect that, as service group sizes shrink, alignment practices improve as bonding becomes prominent, and other architectural changes take place (like a point-of-entry home gateway architecture to be discussed), the quality of the upstream will improve and the margin required to support a particular modulation profile reduced. We do not make any of assumptions about any new dB associated with those "what ifs" here.

Upstream traffic typically comes in small chunks, and therefore can only be supported by smaller block-sized LDPC codes. This leads to less coding gain for a given code rate compared to downstream. We round this difference up to an even 1 dB offset compared to downstream, or 4 dB of new upstream gain from LDPC. For upstream, then we use the assumptions shown in Table 3.

We will go forth with these SNR values as we investigate the implications to key components of the HFC architecture. Note that whether we are discussing legacy single carrier QAM or MCM systems, the SNR thresholds established in these tables are the same. These are based on AWGN performance. Thus, for both legacy QAM style and MCM, the link budget analysis below is applicable. However, it can be argued, in particular for the upstream, that use of MCM offers the opportunity to eliminate some of the margin currently allotted in Table 3, since this is based on experience with today's single carrier upstream channels. The reasoning here is that multi-carrier could be more resilient to some of the things that go into setting the upstream margin.

Table 3 - Upstream SNR Assumptions for M-QAM Profiles

SNR Requirement Assumptions, U/S

	No FEC, Theory 1.00E-08	DOCSIS w Upstream Margin Reed-Solomon	New FEC LDPC @ 4 dB
64-QAM	28	30	26
256-QAM	34	36	32
1024-QAM	40	est. 42	38
4096-QAM	46		

DOWNSTREAM HFC MIGRATION

Table 4 quantifies the delivered performance at the end of line for an HFC network based on a classic 1310 nm linear optical link as a function a modern RF cascade, such as GaAs-based RF. It is a typical mix of bridger (multi-port) amplifiers and line extenders where a cascade ensues. Table 4 includes an assumption of partial analog reclamation – a total of 30 analog carriers remain. Most MSOs have an analog reclamation plan, though many anticipate that they will leave a basic tier in place for a long time. We will analyze a full analog reclamation case as well.

Table 4 - Downstream Performance vs Cascade

	1 GHz, 30 Analog Carriers				
	CNR	CSO	CTB	CCN	QAM CCN
N+6	51	60	65	49	43
N+3	55	62	67	52	46
N+0	57	65	69	55	49

These numbers all relate to analog levels, of course, so in reference to digital they must be lowered by the amount of the digital derating. Mathematically, it is straightforward to show [7] that removal of analog frees up some RF power from the total load that could be re-allocated to digital loading. This varies from approximately 1-3 dB, depending on how many analog carriers remain (zero or 30) and the tilt used from RF amplifiers on the coaxial leg. We will not account for new RF power at this stage, but come back to this point as we discuss link budget closure. QAM SNR levels – 6 dB lower than the yellow column – are listed on the far right of Table 4 using a 6 dB de-rate. An important and expected result from Table 4 is the improvement in the CNR and CCN as the cascade becomes shorter.

Note that CCN stands for Composite Carrier-to-Noise, and captures all noise floor components – AWGN and digital distortions. It is the “true” SNR, although that is an imperfect label technically because of the contributions of distortion. However, the digital distortion contributors are many and largely independent, so a Gaussian assumption is reasonable. On a 6 MHz channel, a white, Gaussian assumption is also reasonable. For wideband channels, the “white” component may deviate, but this is exactly where MCM plays a role. By its nature, it again will make the channel noise, including CCN, “look” white (flat) in a sub-channel.

Using Table 2 requirements and Table 4 performance, and assuming the lower limit of input power is assumed delivered for a QAM channel at -6 dBmV, we can derive what noise performance is needed from the CPE to meet each threshold. This is shown in Figure 3. We have extended the CCN range downward on Figure 3 compared to Table 4 to represent perhaps deeper cascaded architectures than N+6, or systems running somewhat stretched or simply below the performance of Table 4 for a variety of other design reasons.

Also shown in Figure 3 are the above calculated CCN values of 43 dB (N+6), 46 dB (N+3) and 49 dB (N+0), identified and labeled using red vertical lines. Along with the maximum noise figures plotted in Figure 3 are noise figure values (black dashed lines) representative of common CPE platforms in the field, and of today’s vintage, which are lower noise designs. In the case of the “maximum noise figure” curves (color), QAM profiles are supported when the colored line identifying a modulation profile is *above* an example CPE NF threshold in black.

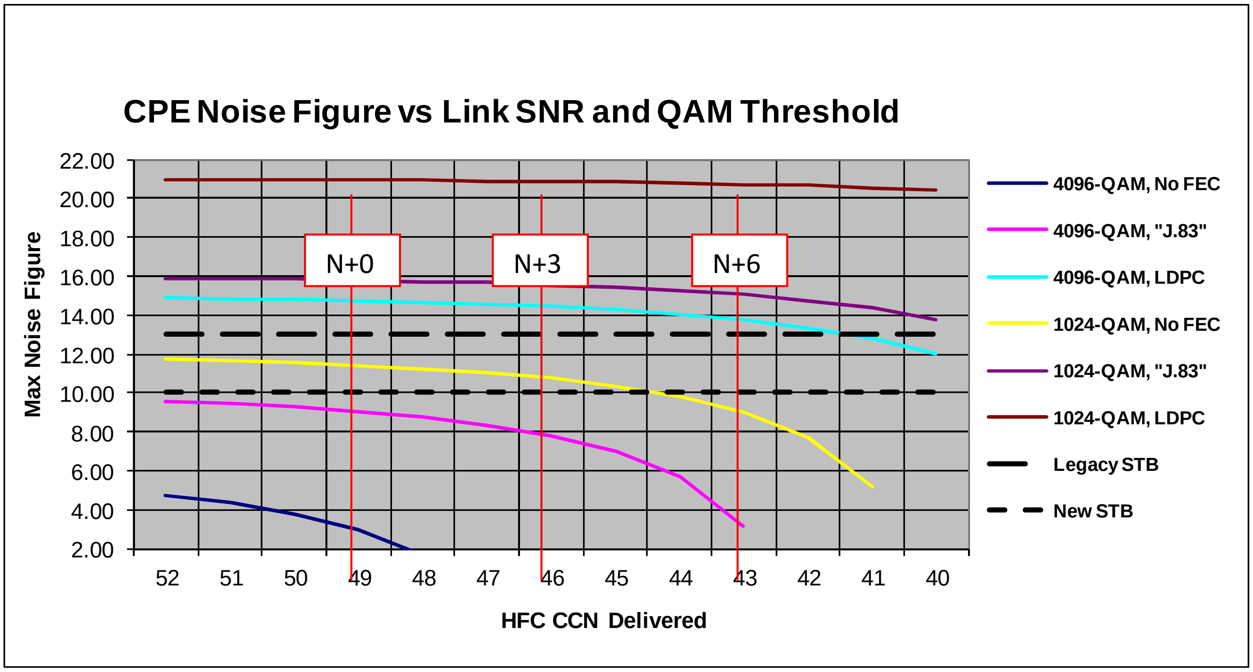


Figure 3 – STB Noise Figure Limit vs. Modulation Efficiency

Apparent from Figure 3 are three things with respect to the access network and home environment:

1) 4096-QAM is not achievable without introducing new FEC, based on today's linear optics and CPE performance. Even with new LDPC FEC, however, it is too marginal to be practical without sensitivity improvements of modern STBs. And, even with those improvements, there is just a little link budget to spare, and only if there is a fiber deep migration. Stretched architectures and high in-home losses could struggle – a QAM input below -10 dBmV to the STB instead of -6 dBmV would be insufficient for N+6, for example. Remember those possible dBs of power allocation gains of analog reclamation we identified earlier? It is cases like this where it becomes obvious that every dB of a link budget becomes critical in some cases for practical margins to be realized.

2) 1024-QAM with a J.83 flavor of FEC is achievable today with legacy STB performance, albeit there is also not much margin. For example, while 256-QAM performance requirements exist down to a -15 dBmV input in DOCSIS, this additional loss would not be able to be absorbed in the 1024-QAM link budget per Figure 3. On the other hand, 1024-QAM with LDPC is the one curve that is clearly and robustly supported – to levels as low as -13 dBmV for even existing STBs and below -15 dBmV for newer class boxes.

3) For HFC migrated to fiber deep architectures such as N+0 or N+(small) – left hand side of Figure 3 – there is little sensitivity of the NF curve to SNR variations for all cases except for a threshold using 4096-QAM without FEC, which is a non-starter.

The fact that 1024-QAM using J.83 is possible is consistent with the conclusions drawn in [9]. That analysis also pointed at the CPE noise as a potential link limiter to 1024-QAM today.

All in all from an SNR standpoint, though, updating the FEC will be instrumental to delivering higher modulation efficiency on today's quality of HFC architectures, and newer CPE will buy important link headroom to enable robustness.

Figure 3 captures access network and home. A missing component of this analysis is that Figure 3 inherently assumes a perfect transmit fidelity. In fact, of course, the DRFI specification governs transmit fidelity today. If we assume that plant linear distortions are properly handled in the receive equalizer (a good assumption), then we can consider the DRFI Equalized MER contribution of 43 dB. There is an implicit assumption that the MER is not dominated by a few discrete spurious components when we are using an SNR analysis. We will consider discrete interference in a separate section. Figure 4 shows the result with the DRFI requirement included.

While there are major differences in Figure 3 and Figure 4, the conclusions previously drawn do not vary very significantly. 4096-QAM is just more impractical than before, and the 1024-QAM "J.83" case has lost some of its margin, but remains on the bubble of workability as long as the HFC link is very good, such as the N+0 case.

Lastly, we point out that while we have captured the DRFI MER requirement in Figure 4, that requirement is obviously a minimum. Although broadband fidelity requirements are among the most difficult to meet, suppliers compete on key parameters and therefore product performance may be better in practice.

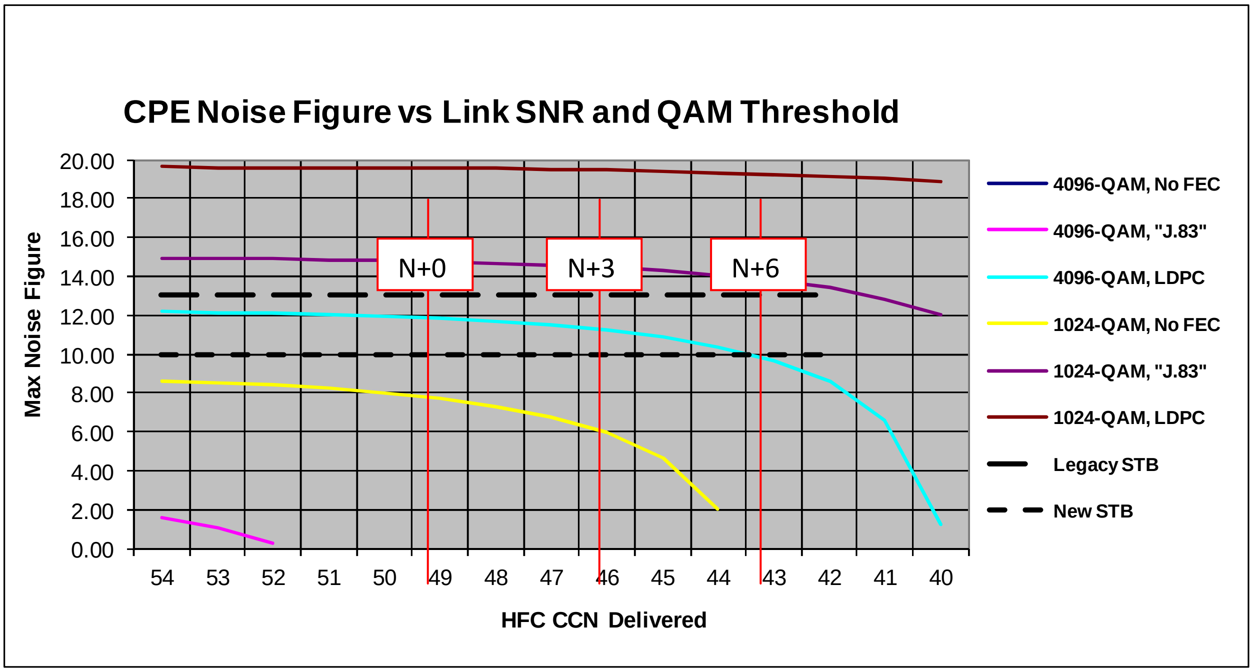


Figure 4 – STB Noise Figure Limit vs. Modulation Efficiency, DRFI MER

Multi-Wavelength Optics

Let's update the architecture now to include full analog reclamation and wavelength division multiplexed (WDM)-based linear optics commonly implemented today in multi-service fiber distribution architectures. WDM tools are becoming very valuable as operators take on Ethernet and EPON-based business services, avoid pulling new fiber where possible, and consider more consolidation of hub locations. With downstream loads moving to more QAM carriage and away from analog, the optical nonlinearities that make heavy analog loads difficult to manage over WDM become less imposing, and reaches can be extended.

Table 5(a-c) show three cases: 750 MHz, 870 MHz, and 1 GHz. In each case, a 1550 nm, ITU-grid-based, Analog/QAM transmitter is shown under nominal link conditions. Performance is also calculated with an 85 MHz upstream mid-split (slightly reduced forward load). This is a likely upstream evolution path for operators looking to exploit the full capabilities of DOCSIS 3.0 while minimizing the imposition on the downstream spectrum. The range of CCN's in Table 5 is within the ranges shown in Figure 3 and Figure 4, so these results are entirely applicable to the cases shown originally using Table 4.

Compared to classic single wavelength 1310 nm delivery, advanced optics such as these are being implemented more often, and across a variety of service scenarios, so extensive effort has gone into characterizing them across load and band variations. The extended calculations of Table 5 offer a few interesting conclusions with respect to variations in performance.

1) As the architecture shortens to N+0, the CCN going from 30 analog to zero analog improves. Accounting for the QAM-relative 6 dB de-rate with analog, for example, would leave 45 dB of QAM CCN for 750 MHz and N+0 (5-42 MHz system), instead is 47 dB. This is indicative of the effect of the analog carriers, which are higher levels than the QAM, to have a major impact on the distortion mix because of the difference in channel power. As a result, without taking advantage of any load, a couple extra dB are available. This does not included any additional dB that may be available from allocating more per-QAM power as the analog load is removed. As previously discussed, there is potential for another 2-2.5 dB on the optical link (flat loading) available that would keep the total power load the same. The tilted RF link would see less benefit.

2) The 85 MHz architecture has minimal impact on QAM CCN (0-1 dB). It removes a small chunk of forward bandwidth, but not enough to have a measurable impact. This may change at 200 MHz or more of upstream bandwidth. That case is not shown here, however, as in that case we would also expect an extended forward band. The net result should benefit CCN, since sliding the entire band results in fewer octaves of coverage, and the number octaves is important for broadband RF distortion characteristics.

3) 750 MHz systems have a 1-2 dB of SNR compared to 1 GHz systems. This is a worthwhile amount of dB gained, but perhaps not a good tradeoff relative to the capacity lost for not having the spectrum available.

Table 5 – Performance vs. Architecture a) 750 MHz, b) 870 MHz, c) 1 GHz

		750 MHz				
		30 Analog				All QAM
		CNR	CSO	CTB	CCN	CCN
Return	N+6	49	61	66	48	41
5-42 MHz	N+3	51	63	67	50	43
	N+0	51	64	67	51	47
	N+6	49	61	66	48	41
5-85 MHz	N+3	51	63	67	50	43
	N+0	52	64	67	52	48

		870 MHz				
		30 Analog				All QAM
		CNR	CSO	CTB	CCN	CCN
Return	N+6	48	61	66	47	41
5-42 MHz	N+3	49	63	67	49	43
	N+0	50	64	67	50	46
	N+6	48	61	66	47	41
5-85 MHz	N+3	50	63	67	49	43
	N+0	50	64	67	50	47

		1 GHz				
		30 Analog				All QAM
		CNR	CSO	CTB	CCN	CCN
Return	N+6	47	61	66	46	40
5-42 MHz	N+3	48	63	67	48	42
	N+0	49	64	67	49	45
	N+6	47	61	66	47	41
5-85 MHz	N+3	48	63	67	48	42
	N+0	49	64	67	49	46

Home Architecture Evolution

New CPE are taking advantage of full band capture (FBC) A/D architectures, which avoid pre-digitizing tuners that can contribute to RF degradation and simplify CPE designs. It will also make them more flexible to evolve moving forward. A low noise amplifier (LNA) precedes the A/D conversion to achieve the necessary levels to efficiently operate an A/D. This architecture is shown in Figure 5.

Analog-to-Digital converters themselves are inherently high noise figure components for typical high-speed bit resolutions because of unavoidable quantization noise. Nonetheless, this architecture does offer added flexibility for front-end sensitivity, and systems are easily optimized by choosing an external LNA, with the effects straightforward to calculate and not frequency dependent.

A quick, nominal, example using Figure 5 illustrates the simplicity:

- Assume a 20 dB gain LNA with a 5 dB NF
- Automatic Gain Control (AGC) amplification to drive A/D converter
- 12-bit A/D (at least 11 effective bits)

A well-design input cascade (LNA + AGC + A/D), can maintain a net NF of 6-7 dB for low input signal levels (where the NF comes into play the most). Thus, with input losses such as diplexers, and design focus on achieving higher modulation efficiency, NF of 8-9 dB could be the next level of sensitivity in new CPE, slightly better than the 10 dB range available today.

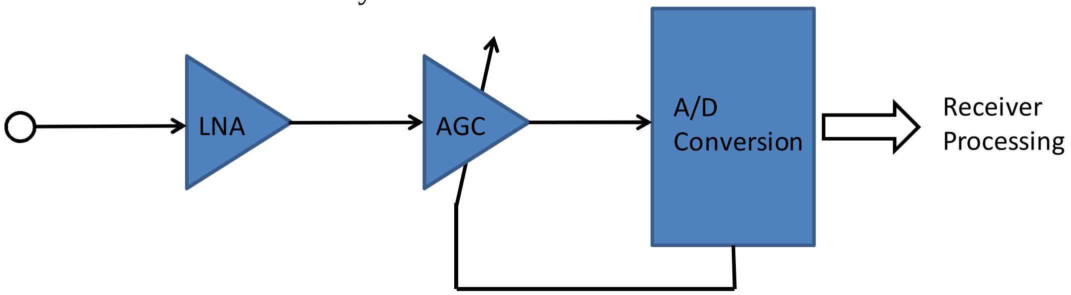


Figure 5 – Full-Band Capture Receiver Architecture

More important than the details of the receiver architecture, however, is that as part of the IP transition, operators are seriously considering investing in the next generation of home gateways based on a point-of-entry (POE) concept. Today, every subscriber is inherently *part* of the HFC access network, which makes for unpredictable results and ultimately money spent on truck rolls. The POE concept would have the cable drop go directly to an IP gateway (legacy support capabilities TBD). This IP gateway would completely abstract the inside of the home from the access network, and use only Home LAN interfaces to deliver content around the home – MoCA™, WiFi, and/or Ethernet interfaces delivering the bits over the last 100 feet. This has valuable benefits to RF losses in and out of the home for QAM receivers and upstream transmitters.

For the downstream, it amounts to benefits to the receiver SNR contribution, which can be substantial and meaningful at low input levels when advanced modulations are considered, as we have seen in Figure 3 and 4. Consider, for example, 20 dBmV tap port levels, 100 feet of drop (RG-6), 4-way splitting in the home, and 50 ft coaxial runs in the home (RG-59). At 1 GHz, we are losing RF power quickly:

STB RFin = 20 – 7 – 7 – 4 = 2 dBmV (virtual) or -4 dBmV.

If a few things break differently, the RF input will drop and the sensitivity of the receiver tested. A secondary splitter (-4 dB), extra drop length (-3.5 dB), and 15 dBmV design levels could challenge this link budget entirely, and house amplifiers may be called into play.

Now let's consider that all of the in-home loss is eliminated except for one splitter (an assumption for legacy considerations), as a POE architecture is apt to look. The loss is now the drop and one splitter, or 11 dB, a 7 dB savings. We can expect receiver NF degradation as the AGC design dials in more attenuation, but this is small relative to the improvement in signal power, so a higher SNR is obtained from the CPE.

Next, we consider the combination of the FBC architecture leading to lower noise CPE, and the POE concept, together as a "next generation" home architecture opportunity. By recalculating Figure 3 with 7 dB more of QAM power, and add a line representing the potential decrease in NF, we can recalculate NF margin to the various modulation profiles. This is shown in Figure 6.

We can quantify an example of possible NF degradation based on the Figure 5 architecture for this increased input level. Using a very simple front end cascade design, the degradation in NF calculated is less than 2 dB. It is probably much less than this, but this offers a boundary for particularly simple Figure 5 architecture. This still means at least a net SNR gain of 5 dB for the CPE contribution to the total SNR. A "degraded NF" case based on the above calculation would be identical to the 10 dB representing today's performance in Figure 6.

On Figure 6, the cascade depth marker lines used are the HFC CCN values taken from the fully loaded 1 GHz case, complete analog reclamation, and an 85 MHz Mid-Split upstream – i.e. Table 5c, orange shaded values.

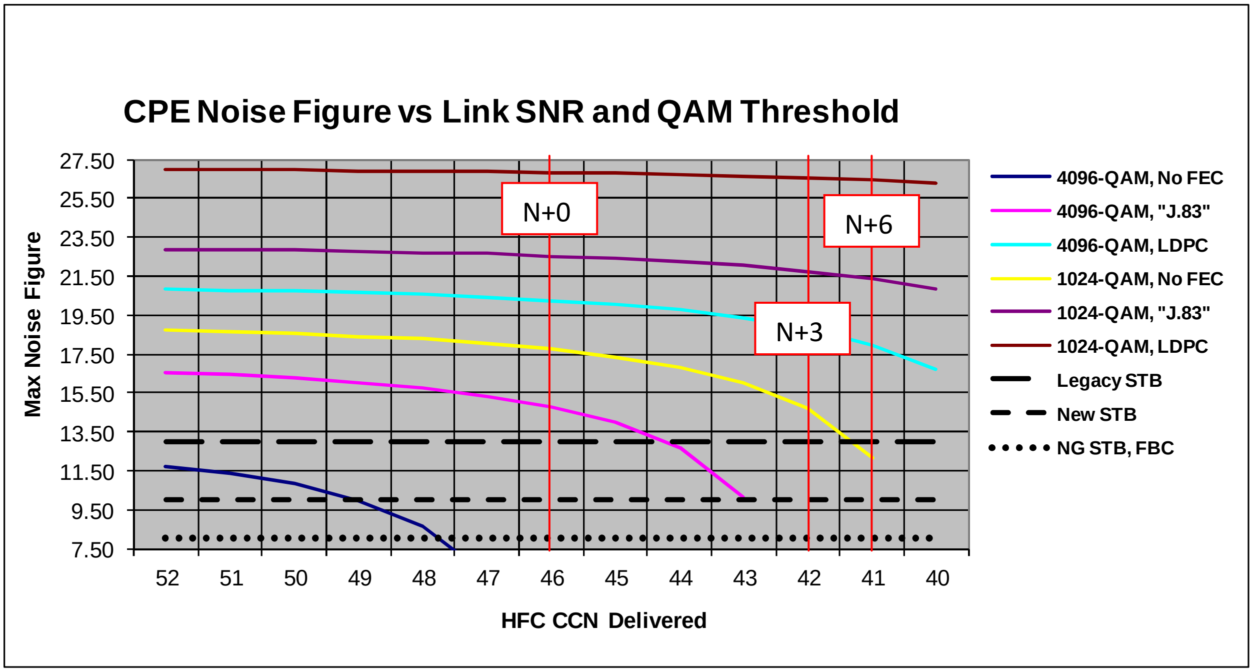


Figure 6 – STB NF Limit vs. Modulation Efficiency, POE Gateway

It would appear in Figure 6 that a tremendous amount of new margin has been created, and this certainly is the case with respect to the access network and home environment's ability to deliver the highest modulation profiles. The results suggest, for example, that "J.83" style 4096-QAM (pink) can be supported, at least on N+0 or equivalently high performing HFC links. Of course, now we are likely to see the biggest impact to the contribution of today's DRFI MER of 43 dB. The impact of this reality is shown in Figure 7.

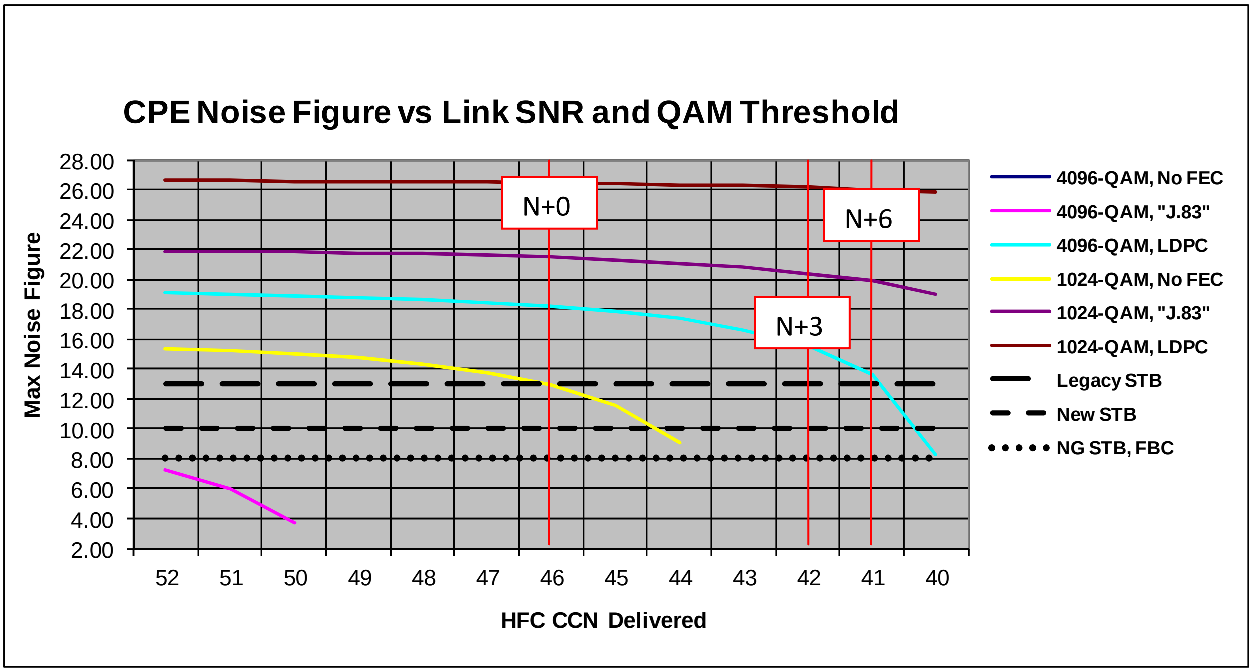


Figure 7 – STB NF Limit vs. Modulation Efficiency, POE Gateway, DRFI MER

We can draw the following conclusions when observing the full picture in Figure 7:

1) 1024-QAM is very comfortably supported from an SNR perspective with old or new FEC, cascade depth or forward band plan
2) Robust 4096-QAM would *require* advanced FEC, and a short HFC cascade – less than N+3 preferably as the curve of support is rapidly becoming sensitive to performance variation and running out of margin as the cascade lengthens.

It seems intuitive that 4096-QAM would require an updated FEC to be operational. Indeed, had it not been so, system architects would have previously considered increasing modulation profiles, as this would have indicated that HFC performance was sufficient. We can come full circle by considering the following:

1) HFC links were designed originally for analog video
2) Analog video, CNR delivered: ~45 dB
3) Digital CNR for a 45 dB analog: 39 dB
4) 4096-QAM with LDPC threshold (Table 2): 37 dB

This illustrates why we are now speaking of 4096-QAM as within range for HFC delivery, and in particular highlights the value of the advanced FEC in making this so. It also illustrates why 1024-QAM in "J.83" style is already on the verge (36 dB) [9].

Figure 8 plots one more example network architecture evolution – in this case removing the linear optical component and distributing the RF generation to the HFC node via digital optics. The assumption in Figure 8 is that this could be accomplished while maintaining DRFI-compliance (43 dB MER). Without the linear optics variation, of course, the "curves" are only a function of the RF cascade depth.

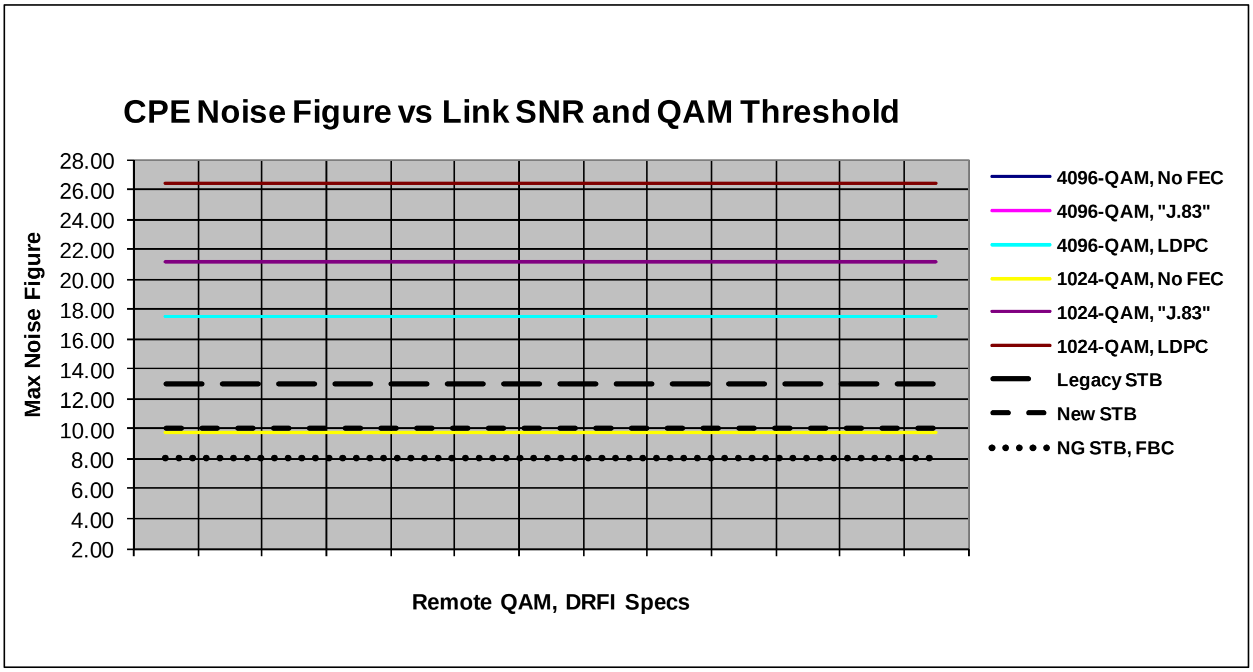


Figure 8 – STB NF Limit vs. Modulation Efficiency, POE Gateway, Remote DRFI, N+3

In Figure 8, we have assumed an N+3 cascade, using the contribution as calculated by comparing the N+0 and N+3 cases in Table 5c, and subtracting the difference. This calculation has some favorable uncertainty in it, because an "N+0" in fact includes the RF chain of the node itself, and the difference between N+0 and N+3 does not capture the effect of these gain blocks. However, Figure 8 gives a ballpark estimate of the performance of a remote QAM with DRFI performance driving an HFC cascade, and the ability this architecture has to support advanced modulation profiles. It also uses the assumptions of Figures 6 and 7 with respect to receiver and home architecture.

HFC UPSTREAM

Turning our attention to the upstream, the SNR threshold assumptions that will be used to illustrate capabilities were shown in Table 3.

In Figure 9, we show today's state of the art for a linear optical upstream. The ability to support 256-QAM upstream over 85 MHz mid-split architectures has been proven in the field, where throughputs of 400 Mbps were obtained [12, 19]. It was shown that DFB optics, coupled with higher sensitivity, higher fidelity DOCSIS 3.0 receivers, could robustly support a fully loaded 85 MHz upstream. A 12 dB dynamic range of sufficient NPR is shown. Dynamic range (DR) in the upstream is much more important than in the downstream. Network design is optimized and aligned precisely in the downstream, while upstream design and environmental variations, alignment techniques, as well as unpredictable RF channel conditions, require an SNR to be met over a range of input levels. Historically, DR on the order of 10 dB has been sought.

Measured packer error rates (PER) were taken at various input levels on the N+3 cascade used in Figure 9. These points are shown in Figure 9 where the yellow marker dots are along the blue noise power ratio (NPR) curve. These are where low PER was observed. The yellow dots on yellow trace are representative of DFB performance of transmitters such as many in the field today. The noise power ratio analysis of Figure 9 makes plain why robust performance was observed. The measured points clearly fall within an area of high NPR and SNR, where good performance would be expected, and with solid 6-7 dB of peak margin extending beyond the 256-QAM threshold identified.

Note that in all figures below, thresholds identified and not labeled as "LDPC" are the assumed "DOCSIS" thresholds of Table 3 (i.e. not the "No FEC" thresholds).

In Figure 10, we introduce the new PHY performance thresholds with advanced FEC from Table 3. In addition, we have included the net performance of the RF portion of the HFC by introducing a deep, combined amplifier cascade (orange). Because the return amplifiers contribute high SNRs individually, the effect here is minor, but we will see how this contribution may also increase as optics improves. Of course, over time, the expectation is that the RF cascade will shorten as well. Or, if the segmentation is virtual (combining in the node is removed), fewer amplifiers will "funnel" upstream. The net effect is the same – fewer amplifiers contributing noise to the return path, reducing the input noise power at the receiver and thereby increasing SNR.

It is clear from Figure 10 that "DOCSIS" coding for 1024-QAM (dashed red) would not be able to be supported from an SNR perspective. This threshold is simply above even the peak NPR. For 1024-QAM enabled with LDPC FEC, however, we can see that the threshold of operation is now exceeded by the combined HFC+CMTS link.

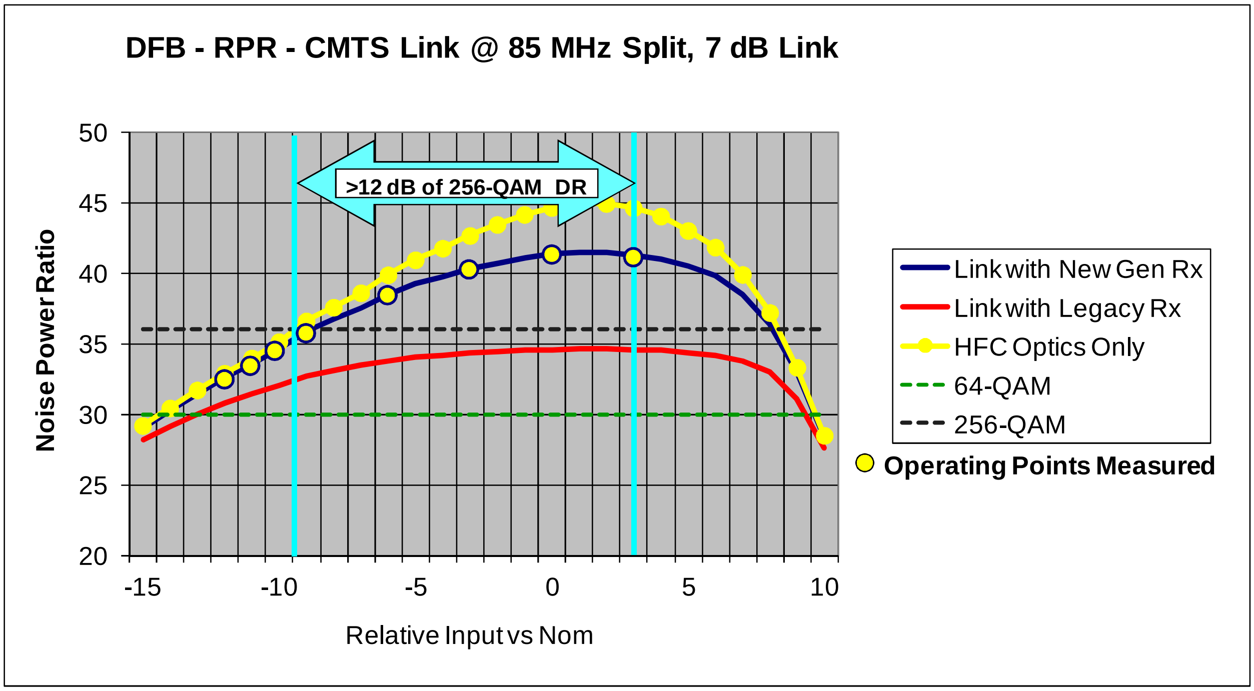


Figure 9 – 256-QAM “DOCSIS” over 85 MHz DFB Return Optics

Unfortunately, the dynamic range is reduced on the left hand side by 3 dB. It is also well understood and documented that, as QAM profiles become increasingly dense, the right hand side (soft distortion components, red arrow) also have a more significant impact, reducing dynamic range from the right. The above reduction from the right is an estimate based on prior work, which only considered up to 64-QAM [23].

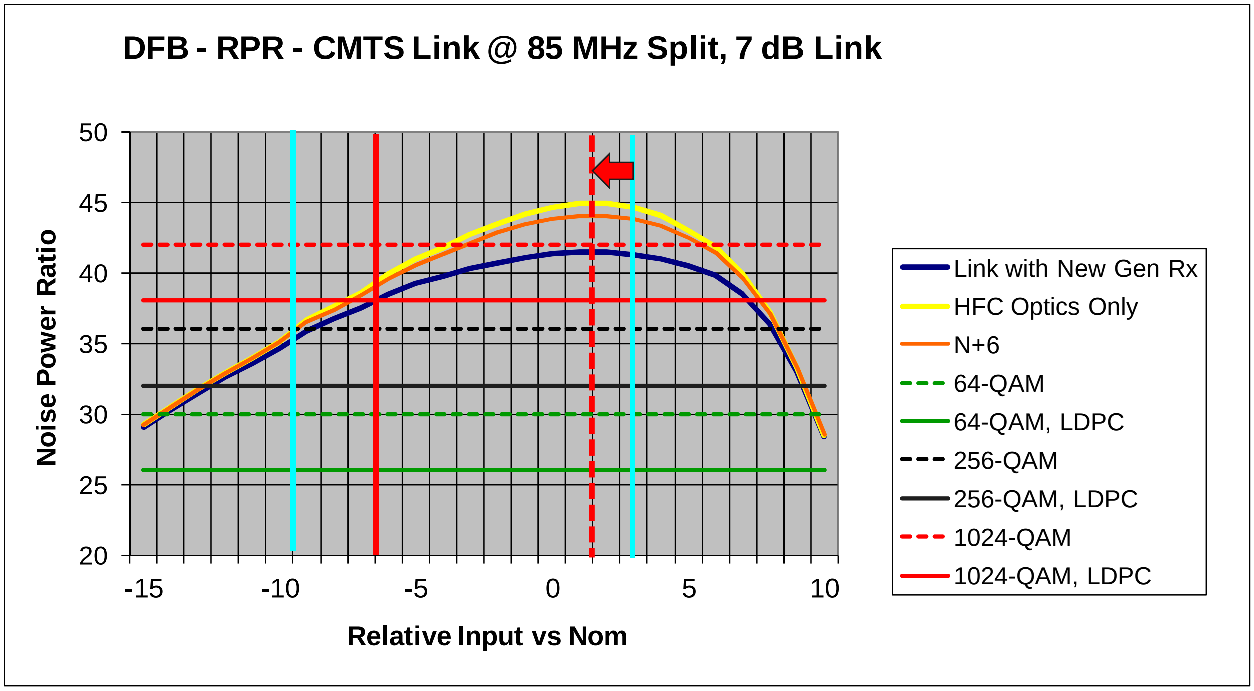


Figure 10 – 64/256/1024-QAM, “DOCSIS” and New FEC, 85 MHz DFB Optics

The net effect of the reduced range (8 dB DR) is that 1024-QAM with LDPC would be marginal in practice. It would likely work in some places, but inconsistently throughout a footprint without further network improvements. An analogous situation is the introduction of DOCSIS 3.0 64-QAM using FP upstream laser technology. FPs exhibit reduced performance compared to DFBs, and thus eat into the DR margin acceptable to run DOCSIS 3.0 64-QAM. Such is the case in Figure 10 for 1024-QAM with LDPC. The operating window is small, and this would likely be reflected in inconsistent performance.

In summary, with LDPC, it would be possible to get 1024-QAM working on well-behaved upstream channels that are aligned properly for laser input level, and only using new DOCSIS 3.0 upstream receivers with higher sensitivity. However, in practice, over a large range of plants, performance would be unreliable (and impossible at all if laser technology has not been updated and legacy receiver performance exists).

In Figure 11, we extend the Figure 10 case to a 200 MHz "high split" return – a long term approach under consideration by the industry to deliver more upstream capacity [2]. An implicit assumption is that the 200 MHz receiver could perform equivalently from a sensitivity standpoint as today's DOCSIS 3.0 receiver. Nonetheless, the pure power loading makes 1024-QAM completely impractical. The 256-QAM case without new FEC is now marginal, as the 1024-QAM case in Figure 10 was – in fact, it exhibits the same DR and signature relative to the performance threshold (highlighted in red). However, it is likely to be somewhat more robust that 1024-QAM simply because 256-QAM will be less sensitive to the types of things that the allocation of DR and margin is meant to protect against.

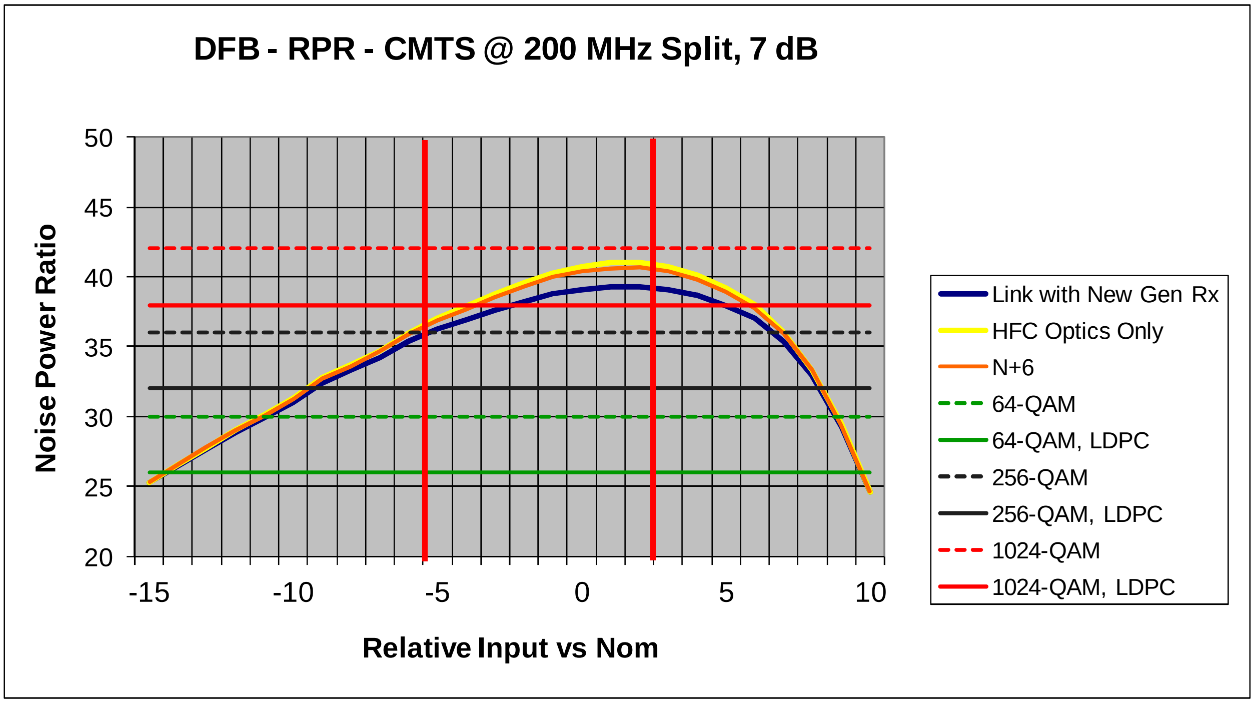


Figure 11 – 64/256/1024-QAM, "DOCSIS" and New FEC, over 200 MHz (Projected)

Enhanced Linear Optical Performance

The analysis in Figures 9-11 uses performance of DFBs like many that may be in the field, perhaps lower power (1 mw), where the upgrade from Fabry-Perot lasers (FPs) had already taken place. Today, because of the rising interest in mid-split architectures and more bandwidth efficient modulation, new development activity is focusing on analog and digital return solutions that optimize performance for extended bandwidths. Continued improvements in noise performance of both transmitters and receivers have also occurred over time. Recently measured performance of a mid-split bandwidth "DFBT3" (Motorola model, temperature compensated, 2 mw) to a return path receiver is shown in Figure 12.

In Figure 12, the performance curves from Figures 9 are also shown. A notable improvement over time in peak NPR is observable, and an associated dynamic range increase for a given threshold. Peak performance of 50-51 dB is observed – again making the point that a good DFB return path transmitter-receiver link looks very much like a 10-bit digital return link [16].

Also shown is this improved performance when combined with a DOCSIS 3.0 receiver (dashed blue). Here, it becomes clear that as the optics improves, the influence of receiver SNR contribution begins to have a larger effect. We will discuss this further later in this section.

In Figure 13, the DOCSIS and "next gen" thresholds using advanced FEC are evaluated against the improved mid-split performance identified and measured in Figure 12. Note that the performance shown is the linear optical return (yellow) only, along with the DOCSIS 3.0 receiver (blue). Thus, this is equivalent to an N+0 case, since there are no RF noise contributions from the plant in Figure 13.

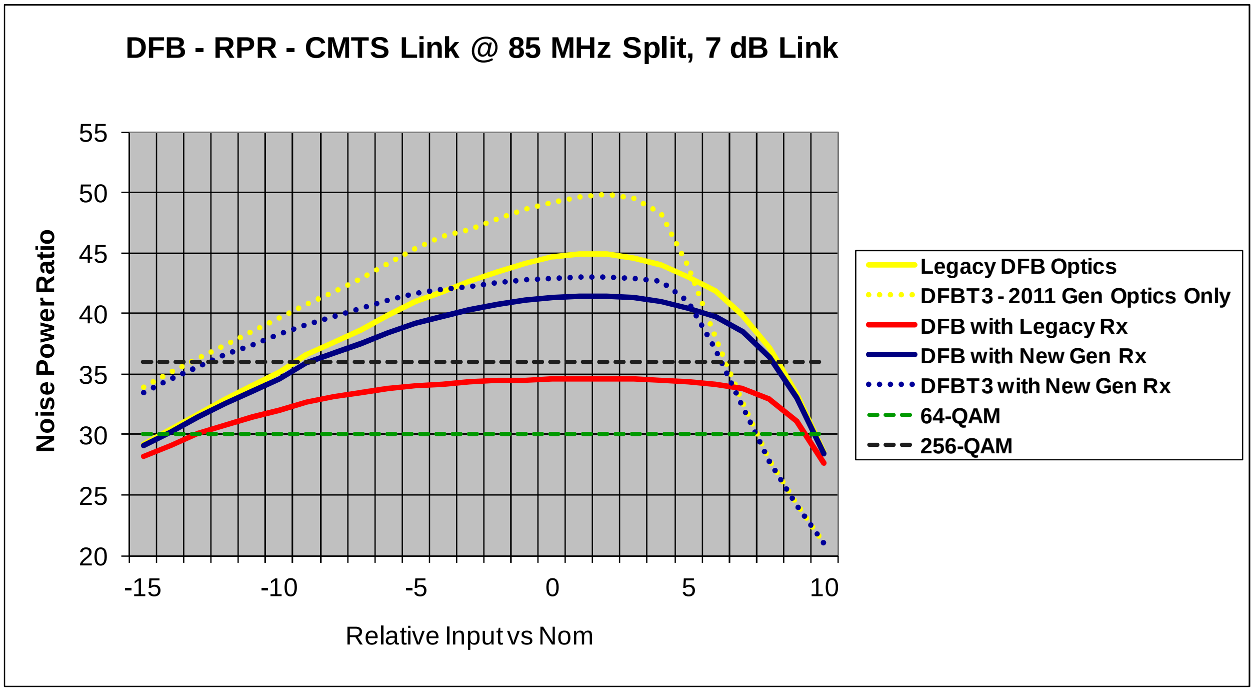


Figure 12 – Measured Mid-Split Performance, Modern DFBT3-RPR Link

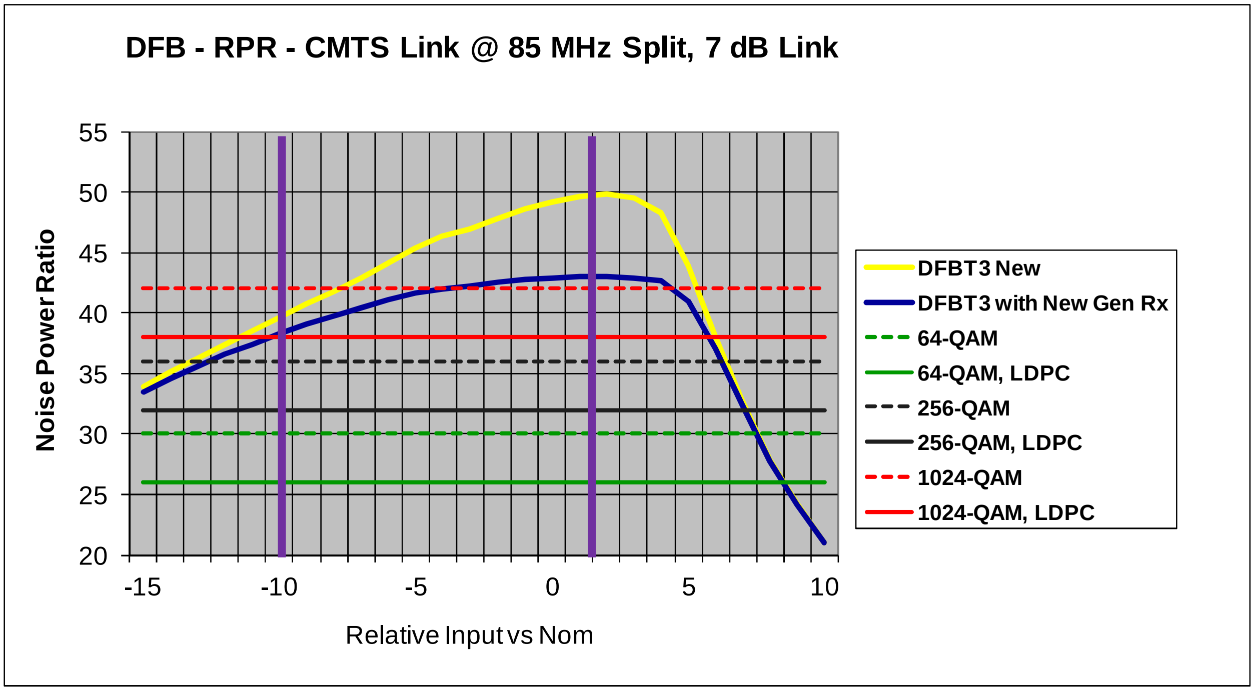


Figure 13 – High Performance DFBT3-RPR Mid-Split, All Thresholds

All of the thresholds are met in Figure 13, and the 64-QAM through 256-QAM cases comfortably supported. This is to be expected since this has been proven to be the case for 256-QAM today without any new FEC applied or improved HFC return performance.

For 1024-QAM, however, clearly "DOCSIS" PHY coding will not be sufficient using the Table 3 assumptions. The advanced FEC applied to 1024-QAM, however, does show promise. Indeed, it shows 11 dB of DR (purple markers) above the threshold identified for 1024-QAM. However, note that the margin above the threshold never exceeds 5 dB, and for half of the DR it is 4 dB or less. The receiver noise contribution, ably supporting 256-QAM (against a 64-QAM-maximum DOCSIS requirement, it should be added), comes into play here.

While the dynamic range appears robust, the low margin across the full DR range may make performance less robust than such a DR would otherwise indicate considering we are dealing with a more sensitive QAM profile. In other words, "peak" margin may need consideration as we move to more complex QAM profiles. Return path alignment is based on setting composite signal levels around a "sweet spot" close to where NPR is at its peak, but allowing for some margin of back-off to avoid the clipping and distortion region on the right hand side. At the very least, it would seem reasonable that the same peak margin available today for 256-QAM (6-7 dB, Figure 9) would be a good starting point objective for higher order scenarios like 1024-QAM.

Today's expectation of DR is built around volumes of 16-QAM and 64-QAM deployments, exclusively. Since the DR for 1024-QAM with LDPC in Figure 13 is about the same as we observed in Figure 9 for 256-QAM, we might expect 1024-QAM with LDPC to be operational under good upstream conditions. Similarly, since the peak margin available is reduced across its DR compared to Figure 9, and because other contributions not captured by NPR will effect 1024-QAM worse than 256-QAM, its performance is likely to be less rugged than the 256-QAM case in Figure 9.

Figure 14 adds an RF noise contribution assuming a deep cascade (N+6) and combined four ways. The degradation due to the quantity of RF amplifiers is seen (orange). However, it is clear in observing the blue curves – (optics + receiver) – with and without RF noise contributions from the cascade, that the limitation about 256-QAM is in the noise contribution of the upstream receiver, at least for nominal architectures and levels as they are implemented today. And, again, we are quantifying a receiver for 1024-QAM that had a requirement to meet 64-QAM performance objectives.

In summary, from an NPR/SNR perspective, with new LDPC-based FEC, 1024-QAM is possible on high performance upstream optical links. However, it may be operationally less robust when considered across a range of possible channel environments given the decreased peak margin and exaggerated sensitivity of 1024-QAM to other link impairments not captured by NPR analysis. It is hard to be certain, but we can get a window into the performance consistency through the ruggedness with which 256-QAM becomes implemented. 1024-QAM with LDPC should be “like” this but somewhat less robust. As might be expected, we will need to ask more of next generation receivers than sensitivity supporting 64/256-QAM if implemented over linear optical returns or today’s vintage of digital returns.

In Figure 15, we apply the Figure 12-14 performance across a 200 MHz upstream. The assumptions are that identical optical noise performance can be achieved (shared over a wider bandwidth) and identical receiver sensitivity can also be achieved.

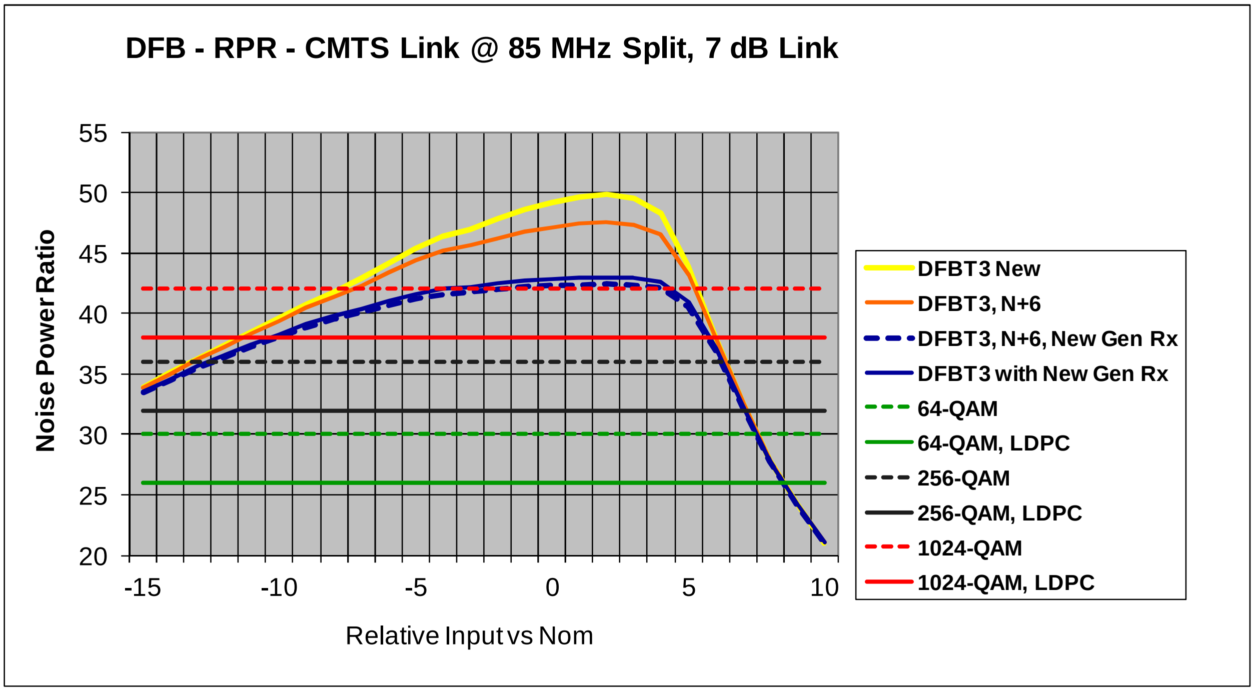


Figure 14 – High Performance DFBT3-RPR Mid-Split, N+6, All Thresholds

An encouraging conclusion is that, for this extended bandwidth case, 256-QAM of the "DOCSIS" variety appears to be operational at acceptable DR and a reasonable peak margin, at least under these assumptions of the wider band design. Since we do not have extensive field lessons for 256-QAM, the peak margin question raised for 1024-QAM could apply in this case as well. However, to be at least within reach of 256-QAM already as it exists today on an extended bandwidth return, based on today's linear optics, is an excellent side for the future of new broadband capacity for the upstream.

The 1024-QAM case now clearly has insufficient DR (purple) as well as small peak margin from the threshold – 4 dB maximum and less than that across the DR.

Figure 16 shows an enhanced return performance example, in this case based on digital return. Key advantages of digital return include ease of setup, and, from the perspective of this paper, NPR performance independent of link length. The return path performance of the digital return approach is almost entirely determined by the A/D converter resolution. Finding A/D converters that provide a high *effective* number of bits (ENOB) as bandwidths expand is the limiting performance component. The potential to require redesign for each bandwidth increasing increment is a disadvantage [16] of this approach.

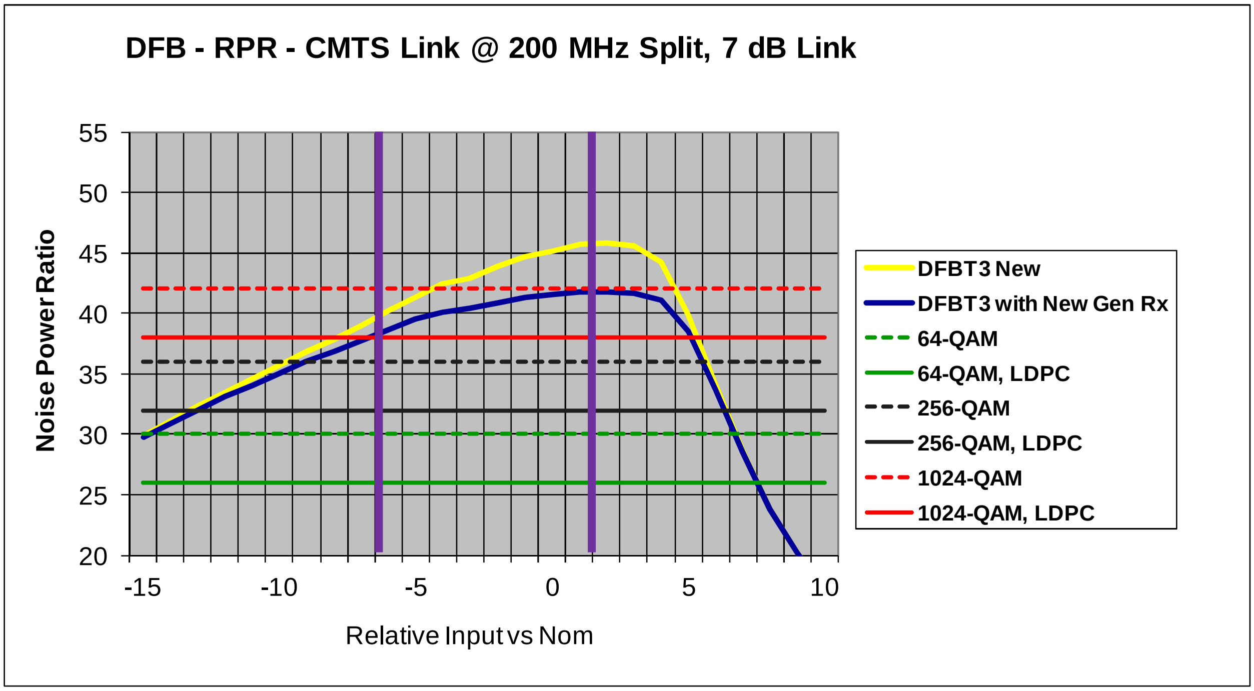


Figure 15 – High Performance DFBT3-RPR "High" Split, All Thresholds

Figure 16 uses measured performance of a solution that behaves like an ideal 10-bit A/D, or, equivalently, has an ENOB of 10 bits.

Three composite NPR curves that include RF and receiver contributions are shown with the digital return-only NPR performance:

1) Digital return, N+6 RF cascade, receiver (purple dash)
2) Low noise DFBT3-RPR link, N+6, receiver (yellow)
3) Original DFB-RPR (256-QAM analysis), N+6, receiver (blue)

The comparison indicates that the digital solution contributes a couple more dB of DR to the 1024-QAM with LDPC threshold on the left side of the NPR curve (the SNR side), which itself adds 4 dB of DR over the legacy solution. The extra couple of dB from the digital return could be the difference between robust or less robust. However, since the peak margin has not changed between the two (yellow and dashed purple), and is limited by receiver sensitivity, their performance will likely be similar – on the bubble of robust-enough margin.

A clear conclusion from these analyses is that support of 1024-QAM would require new FEC, and be aided by an improved receiver sensitivity. This is readily illustrated by observing the effect of a hypothetical receiver designed to support 1024-QAM (as opposed to DOCSIS 3.0, 64-QAM), with a receiver sensitivity improved 3 dB in so doing.

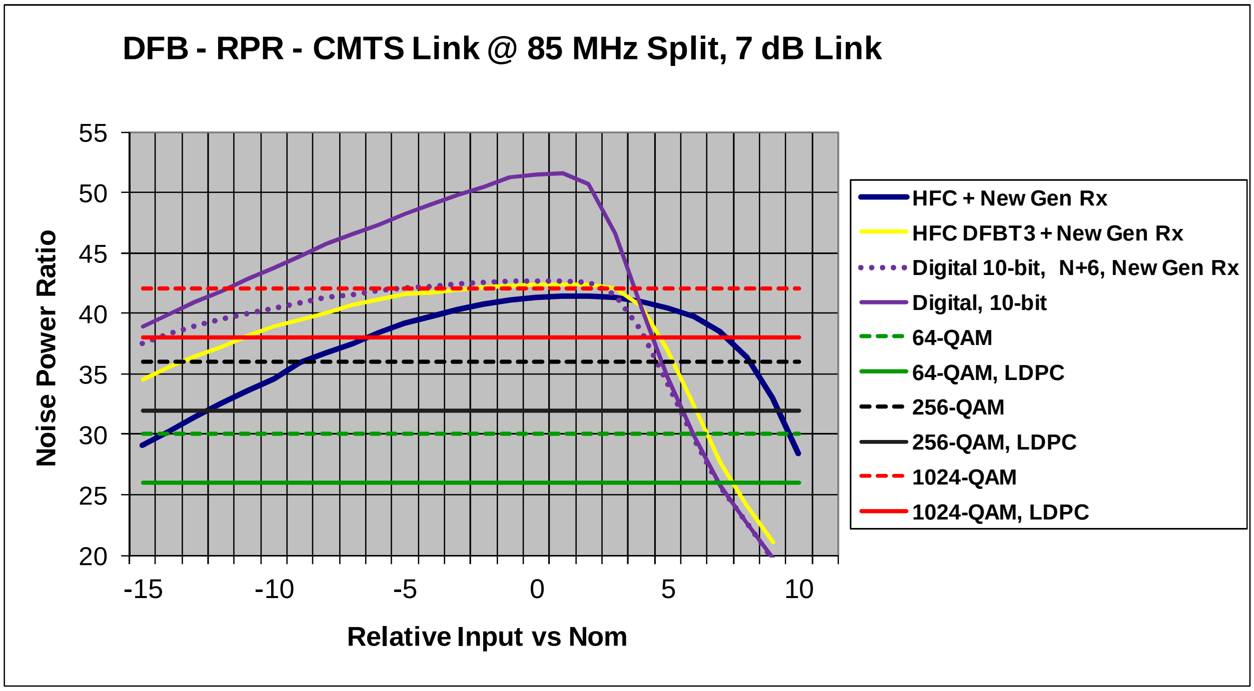


Figure 16 – 10-bit Digital Return & Linear Optics Cases, N+6, All Thresholds

Figure 17 shows the same composite NPR curves, minus the digital return-only one for this case of improved sensitivity. Figure 17 illustrates that for linear optical links or digital returns, it is clear that these would now have margin to support 1024-QAM with adequate DR. This could only be so if the 1024-QAM was accompanied by new FEC. And, there are still some open questions about whether peak margin standards should be considered, much of which may come as 256-QAM deploys under "DOCSIS" threshold conditions.

This leads to an overall conclusion that, for upstream, attention will need to be paid to upstream optical architectures, receiver performance, and possibly Headend architecture on the whole given that the optical receiver-CMTS connection today is a simple, almost forgotten pipe that can create low level inputs to CMTS ports and challenge their sensitivity. For all cases desiring to support 1024-QAM, LDPC-based FEC is a must-have, though itself not a sufficient condition.

Remote Demodulation

Future architectures may take advantage of distributed physical layers. The analogous concept in today's world is "CMTS in the Node." For a transition into new RF and IP technology or extension of DOCSIS, this may be easier to consider than it has historically been with DOCSIS. And, with Gigabit Ethernet and EPON optics available at low cost, it become very attractive to consider taking advantage of these standard interfaces, potentially eliminate linear optics from the plant, and improve performance all at the same time. Modular node platforms are now built to handle various plug-in optical and RF modules, so this approach is consistent with HFC node evolution.

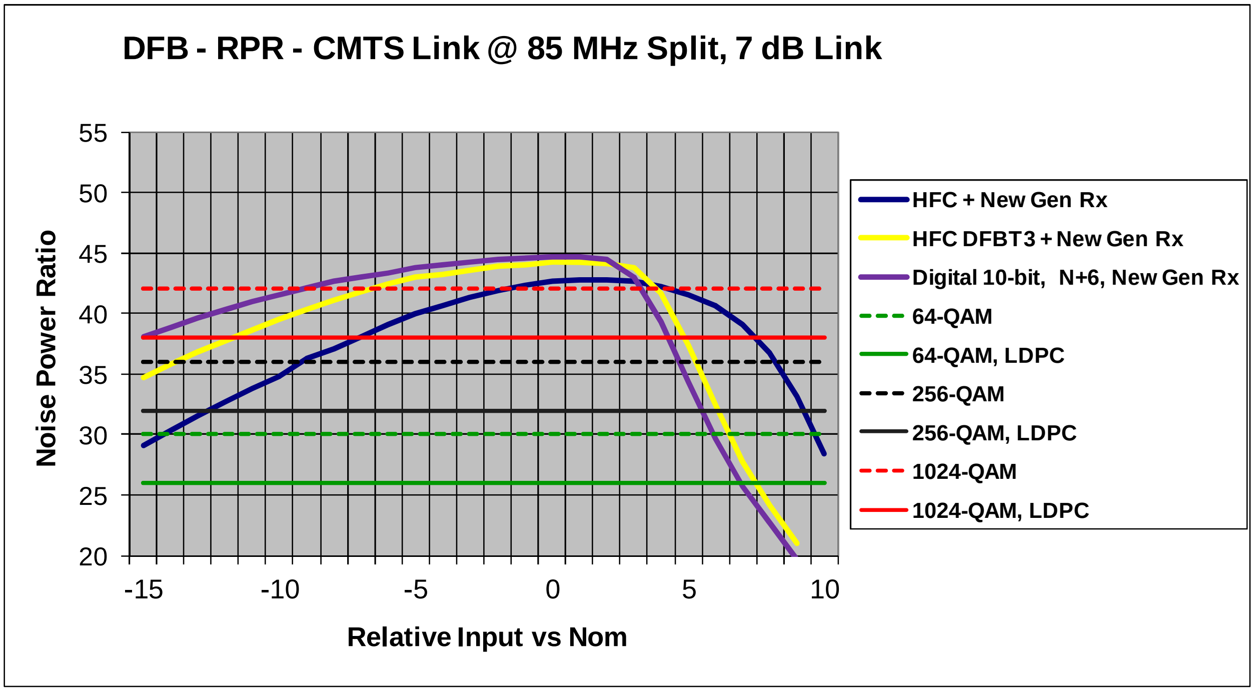


Figure 17 – All Cases of Return Optics, N+6, Improved Rx Sensitivity (3 dB)

If a remote receiver is placed in the plant, then we can simplify the analysis by removing the optical links from the equation. This can be shown using the same NPR curves as before, except now only the softly distorting amplifier limits the right hand side of the curves. While these can be characterized and specified, this is not usually done. It is nominally assumed that the upstream signal transmissions stay comfortably within the range of return amplifier linearity. The alignment of the total level of the load is critical at the upstream optical interface, but in the plant there is not the constraint of total power load to the extent that there is in the optics. Coupled with common noise figures of return amplifiers, the result is that very high SNRs are possible for a single amplifier relative to its noise contribution to the upstream cascade.

Instead of NPR, Table 6 shows a range of required Noise Figures of a module installed in node. We have made an assumption, based on our above discussion of peak margin above threshold for advanced QAM profiles, that a net SNR of 45 dB (7 dB of margin to 1024-QAM with LDPC) is the objective.

A range of port levels (low, mid, high) are shown, and the impact these levels would have on the remote receiver's NF requirement. The path loss between the coaxial input port and node module is accounted for, so this is the noise figure performance at the input to the receiver module. Provided low input levels do not reign, these are not particularly challenging. And, even at the lowest input level identified here, the NF is achievable with good design practices.

Table 6 – Calculating "Remote" Receiver Noise Requirements

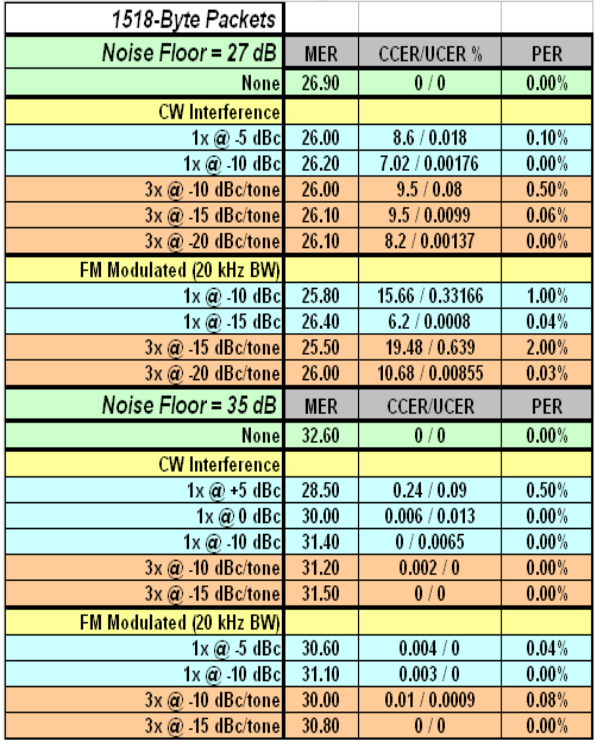
Signal Level		**Signal Level**		**Signal Level**	
Ports	10.0	Ports	15.0	Ports	20.0
Node Path Loss	5.0	Node Path Loss	5.0	Node Path Loss	5.0
Node Combine	6.0	Node Combine	6.0	Node Combine	6.0
Noise		**Noise**		**Noise**	
Amplifier (NF=8)	50.2	Amplifier (NF=8)	-50.2	Amplifier (NF=8)	-50.2
Cascade	6.0	Cascade	3.0	Cascade	3.0
Combine	4.0	Combine	1.0	Combine	1.0
RF Noise	36.4	RF Noise	-45.4	RF Noise	-45.4
SNR Req'd	**45.0**	**SNR Req'd**	**45.0**	**SNR Req'd**	**45.0**
Terminating NF	**6.3**	**Terminating NF**	**16.9**	**Terminating NF**	**22.0**

SIGNAL-TO-INTERFERENCE

Single carrier techniques to combat narrowband interference amount to attempting to notch out the offender's band through an adaptive filtering mechanism, and recover the modulated carrier around it as effectively as possible. Because removing the interference involves removing signal spectrum that subsequently must be equalized and detected, the effectiveness of the process is reduced as the interference becomes wider band, or, for a fixed signal-to-interference energy (S/I), if there are multiple interferers to handle.

Test results have been observed and reported with respect to the A-TDMA DOCSIS upstream in separate studies in recent years [13,22]. Table 7 shows thresholds of uncorrectable codeword errors observed for 64-QAM.

Table 7 – Ingress Thresholds for 64-QAM A-TDMA Upstream



1518-Byte Packets			
Noise Floor = 27 dB	MER	CCER/UCER %	PER
None	26.90	0 / 0	0.00%
CW Interference			
1x @ -5 dBc	26.00	8.6 / 0.018	0.10%
1x @ -10 dBc	26.20	7.02 / 0.00176	0.00%
3x @ -10 dBc/tone	26.00	9.5 / 0.08	0.50%
3x @ -15 dBc/tone	26.10	9.5 / 0.0099	0.06%
3x @ -20 dBc/tone	26.10	8.2 / 0.00137	0.00%
FM Modulated (20 kHz BW)			
1x @ -10 dBc	25.80	15.66 / 0.33166	1.00%
1x @ -15 dBc	26.40	6.2 / 0.0008	0.04%
3x @ -15 dBc/tone	25.50	19.48 / 0.639	2.00%
3x @ -20 dBc/tone	26.00	10.68 / 0.00855	0.03%
Noise Floor = 35 dB	MER	CCER/UCER	PER
None	32.60	0 / 0	0.00%
CW Interference			
1x @ +5 dBc	28.50	0.24 / 0.09	0.50%
1x @ 0 dBc	30.00	0.006 / 0.013	0.00%
1x @ -10 dBc	31.40	0 / 0.0065	0.00%
3x @ -10 dBc/tone	31.20	0.002 / 0	0.00%
3x @ -15 dBc/tone	31.50	0 / 0	0.00%
FM Modulated (20 kHz BW)			
1x @ -5 dBc	30.60	0.004 / 0	0.04%
1x @ -10 dBc	31.10	0.003 / 0	0.00%
3x @ -10 dBc/tone	30.00	0.01 / 0.0009	0.08%
3x @ -15 dBc/tone	30.80	0 / 0	0.00%

In the study summarized in Table 8, recent results for 256-QAM for a fixed PER objective of 0.5% and 1% for a high SNR condition are derived through testing. The SNR condition applied (SNR = 36 dB) is consistent with the Table 4 assumption on a robust SNR threshold to use for 256-QAM in the upstream, and the analysis in [22] further identifies this high SNR as one that increasingly enables ingress cancellation.

Table 8 – Ingress Thresholds for 256-QAM A-TDMA Upstream

			256-QAM		
	Level (dB, dBc)	**UNCORR%**	**CORR%**	**PER%**	**MER (dB)**
Baseline - AWGN	**36**	0.000%	0.000%	0.000%	37
Single Ingressor Case					
QPSK 12kHz 0.5%	3	0.254%	0.435%	1.060%	34
QPSK 12kHz 1.0%	1	0.447%	0.944%	2.300%	34
FSK 320ksym/s 0.5%	29	0.278%	0.032%	0.110%	35
FSK 320ksym/s 1.0%	27	0.633%	0.230%	0.810%	35
FM 20kHz 0.5%	2	0.128%	0.295%	0.750%	34
FM 20kHz 1.0%	1	0.187%	0.554%	1.260%	34
Three Ingressor Case					
CPD 0.5%	28	0.297%	0.041%	0.190%	34
CPD 1.0%	27	0.698%	0.144%	0.750%	33

Tables 7 and 8 are valuable indicators of the performance of ingress cancellation (IC). However, to understand how well it is working, it helps to know how robust the individual QAM profiles are to signal-to-interference ratio (S/I) to begin with. A system simulation was performed to evaluate this sensitivity. An example of 64-QAM with a single CW interferer is shown in Figure 18, while a 256-QAM example is shown with three non-coherent interferers in Figure 19. The familiar CW "donut" pattern is clear in Figure 18.

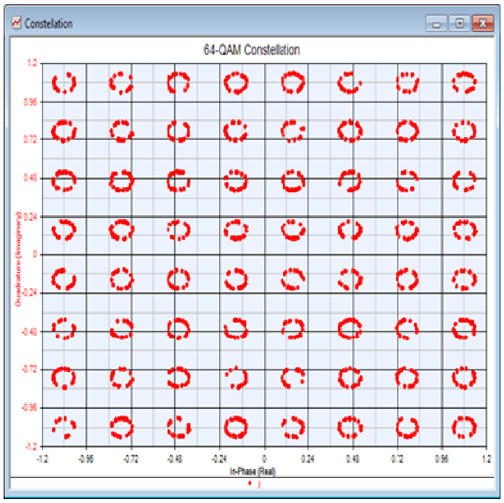


Figure 18 – 64-QAM with a Single CW Interferer

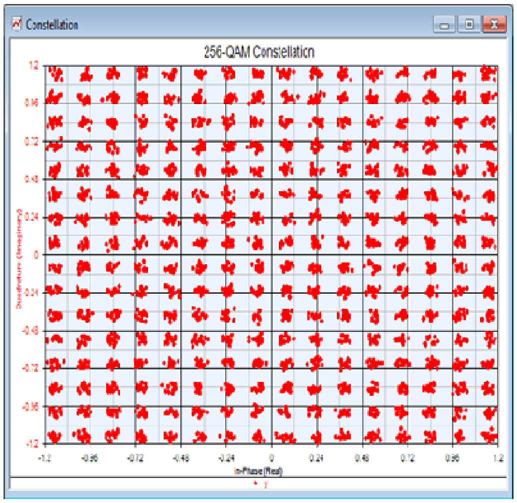


Figure 19 – 256-QAM with 3 CW Interferers

Both cases were evaluated to find thresholds of correctable low error rate. We use the more challenging three-interferer case as a reference. A subset of these simulation results are summarized in Tables 9a-d for 64/256/1024/4096-QAM. The modeling tool is described further in the Appendix.

Obviously, more dense profiles are more sensitive to S/I. The relationship of interest is to note that, for a given "DOCSIS" Table 3 SNR threshold, high enough to have robust performance and allow the IC to work, the relative S/I difference across profiles is also approximately 6 dB for when errors begin to be counted.

Now consider the IC performance based on Table 7 MER before and after IC. It can be calculated as providing roughly an effective 26-28 dB of cancellation for the case of multiple 20 kHz interferers. An estimate for the Table 8 case for 256-QAM and 0.5% PER using 256-QAM data at SNR = 36 dB, from the analysis table shown in Table 9b (est. 28 dB S/I), is that the IC is providing about 26 dB (S/I =2 to S/I = 28) of IC for a single 20 kHz interferer. Table 7 suggests that if the total power of the interference is the same, and it is narrowband, IC performance is close to the same for one interferer or three.

Table 9a-d – Ingress Thresholds for Uncoded QAM
a) 64-QAM b) 256-QAM c) 1024-QAM d) 4096-QAM

	64-QAM		S/I, N=3 Interferers									
		20	21	22	23	24	25	26	27	28	29	30
	25	1.80E-02	1.20E-02	9.10E-03	4.31E-03	3.45E-03	1.72E-03	9.35E-04	7.46E-04	5.01E-04	2.88E-04	2.09E-04
	26	1.90E-02	1.30E-02	7.08E-03	4.20E-03	1.63E-03	1.18E-03	5.63E-04	3.12E-04	2.07E-04	1.09E-04	7.90E-05
SNR	27	1.90E-02	1.00E-02	4.29E-03	2.19E-03	9.51E-04	3.26E-04	2.17E-04	1.22E-04	3.70E-05	2.00E-05	1.90E-05
	28	1.90E-02	9.70E-03	3.17E-03	1.03E-03	5.22E-04	1.36E-04	1.01E-04	2.20E-05	2.00E-05	1.10E-05	5.00E-06
	29	1.20E-02	4.47E-03	2.25E-03	1.31E-03	7.90E-05	1.30E-05	3.20E-05	1.00E-06	0	0	0
	30	6.32E-03	2.58E-03	1.17E-03	2.32E-04	1.78E-04	4.50E-05	7.00E-06	0	0	0	0
	31	1.40E-02	6.29E-03	2.10E-03	3.75E-04	7.00E-06	8.00E-06	0	0	0	0	0
	32	1.60E-02	1.13E-03	2.81E-04	3.90E-05	4.10E-05	2.00E-06	0	0	0	0	0
	33	2.69E-03	5.27E-04	1.05E-03	1.57E-04	2.00E-06	0	0	0	0	0	0
	34	2.36E-03	4.48E-03	3.10E-05	2.30E-05	1.00E-06	0	0	0	0	0	0
	35	9.62E-03	5.78E-04	6.60E-05	0	0	0	0	0	0	0	0

	256-QAM		S/I, N=3 Interferers									
		30	31	32	33	34	35	36	37	38	39	40
	30	4.02E-03	2.68E-03	1.90E-03	1.36E-03	8.94E-04	6.77E-04	5.22E-04	4.48E-04	3.53E-04	2.95E-04	2.86E-04
	31	2.86E-03	1.30E-03	9.74E-04	5.69E-04	4.25E-04	2.36E-04	2.20E-04	1.34E-04	1.11E-04	9.10E-05	8.60E-05
SNR	32	1.40E-03	9.41E-04	5.01E-04	2.03E-04	1.50E-04	7.90E-05	6.50E-05	3.50E-05	1.90E-05	1.50E-05	1.30E-05
	33	5.79E-04	3.93E-04	9.90E-05	1.18E-04	3.50E-05	2.00E-05	1.70E-05	1.20E-05	6.00E-06	4.00E-06	2.00E-06
	34	2.89E-04	2.76E-04	3.80E-05	2.00E-05	2.60E-05	2.00E-06	2.00E-06	0	0	0	0
	35	8.70E-05	1.09E-04	3.00E-05	1.10E-05	8.00E-06	0	1.00E-06	0	0	0	0
	36	3.80E-05	5.00E-06	7.00E-06	0	0	0	0	0	0	0	0
	37	3.70E-05	5.00E-06	1.00E-06	0	0	0	0	0	0	0	0
	38	1.50E-05	4.00E-06	0	0	0	0	0	0	0	0	0
	39	5.00E-06	0	0	0	0	0	0	0	0	0	0
	40	0	0	0	0	0	0	0	0	0	0	0

	1024-QAM		S/I, N=3 Interferers									
		40	41	42	43	44	45	46	47	48	49	50
	35	1.99E-03	1.60E-03	1.38E-03	1.10E-03	1.01E-03	8.36E-04	7.91E-04	8.35E-04	7.51E-04	6.92E-04	6.09E-04
	36	7.18E-04	6.75E-04	5.39E-04	4.05E-04	3.64E-04	3.26E-04	2.61E-04	2.13E-04	2.23E-04	2.04E-04	1.68E-04
SNR	37	4.34E-04	1.94E-04	2.18E-04	1.14E-04	8.40E-05	8.40E-05	4.60E-05	4.60E-05	6.20E-05	5.10E-05	5.40E-05
	38	1.33E-04	1.10E-04	5.00E-05	3.30E-05	2.20E-05	1.40E-05	1.40E-05	1.10E-05	7.00E-06	3.00E-06	7.00E-06
	39	2.60E-05	1.60E-05	1.50E-05	0	4.00E-06	0	0	4.00E-06	0	0	0
	40	5.00E-06	5.00E-06	2.00E-06	0	1.00E-06	0	0	0	0	0	0
	41	1.00E-06	0	0	0	0	0	0	0	0	0	0
	42	2.00E-06	0	0	0	0	0	0	0	0	0	0
	43	0	0	0	0	0	0	0	0	0	0	0
	44	0	0	0	0	0	0	0	0	0	0	0
	45	0	0	0	0	0	0	0	0	0	0	0

	4096-QAM		S/I, N=3 Interferers									
		40	41	42	43	44	45	46	47	48	49	50
	40	0.014	0.01	7.92E-03	6.40E-03	4.87E-03	4.24E-03	3.48E-03	2.83E-03	2.71E-03	2.28E-03	2.16E-03
	41	0.011	7.55E-03	5.34E-03	3.79E-03	2.63E-03	2.18E-03	1.56E-03	1.31E-03	1.11E-03	8.47E-04	7.26E-04
SNR	42	8.11E-03	4.64E-03	2.87E-03	1.98E-03	1.23E-03	9.98E-04	6.45E-04	5.35E-04	4.13E-04	3.17E-04	2.77E-04
	43	5.34E-03	2.49E-03	1.43E-03	1.26E-03	5.54E-04	4.43E-04	2.91E-04	1.32E-04	1.16E-04	6.60E-05	7.10E-05
	44	4.91E-03	2.18E-03	9.43E-04	3.59E-04	2.24E-04	1.02E-04	8.71E-05	3.60E-05	3.50E-05	2.50E-05	1.30E-05
	45	3.48E-03	1.53E-03	3.76E-04	2.03E-04	1.01E-04	2.50E-05	1.60E-05	8.01E-06	1.80E-05	5.00E-06	1.00E-06
	46	1.83E-03	1.04E-03	3.52E-04	1.55E-04	4.80E-05	1.20E-05	8.01E-06	1.20E-05	0	0	0
	47	2.04E-03	1.81E-04	1.68E-04	9.01E-06	1.10E-05	1.00E-06	0	1.00E-06	0	0	0
	48	9.81E-04	2.89E-04	2.70E-05	8.01E-06	1.00E-06	0	0	0	0	0	0
	49	7.19E-04	9.21E-05	1.00E-05	0	0	0	0	0	0	0	0
	50	4.51E-04	8.31E-05	6.00E-06	0	0	0	0	0	0	0	0

Based on the above, then, the IC is providing about the same amount of cancellation in these two cases tested, although there is much more SNR headroom in the 64-QAM case. However, the 64-QAM case also does not *need* additional suppression, so, while it may be available, no further IC adaption is required to deliver low error performance.

For the upstream, we are interested in extending what we have learned for 64/256-QAM to 1024-QAM as an advanced profile. Simulation results for the "DOCSIS" SNR threshold condition for 1024-QAM of Table 3 and a 27 dB S/I easily shows that uncoded error rates are horrendously high (0.1 to 0.01 range). They are so high as to possibly be unable to be corrected adequately if at all by FEC, or it would be not desired to rely so heavily on it. Better than 26 dB of IC would be required, and based on the above mentioned relationship, probably 6 dB better. It is unclear if today's IC can accomplish this for multiple interferers with bandwidth, but it likely will be necessary.

It is worthwhile to point out that for a single CW interferer, effective cancellation of about 35 dB was obtained based on Table 7. So, at least for this friendlier case, the IC function can be stronger.

Threshold Values with New FEC

We have already concluded in prior analysis that a 1024-QAM downstream will require LDPC, so let's consider how this plays out relative to S/I. Our revised threshold of 38 dB is for AWGN. For the upstream, where we allocate more margin, this is still a low error rate condition, as shown in Figure A-5 (appendix). A 40 dB SNR is the 1e-8 BER case for 1024-QAM.

The same modeling table above used for 256-QAM estimation indicates that for the 38 dB SNR case, an S/I of approximately 37 dB will leave a very correctable error rate in the 1e-3 to 1e-4 range. A 1e-5 threshold, based on Table 9c, suggests instead a 42 dB S/I. For the 37 dB objective, -2 dBc of interference successfully suppressed for 256-QAM in Table 9 then requires 35 dB of IC applied. This is *more* than 6 dB (by 3 dB) of additional IC given the 26 dB 256-QAM example. The same exercise for 4096-QAM, were it an upstream mode (using 44 dB SNR), would suggest a 42-45 (Table 9d) S/I, or 40-43 dBc of IC. This is 5-8 dB different than 35 dBc.

So, there is at this point only a range of additional IC expectations that seems to follow from these results. Intuitively, not being an AWGN impairment around which detection (and FEC) is optimized, we should expect that once the impairment is large enough to contribute to errors, the relationship would exceed 6 dB per modulation profile. Because of that and the sensitivity of FEC error rate curves to fractions of dB of SNR, it is probably a good starting point to consider a relationship such as 8 dB more IC capability until more granular modeling over a range of interference patterns and a larger sample size can be established.

Downstream Interference

In the downstream, a significant amount of study has already been performed to quantify the impact of CSO and CTB analog beat distortions on QAM. These distortions look like narrowband interferers, but the nature of their make-up (many independent distortion contributors falling close to one another) is that they have noise-like qualities, including an amplitude modulation component. This is meaningful for BER degradation. An example of a 256-QAM signal with analog distortion components from a 79-channel load is shown in Figure 20.

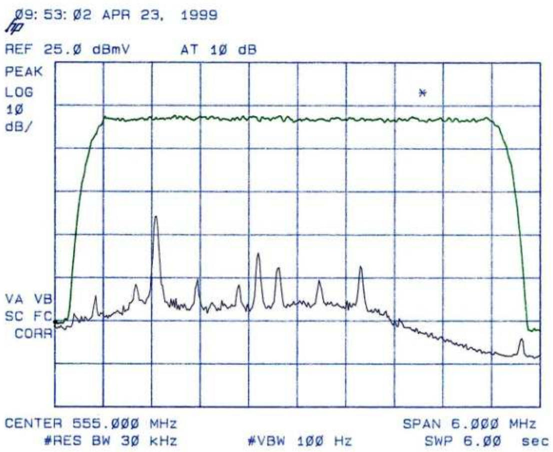


Figure 20 –256-QAM with CSO and CTB Distortions

The "donut" constellation we saw previously for 64-QAM in Figure 18 gives insight into why the beat distortion phenomenon when analog loads are prevalent matters from a QAM perspective. A CW interference example for 256-QAM at very high SNR is shown in Figure 21.

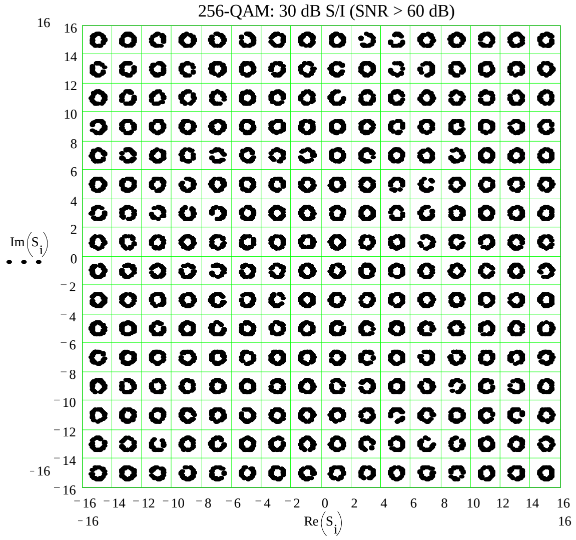


Figure 21 –256-QAM @ 30 dB S/I

In Figure 21, the SNR is 60 dB – an error free region in an AWGN-only environment, where 1e-8 occurs at SNR = 34 dB. Clearly, with only this CW carrier imposed on Figure 21, we still have an apparent error free environment. The peak-to-average ratio of a sinusoid is, of course, 3 dB – it has a constant envelope. If that envelope is not large enough to cause a decision error, then without noise there will be no decision errors even with the interferer.

The concerns with respect to CSO and CTB distortion beats are that, unlike CW or FM interference, they have a noise-like peak-to-average quality to them [20]. In fact, the envelope has a Rayleigh-like fit, which is representative of the detected envelope of a Gaussian process. This is shown in Figure 22 [20]. This amplitude modulating effect can be applied to the "donut" of Figure 21 – envision the circular symbol point breathing in and out. The nature of the distortion beat degradation is also that it is a narrowband process (10's of kHz) relative to the QAM bandwidth. Thus, a distortion beat sample will extend over many symbols in a row, and if it is high enough to induce decision errors, there is likely to be a burst of them.

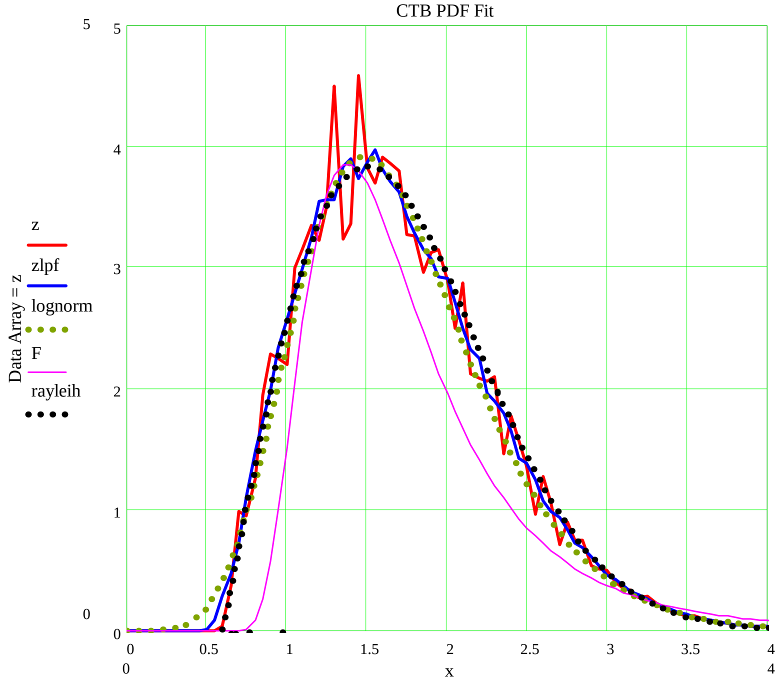


Figure 22 –Amplitude PDF of CTB

Fortunately, as we shall see and describe in more detail in the case of phase noise in the next section, in the downstream a powerful interleaver is available. It is capable of randomizing the errors, allowing the FEC to do its job better. This is quantified for phase noise, but the same dynamics apply in this case. A 20 kHz process has a 50 usec "time constant" of error generation, and a common (I=128, J=4) interleaver setting exceeds this by a factor of more than 5, easily distributing the errors into correctable codewords.

Analysis and test of the CSO/CTB impact for 1024-QAM has been performed [9]. A summary table of the results relative to these analog distortions is shown in Table 10. One of the key take-aways from that analysis – RF cascade depth as a function of analog carriers and amplifier performance – is shown in Figure 23.

For Figure 23, perhaps the single key result for purposes of this paper is that if analog video is reduced to 30 carriers, then the cascade depth that can be tolerated, under assumptions of 20 Log(N) degradation (which is overly pessimistic) is, for all practical purposes, unlimited for 1024-QAM using amplifier performance commensurate with today's plant equipment.

Table 10 – 1024-QAM Downstream Interference and Thresholds

	CW Interference		CTB Interference			
SNR	Pre-FEC Error Threshold	Post-FEC Error Threshold	Pre-FEC Error Threshold	Post-FEC Error Threshold	Post FEC > 1E-6	Post FEC Broken
50 dB	34 dB	33 dB	55 dB	55 dB	55 dB	45 dB
45 dB	35 dB	33 dB	55 dB	60 dB	55 dB	46 dB
40 dB	36 dB	34 dB	60 dB	60 dB	55 dB	49 dB
37 dB	38 dB	37 dB	60 dB	60 dB	55 dB	50 dB

Based on the CSO and CTB values we observed in Table 5, this is not surprising. The historical minimum acceptable value of 53 dBc works out to 47 dBc for 256-QAM, and this proved to be manageable with 256-QAM. With the 30-analog values of CSO and CTB ranging from 61-67 dB, 8-14 dB of better performance is occurring while a modulation profile increase of 6 dB is taken on for 1024-QAM. Also, Table 5 represents the performance at the worst case frequency. For CSO, these components tend to pile up at the low end of the band (analog). CTB tends to pile up in the middle, where QAM spectrum will be allocated – thus the focus on CTB below.

If we consider that 1024-QAM under an LDPC FEC has a threshold SNR of 31 dB based on our Table 2 downstream assumptions, then 60-something dB of CTB distortion, even with a noise like amplitude variation, would have little impact on an architecture delivering this SNR, or an SNR higher but not high enough to advance the modulation profile to 2k/4k-QAM. Basically, if 1024-QAM is workable with today's PHY, then better FEC can only make it better – it just may not achieve all of the new FEC gain of an AWGN-only noise. However, the interleaver, the relationship of the distortion levels to the CCN values in Table 5, the low pre-FEC thresholds observed in [9], suggest that it will achieve full benefits of the FEC.

We had already concluded in [9] that 1024-QAM was possible, and much more so at 30 analog carriers. What can we say about 4096-QAM?

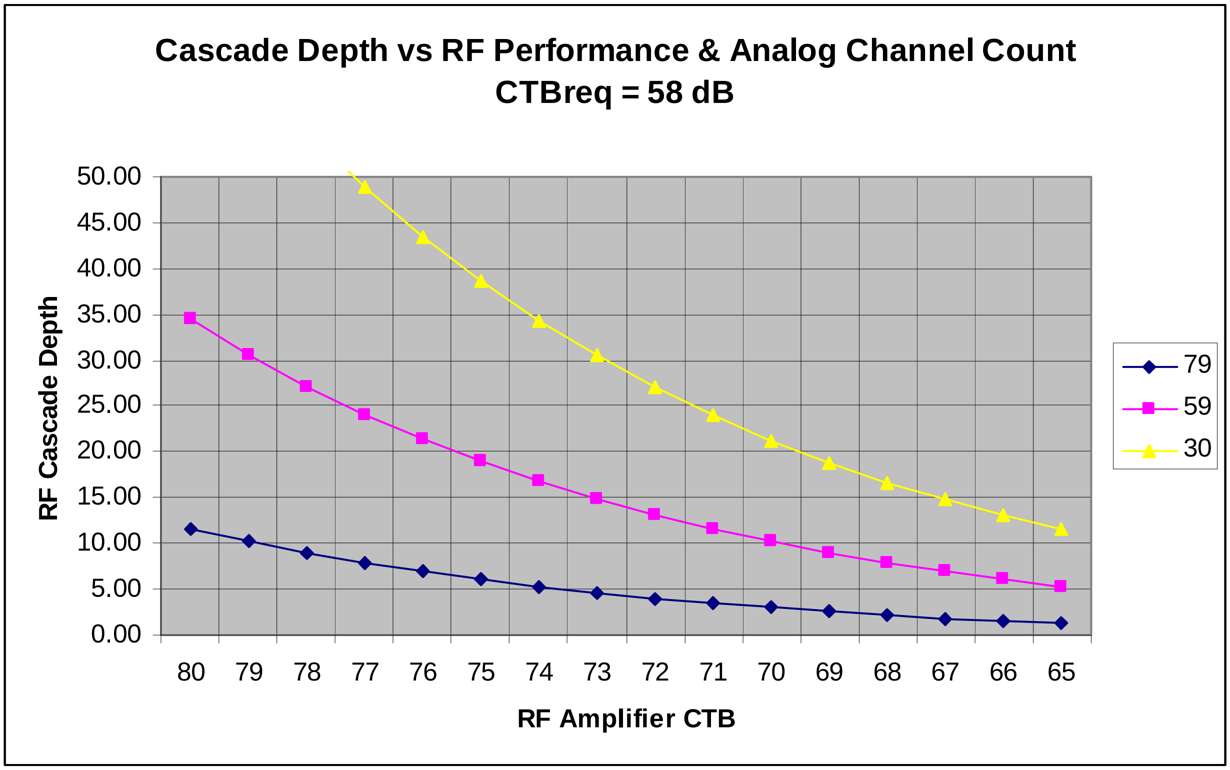


Figure 23 –RF Cascade Depth Limitations vs. Amplifier CTB for Analog Reclamations

First, a few items that may make us not care very much about that situation:

- When analog is fully removed, there are no longer any beat distortion components to worry about as interferers. We can then simply follow the CCN and the SNR analysis previously discussed.
- In the case of narrowband interference in the downstream, we could call on the upstream interference analysis above. It can only be conservative since the upstream must handle burst reception and adapt accordingly on a burst by burst basis. In the downstream, the receiver has the luxury of a constant input signal, a much simpler problem.
- Lastly, by the time we are deploying 4096-QAM, it is very likely that it is part of a multi-carrier downstream, and so narrowband interference analysis applies completely differently. We will discuss that later in the paper.

Figures 3-8 pointed out why the 4096-QAM case will require LDPC to be robust. It will provide the margin necessary to maintain performance against the 37 dB threshold established in Table 2. A high pre-FEC error count would ensue for 79-channel analog system at the 53 dBc minimum: The resulting narrowband interference would be 47 dBc on average, and peak to 33-35 dB, causing pre-FEC error rates worse than .01 (Figure A-5). They might be fixable, but this would not the ideal way to consume the FEC budget.

Therefore, 4096-QAM and full analog loading, we would suggest, is not a good combinations. 4096-QAM ought to be reserved for reduced analog loading or no analog loading. A caveat is that, as part of a transition plan of spectrum, such as Figure 1 indicates, may include new PHY above today's forward band, and beat mapping analysis would likely treat this region favorably in dBc of distortion reduction.

For reduced analog loading, CTB values of 66 dB (60 dB to QAM), peaking amplitudes would instead be 46-48 dB. The 1e-8 value for 4096-QAM without coding is 46 dB. So, it is likely errors will be counted, and corrected. They will be bursty, but the interleaving will arrange them nicely for the FEC. The situation is analogous to 256-QAM measurements with distortion in 2002 [20]. We now have 12 dB more of modulation profile sensitivity, and 13-14 dB better distortion values due to analog reclamation assumptions. Because they are noise like in effect on the constellation (MER is clouded), the similar relative relationship should yield similar results. This is also consistent with the expectation that 1024-QAM will perform very well in a reduced analog system from a distortion perspective.

PHASE NOISE

When QAM signals are put into the RF domain, they inherently have a phase noise mask applied to them. Phase noise is a measure of the spectral purity of the carrier signal itself. It is commonly measured by turning off the modulation on a waveform, and observing the level of noise surrounding the carrier at very close offset frequencies. At its most simplest, a perfect CW tone would be a single line in the frequency domain. In practice it cannot be perfect, and the amount that it is not perfect is quantified by this random phase modulation imposed at frequency offsets from the carrier frequency itself. The shape of the noise around a carrier is well understood, following many classic behaviors of semiconductor-based oscillators and the circuits that perform frequency synthesis to put modulated signals somewhere in the RF band. While there are many variables, in general, the higher the frequency, the worse the phase noise will be, the broader the tuning range, the higher the phase noise will be, and the finer the increment of tuning, the higher the phase noise will be. The shape is also known as a phase noise "mask."

An example phase noise mask is shown in Figure 24 [10]. It illustrates some common characteristics – a close to carrier flat region and small peaking, and a region of 20-30 dB/decade roll-off. The flat region is often much wider than the example shown here. The "0" of the x-axis represents the carrier itself, and the data points plotted indicate offsets from the carrier where noise density is measured. The values at offset frequencies are given in dBc/Hz, and the total noise in a bandwidth is then the area under the curve of the mask, usually recorded as a dBc value or degrees rms.

Signal-to-Phase Noise Relationships

In this section we introduce some simple-to-understand M-QAM-phase noise relationships based on the qualitative descriptions above, some nomenclature, and a deeper understanding of the processes involved in determining its effects.

For converting dBc values of phase noise, or signal-to-phase noise ratios, to degrees rms, we have:

deg rms of phase noise = (180/π) sqrt[10^(-dBc of phase noise)]

The use of degrees rms is often very illustrative when we think about QAM constellations, as we shall see. There is a simple rule of thumb that keeps us from having to rely on the above equation. It is based on the recognition that a 35 dB signal-to-phase noise ratio is the same as 1° rms. Also, rms is a linear quantity, so doubling it is 6 dB.

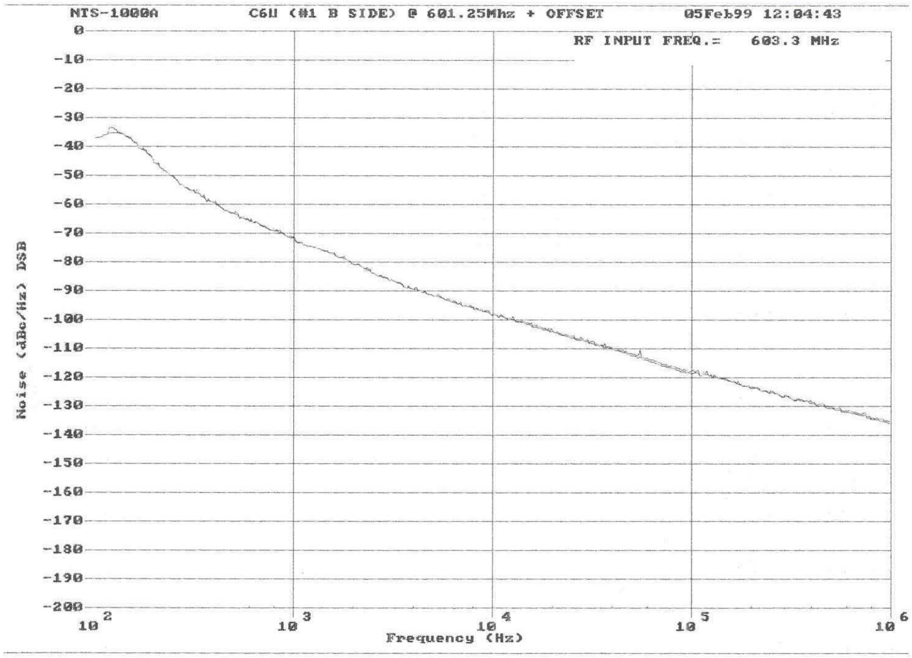


Figure 24 – Example Phase Noise Mask – RF Upconverter @ 601.25 MHz

For example, if -35 dBc = 1° rms, then

-23 dBc = 4° rms
-29 dBc = 2° rms
-35 dBc = 1° rms
-41 dBc = 0.5° rms
-47 dBc = 0.25° rms, etc.

Because of its rotational effect, phase noise affects QAM constellation points non-uniformly. Figure 25 [6] shows an example of essentially noise-free 64-QAM with 1° rms phase noise, using a mathematical tool on the left and a simulation environment (for error rate analysis) on the right. While the angle of rotation is the same for every symbol point, it is apparent and geometrically expected from polar coordinate mathematics how this impacts the outermost symbol points the most relative to breaching decision boundaries.

Compare Figure 25 with Figure 26, which shows the same 64-QAM symbol with only AWGN impairment, set at a 1e-8 BER. Note how the degradation due to additive noise is randomly distributed in I and Q dimensions, whereas the phase noise impact is exclusively angular.

Furthermore, the sensitivity of M-QAM gets worse with increasing M because of this non-uniform rotational effect. The shrinking of the distance to the decision boundaries for increasing M for a fixed average power puts makes the same amount of rotation more deleterious for higher M-QAM profiles. Figures 27 and 28 show a noise-free 256-QAM constellation with just 0.5° rms phase noise imposed, and a 1024-QAM constellation with a .25° rms phase noise imposed, respectively. The similarity in Figures 25, 27 and 28 of the relative rotation to decision boundaries, for the outer symbols in particular, is clear.

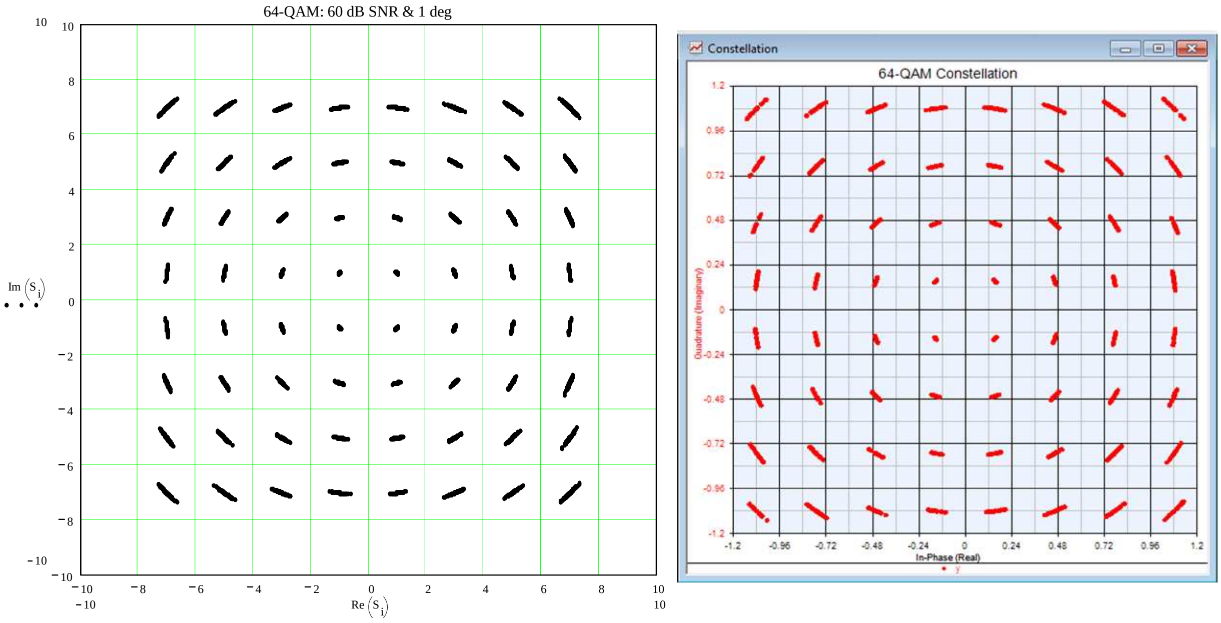


Figure 25 – 64-QAM, 1° rms Phase Noise (Analysis Tool, Simulation Tool)

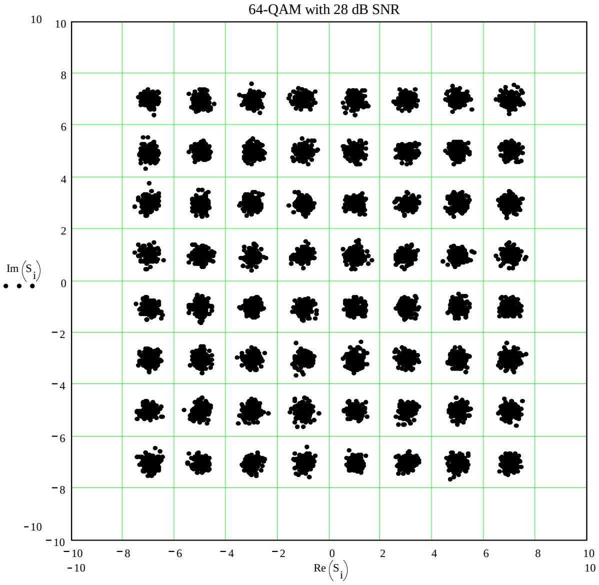


Figure 26 – 64-QAM @ 1e-8 Noise (AWGN) Level

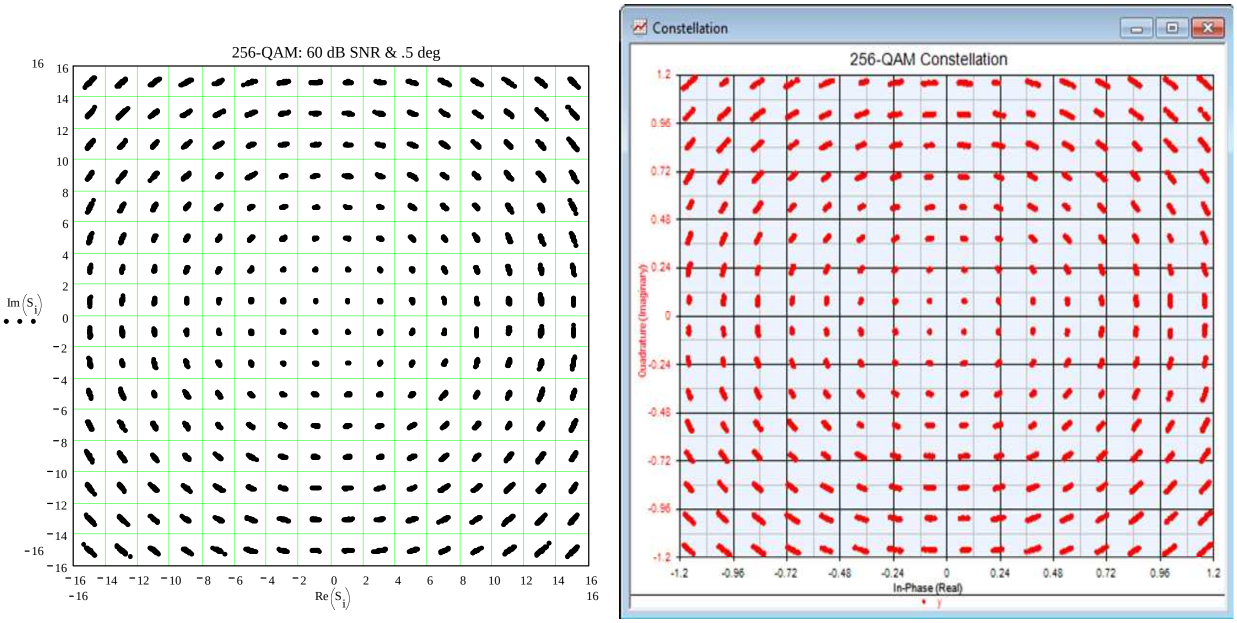


Figure 27 – 256-QAM, 0.5° rms Phase Noise (SNRφ = 41 dB), (Analysis, Simulation Tool)

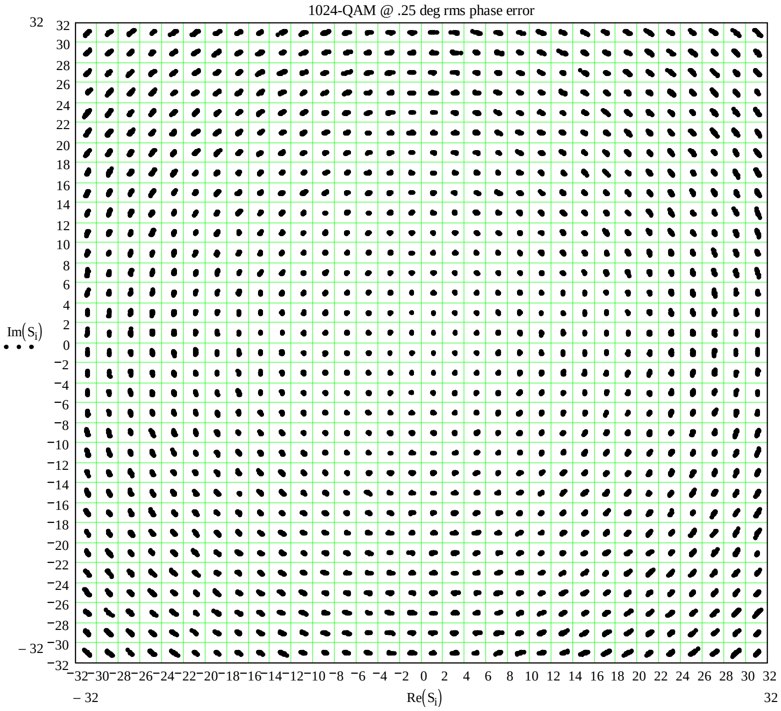


Figure 28 – 1024-QAM, 0.25° rms Phase Noise (SNRφ = 47 dB)

The nature of untracked phase noise is that it can lead to error rate floors at detection, because even without AWGN, if there is enough untracked phase noise after carrier recovery, it alone can cause symbols to cross boundaries. This is most likely for the outermost symbols, and these points can thus be used to determine limitations of phase noise necessary to eliminate flooring. More complex expressions are required to set thresholds associated with minimizing BER degradation [5, 6, 10].

It is a simple trigonometric matter to determine the rotational distance to a decision boundary as a function of M for M-QAM:

$$\varphi \text{ (decision boundary)} = \arcsin[(\sqrt{M} - 1) / M\sqrt{2}\,]$$

SNRφ Thresholds, M-QAM, and BER

Table 11 summarizes the phase error analysis across the modulation profiles of interest. It also identifies recommended levels of phase noise for minimal BER degradation, and levels beyond which the degradation curve shifts from a simple offset from theory to a more severe break from the normal steepness of descent of the BER waterfall curve.

Table 11 – Untracked Phase Noise Limits vs. M in M-QAM

	φ	dBc φ thresh	BER < 0.5 dB	SNR φ	BER on the Brink	SNR φ
16-QAM	16.8°	-10.7	1°	35.0	2°	29.0
64-QAM	7.7°	-17.4	.5°	41.0	1°	35.0
256-QAM	3.7°	-23.8	.25°	47.0	.5°	41.0
1024-QAM	1.8°	-30.1	.125°	53.0	.25°	47.0
4096-QAM	0.9°	-36.1	.0625°	59.0	.125°	53.0

The relationships shown can be deduced in part by recognizing that, since we are using a Gaussian statistical model for the jitter, the boundary merely represents a threshold on a normal curve that we can scale the rms (σ) to calculate its probability of threshold crossing. For example, a 16-QAM floor of about 1e-6 occurs for 4° rms, while for 64-QAM, a similar floor exists for 2° rms.

To exactly quantify allowable degradation with phase noise, AWGN and now phase noise can be combined together to create a composite "σ." However, they do not impact BER in a uniform fashion, as Figure 27 and 28 make apparent. Nonetheless, it is common approximation for lower order modulation formats to sum the AWGN noise and phase noise come up with a composite SNR. This simplification tends to understate the impact for high M, however.

Figure 29 shows a BER analysis that includes both phase noise (.25°, .35°, .5° rms) and AWGN contributions for 256-QAM [6]. The thresholds identified for 256-QAM in Table 11 are shown clearly on this chart by referencing the 1e-8 error rate threshold. The 0.25° rms value represents < 0.5 dB of degradation, while the 0.5° rms value has clearly is losing the characteristic waterfall shape, and on the verge of an error rate disaster.

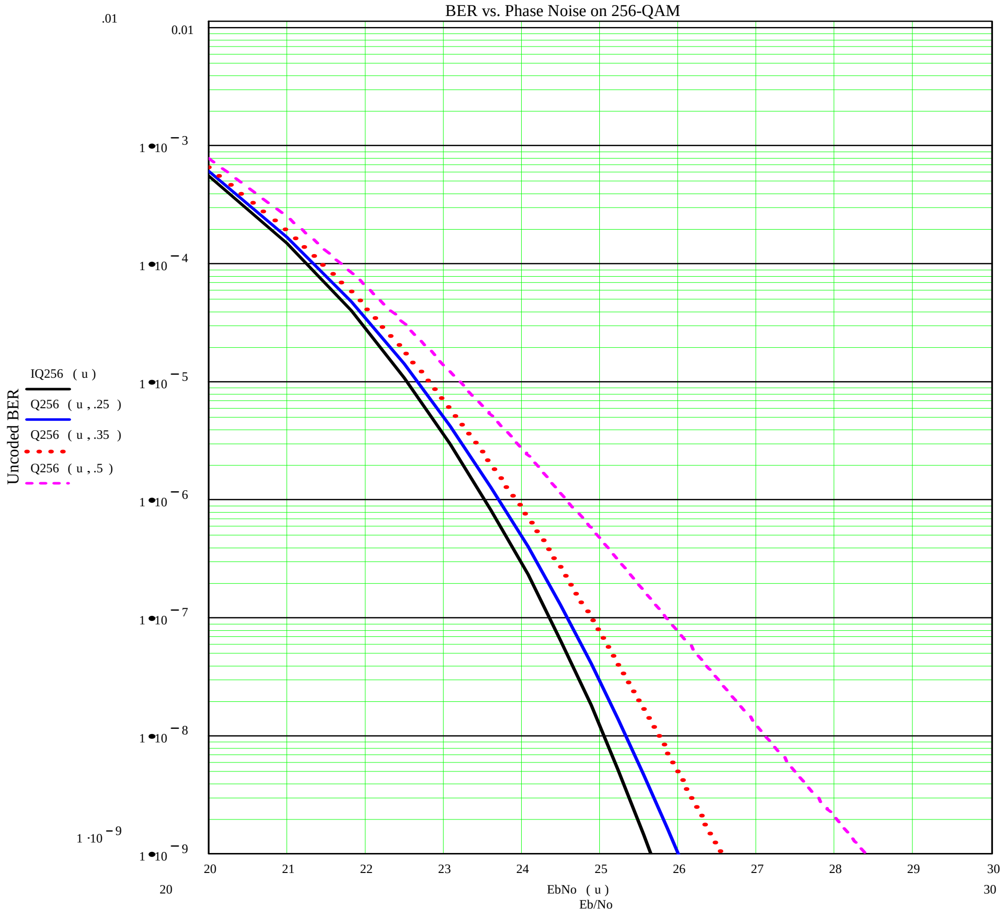


Figure 29 - 256-QAM BER with Phase Noise: .25°, .35°, and .5° rms

Note in Table 11 how the SNRφ required increases with increasing M. This relationship is similar to the relative relationship to AWGN. However, while a network architecture and new FEC may enable an AWGN performance improvement, the RF portion of the architecture that contributes to phase noise tends to remain in place and can be affected mostly in smaller ways by the tracking process. Redesign of RF equipment to achieve the same frequency agility objectives with improved phase noise is no minor proposition.

Offset M-QAM modulations and adjustments to decision boundaries in the face of a dominant phase noise impairment have been explored and this is addressed in [4] for the interested reader.

HFC Equipment Calculations

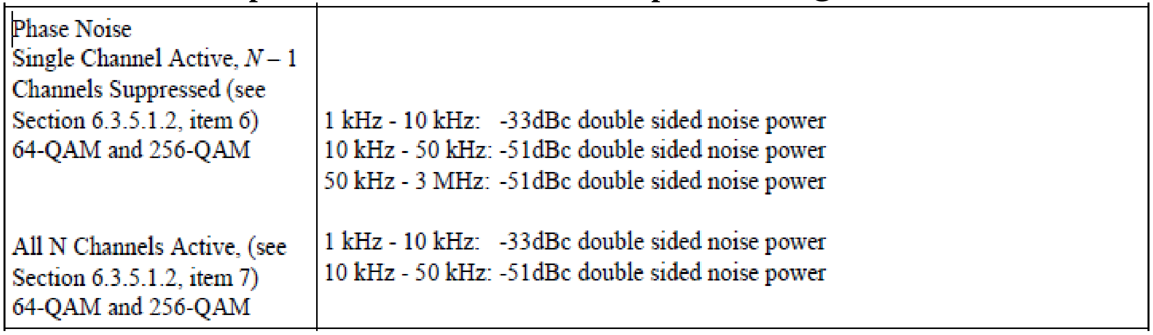
Phase noise is important because QAM, of course, encodes information in the phase of the symbol. A QAM signal contains “I” and “Q” orthogonal components, and the amplitude and phase applied to these identifies a point on a QAM constellation. This is why phase *noise* matters to M-QAM transport – noise in the phase domain translates to constellation position error or MER degradation in the signal space as we have seen in Figures 27 and 28.

The Downstream RF Interface Specification (DRFI), part of the DOCSIS portfolio of requirements, recognizes this and has a phase noise requirement, shown in Table 12.

All RF frequency synthesis or frequency conversion functions along the way contribute to the phase noise mask. The other typical major contributor in cable is the tuning function in the CPE. Though this function has been replaced in the RF circuitry sense by FBC technology discussed previously (wideband A/D conversion front ends), the clocking function of the A/D instead imparts the phase noise.

A modern wideband tuner built for digital cable and designed for compliance with ITU J.83A-C, is the Microtune MT2084. It has a specified phase noise requirement that serves as an excellent reference. These are shown in Figure 30.

Table 12 – Example Phase Noise Mask – RF Upconverter @ 601.25 MHz



Phase Noise Single Channel Active, *N* – 1 Channels Suppressed (see Section 6.3.5.1.2, item 6) 64-QAM and 256-QAM	1 kHz - 10 kHz: -33dBc double sided noise power 10 kHz - 50 kHz: -51dBc double sided noise power 50 kHz - 3 MHz: -51dBc double sided noise power
All N Channels Active, (see Section 6.3.5.1.2, item 7) 64-QAM and 256-QAM	1 kHz - 10 kHz: -33dBc double sided noise power 10 kHz - 50 kHz: -51dBc double sided noise power

Phase noise (SSB)
1 kHz offset -91 dB/Hz
10 kHz offset -92 dB/Hz
20 kHz offset -93 dB/Hz
100 kHz offset -105 dB/Hz
1 MHz offset -125 dB/Hz

Figure 30 – Sample Phase Noise Mask for RF Tuner in CPE

Since coherent QAM is used – meaning the carrier frequency and phase are recovered at the receiver in order to demodulate the signal and select which of the constellation points was transmitted, the final stage of "processing" of the phase noise mask occurs in the receiver. Carrier synchronization is performed by the carrier tracking subsystem. Modern designs use a decision-directed approach, which has been shown of the alternatives to have better noise performance, at least under low error rate conditions [18].

By tracking the carrier, a carrier recovery function is inherently also tracking the phase noise imposed on the carrier up to that point. However, it is a closed loop feedback system, and cannot track all of it without risk of other noise contributors, thermal and self-generated, from disturbing the stability of the recovery process. A feedback loop is in place which creates an error signal that is constantly adjusting the tuning oscillator to keep it aligned to the incoming signal. The feedback loop has a response time set by its loop bandwidth.

Without going into great detail about specific receiver architectures, the tracking occurs roughly up to the point of the loop bandwidth, and any RF-imposed phase noise beyond that is not tracked. There is an optimum bandwidth selection that considers these factors, input noise, and self-noise, among others. It is the total of untracked phase noise that contributes to MER degradation and possible symbol error. Note that DOCSIS specifications, in order to encourage innovation and competitive advantage among suppliers, allow flexibility on the receiver functions, where most of the complexity and sophisticated processing lie. As such, things such as loop tracking architectures or requirements are not defined, only end performance objectives under assumptions on channel conditions and other system assumption.

For the receiver designers, it is important to understand the associated transmit phase noise as defined in DRFI (CMTS or EQAM transmitter) and, for the upstream, in the DOCSIS PHY specification for cable modems:

-46 dBc, summed over the spectral regions spanning 200 Hz to 400 kHz
-44 dBc, summed over the spectral regions spanning 8 kHz to 3.2 MHz

Recall, the upstream has a range of symbol rates, beginning with 160 ksps (200 kHz) and increasing to 320 ksps (400 kHz) and so on in octaves up to 5.12 Msps (6.4 MHz). This range of symbol rates is reflected in the two requirements. We will assume wider symbol rates and thus the value of -44 dBc applies. Because of receiver design variations and the phase noise contributions from the receivers themselves, it is difficult to further quantify contributors to the phase noise process. We can say more will be added, some will likely be tracked out, and that the untracked jitter will have a lowpass structure to it.

We can at least, however, estimate what the specified requirements would mean to our M-QAM constellations, and estimate implications to receiver architectures using the requirements that are in place. Let's examine the effect of the combined DRFI specification and above tuner mask on the QAM profiles of interest. We have used the DRFI requirement in Table 12, and the tuner mask shown in Figure 30 to create a composite mask. This is shown in Table 13.

Table 13 – Composite Mask: DRFI + Tuner

Composite Mask (dBc/Hz)	
50 kHz	-90
100 kHz	-110
1 MHz	-130
3 MHz	-139
Total, dBc	-47

Assume that all of the mask beyond 50 kHz is untracked, and assume it is the dominant contributor to untracked phase noise after carrier recovery. It extends out to the symbol rate edge of 3 MHz (6 MHz double sided). This suggests a tracking bandwidth in the 50 kHz range, tied to other parameters [14, 15], and high SNR conditions in the carrier recovery architecture from self noise and input SNR.

As shown in Table 12 and 13, the composite mask is a 47 dB SNRφ, or .25° rms. The mask in Table 13 is shown on 64-QAM, 256-QAM, 1024-QAM, and 4096-QAM in Figure 31a-d.

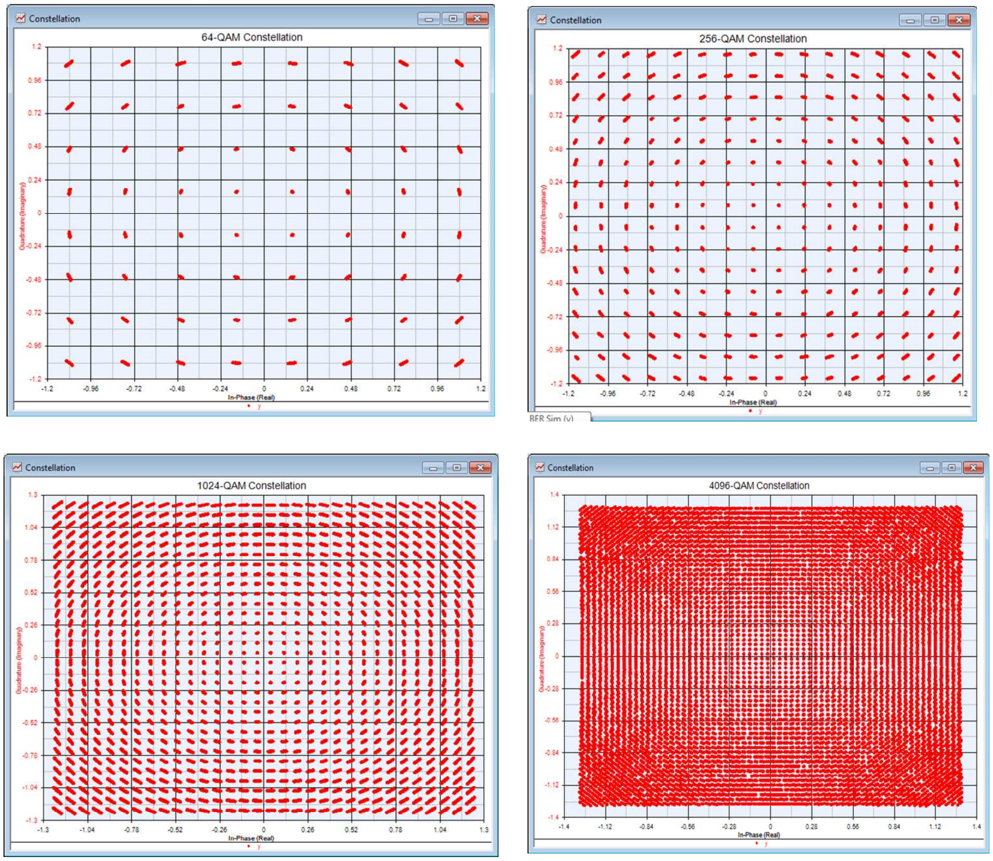


Figure 31(a-d) – Table 13 Mask Applied, Clockwise from Upper Left: a) 64-QAM b) 256-QAM c) 1024-QAM and d) 4096-QAM

In Figure 32, we have evaluated the uncoded BER for M-QAM profiles for M=64, 256, 1024 and 4096-QAM under the 47 dBc SNRφ conditions of Figure 31.

As Table 11 indicates, SNRφ = 47 dBc is the breakpoint between small degradation for 256-QAM, and the BER being on the brink of large degradation for 1024-QAM. The 4096-QAM case is untenable with this amount of untracked phase noise. This is all verified in Figure 32.

For 4096-QAM, which is clearly suffering and in practice would not be able to effectively hold the receiver locked for demodulation, it is difficult to see much in Figure 31 other than clouds of impossible-to-discriminate symbols. This case is shown again by itself in Figure 33, where you can begin to see some daylight between symbol points, mostly inner points. But, the symbol clouds at the edges are still massively intruding on each other's space to the point that they are becoming indistinguishable. This yields the very high symbol error rates, likely to overwhelm an error correction mechanism or carrier recovery subsystem.

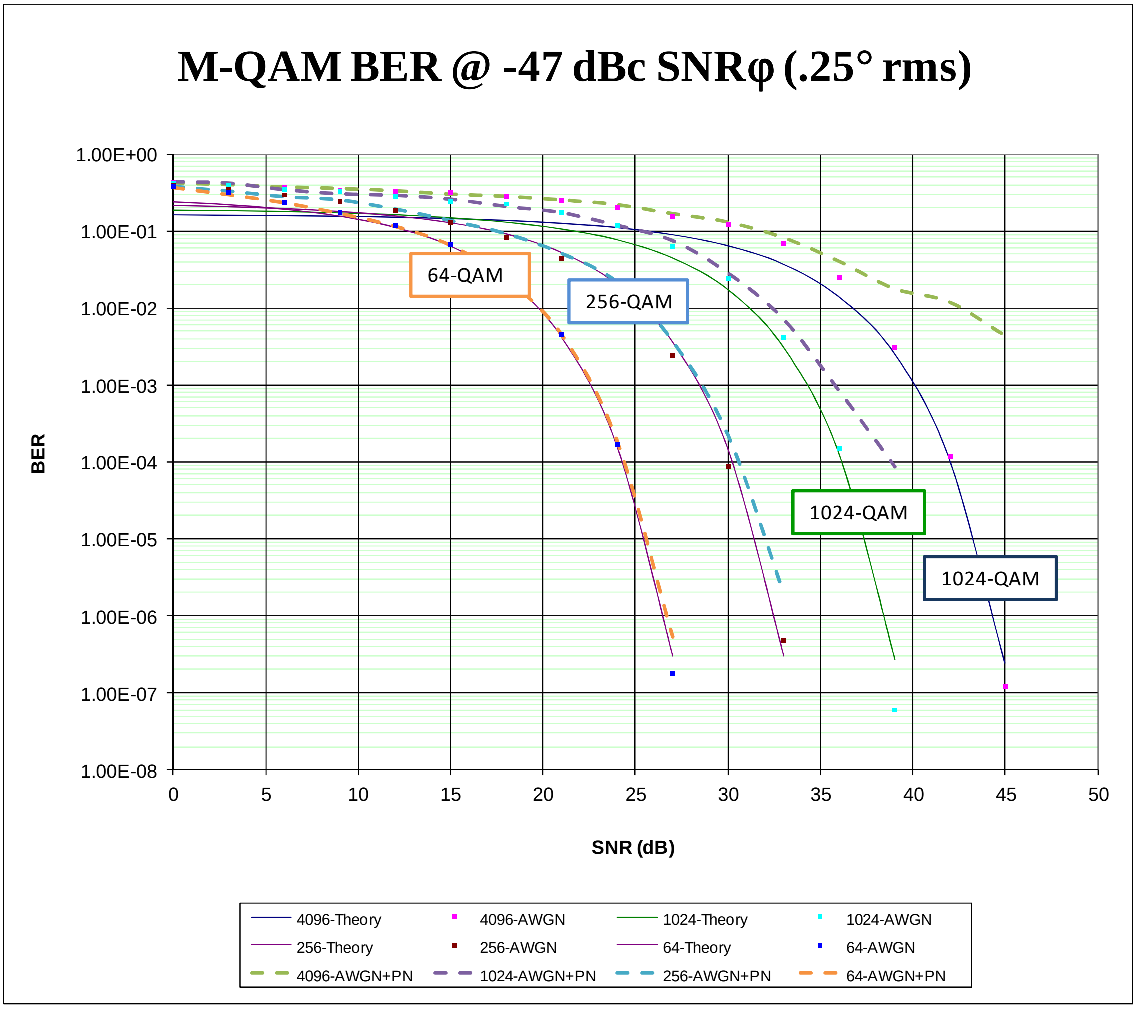


Figure 32 – M-QAM BER for SNRφ = 47 dBc (DRFI + Tuner)

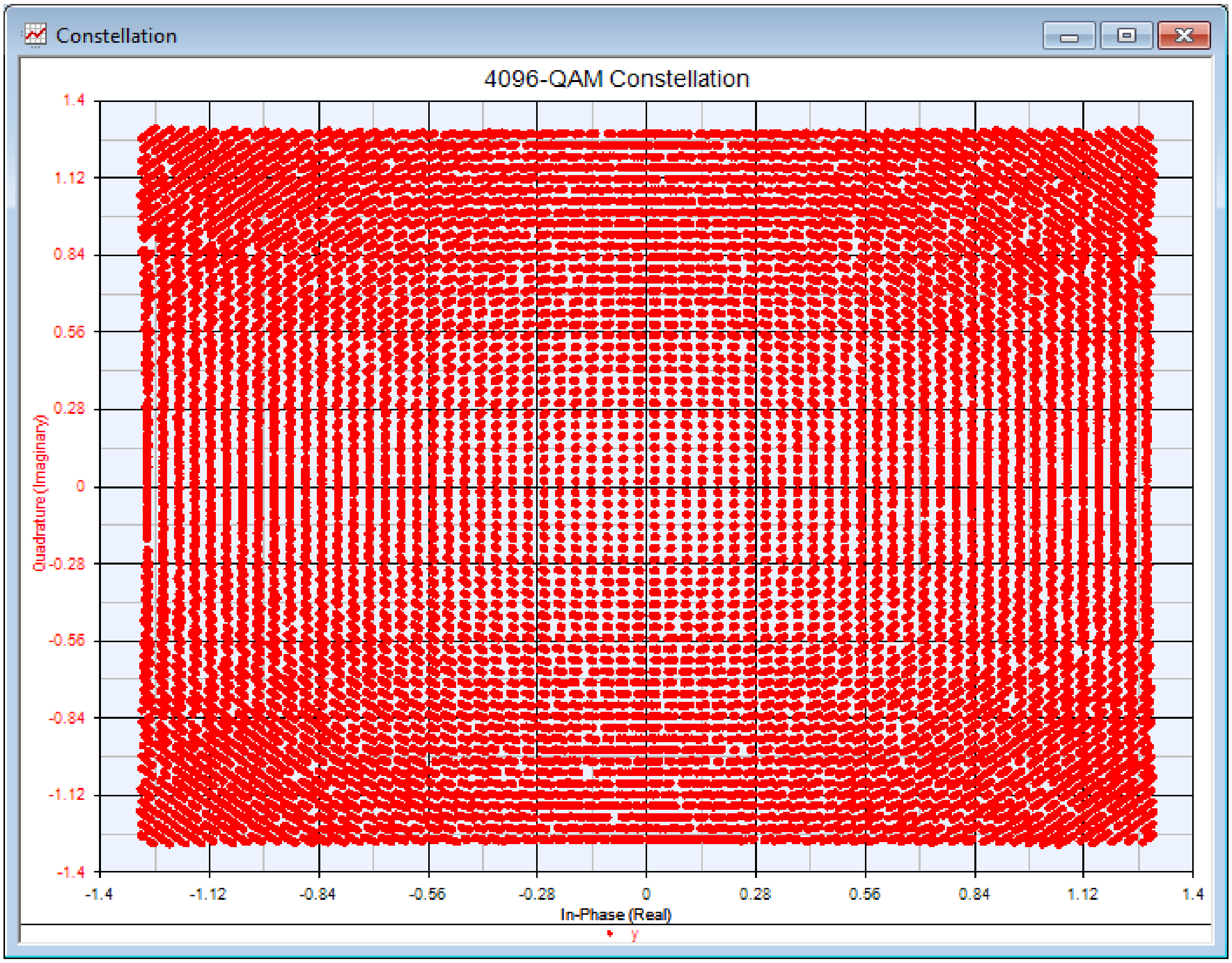


Figure 33 – 4096-QAM@ .25° rms (SNRφ = 47 dB)

According to the guidelines of Table 11, SNRφ = 53 dB is the brink of trouble for 4096-QAM, and a reasonable guideline for 1024-QAM that limits the degradation.

This 1024-QAM case, with SNRφ = 53 dB, is shown in Figure 34. The similar, relative MER characteristic compared to Figure 31b (256-QAM @ 47 dBc) is apparent. The 4096-QAM case for SNRφ = 53 dB is shown in Figure 35.

The BER evaluation for SNRφ = 53 dB is shown in Figure 36.

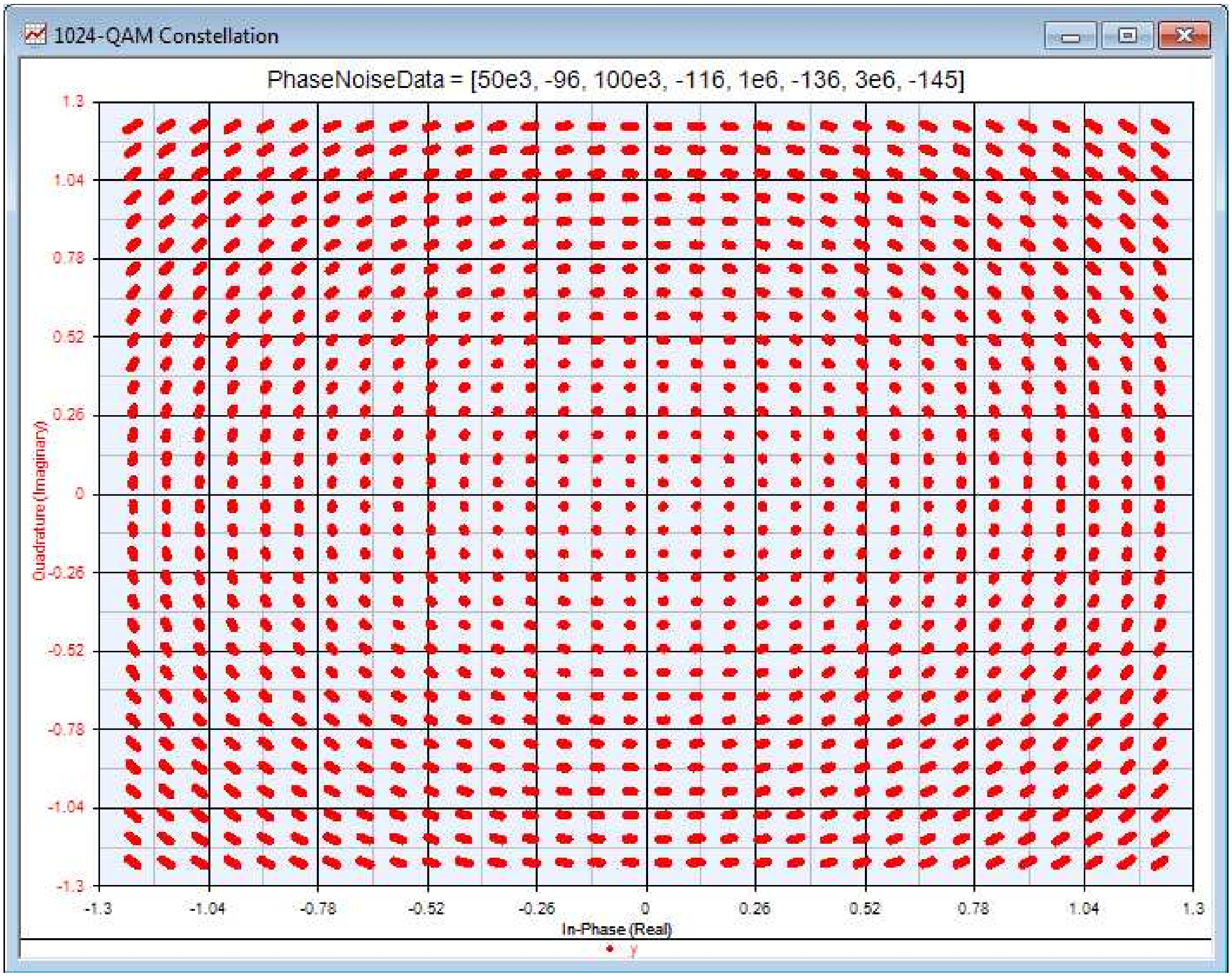


Figure 34 – 1024-QAM@ .125° rms (SNRφ = 53 dB) (Recommended)

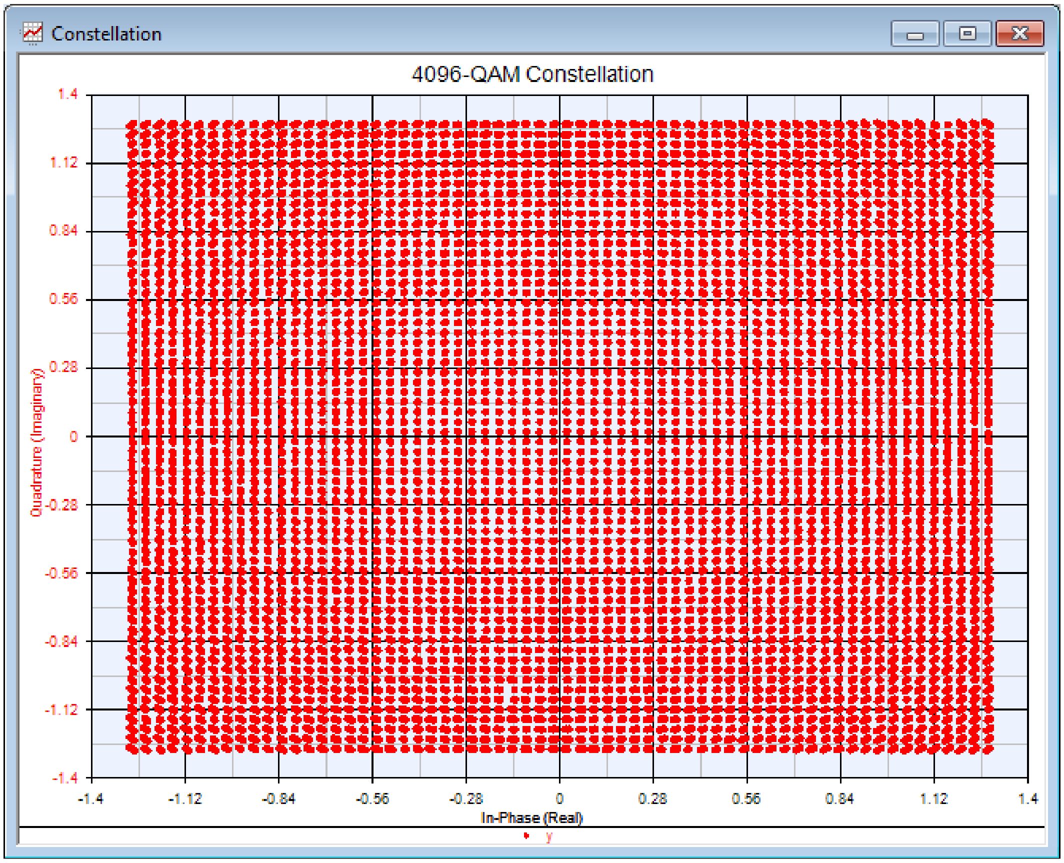


Figure 35 – 4096-QAM@ .125° rms (SNRφ = 53 dB)

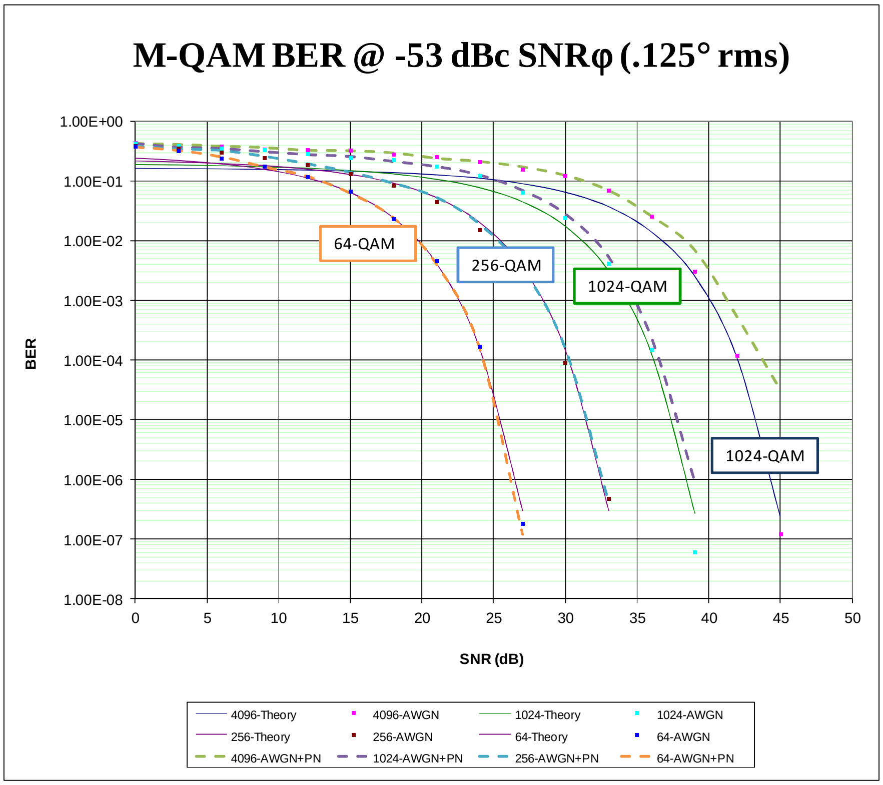


Figure 36 – M-QAM BER for SNRφ = 53 dBc

In Figure 36, we can see that 1024-QAM is now under control with modest degradation, and that 4096-QAM is on the edge of major BER performance degradation. This again is consistent with the recommendations in Table 11.

Finally, Figure 37 shows the constellation impact to 4096-QAM with the recommended maximum phase noise of SNRφ = 59 dBc, or .0625° rms. Of course, link phase noise is not going to adjust for the modulation profile, so the RF and tracking subsystem must be architected for the most sensitive modulation anticipated. The improved fidelity in Figure 37, in particular of the outer symbol points, illustrates why SNRφ = 59 dB is recommended for minimizing degradation against the theoretical performance curve.

Figure 38 shows the BER evaluation for the SNRφ = 59 dBc case, where it becomes clear that the 4096-QAM untracked rms phase noise recommendation of Table 11 is sufficient.

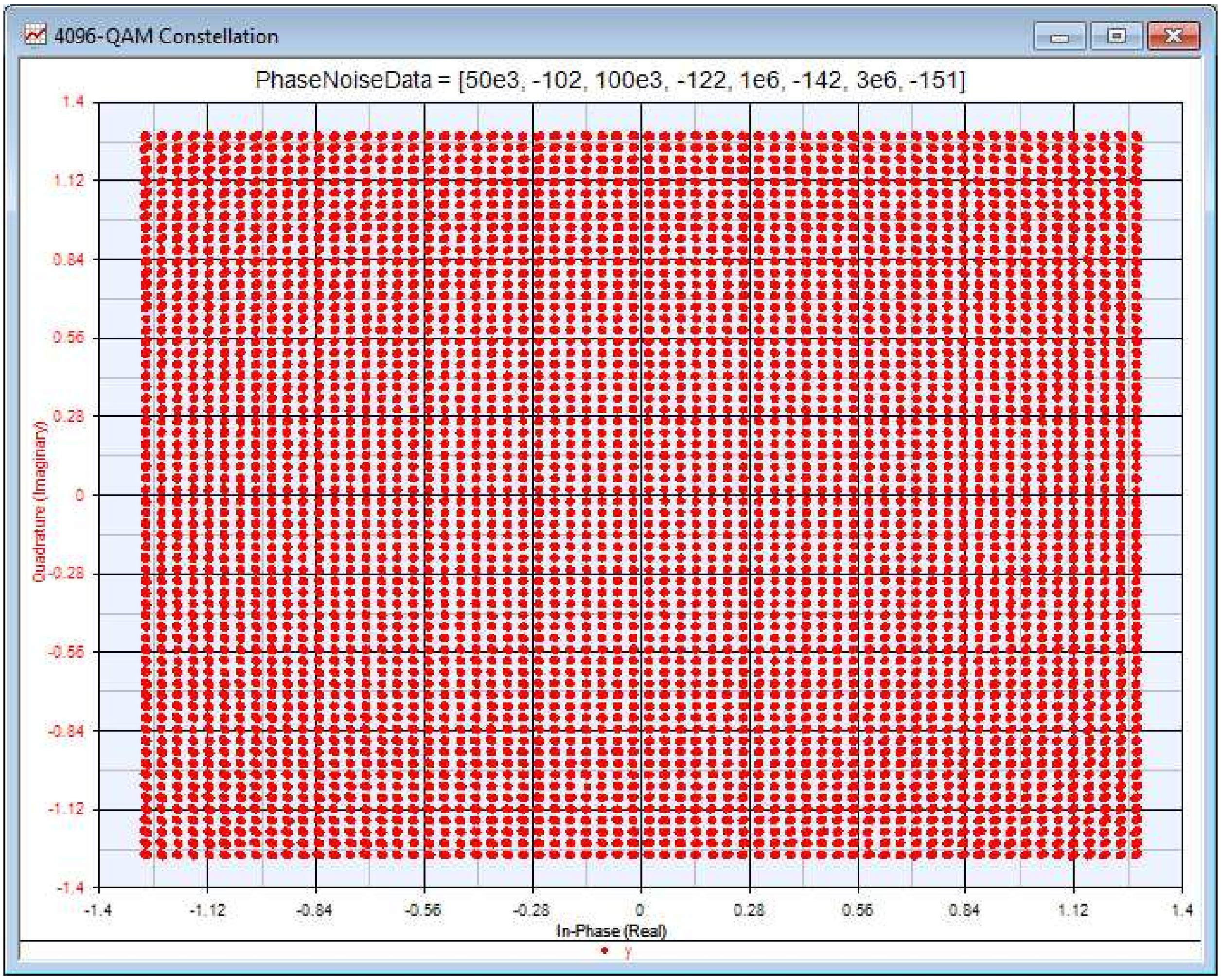


Figure 37 – 4096-QAM@ .0625° rms (SNRφ = 59 dB)
(Recommended)

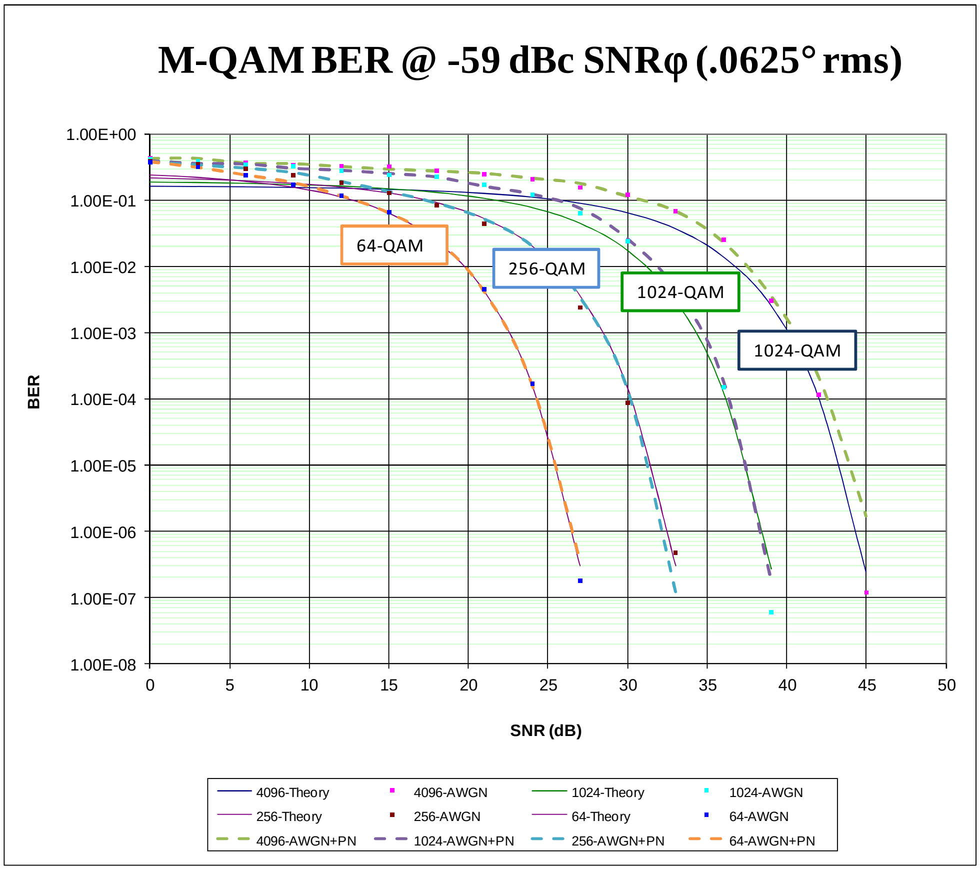


Figure 38 – M-QAM BER for SNRφ = 59 dBc

Post-Detection Processing

For high SNR systems, the loop bandwidth can be, relatively speaking, quite wide. However, it is nonetheless narrow compared to the symbol rates of single carrier QAM signals used in cable. This is important because it means that if there is enough phase noise to contribute to misplacing a symbol in the constellation, it will misplace potentially a large consecutive set of them for single-carrier systems. Because the loop bandwidth is much lower than the symbol rate, a sample of phase noise will be in about the same relative phase location for many symbols in a row – including when the sample is near a decision boundary or across one altogether. This is often referred to as the “slow” phase noise assumption, and is a common characteristic of single carrier QAM systems. The result is that phase noise, as an error mechanism itself, is bursty in nature. This puts pressure on the receiver to have burst correction either via FEC and/or interleaving. Reed-Solomon encoding is burst correcting, but the encoder in the J.83B downstream is only a t=3 symbol correcting design. It instead relies on the interleaver to provide a randomization of the symbol errors to make the RS decoding more effective, spreading out a burst of errors across codewords.

Fortunately, at least in the downstream, J.83B defines a very powerful, configurable, interleaver. It can configure burst protection from 66 usec to 528 usec (Level 2 mode with I = 128) at the expense of introducing latency. The lowest latency value is (I = 128, J = 1), where I and J describe the register structure used to feed Reed-Solomon codeword bits in and out. This setting provides 66 usec of burst protection at the cost of 2.8 msec of latency. Real time voice is the service that is typically most carefully watched for the latency budget, and 2.8 msec can be accommodated easily in a budget that targets around 50 msec typically one-way. I-128, J=4 is a recommended setting, contributing 11 msec of latency in exchange for 264 usec of burst protection.

The symbol rate of 5.36 Msps (256-QAM) works out to 187 nsec symbol periods. Using 50 kHz to represent the rate of the phase noise process, its “period” (it’s a noise process, so period is loosely used) is about 20 usec, or 107 QAM symbols for 256-QAM. The interleaver spreading exceeds 20 usec even for the lowest latency setting. Therefore, the interleaver is a very powerful helper against phase noise impairment - provided the native error rate is low to begin with.

The right-hand side column of Table 11 identifies rms phase noise thresholds that are at the edge of the native BER curve remaining stable. Because of the interleaving downstream, this column could be considered a target objective for the maximum allowable phase noise if it is within the budget of the FEC to support the error contributions from phase noise in addition to other channel impairments it may have been designed to protect against.

Measured performance is available for 1024-QAM in a pseudo “J.83” mode [9]. As shown in Table 14, pre-FEC errors are measured, and these are associated primarily with clipping and phase noise. In each case, however, post-FEC error rate is zero – meaning that the combination of the interleaver and RS FEC was able to completely eradicate any burst errors that may have been caused by the introduction of phase noise.

Unfortunately, in the upstream, we have potentially higher phase noise contributions specified, although it is specified over different ranges that may allow more of the transmit contribution to be tracked. This cable modem requirement was for SNRφ = 44 dBc. Upstream is likely to rely on lower orders of modulation, however, such as being limited to 1024-QAM. However, we have identified the 1024-QAM SNRφ threshold as 53 dBc, or 9 dB better than the cable modem requirement. This has important implications to the carrier recovery requirements in the burst receiver.

The upstream Reed-Solomon FEC is more powerful than the downstream, but still would not be capable of spanning a phase noise induced degradation of 50 kHz of noise bandwidth, much less as low as 8 kHz. This is more than five times the span, so represents about 5 times the number of symbols in error in a row at the highest upstream symbol rate – this likely outlasts the average burst size upstream entirely.

Table 14 – Pre-FEC 1024-QAM Error Rates with Zero Uncorrected Codewords

			1024-QAM Carrier Frequency		
			603 MHz	747 MHz	855 MHz
QAM @ -4 dB to Analog		MER	39.6	39.2	38.9
		BER	6.1E-08	1.12E-07	3.76E-07
QAM @ -6 dB to Analog		MER	39.0	38.9	38.6
		BER	1.5E-07	2.6E-07	2.5E-07
QAM @ -8 dB to Analog		MER	38.3	38.2	37.7
		BER	4.30E-07	2.02E-06	3.48E-06

As such, if impaired by phase noise, post-FEC results should register some low level of uncorrectable codewords. Since FEC is not a source of burst protection from a phase noise perspective, nothing is lost in moving from a RS-based FEC scheme to LDPC.

Note that SNRφ = 44 dBc is about .35° rms, which is plotted in Figure 29 for a 256-QAM – the state-of-the-art throughput available today [12]. In theory, this amount of untracked noise would lead to slightly less than 1 dB of degradation in an uncorrected BER curve. Based on the burst dynamics above, a post-FEC result would register some low level of uncorrectable codewords if there was a phase noise-induced BER contribution measureable. However, in [22], it is shown that there is error-free pre-FEC and post-FEC performance of 256-QAM upstream with SNR = 36 dB. However, this is consistent with the curve for .35° rms even if the SNRφ = 44 dBc is *all* untracked. Thus, it is not possible to learn whether or how much of an rms error reduction takes place in the carrier tracking process.

For purposes of upstream evolution, then, such as beyond 256-QAM, it is impossible to tell from 256-QAM performance, without additional measurements, whether there is adequate margin in the untracked rms phase noise to support 1024-QAM. However, without question, the current CM specification of -44 dBc over the specified bandwidth would be wholly inadequate without the ability to remove substantial induced phase noise in the carrier recovery process. This suggest these requirements may need to be updated to go beyond 256-QAM. The BER curve for 0.5° rms for 256-QAM in Figure 29 would be a reasonable approximation to the trajectory that the 1024-QAM BER would take for 0.25° rms, and this would of course get worse for 0.35°, meaning it would induce more than 2 dB of degradation at low error rates. Without interleaving, there would not be an opportunity to correct for this degradation in the upstream, so this bears consideration. Upstream phase noise for 1024-QAM may create the need for updated requirements.

In summary, to advance the modulation profiles, the phase noise requirements identified in DOCSIS and DRFI may need to be reconsidered to provide the spectral fidelity necessary to support very bandwidth efficient QAM transmissions such as 1024-QAM upstream and 4096-QAM downstream. In the upstream, 256-QAM has been shown to be supported today. It is inconclusive whether or not the phase noise margin contribution that is today adequate for 256-QAM is sufficient also for 1024-QAM. There is no mechanism in place to handle the burst noise environment phase noise-induced errors can create.

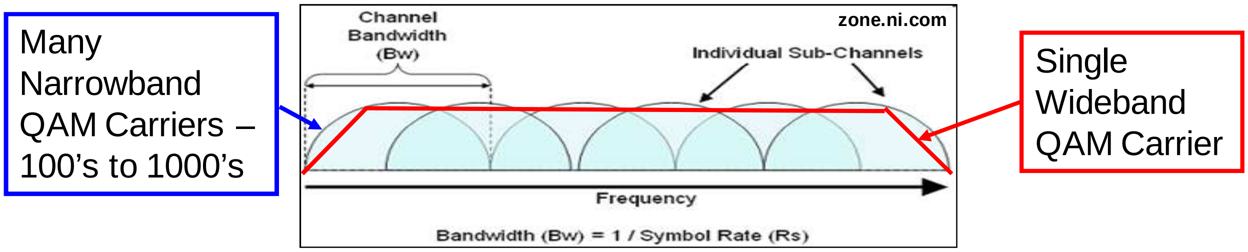
In the downstream, a measured post-interleaver, pre-FEC error floor suggests that there is some residual phase noise impact that is being handled well enough by these burst correcting mechanisms. Additionally, as shown in Figure 32, the error rate performance against the combined DRFI and tuner mask would be very poor for 4096-QAM – so badly so that the ability to manage an effective decision-directed tracking loop and successful decoding process is likely to be compromised. While the interleaver is very effective, there is an inherent assumption of low error rate to avoid overwhelming the interleaver-dispersed errors and the carrier recovery subsystem. Performance recommendations for total untracked rms phase noise for 1024-QAM and 4096-QAM are shown in Figure 34 and 37. Under these conditions, there would not be a heavy reliance on the interleaver, FEC budget, or concern about the sensitivity of decision-aided tracking robustness.

MULTI-CARRIER MODULATION

OFDM Applications to HFC

The industry is considering, as part of the IP transition, adopting a new RF waveform and fundamentally changing the access method away from a line-up of 6 MHz frequency domain multiplexed (FDM) slots. Wideband, scalable MCM or OFDM is being considered for the next generation of RF over HFC, for many of the reasons discussed previously about expanded bandwidths and RF channel uncertainties. There are many acronyms in use that describe an implementation of the same fundamental core concept: lots of narrowband carriers instead of one wideband carrier. Figure 39 illustrates the OFDM concept.

Historically, OFDM applications have been linked by a common thread – unknown or poor RF channels. Virtually all modern RF systems implement some form of MCM – 4G Wireless, MoCA, G.hn, HomePlug AV, 802.11n, and VDSL. The differences are based on the medium and channel conditions expected affecting the band of operation, subcarrier spacing, modulation & FEC profiles, bit loading dynamics, and whether the system is multiple access in the sub-channel domain (OFDMA). Table 15 lists some common Pros and Cons of OFDM.



Many Narrowband QAM Carriers – 100's to 1000's

Single Wideband QAM Carrier

Figure 39 – Fundamental Concept of Multicarrier vs. Single Carrier

Table 15 – Pros and Cons of Orthogonal Frequency Division Multiplexing

Pro	Con (*or Comment*)
Optimizes Capacity of Difficult Channels	High Peak-to-Avg (CPE issue); (*PAR reduction schemes exist - adds OH*)
Simplified Equalization against Frequency Response or Multipath	(*Cyclic Prefix = Guard time OH*)
Robust to Narrowband Interference	Avoidance Approach – Throughput Penalty by Deletion or Mod Profile
Robust to Impulse Noise	*(Similar Principles as S-CDMA - Time Spreading and Parallel Transport)*
Modern Ease of Implementation – IFFT/FFT DSP functionality	Complexity Increase for Shaping and Wavelet schemes – trade-off C/I vs. ISI
Simple Co-Existence via Flexible Subcarrier Allocation (and Power)	Backward Compatibility with DOCSIS
More Spectrally Efficient Wideband Channel than FDM Can be Multiple Access (OFDMA)	Potentially More Sensitive to Synchronization Noise Such as Carrier Phase Jitter (loss of orthogonality)

The most powerful advantage of OFDM has been that it shines in difficult or unpredictable channel environments. With the increasing ability to do computationally complex operations in real time, OFDM implementation – once an obstacle – has become a strength through simple IFFT/FFT functionality that forms the core of the transmit and receive operations.

For HFC, of course, this primary advantage is worth a closer look. The HFC downstream is one of the highest quality digital RF channels available – it is very low noise, and very high linearity. Such channels benefit very little in performance from OFDM, and probably not enough to justify introducing a new waveform if modulation efficiency of today's forward band was the only thing at stake.

However, as discussed, operators are looking for places to exploit more spectrum, and the channel quality of extended coaxial spectrum will be less predictable. This makes OFDM well-suited to be introduced in this part of the downstream band above 1 GHz as shown in Figure 1. Then, as the IP transition moves ahead and legacy 6 MHz slots are eliminated, the spectral flexibility of OFDM through allocation of its subcarriers becomes an especially valuable transition tool.

The HFC upstream, of course, does have a troublesome part of the band at the low end of the spectrum to which OFDM is a good fit. Today, the solution available to exploit capacity here is S-CDMA, which is just now seeing growth in interest and field deployment as upstream spectrum become congested and there is nowhere else to go, but down (in frequency). Like S-CDMA, OFDM should be robust at the low end of the band if properly designed for the impulse and narrowband ingress environments in that region of the return.

Unlike the downstream, above the low end of the band, as the upstream is extended above 42 MHz, there is likely to be steadily *improving* channel conditions, at least up to the FM radio band of 88-108 MHz. As in the downstream, this part of the upstream band – the extended upstream – may have a less obvious need for the primary poor-channel performance value of OFDM. But, there are some unknowns, and the FM band looms. And, since DOCSIS carriers will exist for many, many years, the flexibility of spectrum allocation once again makes OFDM worthy of consideration in this changing environment.

A second well-earned "pro" for OFDM is the simplicity with which poor frequency response can be combated. We discussed this as part of the capacity discussion in the beginning of the paper. However, OFDM also makes difficult multi-path channels more manageable. The HFC network is prone to "multi-path" in the form of micro-reflections associated with impedance mismatches that occur naturally over time and unnaturally through the fact that, as discussed in the section on the POE home gateway, every home in the plant is also part of the access network. Unlike the mobile application, the "multi-path" is static or nearly so. Nonetheless, because the upstream is burst mode from a randomly located source, it has dynamic characteristics associated with the allocation of time slots to modems that are basically on a single frequency but have individually dependent, but unique channel characteristics. OFDM enables the simplification of the equalizer function in these cases.

Perhaps the most talked about disadvantage of OFDM is its inherently high peak-to-average-power ration (PAPR). An OFDM signal is a collection of independently modulated carriers, all sent at once. As such, the composite waveform has noise-like qualities. This is a potential RF concern, as it requires more linearity, or higher P1dB, in the transmit power amplifier stages compared to single carrier signals to ensure the waveform does not get clipped and distorted. PAPR is primarily an issue for CPE – more transmit power headroom translates to more hardware cost. The same is true in principle (the need for more headroom) for the OFDM receiver, but the receive side is rarely tested from a distortion standpoint, processing very low level signals such that the dB differences have much less impact to design and cost. Schemes that encode subcarriers in a way that reduce PAPR have been developed.

OSI Layer 1 Standard?

An emerging analogy for OFDM is to liken it for OSI Layer 1 what Ethernet and IP are for OSI Layers 2 and 3. A complete, modern Layer 1 PHY is emerging as Multi-Carrier QAM with LDPC-based Block Code. The combination of the two drives implementation very close to theoretical capacity, so there is little else to optimize. There is a natural convergence of solutions towards this combination to yield the highest throughput efficiencies for a given channel. System parameters around the OFDM implementation would vary by application as a function of channel characteristics, as do the block sizes used for the LDPC code. The large number of modern systems based on OFDM is another important factor driving towards a PHY layer "standard" approach.

HFC is no different in this regard – looking for optimal ways to extract capacity on channels that are ill defined or know to be potentially troublesome, while adapting around legacy signals. OFDM or MCM is an alternative suited to these objectives, and some variant is a likely final evolution phase of the coaxial last mile, as introduced in [7].

Channel Impairments and OFDM

As discussed, for AWGN channels, the results relative to SNR and architectures above applies directly. While the HFC channel is never "only" AWGN, this assumption applies well to the HFC spectrum that generally represents the "good" part of the spectrum. There are often modest linear distortions comfortably handled by straightforward time domain equalizer structures. OFDM may achieve high performance with less implementation complexity in these cases, but single and multi-carrier systems would otherwise perform very similarly. The same can be said for the upstream, although the adaptive equalizer complexity in the upstream is much greater, so the weight of a simplifying architecture may be of more value.

However, it is interesting to point out that the maximum attainable bit rate expression on a channel for a given SNR is approximately the same for multi-carrier and single carrier when equalized by a DFE – the approach used today for the DOCSIS upstream [1]. The key difference, again, is that as the channel gets more difficult in terms of frequency response, the theory holds up well, but scale of implementation favors multi-carrier, and more so the worse the channel conditions become. In both cases, feedback from receiver to transmitter fully optimized the bit rate attainable.

Let's take a look at some of the impairment scenarios we quantified for single carrier and discuss how they relate to multi-carrier.

Signal-to-Interference

Single carrier techniques combat narrowband interference by notching the band through adaptive filtering mechanism, as previously discussed. OFDM, on the other hand, deals with narrowband interference by avoidance. Subcarriers that are imposed upon by an interferer are notched out our dialed down to a more robust (less bandwidth efficient) modulation profile that can be supported, all as part of an adaptive bit loading algorithm. The effect is a capacity loss, but generally a modest one. Note that single carrier ingress cancellation requires overhead itself (lost capacity) in order to operate.

CW Interference

A simple example of the capacity loss for OFDM can be calculated by recognizing the sub-channel spectrum for OFDM when implemented in a pure FFT-based architecture, the simplicity of which being one of the reasons it has become so attractive. The spectrum of a single FFT-based sub-channel is shown in Figure 40. The roll-off of the Sin(x)/x response is slow – 6 dB per octave – so the "in-band" rejection can as a result be low when an interferer falls onto a sidelobe of a particular sub-channel. The first sidelobe has the commonly referenced 13 dB relationship to the main lobe response

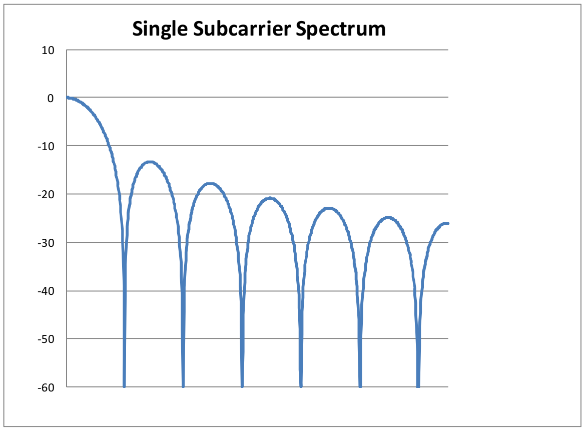


Figure 40 – FFT-Based OFDM Subchannel Spectrum

In Table 7, we noted that the A-TDMA ingress cancellation function had a limit of S/I = 10 dBc for zero corrected codeword errors for single CW ingress signal at an SNR = 35 dB. Three FM interferers could be as high as -15 dBc each (total S/I is still about 10 dBc). How would FM interference at -10 dBc affect the OFDM capacity? Figure 41 shows how several adjacent subcarriers appear in the midst of a CW interfering tone (not to scale of the numerical example).

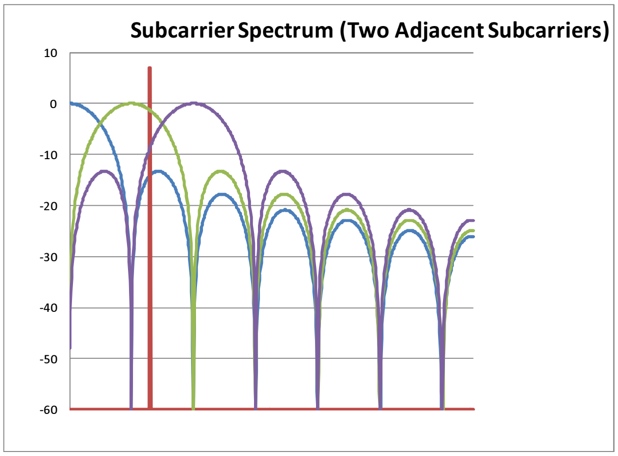


Figure 41 –CW Interference in an OFDM Channel

Let's assume channel SNR conditions are high, such that 64-QAM can be deployed. Going back to our 8k point FFT, each sub-channel is (1/8192) of the total, so the S/I on a per-sub-channel basis is (10-39) = -29 dBc on a subcarrier. Simulations performed such as were shown in prior results for 64-QAM indicate that a required 25 dB S/I for error free BER in this high SNR condition (35 dB). This would require 54 dB of rejection. If the interference coincides with a sub-channel frequency, then it does not interfere with adjacent sub-channels because of the same orthogonality properties that ensure that the sub-channels do not interfere with one another. However, this is unlikely. The worst case is it is just off center of a sub-channel so exposed to the envelope of Sin(x)/x roll-off of the spectrum in Figure 41. For rejection of 54 dB, this occurs at about 160 subcarrier indices away (320 total). If these sub-channels are all nulled, and all FFT sub-channels are used for payload, then the lost capacity is about 3.9%.

If instead of muting, for example, the adaptive bit loading tries to implement 16-QAM where possible, requiring only 20 dB S/I, then only about 160 sub-carriers *total* are lost, or 1.95% of capacity is lost to muting. There are then (320-160) 160 new subcarriers carrying 16-QAM, which works out to 0.65% of lost capacity, for a total of 2.6%. This is an improvement over muting all of them. Another subset could use 8-QAM, QPSK, etc. This is precisely how OFDM is handy for optimizing under varying channel conditions.

Alternatively, all of the above analysis was performed without considering new error correction. Our "new" SNR requirement is 26/32/38 dB for 64/256/1024-QAM, respectively. Simulations like those already discussed show that for these SNRs, we can arrive at S/I conditions that leave codeword error rates that are easily correctable (1e-4 or lower), for delivering low PERs. This would be a different set of dBc values to meet, but the same approach to the calculation of the carrier indices effected. This approach allots FEC "budget," built around AWGN performance, to correcting for interference induced errors, which would come at the expense of SNR to some degree. Error rate curves are very, very steep compared to classic uncoded waterfall curves, so this type of analysis trade-off would require careful simulation and test.

Modulated Interference

Let's assume the interference is FM modulated. Again referring to Table 7, zero correctable codewords required a 15 dB S/I for 64-QAM with high SNR. Now, however, at 20 kHz wide, its bandwidth is roughly that of an entire sub-channel, so looks like a noise floor increase.

Figure 42 shows the implications of an additive "narrowband" modulated interferer when applied to OFDM (not to relative dBc scale of S/I = 15 dB). On a per-sub-channel level, it represents -24 dBc. The main difference in the analysis approach is that we no longer need to refer to S/I behaviors to quantify how much rejection is needed.

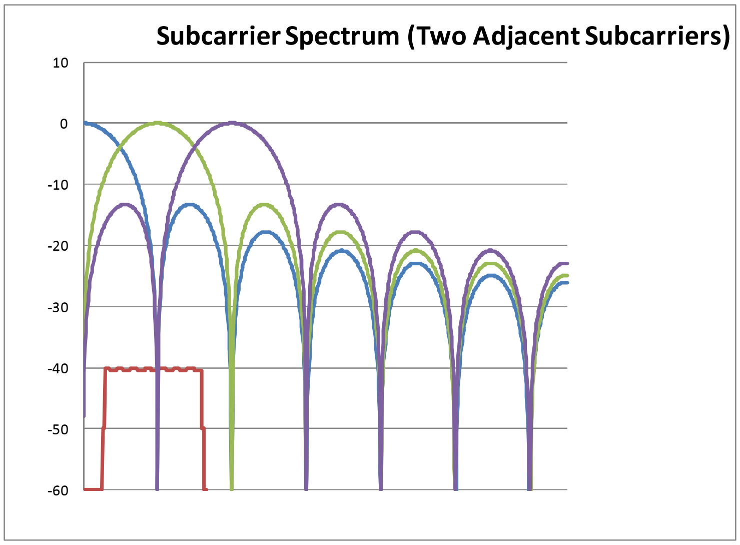


Figure 42 – Modulated "Narrowband" Interference for OFDM

We can treat it instead as a noise floor addition and refer to QAM profile performance against SNR. Modulation that creates a broad noise floor (relative to the sub-channel) and AWGN would not have the same precise effect, but that model is more relevant than a CW S/I model. Based on the thresholds used in Table 3 for upstream and 64-QAM (26 dB), the lost capacity would be close to the CW case. The differences would be in the S/I of a carrier index needing to reach 26 dB vs. 25 dB S/I, and the weighting that would apply for a broad spectrum applied versus a narrow carrier when passed through an FFT receiver.

As previously discussed, a reasonable argument can be made that the margin allotted to the QAM profiles in Tables 2 and 3, which are based on today's single carrier upstream channels, can be decreased *because* of a multi-carrier technique, as OFDM would be naturally more resilient to some of the items that contribute to the margin allotted.

Downstream Distortion Beats

We have noted that CSO/CTB interference has been a cause for concern for downstream QAM. We have also noted that, under the assumptions of analog reclamation, the CSO/CTB levels decrease dramatically. And it was noted that the bandwidth of these distortions is on the order of tens of kHz [20]. So, like the modulated interference previously described, this type of distortion is on the order of a sub-channel bandwidth for OFDM.

For full analog reclamation, the only number that matters becomes CCN, as all the distortions themselves become digital and spread across the spectrum in a noise-like fashion. Unlike the case of the FM interference above, beat distortion have an amplitude modulation component. Under the assumptions in Table 5, the worst case CTB indentified is 66 dBc. On a per-sub-channel basis, this becomes 27 dBc. Considering the noise-like peak-to-average in analysis makes sense for single carrier QAM because the distortion is "slow" in relation to the bandwidth, so peak samples exist for symbol after symbol. In the case of OFDM, the distortion bandwidth is on the same order of the sub-channel width (in this example), so noise averaging takes place just as symbol detection averaging does. As in the modulated interference case, we now can compare this 27 dBc to the modulated thresholds in noise environments to arrive at the impact to OFDM subcarriers.

Referring to Table 2, then, 256-QAM on a sub-channel interfered by this level of CTB would be supported in high SNR conditions (AWGN + CSO/CTB do not exceed the 25 dB shown). No capacity is lost in this case due to CTB for 30 analog carriers and 256-QAM. For 1024-QAM and 4096-QAM, however, threshold SNRs were identified as 31 dB and 37 dB. In both cases, only the sub-channel or two where the CTB falls will be impacted, since the spectral roll-off provides enough rejection (13 dB minimum) to meet these two SNR requirements.

This number of effected channels, too, can be calculated, as distortion noise "lumps" occur at periodic increments – two CSOs and two CTBs every 6 MHz (see Figure 20) – so there will be over 100 sub-channels imposed upon in the 192 MHz example discussed. This would be a maximum of 3.1% of the channels (128 = (192*4/6)). However, since we have shown that these channels could support 256-QAM, the capacity impact is only 0.62% for 1024-QAM and < 1.1% for 4096-QAM (less than 1.1% because some of the sub-channels could likely use 1024-QAM).

The same argument previously made about the FEC budget can be made here, which applies in particular for 4096-QAM. The SNR thresholds are based on AWGN, so it is the combination of AWGN and new noise contributors like CTB that should meet these thresholds. The "high SNR" assumption would then assume that this beat distortion is then the dominant effect in the sub-channels where it appears. When this is not the case, the offset for the addition of noise power must be made to guide the modulation profiles that can be supported.

For example, adjacent channel rejection of 13 dB from the 27 dBc example is 40 dBc. If our AWGN performance is 40 dB, supporting 4096-QAM with new FEC, then the two combine to 37 dBc, the 4096-QAM threshold identified in Table 2, and no capacity is lost in the adjacent channel. If instead we were already maximized at 4096-QAM with a 37 dB SNR, then the 40 dBc pushes the composite SNR closer to 35 dBc. In this case, a 2048-QAM profile may be required to have a robust channel performance.

Lastly, again, this combination of impairments, when coupled with sharp error rate functions that swing orders of magnitude on a dB of SNR difference, requires robust simulation to quantify precisely.

We nonetheless conclude that, in the case of a partial analog reclamation, OFDM should support the advanced modulation formats with little capacity degradation to do distortion interference.

Other Multi-carrier Approaches

Various shaping techniques and use of wavelets for orthogonality have been studied to reduce the effect of narrowband interference. However, these add complexity, and waveforms that provide narrower frequency response are inherently creating longer symbols in the time domain, so negatively affect performance on dispersive channels. By defining expected channel conditions, an optimum balance of time domain and frequency domain robustness can be implemented.

For an OFDM system for HFC, there is still some homework to be done on the system optimization side to determine the sub-channel spacing, shaping, and channel conditions anticipated and specified as new HFC bands become part of the RF channel definitions. We can count on improvements in SNR and downstream distortions, but updated frequency response and impairment models need a careful examination to ensure the HFC flavor of OFDM is optimized for its channel the way other OFDM-based systems in the wireless and wireline world have been optimized in their applications.

Phase Noise

OFDM creates an interesting scenario with respect to phase noise degradation. A core component of the analysis for the single carrier case discussed previously was built around recognizing that symbol rates for single carrier QAM generally exceed the frequency offsets where phase noise is prevalent. We used the example of 50 kHz of phase noise "bandwidth" to point out that the assumption for degradation is "slow" phase noise. A phase noise sample that rotates a constellation tends to be in the same place over many symbols in a row. If that phase error is close to a decision boundary or across it, there is likely to be a consecutive burst of errors.

Of course, with OFDM, we are using many, many narrow sub-carriers. For example let's use 192 MHz as a maximum OFDM bandwidth – consistent with coexisting with 6 MHz and 8 MHz forward path channel line-ups. An 8k FFT implementation for OFDM would mean that subcarriers are approximately 23.4 kHz apart. For a 16 k FFT, it would be 11.7 kHz.

Compare these to the 50 kHz phase noise “bandwidth” we were using earlier. This phase noise mask then extends *beyond* the sub-channel QAM symbol rate and beyond the main lobe of the OFDM spectrum implemented via FFT. Clearly, this is no longer a case of a phase noise process that is “slow” compared to the QAM bandwidth.

Let’s look at an example phase noise spectrum after carrier recovery. We saw a sample spectrum in Figure 24 of phase noise that would be imposed on a QAM carrier by an RF frequency conversion. For slow phase noise, the exact shape is not as important – it is the total rms noise that matters, because the phase noise power is dominated by “slow” or low frequency energy. As such, we referenced “dbc” values of total phase noise based on requirements currently in place. Some of the may be tracked out at the receiver, but in all cases for single carrier it can be characterized as “slow” except perhaps for the lowest upstream symbol rates, which are rarely implemented today.

For OFDM, however, the spectral content of the phase noise mask matters. Consider an example post-carrier recovery mask shown in Figure 43. This is the characteristic lowpass shape of untracked phase noise. It has components associated with RF phase noise imposition, additive noise, and self-noise of the carrier recovery process itself.

Now let’s take a look at how this type (two examples) of mask might look against an OFDM sub-channel spectrum. This is shown in Figure 44.

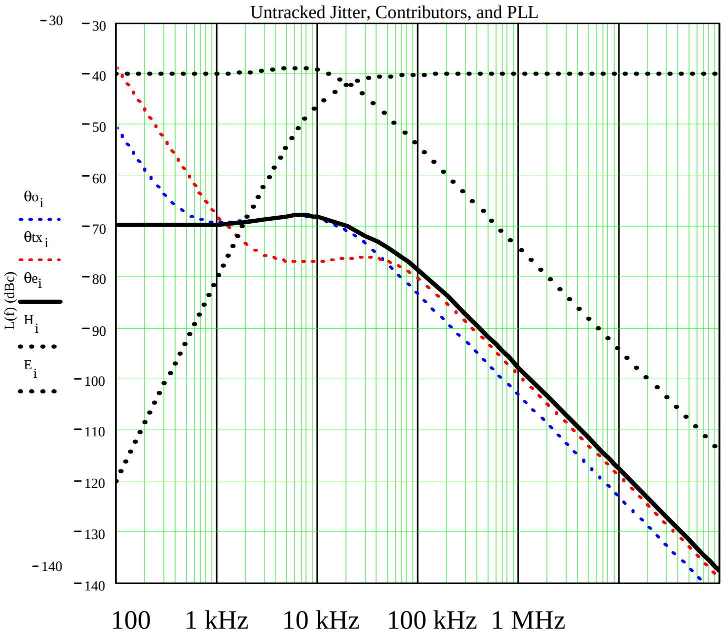


Figure 43 – Example Untracked Phase Noise Spectrum

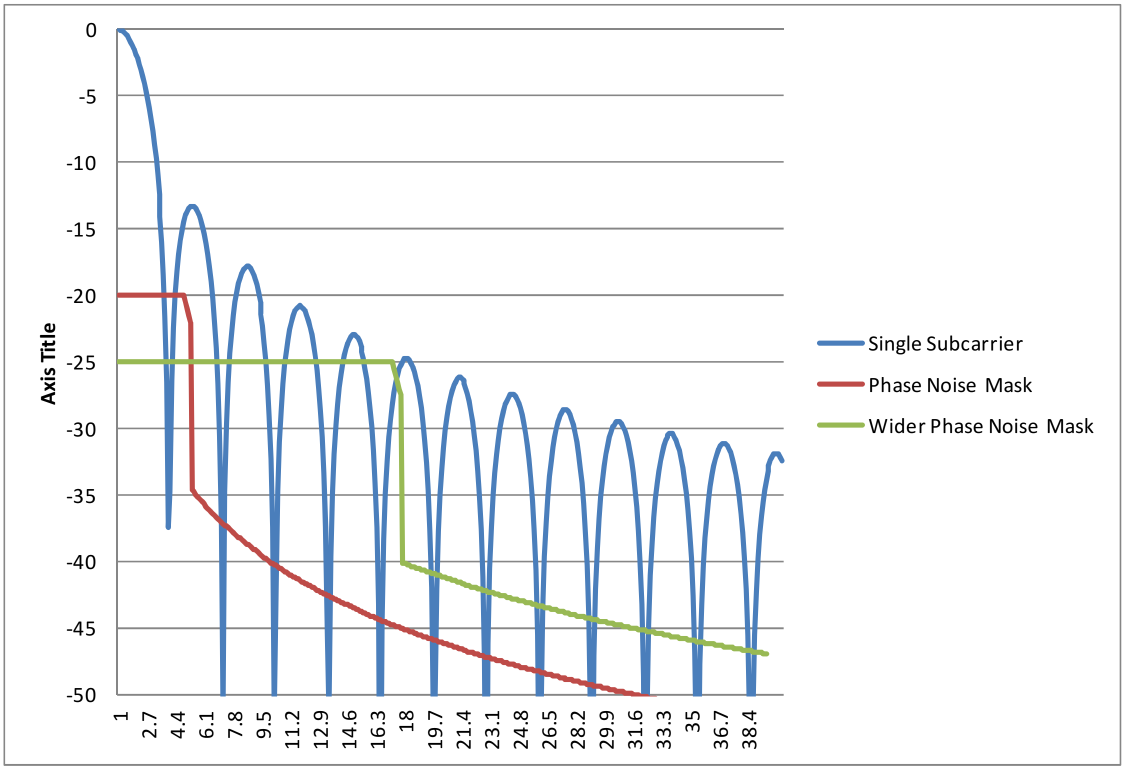


Figure 44 – Untracked Phase Noise vs. OFDM Sub-Channel

Two examples are shown. The red mask, for example, represents how the spectrum of Figure 43 relates to the 16 FFT discussed previously, with its roughly 12 kHz sub-channel spacing. Under this scenario, the 50 kHz of phase noise bandwidth we used earlier to discuss single carrier degradation is shown in green. Every OFDM sub-channel is effectively demodulated with the noise imposed by the phase noise mask.

However, as Figure 44 reveals, phase noise contributes two kinds of degradation to OFDM. There is an error common to all subcarriers related to the "in-band" effects – what for single carrier is the "slow" phase noise. Only, for OFDM, there is a good chance that the slow assumption is no longer valid, advantageously so in fact. It depends on the untracked mask – the shape or spectral occupancy of the phase noise is now important. Moderately varying or "rapid" phase noise allows some averaging over the bandwidth that the symbol is integrated over, and an average of a zero mean process is a better scenario than a single amplitude phase error sample.

However, there is also a component of phase noise that contributes to Interchannel Interference (ICI) as the masks cross into other sub-channel bands *because* of the relative relationship of phase noise mask to subcarrier spacing. This phase noise effect is additive looks nature, just as AWGN (mathematically easy to demonstrate sing the small angle assumption: exp(jφ) ≈ 1+jφ). In Figure 44 it is also clear that the noise in an adjacent band is a function of the phase noise spectrum itself (the shape) weighted by the Sin(x)/x response it leaks into. Not obvious from Figure 44 is that all of the subcarrier phase noise spectra combined create the full ICI effect. Because of this, and because the noise level is monotonically decreasing, the middle sub-carriers are the most effected by ICI due to phase noise.

The SNR degradation due to phase noise for OFDM has been calculated in many papers, and is simplified in [17] for a basic coherent receiver architecture as

$$\mathrm{SNR}(\mathrm{penalty}, \varphi) = 1 + \mathrm{SNR} * \varphi_{rms}$$

For less than 0.5 dB of degradation, this simply reflects that the rms phase noise would be about 10 dB better than the SNR itself – a common relationship when analyzing additive impairments, again verifying the ICI component of phase noise degradation for OFDM.

Comparing this to the assumptions in Table 11, we see that this is about 3 dB better per modulation when compared to the single carrier, no error correction, slow phase noise case. As discussed, slow phase noise and its angular rotation effect is more painful than an averaging of that noise or an additive effect such as in OFDM. The common phase error on all channels is less degrading when some of the energy is outside of the symbol bandwidth, and the energy that contributes to angular rotation is now no longer slow by definition. This appears to overcome the effect of additive noise contributions that leaks across and create ICI.

The study of these two effects has led to substantial research on the sensitivity of OFDM and studies of the proper carrier recovery approach for OFDM frequency and phase synchronization and manipulation of phase noise processes by transmission and tracking systems. In fact, you can find literature that indicates OFDM is *less* sensitive to phase jitter, or *more* sensitive to phase jitter. And, in fact they can both be correct because the nature of the relationship of phase noise to the QAM carrier has changed, and new variables come into play. The assumptions about those relationships affects the results, and a comprehensive analysis for an HFC version against phase noise mask requirements such DRFI would be necessary for 1024-QAM upstream and 4096-QAM downstream to be effectively deployed.

AND THAT'S NOT ALL FOLKS

We have offered some guidance here on requirements and impairments for new modulation profiles and access techniques, but also recognize that more information is required to make solid requirements and recommendations in many cases. Even so, we have not covered all of the potential angles of the analysis. As more work goes into defining advanced PHY profiles for HFC, we will consider yet the next level of details. This includes items such as new isolation requirements for new service on old and old service on new. And, discussion of the equalizer complexity issues of single carrier, or the pro-con trade-offs of different multi-carrier approaches. We have discussed carrier synchronization in depth, but not timing synchronization. Symbol degradation is a quantifiable problem by quantifying timing jitter relationship relative to the eye diagram and pulse shaping used. We also have not discussed timing requirements that become complex in OFDMA. All of these are important topics for future discussion, along with new depth and insights on what we have discussed here as more variables become known and information complete.

SUMMARY

In this paper, we have discussed HFC architectures and key variables for downstream and upstream in order to allow an increase in spectral efficiency and maintain robust performance. We have provided guidelines for system parameters and discussed specifications of equipment today and the implications of the requirements to support for long term bandwidth efficiency objectives. We have investigated the component parts from optical links to RF links, to CPE, and into the home itself. We have explored options for network architectures that extend beyond today's bandwidth and carrier access methods, and quantified how such shifts in network design may affect these choice. We have analyzed how these modern multi-carrier methods may be affected by HFC conditions today and moving forward. We have broken them down into downstream scenarios and upstream

All in all, the outlook is hopeful for cable operators to be able to exploit modern tools and enable more spectral efficient use of the network, prolonging its already healthy lifespan. However, indications are that some important changes to business-as-usual may be in store to ensure the required robustness on the most advanced modulation profiles – the silicon itself of course, but also outdoor plant architectures, potential requirements changes to create the fidelity conditions that support the modulation efficiencies of interest, changes in service delivery at the interface to the home, and comprehensively defining RF channels that heretofore have not been used by cable, but would necessarily be so to deliver on the capacity needs of the future. Indeed, there is much to do, and based on lifespan projections of service mixes ahead, now is proper time to be game-planning the transition.

APPENDIX – SIMULATION TOOL

Modeling Environment

Agilent's SystemVue was used to conduct system-level analysis of M-QAM with combinations of AWGN, static narrowband interference, and phase noise. The SystemVue model was comprised of random data source, transmitter, channel, receiver, and a data sink. Data streams at the source and sink were compared bit-for-bit to approximate system impact in terms of Bit-Error-Ratio (BER). Parameter sweeps were conducted for the relative RF levels for both AWGN and interference contributions within the channel.

Modeling QAM

Construction of the baseband signal first required a random sequence generator, such as a PN15 data pattern, operating at the system bit rate. As an example, the system bit rate for 4096-QAM is 12 bits/symbol multiplied by the symbol rate, in this case 5.360537 symbols/second, resulting in a system sample rate of approximately 64.33 Mbps. The random bit sequence was then fed into a symbol mapper, whose states were organized such that errors associated with misinterpreting a received symbol as an adjacent symbol would result in only 1 bit error in the symbol. After mapping, both the In-Phase and Quadrature (I and Q) baseband signals were filtered using root-raised-cosine (RRC) filters. Below are the impulse and frequency response plots associated with the RRC filters with an alpha of 0.12 (downstream excess bandwidth).

Figures A-1 through A-4 show the time domain pulse, the baseband pulse spectrum, and the RF spectrum (SNR = 28 dB) and constellation (SNR = 28 dB), respectively.

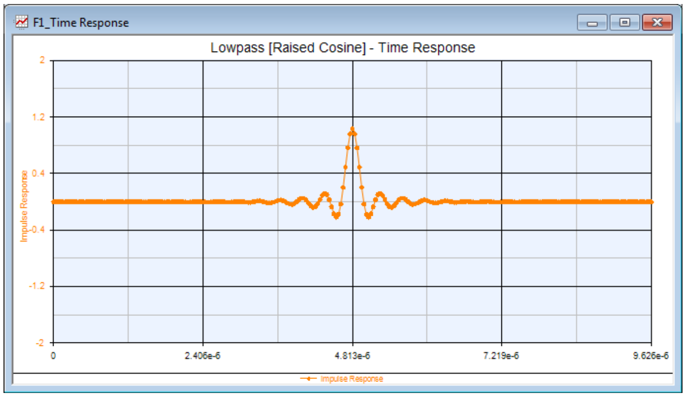


Figure A-1 – Time Domain Root Raised Cosine (RRC) Pulse Shape

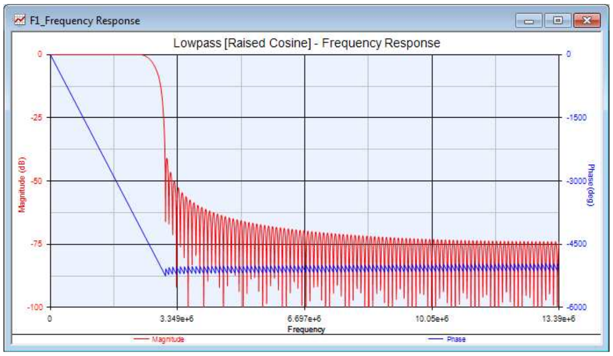


Figure A-2 – Root Raised Cosine (RRC) Spectrum

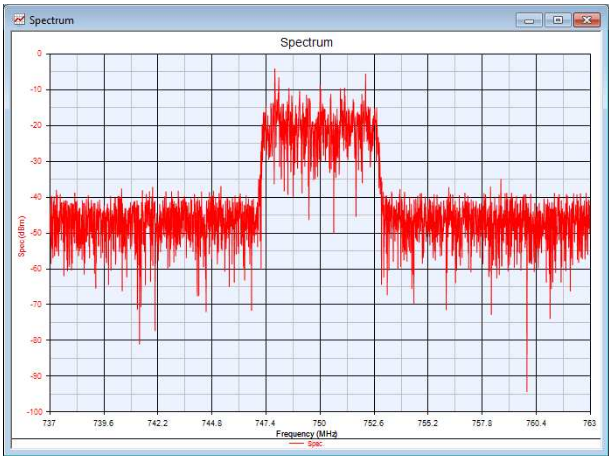


Figure A-3 – RF Spectrum with AWGN – SNR = 28 dB

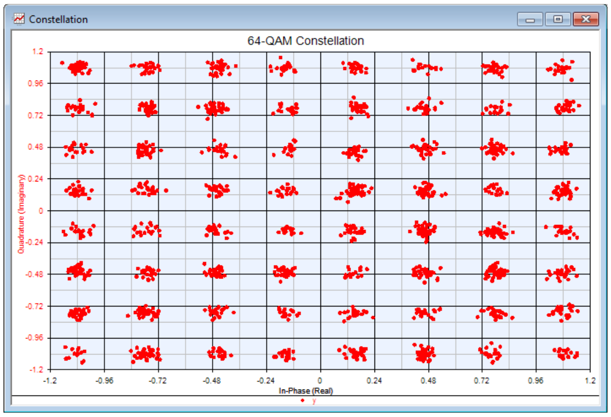


Figure A-4 – 64-QAM Constellation with AWGN – SNR = 28 dB

At the receiver, the signal is demodulated, RRC filtered and de-mapped using the same structures described above in reverse order. The response of 64, 256, 1024, and 4096-QAM to varying SNR measured against theory for uncoded transmissions is verified in Figure A-5. It can be seen that the simulated results track closely with theoretical expectations. This exercise provides the model basis for now extending channel impairments to items such as phase noise and narrowband interference.

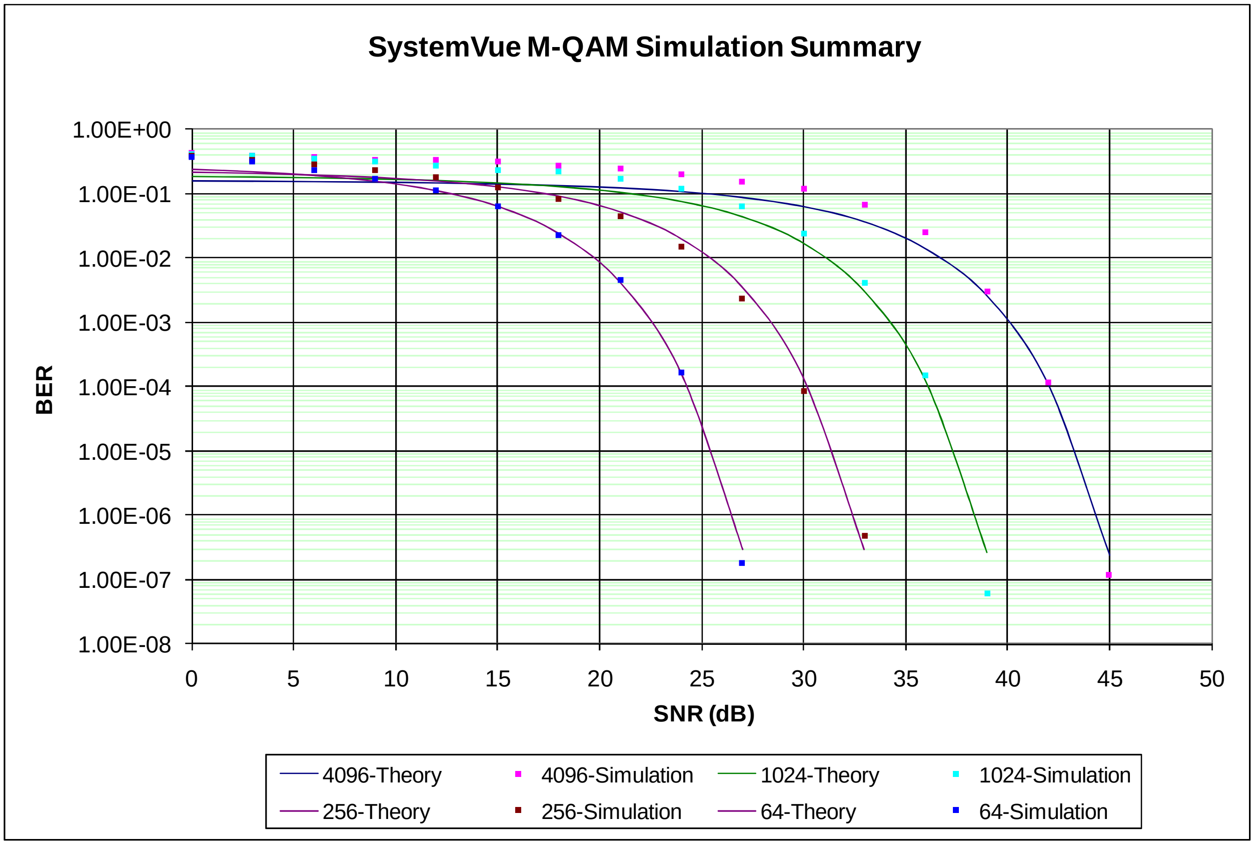


Figure A-5 – Simulated BER vs. Theoretical

REFERENCES

[1] Bingham, John C, *Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come,* IEEE Communications Magazine, May 1990.

[2] Chapman, John, Mike Emmendorfer, and Dr. Robert Howald, *Mission Is Possible: An Evolutionary Approach to Gigabit-Class DOCSIS*, 2012 Cable Show, Boston, MA, May 23-25.

[3] Howald, Dr. Robert, *Boundaries of Consumption for the Infinite Content World*, 2010 Cable-Tec Expo, sponsored by the Society for Cable Telecommunications Engineers (SCTE), New Orleans, LA, October 20-22, 2010.

[4] Howald, Dr. Robert, The Communications Performance of Single-Carrier and Multi-Carrier Quadrature Amplitude Modulation in RF Carrier Phase Noise, UMI Dissertation Services, 1998.

[5] Howald, Dr. Robert, *The Exact BER Performance of 256-QAM with RF Carrier Phase Noise*, 50th Annual NCTA Convention, Chicago, IL, June 10-13, 2001.

[6]Howald, Dr. Robert, *Fueling the Coaxial Last Mile*, 2009 Society for Cable Telecommunications Engineers (SCTE) Emerging Technologies Conference, Washington, DC, April 3, 2009.

[7] Howald, Dr. Robert, *Looking to the Future: Service Growth, HFC Capacity, and Network Migration*, 2011 Cable-Tec Expo Capacity Management Seminar, sponsored by the Society for Cable Telecommunications Engineers (SCTE), Atlanta, GA, November 14, 2011.

[8] Howald, Dr. Robert**,** Michael Aviles, and Amarildo Vieira, *New Megabits, Same Megahertz: Plant Evolution Dividends*, 2009 Cable Show, Washington, DC, March 30-April 1.

[9] Howald, Dr. Robert, *QAM Bulks Up Once Again: Modulation to the Power of Ten*, SCTE Cable-Tec Expo, June 5-7, 2002, San Antonio, TX.

[10] Howald, Dr. Robert L. and John Ulm, *Delivering Media Mania: HFC Evolution Planning*, 2012 SCTE Canadian Summitt, March 27-28, Toronto, ON, Canada.

[11] Howald, Dr. Robert and Phil Miguelez, *Upstream 3.0: Cable's Response to Web 2.0*, The Cable Show Spring Technical Forum, June 14-16, 2011, Chicago, IL.

[12] Howald, Dr. Robert L., Phillip Chang, Robert Thompson, Charles Moore, Dean Stoneback, and Vipul Rathod, *Characterizing and Aligning the HFC Return Path for Successful DOCSIS 3.0 Rollouts*, 2009 SCTE Cable-Tec Expo, Denver, CO, Oct 28-30.

[13] Lindsey, William C. and Marvin K Simon, Telecommunication System Engineering, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[14] Mengali, Umberto and Aldo N. D'Andres, Synchronization Techniques for Digital Receivers, Plenum Press, New York, 1997.

[15] Miguelez, Phil, and Dr. Robert Howald, *Digital Due Diligence for the Upstream Toolbox*, 2011 Cable Show, Chicago, IL, June 14-16.

[16] Piazzo, L and P. Mandarini, *Analysis of Phase Noise effects in OFDM Modems*, Technical Reprt No. 002-04-98, INFOCOM Dept. University of Rome "La Sapienza", May 1998.

[17] Proakis, Dr. John G, Digital Communications, McGraw-Hill, New York, 2001.

[18] Robuck, Mike, *Cox, Motorola lay claim to new return path speed record*, CedMagazine.com, March 01, 2011.

[19] Stoneback, Dean, Robert Howald, Tim Brophy, and Oleh Sniezko, *Distortion Beat Characterization and the Impact on QAM BER Performance*, 1999 NCTA Convention, Chicago, IL, June 13-16.

[20] Stott, J., *The Effects of Phase Noise on COFDM*, EBU Technical Review, Summer 1998.

[21] Thompson, Robert, *256-QAM for Upstream HFC Part Two*, 2011 SCTE Cable-Tec Expo Atlanta, GA, November 15-17, 2011.

[22] CableLabs, Inc., *Return Laser Characterization Techniques - Results and Recommendations*, Engineering Report, September 21, 1999.

[23] Data-Over-Cable Service Interface Specifications Physical Layer Specification (DOCSIS-PHY), CM-SP-PHYv3.0-I08-090121, January 21, 2009, Cable Television Laboratories, Inc.

[24] DOCSIS Downstream RF Interface Specification (DRFI), CM-SP-DRFI-I10-100611, June 11, 2010, Cable Television Laboratories, Inc.

ACKNOWLEDGEMENTS

The authors would like to thank Stuart Eastman for his invaluable contributions to this paper.

TRANSFORMING CABLE INFRASTRUCTURE INTO A CLOUD ENVIRONMENT

Gerry White
Motorola Mobility Network Infrastructure Solutions

Abstract

The paper outlines a methodical evolution strategy from today's RF centric, headend-based infrastructure to a digital-centric one, taking advantage of mature Internet, cloud computing and data center technologies. It presents a phased approach, identifying incremental steps leading to the ultimate goal of an efficient delivery infrastructure, and most importantly one that is aligned with Internet and data center technologies, and henceforth is able to leverage their continued development. Each transitional step is evaluated in the context of current and expected changes in technology, products and services. For each step the advantages provided are highlighted and potential risks are noted.

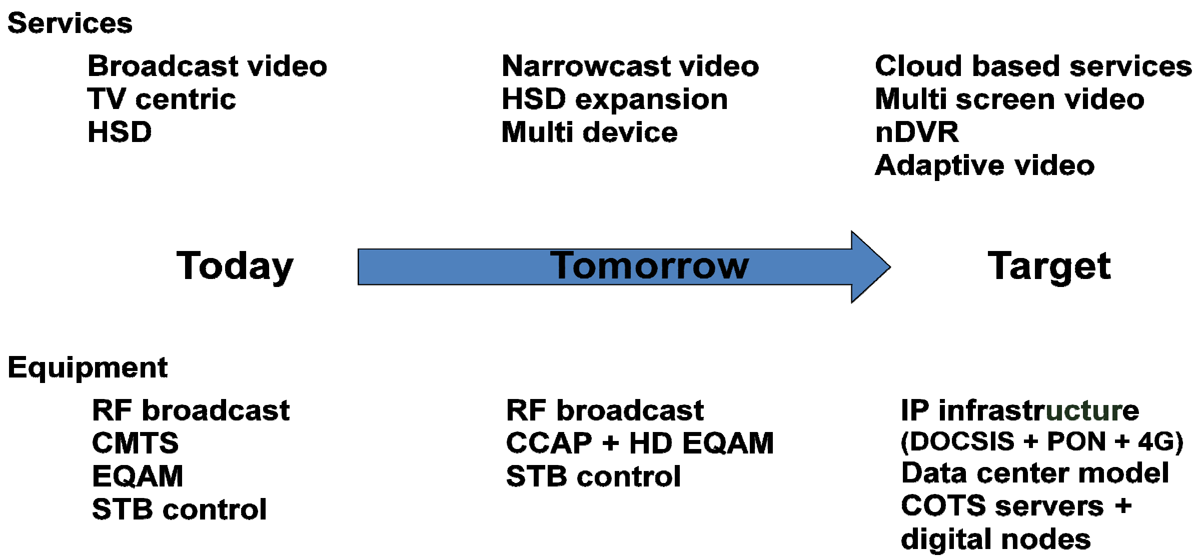
A number of practical options to deploy subsets of the phases are provided, depending on the individual circumstances of an operator, such as service needs and timing, network characteristics, and risk tolerance.

INTRODUCTION

One of the most significant developments in service delivery in the last few years has been the evolution of cloud computing and the massive data centers used to deliver it. The combination of high bandwidth network connections together with low cost computing and storage platforms has enabled companies such as Amazon and Google to deliver sophisticated services at very low cost points. To date, cable operators have used some of this technology for services such as TV Everywhere but in general it has been competitors such as over the top (OTT) video providers who have best leveraged the new technology. This paper proposes a way for the cable industry to take better advantage of data center technology and outlines a number of steps to achieve the transition.

EVOLUTION

In order to speculate on the evolution of the cable network infrastructure we need to consider the evolution of the services which it must support. Figure 1 shows these parallel evolutionary paths.



Services

Broadcast video
TV centric
HSD

Narrowcast video
HSD expansion
Multi device

Cloud based services
Multi screen video
nDVR
Adaptive video

Today → Tomorrow → Target

Equipment

RF broadcast
CMTS
EQAM
STB control

RF broadcast
CCAP + HD EQAM
STB control

IP infrastructure
(DOCSIS + PON + 4G)
Data center model
COTS servers +
digital nodes

Figure 1: Service and Equipment Evolution

<u>Service Evolution</u>

Current services are heavily focused on delivering linear video programming to a broadcast audience via an STB/TV combination.

In the immediate future we expect to see significant expansion of this service set as narrowcast video services delivering video on demand (VOD) and network based DVR to STB/TV platforms are deployed in parallel with the broadcast service. At the same time high speed data service expansion with compound annual growth rates in the order of 50% is being driven by over the top video services delivered to both TVs and other screens [SAND].

Looking further into the future operators will continue to expand on demand and nDVR services increasingly moving from broadcast to narrowcast services. At the same time competition with OTT video providers will require MSO's to deliver equivalent or better services to multiple types of CPE devices, in the home and on the road. Thus cloud based adaptive video services to multiple devices will become a key component of the service mix.

<u>Infrastructure Evolution</u>

The cable infrastructure must evolve in parallel with the services. Today video services are delivered over an RF broadcast infrastructure, primarily to set top boxes using a proprietary control system. High speed data services are delivered via a parallel CMTS based infrastructure sharing the same physical HFC network as video services but little else.

In the immediate future, as high speed data and narrowcast video expands, existing CMTS and EQAM equipment will be augmented or replaced by higher density platforms to add more narrowcast channels. Systems based on the CCAP specifications will combine CMTS and EQAM functions into a single edge platform but retain the same frequency division multiplexing to share the HFC network and data and video will continue to use independent control systems.

As multi screen video delivery expands the rapid deployment and short lifetimes of the CPE devices will require that service delivery to these devices be based on standard internet protocols with minimal or no changes for the cable infrastructure. The infrastructure must evolve to support IP video delivery to these devices. Thus over time more video will move to an IP delivery mechanism sharing the same resources and technology as high speed data services. This will use existing IP backbone technology for distribution to access networks. The access networks will initially be based on DOCSIS but will include PON and wireless alternatives. This move to a standard IP solution enables the use of standard data center and cloud based services to reduce costs and increase service velocity.

CLOUD SOLUTION

Currently the cloud based environment provides a platform for applications such as OTT video which run over the cable operator's IP broadband service. OTT vendors have taken advantage of cloud services to improve efficiency. Operators have followed suit to some extent with their own OTT like offerings but in terms of leveraging cloud technology are at best on a par with their competition. The remainder of the paper examines how the operator can gain additional advantages from cloud technology by migrating some components of the broadband service itself into this same cloud environment. The technology steps required and the benefits and impacts it may have will be reviewed.

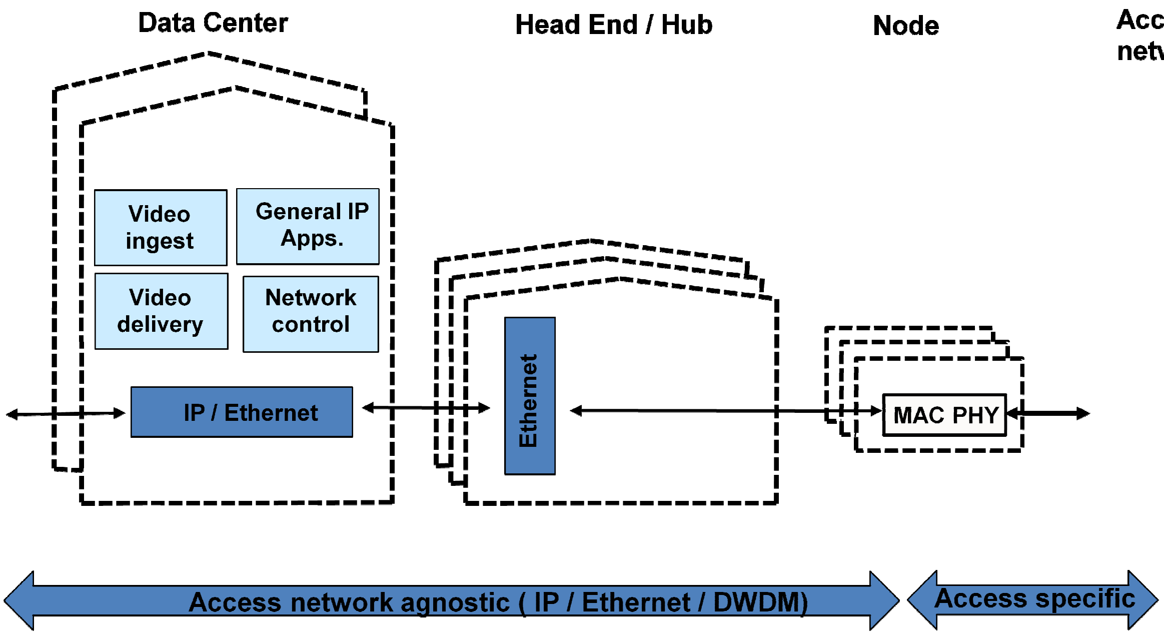


Figure 2: Cloud Centric Solution

Figure 2 shows a possible implementation of this type of architecture illustrating the major components and their location in the network. The philosophy behind this approach is simple; to put as much functionality in the data center as practical, to leverage Ethernet and digital optics where possible and to keep the node simple. The reasons to migrate functions to the data center are to leverage off the shelf hardware and software for reduced cost, to leverage virtualization for redundancy and scaling and to centralize complexity for simpler operation. Ethernet and digital optics are used to provide low cost and long distance options. The node is kept simple for low cost and ease of operation. As a consequence of the migration of functions to the data center and node the head end and hubs become much simpler.

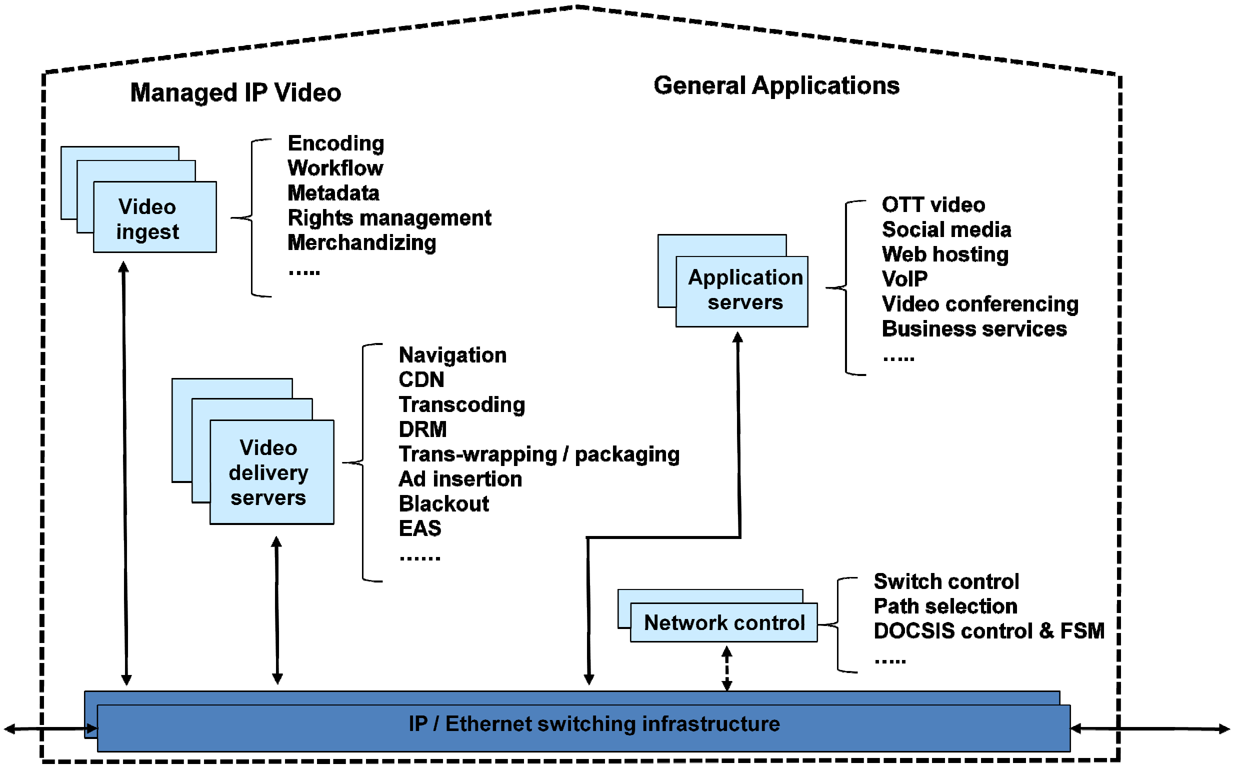


Figure 3: Data Center Functions

Data Center

As with existing cloud based services the data center is based on off the shelf servers running general applications software in a virtual environment as shown in Figure 3. These include general applications software such as OTT video, social networking and Internet access.

Additional servers provide the functions needed to create a multi-screen video service. Video ingest servers receive content from multiple sources and provide encoding, metadata and content management functions.

Video delivery servers provide transcoding and packaging functions to create multiple bit rate program streams along with additional functions such as advertising insertion, and content protection that are required by a full service video provider. A description of a layered architecture for an end to end IP video system can be found in [MSIPD].

This is a simplified description in that these functions may be implemented in a central data center or in a more distributed model based on multiple data centers interconnected with CDN distribution networks. For the purposes of this paper the simple model will suffice as the evolutionary steps proposed for the network are independent of the model selected for data center deployment. The video service architecture described above is in fact deployed currently in both centralized and distributed modes and can be used with both existing HFC delivery networks and the evolved network proposed.

The next step in evolving the network is to use additional servers in the data center to run the access network (e.g., DOCSIS) control plane. In a traditional router or CMTS the data plane typically runs in specialized hardware while the control plane runs in a general purpose CPU embedded in the platform. In this case the general purpose CPU has migrated to a server and standard IP/Ethernet switches provide the data plane forwarding within the data center and from the data center to the head end. A control protocol between the server and the switch such as OpenFlow [OPENF] is used to control the forwarding path.

Head End

The head end in this architecture continues to support traditional analog and MPEG based broadcast video. High speed data and narrowcast video processing has moved to the data center as described above. HSD and IP video traffic passes through the head end via an IP/Ethernet network and is forwarded to the node using the existing fiber links which it shares with the broadcast video using WDM.

Node

In the node the broadcast video is converted from optical to coax media as today. The Ethernet traffic is converted from baseband Ethernet to the protocol to be used for the node to home portion of the network. This may be DOCSIS, other Ethernet over coax (EoC) access technologies, point to point Ethernet or even wireless technologies such as WiFi or LTE. Operation of the node is controlled from the data center (via the head end).

ADVANTAGES

The architecture described above has multiple advantages over a traditional HFC video delivery infrastructure.

Leverage Data Centers

It leverages the work done to provide massively scalable Internet based applications over the last decade to provide cost savings and efficiency:

- The use of COTS technology reduces costs by using general purpose servers as processing engines and standard Ethernet networking equipment for connectivity.
- It leverages standard virtual environments to provide horizontal scaling and redundancy
- It provides a simpler and more robust environment for networking software development.
- It provides a friendly platform for application level software development where familiar toolsets and environments enable software to be created by operators, equipment vendors and third parties to accelerate service velocity.

Head End Simplification

The head end becomes a much simpler environment leading to lower operational costs:

- Complex software functions are centralized in the data center where expertise can be concentrated.
- Power, cooling and rack space needs at the head end are reduced as devices such as CMTS are removed.

- IP services are delivered via a standardized Ethernet network as low level DOCSIS and QAM functions migrate to the node and the need for RF combining in the head end is reduced or ultimately removed.
- Multiple head end / hub locations may be collapsed or run remotely as a “lights out” operation saving operating and real estate costs.

Network Simplification

The use of IP transport to the node simplifies the network in the following ways:

- Standard Ethernet and digital optics can be used between the head end and the node eliminating distance limitations and enabling distribution hubs and small head ends to be consolidated.
- Ethernet switching is used in the data center, head end and hub to reduce costs in the transport network.
- The network from the data center to the hub is independent of the last mile technology from the hub to the home allowing these parts of the network to evolve independently and more cost effectively.

TRANSITION STAGES

From the above it appears that there are significant advantages to moving to the proposed architecture but as always the problem is in how to facilitate the transition at a reasonable cost while continuing to provide service. The following sections of the paper address this transition and break it down into a number of potential stages. The discussion identifies six possible transition stages between today’s HFC network and that shown in Figure 2. These are:

1. Introduction of CCAP
2. Split packet processing from physical media dependent and physical layer (PMD/PHY) processing
3. Move PMD/PHY to the node
4. Move MAC processing to the node
5. Move narrowcast processing from head end / hub to data center
6. Retire broadcast processing and remove from head end /hub

Not all stages are required and each operator can select the most appropriate path based on their existing network, operational needs and competitive demands.

Stage 1

This first phase of the transition, shown in

Figure **4**, addresses the change in services from broadcast to narrowcast. Narrowcast channels for MPEG and DOCSIS are added to the head end using high density CCAP based equipment. This may augment or replace existing EQAM and CMTS equipment. The CCAP platform connects to the core network through the Ethernet distribution network in the head end as for existing equipment but will use 10Gbps rather than 1Gbps Ethernet links. Connectivity to the HFC remains RF based and connects to an optical shelf through the existing RF distribution /combining networks in the head end. The optical shelf converts the signals to analog optics and transmits them to the fiber nodes in the outside plant.

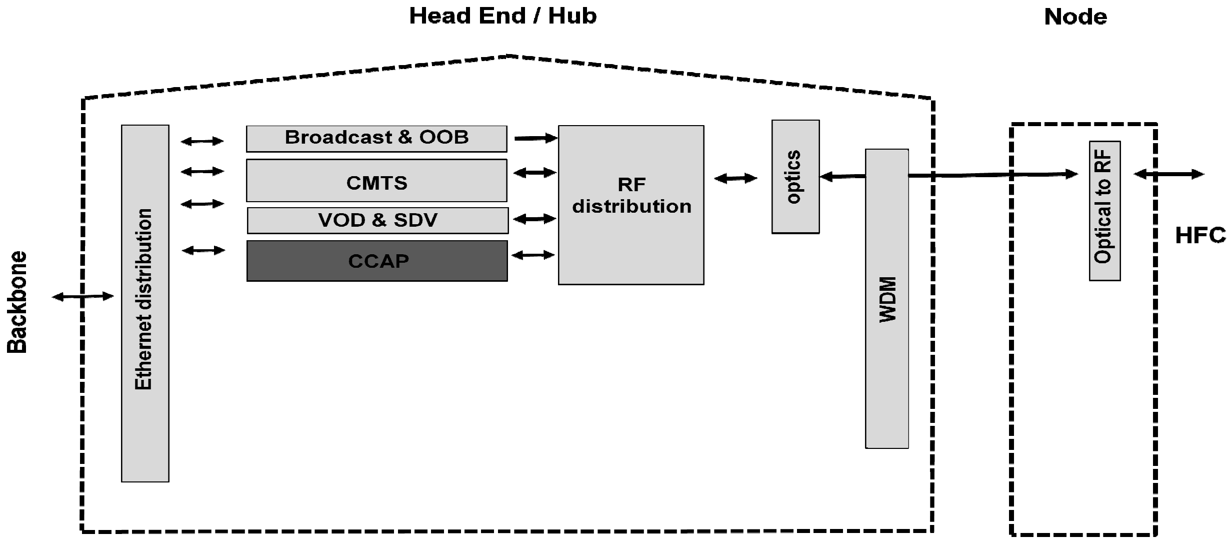


Figure 4: Phase 1, CCAP

The advantages of this transition are the savings in cost, real estate and power consumption provided by the increased density of next generation platforms [CCAP].

The technology changes required and risks associated with this change are those associated with the CCAP program and are well understood at this time.

Stage 2

The next stage of the proposed transition requires a slightly more radical change and is shown in Figure 5. It leverages some of the work done for modular CMTS [M-CMTS] and distributed CCAP architectures [D-CCAP] to move to an Ethernet based distribution system in the head end. The core principal is to decouple the packet processing from the physical media dependent (PMD) portion of the MAC and the PHY. The PHY and PMD dependent functions move out of the core processing engines to be collocated with the optical shelf which may be within the head end or in a distribution hub. The CMTS and EQAM core engines output MPEG streams over Ethernet using standard framing [M-CMTS]. The downstream modulation function is included with the optical transmitters and the upstream demodulation function is included with the optical receivers.

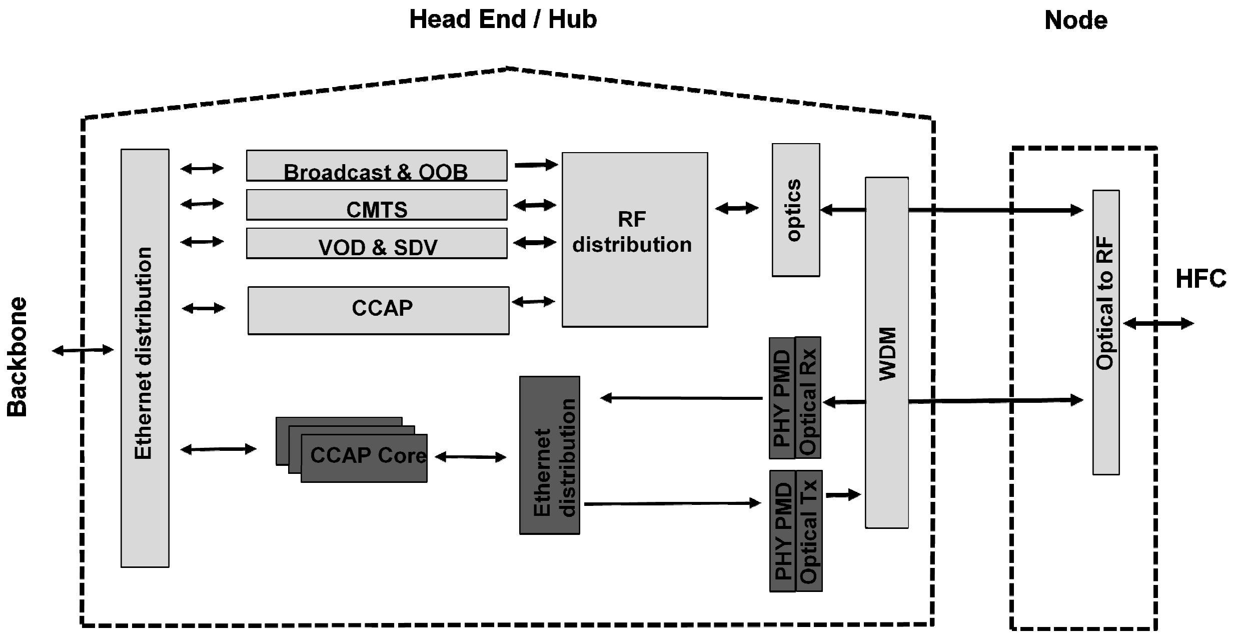


Figure 5: Phase 2 with Local CCAP core

This separation of digital processing from modulation has several advantages. Scaling the system becomes more efficient as CMTS and EQAM core engines scale based on processing needs while the PMD and PHY layer functions scale based on port counts. Thus the core engines scale up as the total number of DOCSIS and EQAM channels to be delivered increases while the PMD/PHY functions scale as the number of serving groups increases. While these factors (channel count vs. serving group count) are certainly related it is typically not a strictly linear relation so there is benefit to the separation.

Redundancy may also be simplified in this case. Ethernet based redundancy can be used for the core processing engines so that there is no need for complex RF switching logic in the core chassis in the data center. RF redundancy can be provided in the optical shelf for large serving groups but for smaller serving group sizes the failure group in the optical shelf may be small enough that RF redundancy is not needed providing further cost reductions.

With the replacement of the RF combining network by Ethernet switching changes such as node splits are made simpler; ports are added to the optical shelf for the new node and processing engines added to the core shelf if required. The operations and management network for the head end can share the Ethernet infrastructure to provide central configuration and monitoring using standard IP tools. Thus path traces and testing across the head end become trivial using tools such as ping and traceroute.

The most interesting feature of the PHY-PMD separation is the increased flexibility it enables for equipment deployment. The standard Ethernet links between the core processing engines and the optical shelves effectively remove any distance restrictions between them. Thus the optical shelves could be deployed in a remote head end or hub while the processing engines reside in a data center. This enables “lights out” operation of the remote facility and significantly reduces power and cooling needs at these locations. Centralizing the complex equipment also reduces the operations skill sets needed at the remote locations.

Figure 5 above shows a potential deployment scenario in which additional services are added via the new model as needed. Existing broadcast, CMTS and EQAM equipment can remain unchanged or be replaced over time.

Figure 6 shows an alternative deployment model with the CCAP core platforms located in a data center remote to the head end.

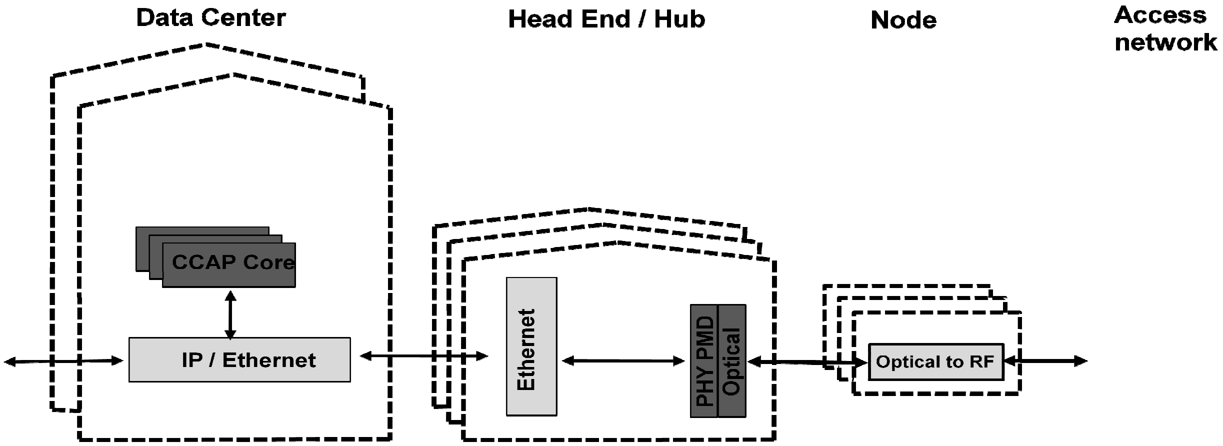


Figure 6: Phase 2 with Remote CCAP core

The technology risks associated with this phase are primarily the timing issues associated with the separation of the DOCSIS MAC and PMD functions.

Stage 3

As mentioned above the use of an Ethernet distribution network between the core processing engines and the optical shelf eliminates distance restrictions between them. The next phase takes advantage of this fact to move the PHY-PMD function out of the head end to the node and extend the Ethernet distribution to the node using standard Ethernet optics as shown in Figure 7 . New services are added using the Ethernet to the node transport while legacy services continue to be supported using an RF overlay. Ethernet and analog wavelengths are multiplexed onto the same fibers using existing WDM equipment.

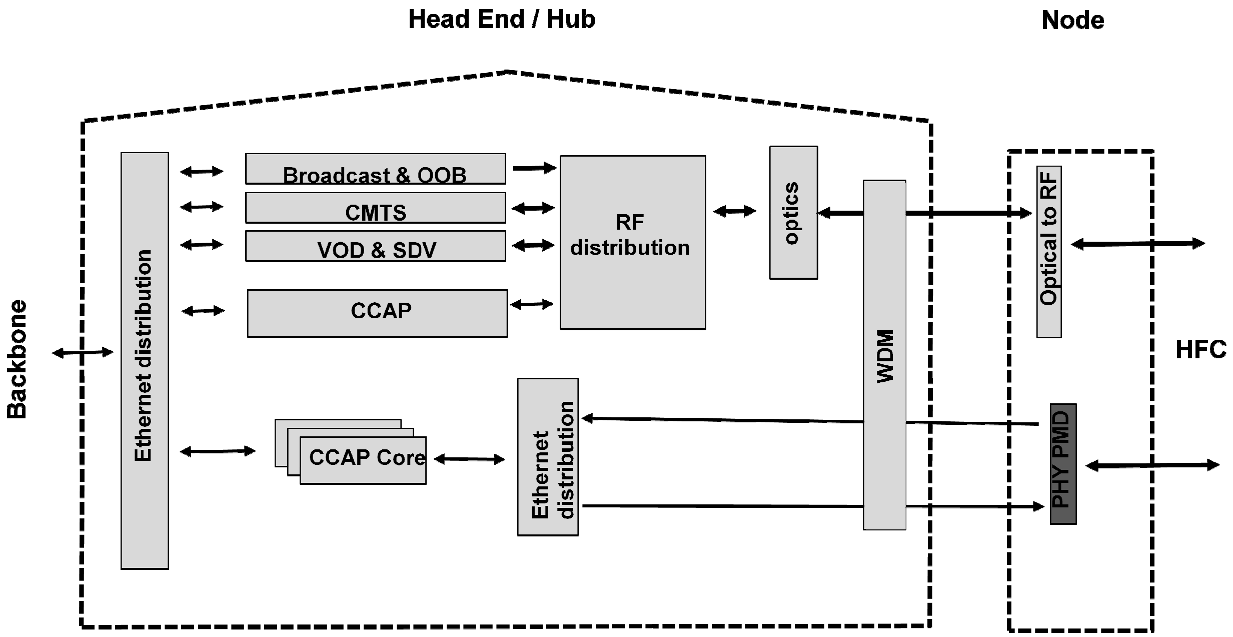


Figure 7: Stage 3, Remote PHY-PMD

Deployment of this phase of the transition plan brings several advantages to the operator. Service expansion is no longer limited by the head end to node transport as the use of digital optics and DWDM provide essentially unlimited bandwidth on this link. The links leverage the costs, speed and density trajectory set by Ethernet systems and standard DWDM platforms.

If the RF overlay can be removed then distance limits between the head end and the node are eliminated. It may then be possible to centralize head end functions and retire some distribution hubs and small head ends providing savings in real estate and operational costs.

The technology risks associated with this phase are the density, powering and cooling of the components in the fiber node and the provision of timing services to the node.

Stage 4

This phase, shown in Figure 8 is a conceptually small change in which the upper layer MAC functions are moved to the node so that they are co-resident with the PMD and PHY functions. As in the previous phase new services are added using the Ethernet to the node transport while legacy services may continue to be supported using an RF overlay

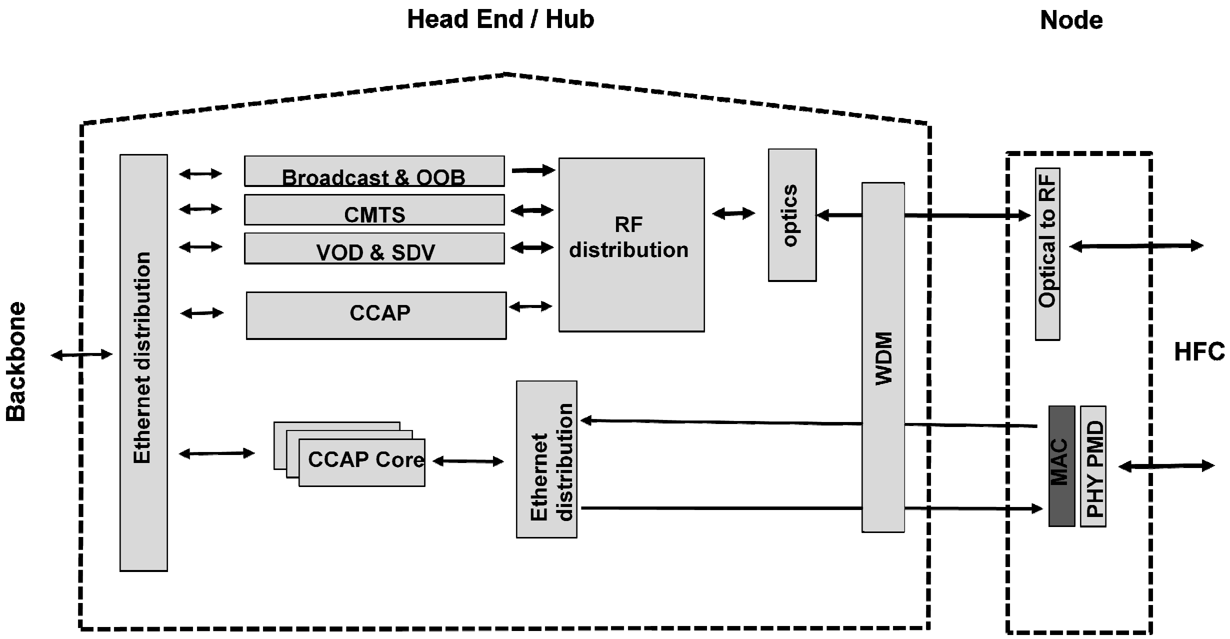


Figure 8: Stage 4 Remote MAC

The advantage of moving the MAC to the node is that it simplifies the timing issues relative to the split implementations described previously. More importantly it decouples the technology for the head end to node transport from that used between the node and the home. This allows the technologies to evolve at their natural and different paces. The head end to node link uses general enterprise networking technology while the node to home link remains specific to the HFC plant. Node splitting and the progression towards an n+0 architecture becomes a process of replacing the DOCSIS MAC/PHY module in the node with an Ethernet switch and moving it further downstream. The impact of transitions to next generation technologies such as PON or EoC is restricted to the node to home portion of the network.

The technology risks are similar to the prior phase as more functions are moved to the node increasing powering and cooling needs.

Stage 5

In this phase, shown in
Figure 9 the legacy narrowcast equipment in the head end is retired and all processing elements migrate to a central data center. The HFC specific MAC elements have migrated to the node and IP/Ethernet transport is used throughout the network from the core to the node. Application and control plane logic are no longer in the head end which is only used for legacy broadcast services. It still acts as a pass through for narrowcast services but this is reduced to an IP/Ethernet switching function.

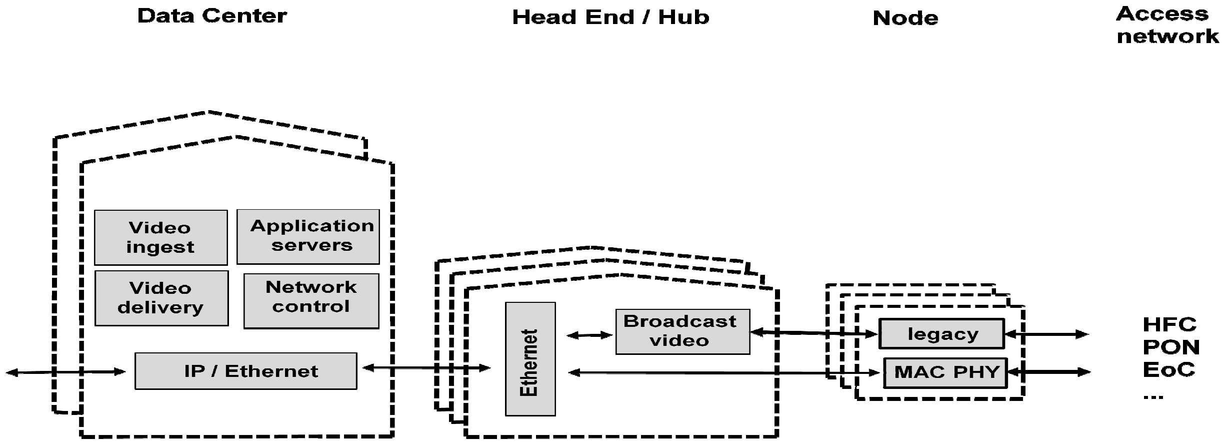


Figure 9: Stage 5 Narrowcast Equipment Removal

Stage 6

With the removal of traditional broadcast the head end in its current form can go away completely and be replaced by a data center, an Ethernet distribution hub and a simple node as shown in Figure 10. At this point the advantage of centralization, standard IP/Ethernet transport and isolation of HFC specific functions described in the earlier phases are now fully realized.

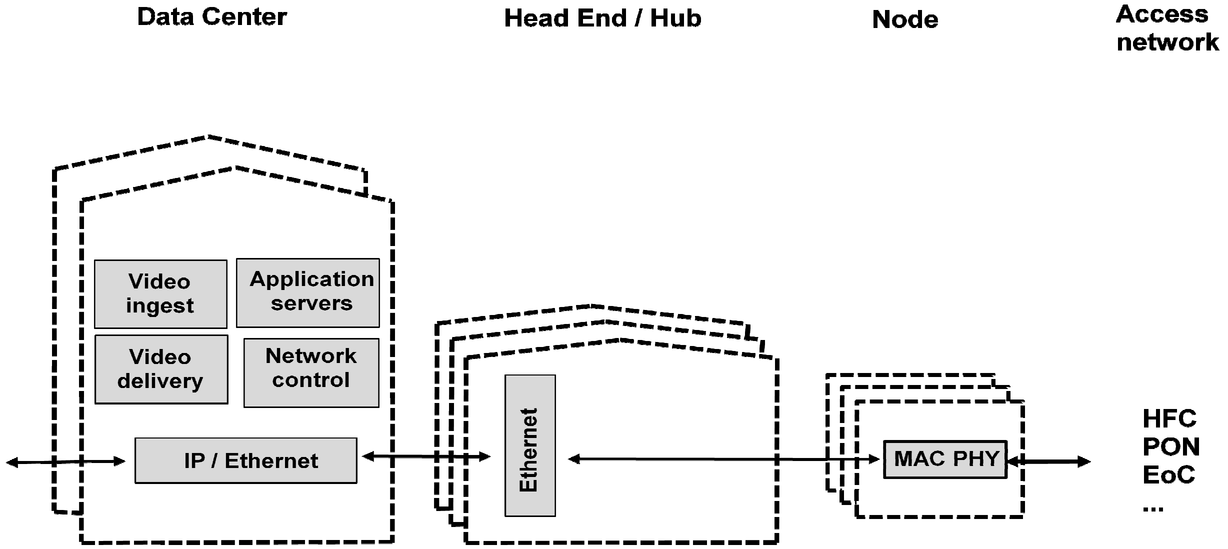


Figure 10: Stage 6 Broadcast Equipment Removal

Moving the MAC and PHY functions from the head end to the node allows the use of standard Ethernet optics and enables distributed processing but results in a more intelligent outside plant architecture. Operators who do not wish to take this step and prefer to keep a simpler outside plant can elect to deploy the MAC-PHY components in the remote hub rather than the node as shown in Figure 11. They still retain the advantages of the move to the data center and a significant reduction in hub complexity. Readers interested in an in depth comparison of traditional and intelligent HFC architectures are referred to [HFCDFC].

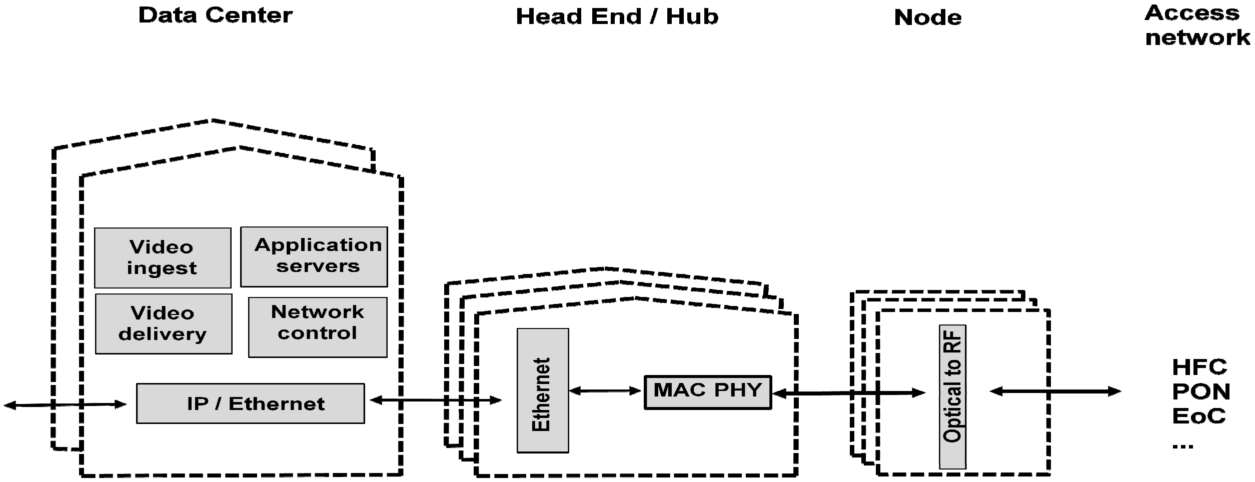


Figure 11: Passive HFC Architecture

DEPLOYMENT

The transition stages previously described illustrate a logical roadmap to the data center architecture. They are based on technology evolution but are essentially independent. Which stages are deployed by an operator will depend on their customer needs, timing, risk profile, budget and network architecture.

Table 1 shows a summary of the risks and benefits associated with each phase together with an indication of when it would be appropriate to be used.

Stage	**Change**	**Benefits**	**Risks**	**When Appropriate**
1	Move to CCAP platform	Increased density, lower cost, simplified combining	Minimal, well understood problem	Need to add high density narrowcast services in HE/hub
2	Decoupled MAC / PHY in head end	Processing and port scaling separated, simpler redundancy, simpler combining	DOCSIS MAC/PHY timing split.	Gain HE/hub benefits without touching node
3	PMD+ PHY move to node	Ethernet Digital optics to node	Power & cooling in node	Consolidate small HE/hub; retain existing core platforms
4	MAC to node	Data center to node links all Ethernet for narrowcast traffic	Power & cooling in node	Standardize on Ethernet transport to the node to set up for stages 5 & 6
5	Narrowcast removed from HE	Head end space savings	minimal	Consolidation from HE/hub to data center for lower OPEX
6	RF broadcast removed from HE	HE becomes simple switching center or is removed	minimal	Consolidate or retire HE/hub

Table 1: Risks and Benefits of Each Stage

CONCLUSION

Moving functions from the head end or distribution hub into a data center has many advantages and has the capability to provide significant capital and operational savings. To transition to a network architecture which can take full advantage of this move is not trivial but can be achieved through a series of stages as technology evolves and service needs demand. The transition stages described are largely independent and any given operator can select the transition path best suited to their specific needs.

REFERENCES

[CCAP1]	J. Salinger, "Proposed Next Generation Cable Access Network Architecture", SCTE Conference on Emerging Technology, 2009.
[CCAP2]	J. Salinger, "Understanding and Planning CMAP Network Design and Operations", SCTE Cable-Tec Expo, 2010.
[CCAP3]	J. Finkelstein, J. Salinger, "IP Video Delivery using Converged Multi-Service Access Platform (CMAP)", SCTE Canadian Summit, 2011
[D-CCAP]	John Ulm, Gerry White New Converged Access Architectures for Cable Services, NCTA 2011 Spring Technical Forum
[HFCDFC]	M. Emmendorfer, S. Shape, T. Cloonan & Z. Maricevic Examining HFC and DFC (Digital Fiber Coax) Access Architectures, SCTE Cable -TEC Expo 2011
[M-CMTS]	Cablelabs DOCSIS® Specifications — Modular Headend Architecture (MHA)
[MSIPD]	John Ulm, Gerry White Arch & Migration Strategies for Multi-screen IP Video Delivery, SCTE Canadian Summit 2012
[OPENF]	www.openflow.org
[SAND]	Global Internet Phenomena Report Fall 2011; Sandvine

ABBREVIATIONS AND ACRONYMS

CCAP	Converged Cable Access Platform
CDN	Content Delivery Network
CMTS	DOCSIS Cable Modem Termination System
COTS	Commercial Off The Shelf
CPE	Customer Premise Equipment
DOCSIS	Data over Cable Service Interface Specification
DRM	Digital Rights Management
DVR	Digital Video Recorder
DWDM	Dense Wave Division Multiplexing
EAS	Emergency Alert System
EoC	Ethernet over Coax
EPoC	EPON over Coax
EPON	Ethernet Passive Optical Network
EQAM	Edge QAM device
FSM	Finite State Machine
Gbps	Gigabit per second
HFC	Hybrid Fiber Coaxial system
HSD	High Speed Data; broadband data service
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
MAC	Media Access Control (layer)
Mbps	Megabit per second
MPEG	Moving Picture Experts Group
MPEG-TS	MPEG Transport Stream
nDVR	network (based) Digital Video Recorder
OTT	Over The Top (video)
PHY	Physical (layer)
PMD	Physical Medium Dependent (layer)
PON	Passive Optical Network
RF	Radio Frequency
STB	Set Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VOD	Video On-Demand
VoIP	Voice over IP
WDM	Wave Division Multiplexing

Video Calling Over Wireless Networks

David Urban
Comcast

Abstract

Video calling over wireless networks has become increasingly popular as wireless networks become faster and more reliable and devices with cameras and video calling applications become more ubiquitous.

Measurement and analysis of video calls over home, outdoor and cellular wireless networks have determined the criteria for making a successful call over a wireless network. Signal strength, packet loss, jitter and round trip delay are critical parameters. Video calling over wireless networks is shown to be practical, provided that the critical parameters are met.

INTRODUCTION

Video calling is a technology that has been around for quite some time but has never caught on as much as one might think. As the saying goes, video calling is the technology of the future and always will be.

Bell Labs began building experimental prototypes in 1956 culminating in the 1964 New York World's Fair demonstration of the Picturephone service [1]. By 1969, the transition from voice calling to video calling appeared to be at the threshold. Looking back at the Picturephone service of the 1970s, it is interesting and instructive to find that many of the standards and specifications are similar to those used today. The analog bandwidth of the black and white video picture was 1 MHz with an interlaced 250 lines refreshed at 30 frames per second. The screen size was 5.5 x 5". When digitized to be transported a distance greater than 6 miles the combined video and audio signal was 6.3 Mbps.

So have things changed appreciably enough to suspect that this might just be the time that video calling really catches on? There is ample reason to think that the answer is yes. Many factors have fallen into place to make video calling more feasible than ever before.

Many people now carry around smart phones with a front and back camera, video calling software and a data connection fast enough for video calling. This means that when initiating a video call, one needn't count on the other side being at their computer with attached web camera and logged into a video calling application.

Televisions can be made into high definition video conferencing solutions with convenient and inexpensive add on products such as video cameras with built in microphones and small computer appliances to run the video calling application. Again, this avoids the inconvenience of having to fire up your computer, plug in your webcam, and open and log in to your video conferencing software. Any time you are watching television you can make or receive a video call.

A video call on a large screen television set can be much more enjoyable than using a computer. Several members of the family can participate while sitting on the couch rather than crouching around a small computer screen. And the 720P resolution of a 32 inch or larger diagonal flat screen television provides a much better viewing experience than a notebook computer or smart phone screen can.

Successful video calling requires a network connection with high data rate, low latency and jitter, and negligible packet loss.

Broadband connections are becoming more common and the performance keeps improving, making video calling more practical. This is true for both fixed and mobile networks. Video codecs have and continue to improve and work is being done specifically for video grade wireless distribution.

While there are still some impediments to video calling such as high cost and the lack of simple, intuitive and reliable user interfaces, many hurdles to successful video calling have recently been cleared and the remaining obstacles are trending toward resolution. In the 1960s and 1970s video calling moved from a laboratory curiosity to an ambitious but ultimately disappointing large scale national project. Then in the 1980s and 1990s video conferencing remained a niche application mainly for big businesses. With the advent of the personal computer and broadband residential network connectivity video calling has become an increasingly popular method to stay in touch with family and friends. The big transition occurring today is the move from video calling on desktop and notebook computers to video calling on smart phones, tablets and televisions. This transition makes wireless home and public network performance and reliability more important than ever. Table 1 shows some video call data rates.

Video Call Type	Tx Mbps	Rx Mbps	Video Quality
1080P WiFi	5	5	excellent
720P WiFi	1.5	1.5	excellent
3-Way WiFi	1	1	good
Smart Phone	0.5	0.5	fair
3G Cellular	0.2	0.2	poor

Table 1. Typical Data Rates of video calls

VIDEO CALLING PARAMETERS

A video call can be at times amazing when it works perfectly while at other times the experience can be frustrating when things go wrong. The elements of a video call include two parties, each with a video camera, video display, audio microphone, and audio speaker. Each party needs a computing device to run a video calling application and the devices need to have a network connection to establish the call, send the video and audio streams, and end the call.

For a two-way video call, a video stream will be sent from the video camera and another video stream will be received for the video display. Likewise, an audio stream will be sent from the microphone and another audio stream will be received by the speaker.

The audio and the video must be synchronized. A delay from the video camera of one user to the video display of the other user can be distracting. For example, if a caller wishes to show an object by putting it in front of the camera but gets no reaction from the other side, this can be confusing. Then the other side finally comments but long after the object has been removed from the camera view. This distracts from the real time interactivity of the video call.

Disconnects, long reconnections, poor video quality, long delays, lack of video and audio synchronization, freezing of the video, brief distortions of the video display, screen refresh and resolution issues; these are the problems that make video calling frustrating. A successful video call requires a good network connection on both ends, good processing power in the CPU running the application, a good video calling application, a good camera and display on both ends.

The key network connection parameters necessary for a successful video call include data rate, packet loss, jitter, delay, and relays. The data rate will be dependent upon the screen size and resolution. A video call on a 1080P LCD television will have a data rate of 10 Mbps whereas a video call on a 4.3 inch diagonal screen smart phone will have a data rate of 1 Mbps.

Video calling applications often report call technical information. Among the reported parameters are jitter, packet loss, send packet loss, receive packet loss, round trip time, and relays. Relays can be used to work around firewalls and other networking issues that prevent a direct UDP connection between the two video callers. Relays in general are undesirable since they often prevent HD video calling. The video and audio streams are sent as UDP packets.

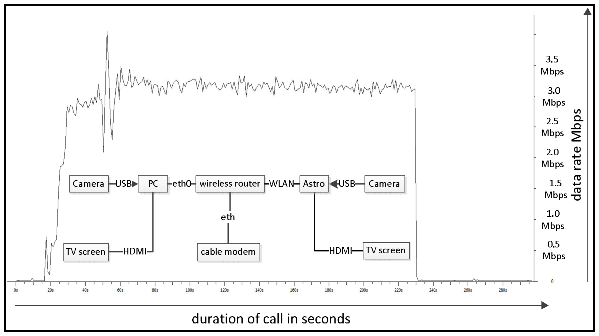


Fig.1 Data Rate and Block Diagram of 720P Video Call

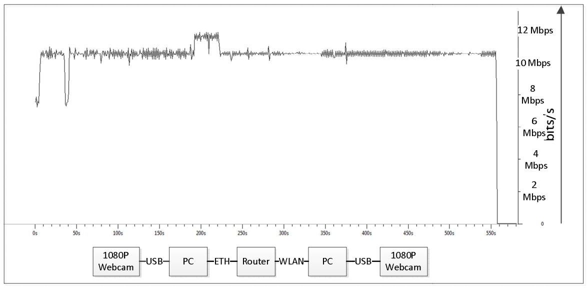
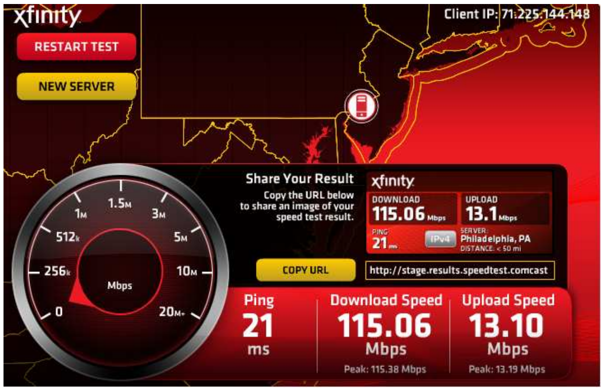


Fig. 2 1080P video call data rate

Wireshark was used to record the packets during a video call. The data rate during the call is shown in Fig.1 along with a block diagram of the test set up. Both the camera and the display were capable of 720P operation. The upstream and downstream data rate was measured to be 1.5 Mbps for a total of data rate of 3 Mbps. The video call quality was excellent. Packet analysis shows that the data protocol was UDP with packet size around 1400 bytes. In this particular test the video and audio streams were sent between devices on the same local area network. Most video calls will span a wide area network adding additional challenges for a successful video call.

Figure 2 shows a video call with 1080P video resolution. In this case the data rate is much higher at 10 Mbps, 5 Mbps for each video stream. Setting up a 1080P video call can be a bit tricky. You’ll need a 1080P video camera and display at both ends, video calling software the supports 1080P resolution at both ends, and network connectivity supporting 5 Mbps UDP traffic in both the upstream and downstream direction. Residential broadband connections that support this high upstream data rate have only recently been offered. Figure 3 shows a speed test for a cable modem connection capable of supporting 1080P video calling.



Client IP: 71.225.144.148

Share Your Result

Copy the URL below to share an image of your speed test result.

xfinity

DOWNLOAD 115.06 Mbps

UPLOAD 13.1 Mbps

PING 21 ms

SERVER: Philadelphia, PA

DISTANCE: < 50 mi

http://stage.results.speedtest.comcast

Ping 21 ms

Download Speed 115.06 Mbps Peak: 115.38 Mbps

Upload Speed 13.10 Mbps Peak: 13.19 Mbps

Fig.3 Broadband Connection speed for 1080P video call.

The user experience of a 1080P video call is remarkable. The picture is clear and sharp on a big screen television and the live fast action response to motion is impressive.

Large screens with high resolution benefit from very high continuous data rates during a video call; however, many video calls involve smart phones which have much smaller screens that do not need such high data rates. Figure 4 show the data rate measured during a video call using a smart phone. The smart phone network connectivity is over a wireless home network.

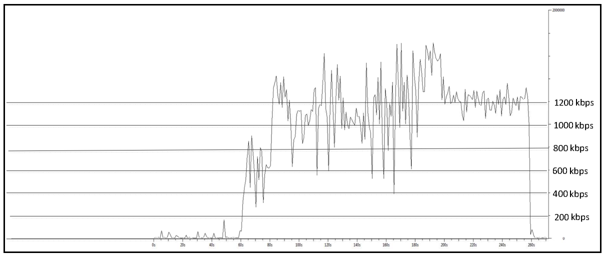


Fig. 4 Video Call using a smart phone with WiFi

The data rate measured about 1.2 Mbps and the quality of the video was good. Notice that the data rate is much less consistent than the previous plots with the bit rate over time being very choppy. This is due to several factors including the wireless home network link, the smart phone CPU processing speed and memory, and the impact of the wide area network. For the video call in figure 4 two cable modems were used so that the video and audio streams had to traverse the HFC network.

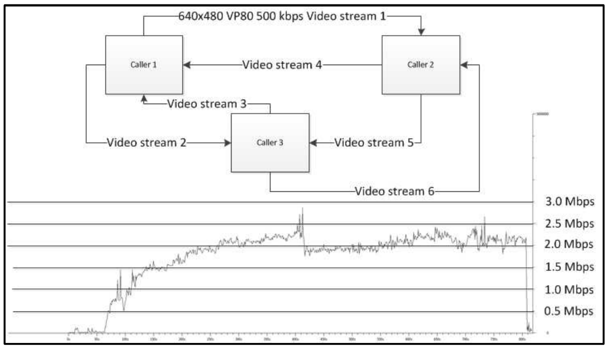


Fig. 5 Three-Way Video Call

A video call can be made between three or more parties. For a three way video call, a video caller sends two video streams and receives two video streams. The video callers' display typically shows the two received video streams side by side on the screen with a small caption of the video send stream. With this format the video caller can see both of the people he is calling as large as possible and still monitor what the other parties see of him. Since two video streams must share the display, the resolution and bit rate of a single video stream is reduced, i.e. one cannot display two 1080P video streams on a single 1080P video display. Testing 3-way video calls, the video send and receive streams were found to be 640x480 with VP80 codec at 30 frames per second and a bit rate around 500 kbps. A video caller participating in a 3-way call will thus send two 500 kbps video streams and receive two 500 kbps video streams for a total data rate of 2 Mbps as shown in figure 5.

VIDEO CALLING OVER WIRELESS NETWORKS

Characteristics of the wireless network

The making of a successful video call requires network connectivity with low packet loss, low latency, and low jitter and must support UDP data rates between 1 and 10 Mbps depending on the screen size and video quality requirements. Several wireless networks were tested to gauge their performance against the demands of video calling.

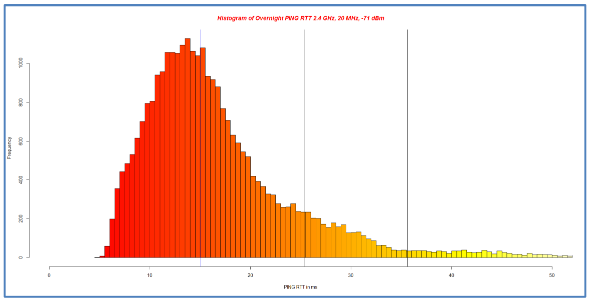


Fig. 6. Histogram of Overnight PING RTT 2.4 GHz, 20 MHz, -71 dBm from 0 to 50 ms

Fig. 6 shows a histogram of the round trip time, RTT, measured while sending PING packets between a wireless client and a wireless access point of a home wireless local area network, WLAN. The x-axis is the PING RTT from 0 to 50 ms and the y-axis is the number of occurrences. As indicated by the vertical lines in figure 6, the median RTT was found to be 15 ms, the first standard deviation above the median was 25 ms and the second standard deviation above the median was 35 ms.

The wireless access point was set to 2.4 GHz channel 8 with a 20 MHz channel width. Both the STA and the AP were IEEE 802.11n with dual stream capability. The wireless access point was set to B/G/N Mixed wireless mode. The beacon interval was set to 100. The RTS threshold was set to 2347. The guard interval was set to 800 ns. The STA and AP were separated by 36 feet and one floor and two walls of a residential home. The PING tests were taken over a 12 hour period. The x-axis of the plot is the PING round trip time measured in ms. There are three vertical lines shown on the graph, from left to right these lines are the median, first standard deviation, and second standard deviation, respectively. The receive level measured by the STA was -71 dBm.

The results indicate that the latency and jitter of a wireless home network have much more variability than a wired network over the course of time. This can be due to signal fading and interfering signal sources. The statistical distribution of the round trip time of packets between the AP and the STA are well within the requirements of a video call. A round trip time of 40 ms will support a high quality video call. On the histogram of round trip time measured in ms, 40 ms is beyond the second standard deviation and thus is a rare occurrence.

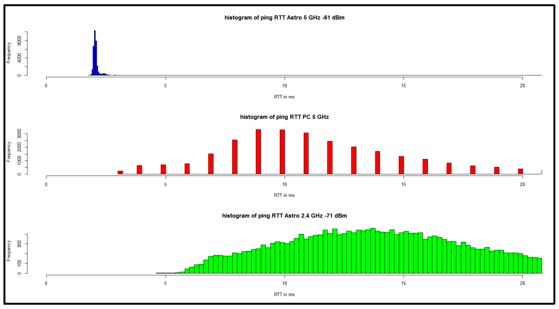


Fig. 7 Histogram of PING RTT for 5 GHz -61 dBm(top), 5 GHz at same location with PC (middle), and 2.4 GHz -71 dBm (bottom) from 0 to 20 ms

Figure 7 shows test results of the distribution of PING round trip time in ms over the course of a 12 hour test period. There are three different test conditions, the top blue graph is the PING RTT distribution over a 12 hour test period of a 5 GHz wireless home network connection with a receive level of -61 dBm. The computer used for this test was a small form factor LINUX device with built in WiFi client. For this test the AP and the STA were separated by one wall and 12 feet. 5 GHz band with the AP and STA in close proximity results in a much lower median round trip time latency of 2 ms with no significant measurements greater than 5 ms.

The middle red distribution of figure 7 shows another 5 GHz test taken in the same location and same time as the top distribution but using a different wireless client station. The middle test results used a notebook computer with built in wireless card and antennas. This RTT distribution shows a median of 10 ms with most RTT measurements fewer than 20 ms.

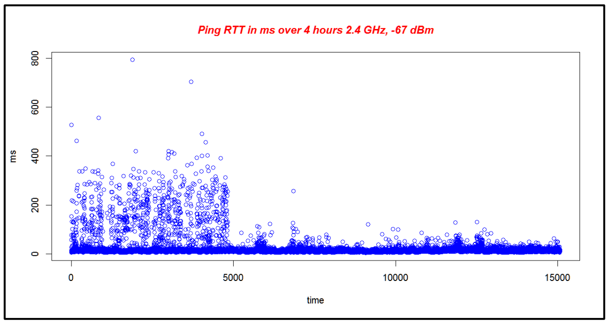


Fig. 8 Plot of PING RTT in ms over 4 hours at 2.4 GHz, -67 dBm, y-axis is RTT in ms from 0 to 800ms.

Both test results are good and well within the requirements for a successful video call. But why would two tests, both wireless clients using the same channel at the same time, both in the same location, give such different results? It turns out that it was not due to hardware differences between the two stations since subsequent overnight tests revealed that by slight manipulations of the antenna positioning one could reverse the results.

As illustrated in figure 8 the PING round trip time can change abruptly in time and these changes can last for hours at a time. This could be due to other applications sharing the spectrum or even to the movement of people or objects within the home.

The antenna patterns of notebook computers and small form factor devices with built in antennas will have nulls due to internal obstructions. By slightly repositioning the computers and devices one can place the nulls in more or less advantageous a location and this can influence the PING RTT results.

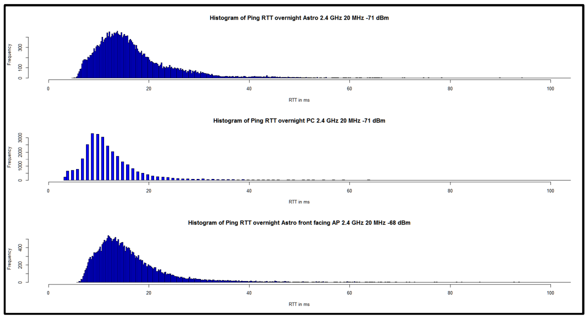


Fig. 9 Histogram of PING RTT overnight at (top) 2.4 GHz, 20 MHz, -71 dBm, LINUX, (middle) 2.4 GHz, 20 MHz, -71 dBm, WINDOWS, (bottom) 2.4 GHz, 20 MHz, -68 dBm, LINUX with antenna facing AP, x-axes are RTT in ms from 0 to 100.

The bottom green PING RTT overnight test distribution in figure 7 was taken using 2.4 GHz with the STA receive level of -71 dBm due to a larger separation distance between AP and STA of 36 feet with one floor and two walls. As expected the PING RTT distribution is much larger than the test using 5 GHz at closer AP to STA separation distance with a median of 14 ms and a significant number of round trip times having latency greater than 20 ms. The 2.4 GHz and -71 dBm receive level overnight PING test shows performance that is well within the bounds for a successful video call but as we will later see this is at the threshold of successful video calling operation.

One can expect variety of latency and jitter distributions for wireless home networks. Other examples are shown in figures 9 and 10. Figure 9 shows three tests taken in the same location, all at 2.4 GHz with 20 MHz channel width. The difference between the three plots is due to slight differences in antenna positioning. Figure 10 shows a comparison of 2.4 GHz performance versus 5 GHz at a 24 foot AP to STA distance with one wall in between. At close range 5 GHz proved to have consistently lower round trip times, however, figure 10 shows that at farther distances and more wall attenuation 2.4 GHz operation can have lower round trip time than 5 GHz.

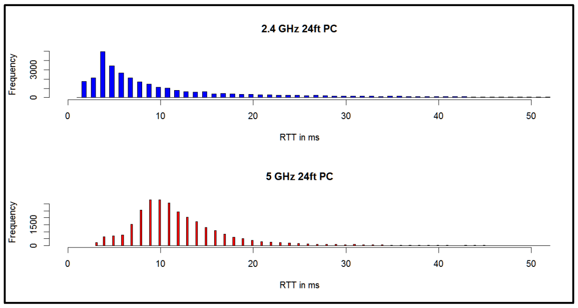


Fig. 10 Histogram of PING RTT 2.4 GHz vs 5 GHz and 24 ft AP to STA distance

In general wireless connections will be worse than wired connections in this regard. It is important to note that some routers handle IP video and peer to peer stream video better than others. As a rule of thumb, using 5 GHz at very close distance will give more consistent performance for video calling than 2.4 GHz at far separation distances as illustrated in figure 7. If you are having trouble making a video call using a wireless home network connection, then slight variations in antenna positioning of either the client station or access point can make significant performance improvements. Changing the RF channel, channel width, guard interval, and mode may also be experimented with to fix problems.

Figure 11 shows the statistical distribution of the call technical information reported by the video calling application. A small form factor Linux computer was used to run the video call application. The video camera and display at both ends of the video call were 720P and the video call data rate was 3 Mbps. The wireless network connection used 2.4 GHz with a 36 feet AP to STA separation distance with one floor and two walls in between. The video calling software has an option to report call quality technical information which includes a measure of network packet loss, roundtrip time, and jitter. These statistics are used to adjust the call quality to account for network connectivity issues. If these parameters degrade, then the video calling application will adjust by lowering the video quality such as adjusting the resolution from 1280x720 to 640x480. Adaptive bit rate streaming is a technique to provide the best video quality for given network limitations.

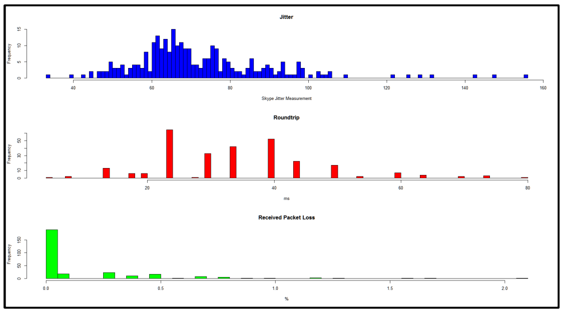
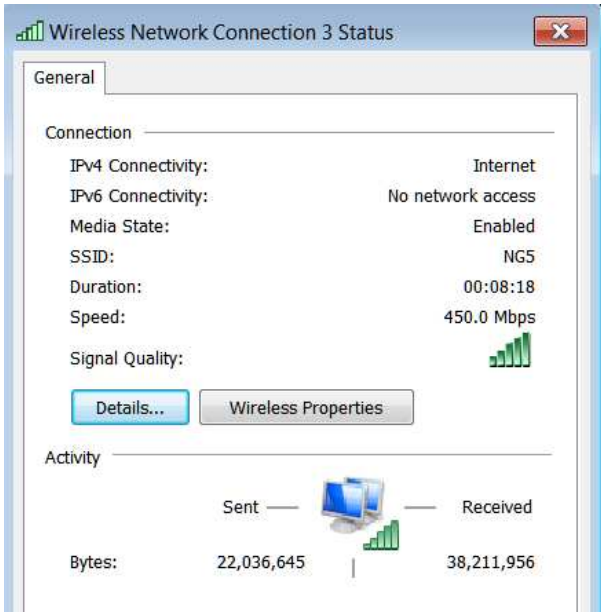


Fig.11 Video Call Quality Technical Information Top Histogram is Jitter from 40 to 160, Middle Histogram is Roundtrip time from 0 to 80 ms, Bottom Histogram is received packet loss from 0 to 2%

The call quality technical information was saved to a text file during the video call. A PERL language program was written to filter out the three parameters of interest into an array suitable for statistical analysis using the R statistical programming language. The analysis shows that the packet loss throughout the video call remained low at less than 0.5%. The round trip time varied significantly with a noticeable amount of measurements as high as 60 ms. The jitter measurement also varied significantly during the course of the call. Despite the variations, test results show that the wireless network was able to support a 720P video call with good reliability.



Wireless Network Connection 3 Status

Connection

IPv4 Connectivity: Internet
IPv6 Connectivity: No network access
Media State: Enabled
SSID: NG5
Duration: 00:08:18
Speed: 450.0 Mbps
Signal Quality:

Activity

Sent — — Received

Bytes: 22,036,645 | 38,211,956

Fig. 12 3by3 AP and 3by3 STA reporting 450 Mbps modulation and coding scheme

Wi-Fi packet analysis of a 1080P video call

A video call was set up between two callers with one caller using a wireless home network. Both the access point and the client station of the wireless home network were capable of three stream operation. The highest data rate of the wireless home network was 450 Mbps. By carefully positioning the access point and the client station in close proximity and applying some tricks such as using cookie sheets to create reflections it was possible to get the client wireless software to report 450 Mbps as shown in figure 12.

However, during the video call the highest data rate achieved was 324 Mbps. The data rate of 324 Mbps has three spatial streams, a guard interval of 800 ns, a 40 MHz channel width, and MCS 21 64-QAM with 2/3 rate binary convolutional coding. The method to calculate the data rate of 324 Mbps is shown in equation 1. The details behind these calculations can be found in [2], [3],[4].

$$R = \frac{3(streams) * 6\left(\frac{bits}{sc}\right) * \left(\frac{2}{3}\right) * 108(sc)}{(3.2 + .8)(\mu s)} = 324\ Mbps\ [1]$$

With three transmit and three receive antennas in the access point and the client station there are nine paths between the transmit antennas and the receive antennas as shown in figure 13.

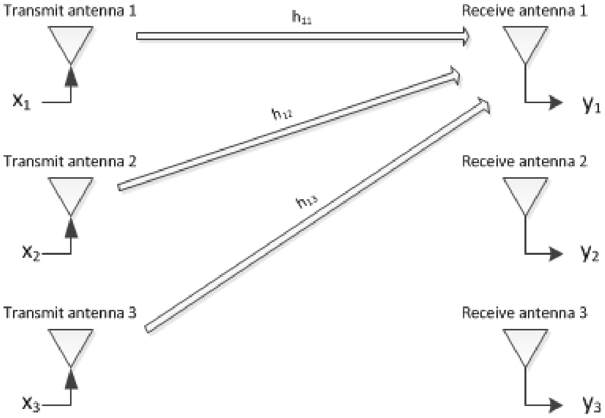


Fig 13. 3x3 MIMO Block Diagram

The output signals of the receive antennas, y_i with i={1,2,3}, is equal to the input signals of the transmit antennas, x_i, times the complex path loss of the nine paths between transmit and receive antennas, h_{ij}, as shown in figure 13. The relationship between the output signal of the three receive antennas and the path loss between the antennas and the input signals to the three transmit antennas can be expressed as a matrix equation [2]. If the H matrix of equation [2] can be inverted then it is possible to calculate the input signals by measuring the output signals and multiplying by the inverse of the H matrix. The determinant of the H matrix is zero if all of the elements are the same. The inverse of the H matrix is proportional to the inverse of the determinant. Thus, if all the elements of the H matrix are identical the determinant will go to zero and the inverse will blow up to infinity and it will not be possible to determine the input signals with knowledge of the output signals and path characteristics. Multiple streams can only work if there are differences, most desirably phase differences, between all of the nine paths between antennas. Spreading the antennas apart spatially is one method to increase the phase differences between the paths. However, with compact access points and particularly with compact client stations the amount of spatial separation is limited. Here, 5 GHz operation has an advantage over 2.4 GHz operation since for a given spatial separation, electrically in terms of wavelengths the separation between antennas is greater at 5 GHz than 2.4 GHz. The most effective and desirable method to create differences between the paths is reflections. A multipath rich environment with many reflected signals is the best for realizing multiple streams of data. In equation [1] the data rate from a signal antenna is multiplied by 3 since each of the 3 transmit antennas are transmitting an independent data stream.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\ [2]$$

Each subcarrier of the OFDM symbol is 64-QAM modulated so that a subcarrier is mapped to 6 bits. The bits are the output of a binary convolutional coder that inputs 2 data bits and outputs 3 coded bits. Thus, each OFDM subcarrier is mapped to 4 data bits as reflected in equation [1].

The channel width observed for this test video call for this packet was 40 MHz. A 40 MHz 802.11n signal consists of 128 subcarriers. Subcarriers at the channel edges and center are nulled to form a guard band and prevent DC offset. Some subcarriers are used as pilots to allow for frequency acquisition and carrier lock. This leaves us with 108 data subcarriers for a high throughput 802.11n data packet as reflected in equation [1]. Equation [1] reveals that each OFDM symbol for the observed packet has 1,296 data bits or 162 data bytes.

Finally, in order to determine the data rate we need to know the symbol time. The 128 modulated subcarriers that comprise the OFDM symbol are converted to a time domain representation using an inverse fast Fourier transform, IFFT. The 128 point IFFT is a transform with 128 complex frequency domain input numbers and 128 complex time domain output numbers. A digital to analog conversion, ADC, is required to turn the complex numbers into a real waveform capable of being upconverted to a carrier frequency to excite an antenna current in order to form an electromagnetic wave that can radiate from the transmit antenna to the receive antenna. The clock of the ADC determines the channel width of the analog time domain waveform. The channel width will be 40 MHz if the sampling interval is 25 ns. The formula is shown in equation [3] with W being the channel width in Hz and τ is the sampling interval of the time domain waveform in seconds.

$$W = \frac{1}{\tau} \quad [3]$$

With 128 subcarriers turned into 128 time domain samples by an IFFT and applying a 25 ns sampling interval, an OFDM symbol can be transmitted in 3.2 µs. This is sometimes referred to as the useful symbol rate. In theory OFDM symbols could be sent every 3.2 µs, however, in practice a guard interval in the form of a cyclic prefix is added so that the OFDM symbols are sent at an interval that is longer than the minimum possible. This eliminates inter-symbol interference that results when the receiver is hit with two different symbols at the same time due to reflections. As long as the guard time is longer than the time delay of the largest reflection then the receiver can ignore the guard time and demodulate the useful symbol time without inter-symbol interference. For 802.11n OFDM symbols the guard interval, GI, can be either 800 ns or 400 ns. A GI of 400 ns is referred to as a “short guard interval.” For the packet analyzed in the example the guard interval was 800 ns or 0.8 µs. The total OFDM symbol time is the sum of 3.2 µs and 0.8 µs for a total symbol time of 4 µs. This is shown in equation [1]. Now the data rate can be calculated by dividing the bits per OFDM symbol by the symbol time, in this case the data rate is 324 Mbps.

The frame length of the QoS data packet was 1395 bytes so that 9 OFDM symbols carrying 162 bytes each are needed to send the packet data payload. A burst of 9 OFDM HT symbols in this example lasts 36 µs since the OFDM symbol time for normal guard interval is 4 µs. Thus, over the 36 µs period of the 9 OFDM HT symbols the data rate is 324 Mbps.

When digital video signals are sent from a cable headend to a receiving set top box, the 256-QAM modulated 6 MHz wide signal transmits 38 Mbps continuously, a 100% duty cycle. This is not the case for wireless local area network transmissions. Since the medium is shared between uplink and downlink transmissions, amongst other users of the wireless home network, amongst co-existing wireless home networks, and amongst other spectrum users such as microwave ovens, cordless phones, remote controls, and sensors, a 100% duty cycle is not possible. Data is sent in bursts with short time frames and these bursts require a preamble in order to be received. The preamble is needed in order for the receiver to acquire carrier lock, understand the basic parameters of the packet, so that demodulation of the payload symbols can be made accurately.

All packets must be preceded with a preamble. The first part of the preamble is a short training field made up of 12 subcarriers. The short training field is 8 µs long. The short training field consists of 12 subcarriers. Figure 15 shows the subcarriers of the short training field measured by a vector signal analyzer. The short training field is followed by and 8 µs long training field and then a 4 µs signal field.

The transmission of a video packet of from the AP to the STA requires a sequence of packets as shown in figures 14 and 15. First, the AP sends a request to send message to the STA. The STA responds with a clear to send message. A QoS Data packet is sent from the AP to the STA. Finally, a block acknowledgement message is sent from the STA to the AP. This process is repeated continuously throughout the video call for both uplink and downlink transmission.

The request to send packet reported a length of 16 bytes and a data rate of 24 Mbps in its signal field, labeled SIG in Fig. 14. This is a legacy packet and thus has a 20 µs preamble consisting of an 8 µs short training field, STF, an 8 µs long training field, LTF, and a 4 µs signal field. The symbol period is 4 µs, a symbol has 48 data subcarriers mapped to 2 data bits so each symbol carries 96 bits or 12 bytes of data. The RTS packet is 28 µs and the request is to transmit a packet sequence with 224 µs duration.

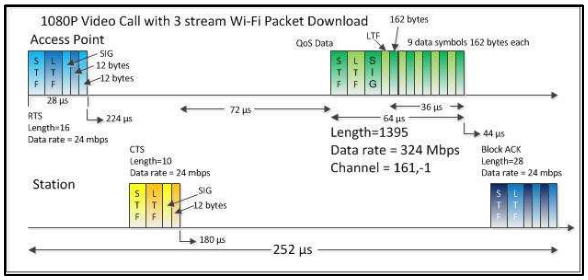


Fig. 14 WiFi Downlink video packet sequence

Following the RTS from the access point to the station will be a clear to send, CTS, response from the station to the access point. The clear to send packet has a length of 10 bytes and a data rate of 24 Mbps with a channel of 161. As with the RTS, the CTS packet has an 8 µs short training field, followed by an 8 µs long training field, followed by a 4 µs signal field, followed by 4 µs OFDM data symbols carrying 12 bytes of data. Since the CTS field length is 10 bytes only one OFDM data symbol is needed.

The CTS packet time duration is 24 µs. The CTS signal field reports that the duration from the end of the CTS to the end of the packet sequence is 180 µs. By taking the difference between the reported duration by the RTS packet and the CTS packet, we calculate the time duration from the end of the RTS packet to the end of the CTS packet of 44 µs. Thus there is a gap of 12 µs from the end of the RTS packet to the beginning of the CTS packet allowing for time for the access point request to be made and the client station response to be sent.

After the access point makes a request to transmit data and the client station responds with a clear to send signal then the QoS data packet can be sent from the access point to the client. Once the QoS packet has been sent by the access point and received by the client station then the client station sends a block acknowledgement back to the access point.

So in this example packet sequence measured during a 1080P video call, 1395 data bytes were transmitted over a 252 µs time period. The data rate accounting for the signaling and overhead is 1395 bytes divided by 252 µs which is 44 Mbps. Since the 1080P video call requires a sustained 10 Mbps data rate, the duty cycle of 324 Mbps data rate QoS data packet sequences during a 1080P video call is 23%.

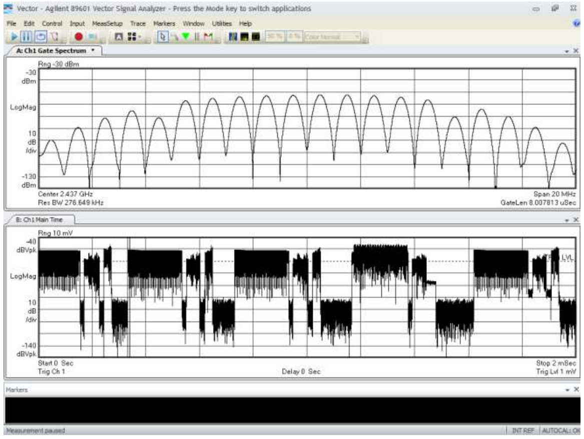


Fig. 15 Spectrum Analysis of WiFi video call

Taking a look at the distribution of the data rate of QoS Data packets reveals that many of the data packets were sent at a lower data rate than 324 Mbps. During the entire 1080P video call 3 stream operation was only utilized a small percentage of the time. All in all, 392,129 packets were analyzed. Of all of the downlink QoS data packets 9.42% had a data rate of 324 Mbps utilizing 3 stream operation. The majority of downlink QoS data packets operated at a 2 stream data rate of 243 Mbps representing 74.92% of the downlink QoS data frames. On the uplink 85.59% of the QoS data frames had 2 stream 270 Mbps while only 0.4% of uplink packets used 3 streams at a data rate of 324 Mbps or higher.

Plotting the histogram of the packet lengths during the video call shows statistically the anecdotal observation made by looking through the packet decodes. The RTS, CTS, QoS Data, Block ACK sequence with a 1400 byte UDP data burst is repeated throughout the video call. This packet sequence dominates the WiFi traffic during the video call. This is illustrated by the histogram shown in figure 16. Packet sizes are either less than 40 bytes or around 1400 bytes. The small byte size packets are signaling messages, RTS, CTS, and Block ACK. The packet sizes concentrated around 1400 bytes are video packets.

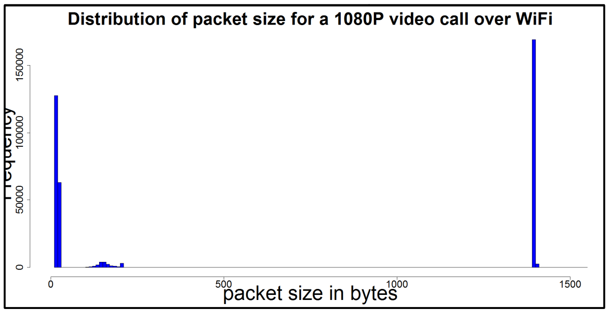


Fig. 16 Video call packet sizes are either small or large

Much of this analysis of the WiFi packets during a video call is focused on allowing the calculation of duty cycle, the percent of the time the application needs to use the RF spectrum. The reason that this is so critical is that WiFi uses unlicensed spectrum and thus any application must be judged based upon how well it will work while sharing the spectrum with other devices and applications. It is not a valid excuse for wireless LAN equipment and applications to claim that poor performance is due to a “noisy” environment. By “noisy” it is meant that other users of the spectrum are preventing the equipment or applications from working. However, equipment and applications using unlicensed spectrum must be designed to work in a shared spectrum environment. Users of unlicensed band equipment and applications must not expect performance levels that can only be realized with unshared spectrum. Even licensed band spectrum suffers considerable interference from adjacent cells and from spectral spillover from harmonically related or adjacent spectrum bands. So even licensed band equipment and applications must be designed to operate in the presence of fading and interference.

WIFI PACKET ANALYSIS OF A 720P CALL

A video call was set up with both callers having a 720P camera and display. One of the computers used a wireless home network connection. The wireless home network used

2.4 GHz channel 8 with a 20 MHz bandwidth and both the AP and the STA had 2 stream capability. The distance between the wireless access point and the wireless client station spanned about 36 feet, two floors and three walls of a residential home. The received signal strength level at the wireless client station ranged between -68 and -72 dBm. The video call lasted about 6 minutes and 30 seconds.

The video calling software reported call quality technical information. The video send stream was 1280 by 720 H.264 at 30 frames per second with a 1522 kbps data rate. The video receive stream was 1280 by 720 H.264 at 30 frames per second with a 1507 kbps data rate. The call technical information reported 0 relays indicating that the UDP traffic flowed directly between the two callers without intermediate nodes. The set up and data rate are shown in figure 17. The video call experience was excellent during this test. One video caller has an Ethernet 1 Gbps connection while the other video call uses a wireless home network connection with challenging RF signal conditions.

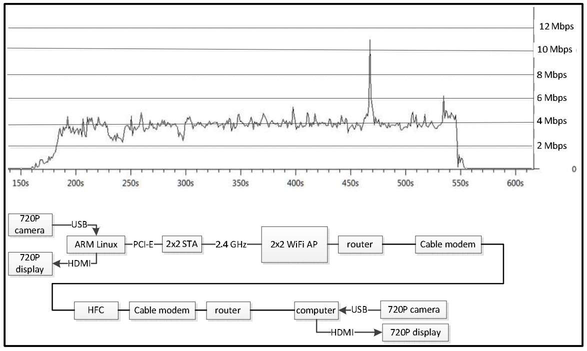


Fig.17 Video Call Set Up and Data Rate 720P 2.4 GHz.

During the video call an Airmagnet WiFi analyzer was used to capture the wireless local area network traffic. In all, 408,803 WiFi packets were captured and used for the statistical analysis of the call. The WiFi analyzer packet capture data file was saved as a Wireshark file and Wireshark analysis was used to create the IO data rate graph. The Wireshark data was then exported to a text file and a Perl program was written to extract and calculate a data array consisting of the burst time of the 408,803 packets. The R statistical programming language was then used analyze the distribution of the WiFi burst times.

The summation of the burst time of all the WiFi packets was 92.208012 seconds. Since the call lasted 6.5 minutes or 390 seconds, the percentage of time that the video call computer wireless station was either transmitting or receiving was 23.6%. In other words, the duty cycle of the 720P video wireless home network was found to be about 25%, one quarter of the time. If four such video calls were made utilizing the same wireless spectrum then we would expect conflicts due to 100% spectrum utilization.

The mean burst duration was calculated to be 225 µs. The median burst duration was found to be 40 µs indicating that many of the bursts were of short duration such as RTS, CTS, and Block ACK signals. The standard deviation of the burst times was 380 µs.

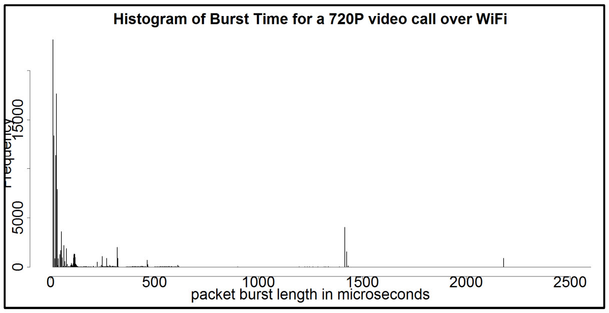


Fig 18. Histogram of the WiFi Burst Duration during a 720P video call.

The histogram of the burst durations is shown in figure 18. The bursts that last longer than 2 ms are beacons.

Figure 19 shows the histogram of packet burst time duration from 20 µs to 1 ms. The spreadsheet in Table 2 shows the percentage of packets for each possible data rate of transmission. With this histogram and spreadsheet it is easy to identify the main data rates used for sending video packets of about 1400 bytes. The three most prominent data rates are 13, 19.5, and 52 Mbps with burst durations of about 950, 640, and 260 µs, respectively.

Data Rate	Burst length	Burst Time	RX	TX	RX	TX
Mbps	bytes	microseconds	frames	frames	%	%
1	16	156	453	9,008	0.19%	5.31%
2	16	92	366	0	0.15%	0.00%
5.5	1495	2204	1	0	0.00%	0.00%
6.5	1495	1868	375	5,633	0.16%	3.32%
11	16	40	90,123	46,470	37.93%	27.40%
12	16	40	36,099	0	15.19%	0.00%
13	1495	948	1,159	27,950	0.49%	16.48%
19.5	1495	644	8,375	52,457	3.52%	30.93%
24	16	36	29,534	14,328	12.43%	8.45%
26	1495	488	7,615	10,199	3.20%	6.01%
39	1495	336	12,127	1,425	5.10%	0.84%
52	1495	260	50,095	1,041	21.08%	0.61%
58.5	1495	236	2	34	0.00%	0.02%
65	1495	212	17	49	0.01%	0.03%
78	1495	184	1,284	676	0.54%	0.40%
104	1495	144	1	338	0.00%	0.20%
117	1495	132	0	10	0.00%	0.01%
			237,626	169,618		

Table 2. 720P video call data rates and burst time

There are a couple points of interest in this analysis. First, although both the wireless access point and wireless station have two transmit and receive antenna chains and are thus capable of dual stream operation, the data rate rarely goes above 65 Mbps and most video packets are being sent at a data rate lower than 65 Mbps. This is significant because a single antenna wireless station lacking dual stream capability will max out at 65 Mbps for 20 MHz channel width and normal guard interval. Under these circumstances, the single antenna client station is at no disadvantage compared with a multi-antenna client. In fact, with only one antenna chain the power consumption is reduced and there is less physical footprint to pick up on board interference. The late Steve Jobs was noted for his passion for simplicity and functionality. He demanded products that worked and were a pleasure to use. Long battery life and comfortable operating temperature trumped the fastest Mbps claim on the outside of the box. This is reflected in mobile products that for the most part use a single antenna design with 20 MHz channel width and normal guard interval.

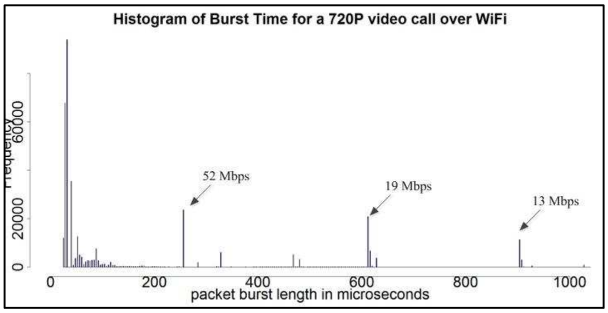


Fig. 19 Histogram of the WiFi Burst Duration up to 1 ms.

The second thing to note is that this test set up is operating at the threshold of a successful 720P video call. A significant portion of packets are operating at 13 Mbps having burst duration of almost a millisecond. This is good enough for a 720P call and as we've seen only a quarter of the RF spectrum is utilized for this application, meaning that co-existence with other applications is reasonable. However, any lower modulation and coding schemes than this and the 720P video call will not work. Once operation goes below the 13 Mbps data rate bursts, the video calling software will reduce the video quality due to packet loss and jitter measurements. And this will be particularly noticeable if any competing traffic or applications are sharing the spectrum.

Video Call with a smart phone over a wireless home network

A video call was set up between a PC and a smart phone. The display of the smart phone had a 4.3 inch diagonal and the video camera was 1080P. The smart phone connected over WiFi 2.4 GHz to a home wireless gateway with integrated WiFi and cable modem. A speed test application run on the smart phone measured a latency of 29 ms, a download speed of 5294 kbps and an upload speed of 7968 kbps. The gateway and the smart phone were separated by 36 feet one floor and two walls.

The other end of the call was a PC with a 1080P video camera and display connected to the Ethernet interface of a router and a cable modem. Two different cable modems were used in this test so that the video call packets would have to traverse the HFC network to the CMTS. The same CMTS terminated both cable modems in this test. The data rate and block diagram of the test is shown in figure 20.

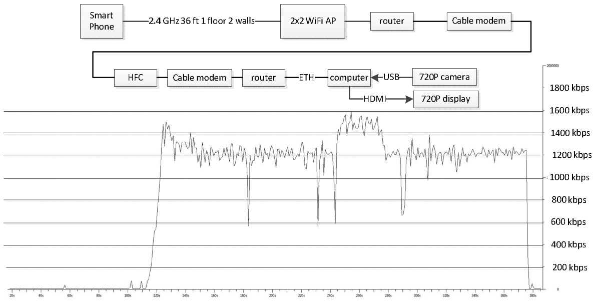


Fig. 20 Block Diagram and Data Rate of video call of smart phone with WiFi network connection.

The call technical information reported by the video calling software was monitored during the call. The number of relays was 0. The roundtrip time was 19 ms. The jitter was 69. The packet loss was 0.1%. The call lasted for 380 seconds or about 6 minutes.

The video send stream was 640x480 at 15 frames per second with H264 coding and 549 kbps bit rate. The video receive stream was 320x240 with H264 coding at 14 frames per second and a 605 kbps bit rate.

The number of packets captured for analysis was 60,348. The traffic protocol was UDP. The average data rate during the video call was 823 kbps and the average packet size was 648 bytes.

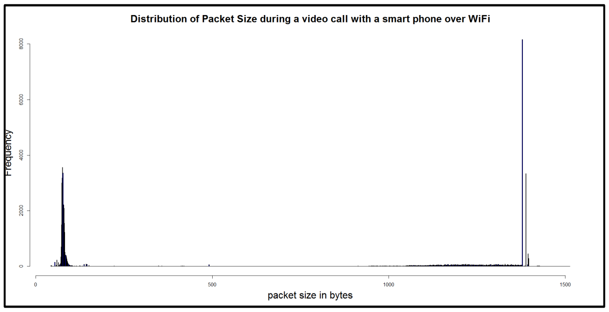


Fig. 21 Distribution of Packet Sizes during a video call using a smart phone with wifi network connectivity, x axis is packet byte size from 0 to 1500

Figure 21 shows the distribution of packet sizes during the video call using a smart phone with WiFi network connectivity. The packets are either very large or very small. The UDP video packets are typically about 1400 bytes whereas the WiFi signaling packets are typically less than 20 bytes in length. This explains the barbell type distribution of packet sizes.

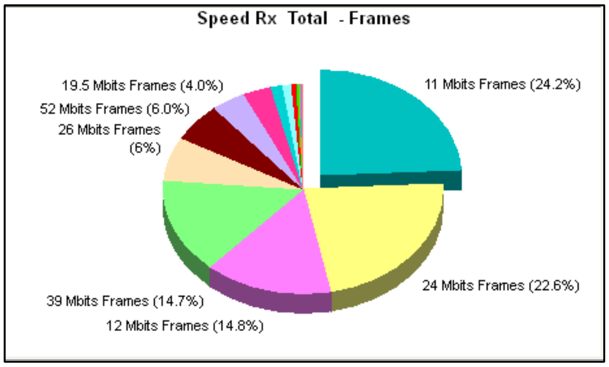


Fig. 22 Smart Phone over WiFi video call

Figure 22 shows the percentage of frame types with various data rates. A WiFi packet analyzer was used to create the pie chart. The majority of the packets had a data rate of 11 Mbps representing 24.2% of all WiFi packets sent. These packets are signaling packets, typically RTS,CTS, or Block ACK packets with short lengths of 16, 10, and 28 bytes respectively. 22.6% of the frames were 24 Mbps which are also signaling frames. The largest percentage of data carrying frames was the 39 Mbps frames representing 14.7 percent of the total number of frames. The 39 Mbps frames carry the large UDP video packets of about 1400 bytes of payload data. 14.8% of the frames were 12 Mbps. 6% of the frames were 26 Mbps. 6% of the frames were 52 Mbps. 4% of the frames were 19.5 Mbps.

During this video call using a smart phone with WiFi connectivity the WiFi analyzer captured WiFi network packets, the output was saved as a text file and a PERL program was written to calculate the burst duration of the 130,226 packets captured based upon the data rate and the byte size. The video call lasted 142.413 seconds and the period of time that the WiFi client was either transmitting or receiving was found to be 27,290,218 microseconds. Dividing the latter number by the former allows us to calculate that the utilization factor of the wireless spectrum during the video call was 19.2%. A histogram of the burst times is shown in figure 23. The predominate data rate of 39 Mbps for video packets of 1400 bytes which has a burst time of 336 microseconds is clearly indicated in the histogram. All in all it has been determined that a video call over a wireless home network using a smart phone has a data rate of 1 Mbps and uses up about one fifth of the wireless channel capacity.

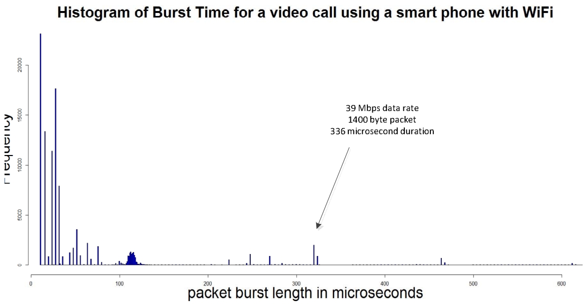


Fig. 23 Smart Phone over WiFi video call

Video Call over Cellular Wireless Networks

Video calls can be made over both 3G and 4G cellular networks. Here 3G networks refer to CDMA based networks and 4G networks refer to OFDM based networks since the characteristics pertinent to video calling varies considerably between these two multiplexing techniques. From a standards body standpoint, and from a service marketing standpoint, the use of the terms “3G” and “4G” is much more complex and nuanced and outside the scope of this paper.

Packet analysis was performed on a video call using a 3G cellular network lasting 1589 seconds or about 26 minutes. The video call quality was poor and the call dropped and re-established many times during the conversation. Still, the 3G end of the call was in the beautiful Florida Keys and the overall video calling experience was satisfying, refreshing to see a warm beach on a sunny day while huddled inside to avoid a cold grey Philadelphia winter. Video calling using a smart phone with a 3G data connection can be quite good at times as long as there is some tolerance for occasional disconnects, screen freezes, and fuzzy video.

Packets were captured with Wireshark on a PC with an Ethernet connection. The PC established a video call with another PC using a 3G cellular data card. The number of packets captured was 116,477. The average packet size was 295 bytes and the average data rate was 173 kbps.

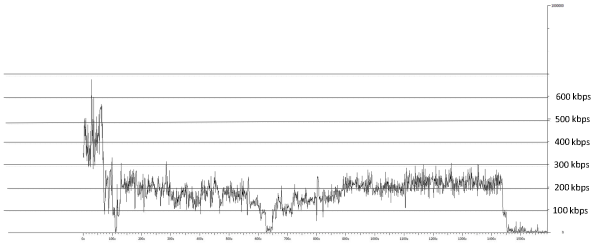


Fig. 24 Data Rate Measured During 3G video call

Figure 24 shows the data rate measured throughout the video call over the 3G cellular network. The data rate peaks at about 500 kbps, shows two steep drop offs where the call was lost, and otherwise runs at about 200 kbps.

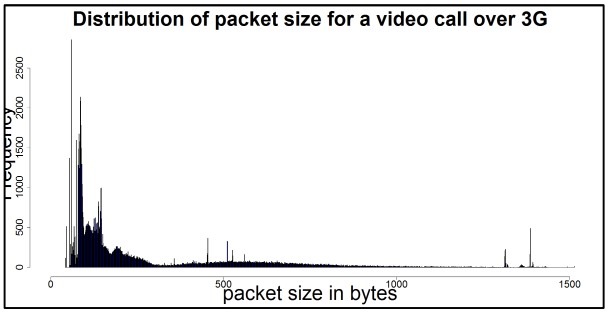
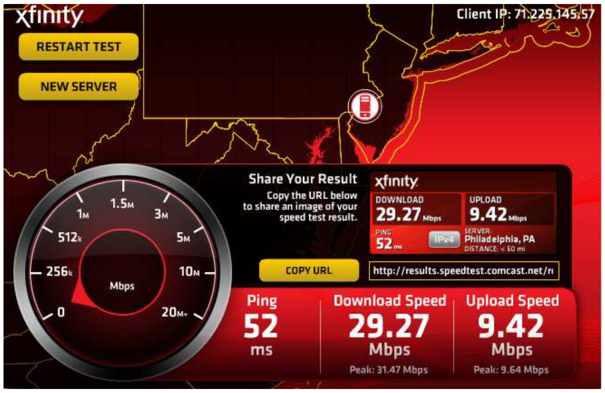


Fig. 25 Distribution of Packet Sizes of 3G video call.

The Wireshark packet analysis was exported to a text file, a PERL program was written to create an array of all the packet sizes for statistical analysis with the R statistical analysis tool. The resulting histogram is shown in figure 25 with the x-axis being the packet size in bytes ranging from 0 to 1500 bytes. By comparing the distribution of packet size between the 3G cellular network with that of the wireless home network, one notices that the packet sizes are generally much smaller when making a video call using the 3G network when compared to using a home WiFi network.



Client IP: 71.225.145.57

Share Your Result

Copy the URL below to share an image of your speed test result.

DOWNLOAD 29.27 Mbps
UPLOAD 9.42 Mbps
PING 52 ms
IPv4
SERVER Philadelphia, PA
DISTANCE: < 50 mi

http://results.speedtest.comcast.net/n

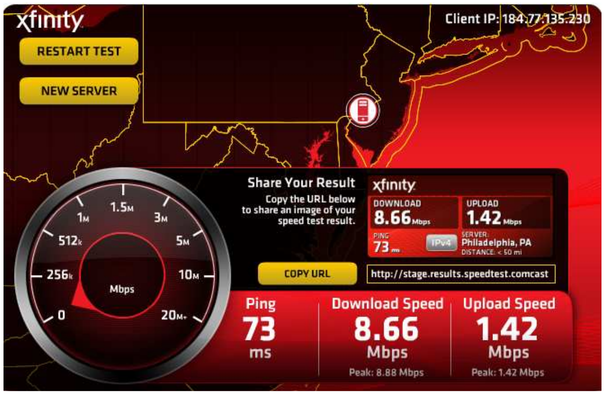
Ping 52 ms
Download Speed 29.27 Mbps Peak: 31.47 Mbps
Upload Speed 9.42 Mbps Peak: 9.64 Mbps

Fig.26 Speed test of 4G 700 MHz 10 MHz FDD pair

With the introduction of 4G networks having much higher data rates, and much lower latency and jitter, video calling over cellular networks will become better and more reliable. Like WiFi, 4G networks use OFDM which has a guard band in time to reduce inter-symbol interference as compared to a rake receiver or some type of adaptive equalizer used in CDMA networks. The adapter equalizer techniques used for single carrier wideband systems work well at times but require knowledge of the channel impulse response and so have difficulty under rapidly changing multi-path conditions. OFDM with a much simpler guard time inter-symbol interference mechanism can work even under rapidly changing multi-path conditions.

As the channel width increases the impulse response gets more complicated, requiring more taps for an adaptive equalizer and more calculations to respond to changes in multi-path conditions. This limits the channel width of CDMA based systems. The channel width used in the 3G video call of figure 24 and 25 has a 2 MHz channel width using CDMA. There are also 3G CDMA networks with 5 MHz channel width.

4G OFDM systems can operate at increased channel widths of 10 MHz, which gives them higher data throughput. Figure 26 shows the speed test results for a 4G network operating in the 700 MHz spectrum band with two frequency division duplexed, FDD, 10 MHz channel width signals. The download data rate is 29 Mbps and the upload data rate is 9 Mbps with 52 ms latency. These data rates, if maintained throughout the course of the call, are sufficient for 1080P video calling. One caution, cellular networks tend to be used in cars, buses, trains, or even when walking around and while moving throughout a geographical area the data rate will vary significantly and even switch from 4G to 3G and 2G coverage areas. So it is unlikely to always maintain these speeds while moving. In the area where video call testing was performed for this paper, 4G coverage was not available so video call testing and analysis was performed using a 3G network.



Client IP: 184.77.135.230

Share Your Result
Copy the URL below to share an image of your speed test result.

xfinity
DOWNLOAD 8.66 Mbps
UPLOAD 1.42 Mbps
PING 73 ms
SERVER: Philadelphia, PA
DISTANCE: < 50 mi

http://stage.results.speedtest.comcast

Ping 73 ms

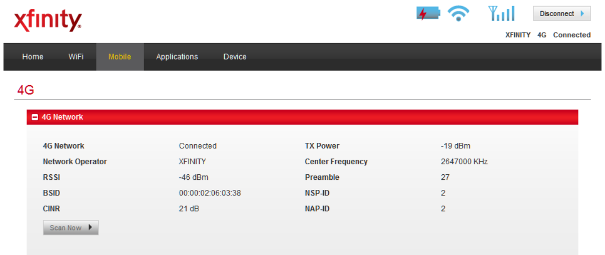
Download Speed 8.66 Mbps
Peak: 8.88 Mbps

Upload Speed 1.42 Mbps
Peak: 1.42 Mbps

Fig. 27 4G network 2.5 GHz, 10 MHz TDD

Figure 27 shows the speed test results of a 4G network operating in the 2.5-2.7 GHz spectrum band with a 10 MHz channel width using time division duplexing, TDD. The measured upload speed was 1.4 Mbps and the download speed was 8.6 Mbps with 73 ms latency. While the upload data rate was not high enough to support a 1080P video call, it was close to the 1.5 Mbps upload speed required for a 720P video call. The upload data rate of 1.4 Mbps is enough for a 500 kbps video send stream of a smart phone video call and the 8.6 Mbps download data rate has lots of room to support a 500 kbps receive video stream from a smart phone video call.

Figure 28 shows the parameters for the speed test results shown in Figure 27. The center frequency of operation is 2.647 GHz. The received signal strength is a very high -46 dBm indicating very good RF signal conditions and probable operation in close proximity to a base station. The carrier to interference and noise ratio was 21 dB. The transmit power was -19 dBm, the transmit power can go up as high as +20 dBm if the attenuation of the RF signal to the base station is high.



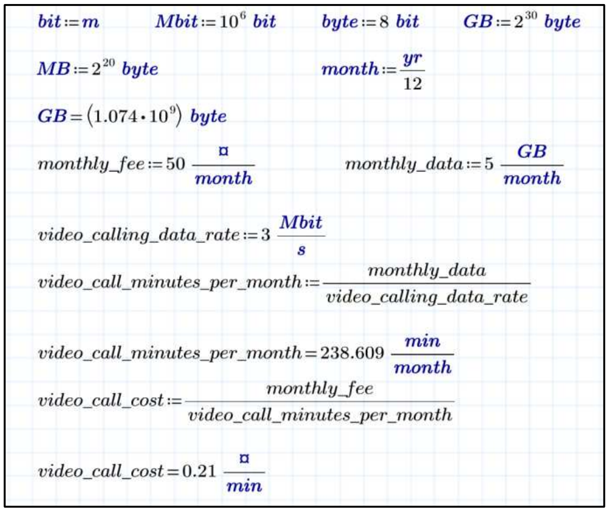
4G

4G Network

4G Network	Connected	TX Power	-19 dBm
Network Operator	XFINITY	Center Frequency	2647000 KHz
RSSI	-46 dBm	Preamble	27
BSID	00:00:02:06:03:38	NSP-ID	2
CINR	21 dB	NAP-ID	2

Fig. 28 4G Network 2.5 GHz, 10 MHz TDD

While 4G networks have the technical capability to make video calls under good RF conditions, will the costs impede usage? Figure 29 shows some calculations that translate a typical 4G data plan into some metrics familiar to many who in the past have bargain shopped for long distance plans based upon cents per minute or cellular phone plans based upon monthly minutes of talk time. The plan analyzed is a cellular data plan with 5 GB for $50 per month. If a video call is assume to have a data rate of 3 Mbps, the data rate measured for a 720P video call, then a video call is 21 cents per minute. At one point in time 10 cents a minute for a long distance plan seemed like a good deal. Cellular data plans with monthly minutes of talk time tend to be priced about 9 cents per minute. In terms of monthly talk time if one was to switch from voice calling to video calling, $50 per month with a 5 GB data cap gives one 238 minutes



$$bit := m \qquad Mbit := 10^{6}\ bit \qquad byte := 8\ bit \qquad GB := 2^{30}\ byte$$

$$MB := 2^{20}\ byte \qquad month := \frac{yr}{12}$$

$$GB = \left(1.074 \cdot 10^{9}\right)\ byte$$

$$monthly\_fee := 50\ \frac{¤}{month} \qquad monthly\_data := 5\ \frac{GB}{month}$$

$$video\_calling\_data\_rate := 3\ \frac{Mbit}{s}$$

$$video\_call\_minutes\_per\_month := \frac{monthly\_data}{video\_calling\_data\_rate}$$

$$video\_call\_minutes\_per\_month = 238.609\ \frac{min}{month}$$

$$video\_call\_cost := \frac{monthly\_fee}{video\_call\_minutes\_per\_month}$$

$$video\_call\_cost = 0.21\ \frac{¤}{min}$$

Fig.29 Calculating the cost of a 4G video call

of talk time with video calls at 3 Mbps. Voice cellular plans in this price range tend to offer about 450 minutes of talk time. So with these parameters, making video calls rather than voice calls tends to cost about twice as much. Of course, the assumed bit rate of the video call is the critical parameter. If the video calls were all 10 Mbps 1080P then it would be very expensive, 70 cents per minute and only 70 minutes of monthly talk time. However, on the other hand many folks may be quite content with making video calls on the road with a smart phone operating at 1 Mbps, in this case the cost per minute is 7 cents with 715 minutes of monthly talk time. These last numbers are roughly equal to the cost of cellular voice calls. So if you have a smart phone and a 4G data plan, live it up, make a video call instead of a voice call.

CONCLUSION

Many things have come together recently to encourage the use of video calling. Broadband connections in the home are faster than ever. Many homes have wireless home networks to connect mobile devices. In the past providing Internet connectivity to your television may have been inconvenient due to the lack of a nearby CAT-5 outlet. The wireless home network takes away the inconvenience. More and more people carry smart phones with WiFi and 3G or 4G network connectivity and these smart phones have front and back cameras and video calling application software.

Tests of video calls have shown that a 1080P video call can run at a symmetrical 10 Mbps, with the video send stream at 5 Mbps and the video receive stream at 5 Mbps. An excellent quality 720P video call on a large screen television set can run at 3 Mbps total data rate. 3-way video calls tend to run at about 2 Mbps. Smart phone video calling with a 4.3 inch diagonal display runs at about a 1 Mbps data rate over a home WiFi network. Video calling using a 3G cellular network runs at about 200 kbps with lower video quality and reliability.

Video signals are sent in packets of about 1400 bytes. Wireless home networks supporting video calls tend to have very concentrated packet size distribution around 1400 bytes and 20 bytes, representing the video packets and signaling packets, respectively. A typical packet sequence of a video call over WiFi lasts about 250 µs and consists of a request to send signal, a clear to send signal, the data packet of 1400 bytes, and a block acknowledgement signal.

Tests were performed of a 720P video call and a 1080P video call whereby all of the WiFi packets were captured. The captured packets contained information on byte size on the wire and data rate of the modulated burst. Accounting for the preamble length, the time length of each packet transmission was calculated and statistically analyzed. With the duration of the video call and the transmission time of each packet during the call determined, the percentage of time that the wireless home network was used by the video calling application was determined. For a 1080P video call under ideal RF conditions, the duty cycle was found to be 25%. For a 720P video call under threshold RF conditions the duty cycle was 20%. This indicates that most wireless home networks could support no more than 4 to 5 simultaneous video calls and that even during a video call over a wireless home network there is still over 75% of the capacity available for spectrum sharing.

Finally, the distribution of packet sizes of a video call using a 3G network was measured and analyzed. The packet size distribution shows that packet size in general was not as large when making a video call over a 3G network. The video calling application adjusted to the higher latency and packet loss of the 3G network in order to make the call while sacrificing quality. The speed of 4G networks was measured and reported indicating that 4G networks do support the data rates required for high quality video calling, at least under ideal RF conditions. The cost of video calls on 4G networks was analyzed and it was found that with today’s pricing plans, video calls of very high quality are more expensive than voice calls but not prohibitively so, while lower quality video calls on a smart phone screen today cost about the same as most common voice plans.

Wireless network connectivity is a crucial factor in encouraging the use of video calling. The signal strength is a critical indicator, wireless home networks should have signal strength indication of -60 dBm or higher for reliable video calling. Signal strength of -70 dBm for a wireless home network connection was found to be at the threshold of operation for a successful video call. The use of 5 GHz band can work better than 2.4 GHz band but only at close proximity. It was found that during a video call with a 3x3 AP and 3x3 client at 5 GHz in close proximity that 3 stream operation was very rare. It was also found that during a video call with a 2x2 AP and 2x2 client at 2.4 GHz at a 36 foot AP to STA separation distance that 2 stream operation was very rare.

REFERENCES

[1] Dorros, Irwin, “Picturephone”, Bell Laboratories Record, Vol. 47, Number 5, May/June 1969, pp. 136-141.

[2] Urban, D., Albano, C., Devotta, D., “Delivering DOCSIS 3.0 Cable Modem Speeds over the Home Network”, SCTE Cable-TEC EXPO’10, October 2010.

[3] Urban, D., Albano, C., Ong, I., Gilson, R.,”Designing a reliable wireless home network in a residential environment to optimize coverage and enhance application experience”, SCTE CABLE-TEC EXPO ’11, November 15-17 2011 Atlanta Georgia.

[4] IEEE Std. 802.11n-2009, “Amendment 5: Enhancements for Higher Throughput.

ABBREVIATIONS

AP Wireless local area network access point
STA Wireless local area network client station
OFDM orthogonal frequency division multiplexing
GI guard interval for OFDM
CDMA code division multiple access
HFC Hybrid Fiber Coaxial Cable network architecture
CM cable modem
RTS request to send WLAN signal
CTS clear to send WLAN signal
Block ACK Block Acknowledgement WLAN signal
WLAN Wireless local area network
LTE Long Term Evolution 4G cellular network
WiMAX type of 4G cellular network
4G OFDM based cellular network
3G High speed CDMA cellular network
RTT round trip time in ms
HT High Throughput WiFi mode
MIMO multiple input multiple output antennas
IFFT inverse fast Fourier transform
FFT fast Fourier transform
TDD Time Domain Duplexing
FDD Frequency Domain Duplexing

ACKNOWLEDGEMENTS

The author wishes to thank colleagues at Comcast and especially Chris Albano, Doug Jones, Erami Botchway, Sam Chernak, Chris Bastian, Jorge Salinger, and Neil Spence for help and support in writing this paper.

VIRTUAL ENVIRONMENT FOR NETWORKING TESTING AND DESIGN

Judy Beningson, Colby Barth, Brendan Hayes
Juniper Networks, Inc.

Abstract

This paper describes the use of virtual environments for the testing, design and modeling of networks. This paper will also explain the architecture and technology behind these virtual networking environments, and will highlight two real world use cases. The paper will also cover the benefits and limitations for cloud-based network modeling and testing to help operators determine the best uses.

INTRODUCTION

Operators who own and run IP transport networks understand that testing new protocols, design changes and/or modeling service introductions can be challenging. Most operators have access to a test lab for such purposes, but these labs have limitations in terms of scale and flexibility. Even the largest test labs do not approximate the size of an actual production network; smaller operators' labs may be non-existent or so small that any realistic control plane scalability testing is simply not feasible.

Due to size, budget availability and space limitations of current physical test labs, it can be difficult to test or design for the same level of scale as an operational network. Additional challenges result from the requirement for physical "racking and stacking". To test different topologies or configurations typically means making changes to physical connections and systems, which can be time-consuming and in some cases can have an impact on the number of test iterations.

Physical labs are also costly to both acquire and maintain. There is typically some level of capital outlay required for new projects, and once equipment is purchased, there are recurring costs associated with power, space, cooling and maintenance.

While physical labs are absolutely a critical part of any operator's test and design toolkit, because of the aforementioned limitations in terms of scalability, flexibility and costs many have considered the possibility of moving some testing and design exercises into the software realm. In fact, there exists several commercial and open-source software-based network simulation tools (e.g., GNS3, Olive), but these introduce another set of challenges and limitations. Generally these solutions are not officially supported by the major network equipment manufacturers, so features, protocol behavior and capabilities vary between what is available in software and what one will see on an actual network. For example, some of the router simulation software options lack forwarding capabilities. Other offline modeling tools can show results that diverge from actual world behavior. While these software solutions certainly have their place, to be able to test and design with confidence, one needs to conduct tests with the actual code that will run in your physical network.

To help fill the gap between physical test labs (realistic but limited scale and flexibility) and traditional software simulation solutions (flexible but limited realism), networking equipment vendors such as Juniper Networks are now offering cloud-based services that enable operators to create and run networks in a virtual environment. These environments enable users to create and operate virtual networks consisting of fully functioning router/switch "stacks" of network equipment operating systems. Some solutions also include virtual machines of the test equipment you would expect to see in a physical lab.

These cloud-based environments have the benefit of using virtual resources—so they are immensely flexible and scalable—and are also fully supported by network equipment vendors. This latter point ensures feature parity across multiple versions of router OSes and protocol consistency across both the virtual environment and physical gear.

Use Case	Virtual environment solution
Scalability	✓
Protocol interop	✓
OSS/BSS integration	✓
What-if scenarios	✓
Alternate Network architectures	✓
Training/Education	✓
Hardware testing	✗
Forwarding performance	✗

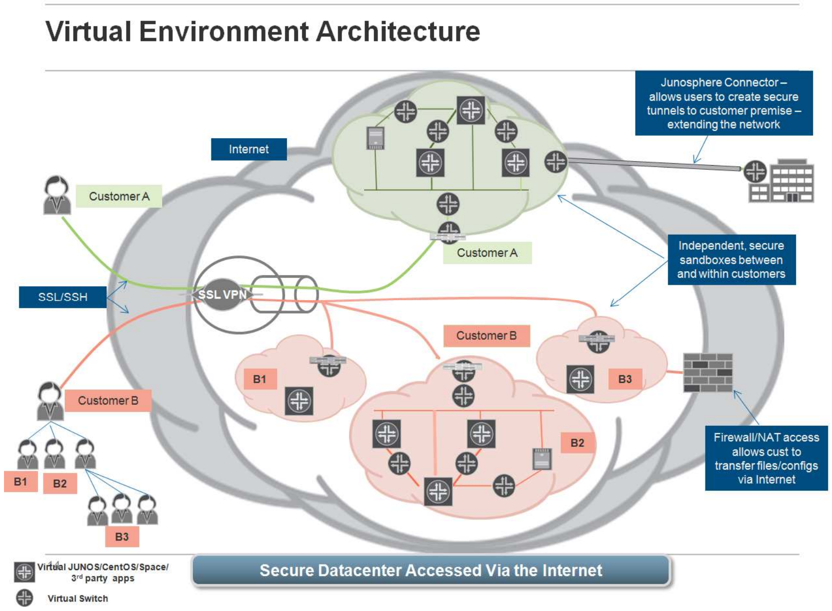
Table 1: Virtual Testing Environment use-cases

Within a virtual environment, operators can essentially replicate their production network and conduct test and design exercises with a level of scale and realism not otherwise possible, along with many other use cases. Refer to table 1. However, because it is a virtual environment, some tests are simply not possible. In this paper, we will outline the technology behind these virtual environments; examine some real-world use case examples; and discuss the benefits and limitations of such solutions.

VIRTUAL ENVIRONMENT

The network virtualization environment used for the tests described in this paper is a Juniper solution (marketed under the name Junosphere), and it is essentially used to create networks in virtual, rather than physical space. These virtual networks can be used for design, test and training exercises without the need for physical gear while providing a true instance of a router operating system (in this case, Junos) along with an emulated data-plane.
The key components of a virtualized networking system are:

- A secure, multi-tenant Data Center, optimized for high-speed networking between servers and network-attached storage
- A virtual machine (VM) management



Virtual Environment Architecture

Junosphere Connector – allows users to create secure tunnels to customer premise – extending the network

Independent, secure sandboxes between and within customers

Firewall/NAT access allows cust to transfer files/configs via Internet

Secure Datacenter Accessed Via the Internet layer customized for creation of network topologies

- A series of VM images, that users can load on demand
- A graphical user interface which allows users to save and store custom topologies as well as control permissions and access to the service

Each of these components is covered in more detail below.

Data Center

Because the virtual environment will be used to create and operate networks, the demands on it are quite different from most cloud environments or services, which traditionally are priced and offered based on compute power and/or storage. It would be very difficult to simulate a network in these environments, so it was necessary to build out an entirely new, next generation, cloud Data Center for the foundation of this virtual networking environment. The data center is a combination of Intel-based servers and network-attached-storage, with all Ethernet ports connected together via Juniper EX Ethernet switches. DC file upload and download protection is provided via high-end firewalls, and end-user topology access is secured via the SSL VPN gateway software. The cloud is located in a high-availability colocation facility that provides rack space, cooling, redundant power and high-speed, redundant Internet access. DC uptime is designed to be 24x7, 365 days per year, with service maintenance windows roughly occurring monthly. Finally, a publicly accessible URL completes the access.

Virtual Machine Manager

The real brain of the solution is the Virtual Machine Manager (VMM) software that handles the virtual machine creation and deletion as well as the unique job of VM inter-working. A purpose-built cloud for this virtual networking environment was required because we are building customer-specified networks of VMs, and not just leasing workload CPU cycles and/or access to storage.

The VMM used is a Juniper-developed KVM/QEMU-based solution that provides the ability to scale according to the size of the computing platform, offering support of complex network topologies as well as hosting a mixture of Junos, Unix and other 3rd-party VMs. VMM takes in via its API an execution script that, in conjunction with the Virtual Distributed Ethernet (VDE) switches, provide emulated Ethernet segments to which virtual machines are able to interconnect. VMs within a user's space are able to communicate over these emulated segments, the interfaces operating in the same way that a Layer2/Layer3 interfaces on a regular physical device would. VMM, thus, creates a "VMM topology" per customer which is a unique instantiation of the VDE Switch process, the number of VMs, and the type of VMs. The spaces are "secure"; VMs from User A are unable to communicate with those of User B.

Virtual Machine Images

During the instantiation of the VM by the VMM software, a personality (image file) is loaded onto the VM. This personality decides the operation of the VM. Within the virtual environment discussed in this paper, the available image files included:

- VJX1000 – a virtual version of a Juniper router/switch – that supports current releases of the Junos operating system. It is a "real" operating system, with an emulated forwarding plane capable of supporting all routing (MPLS, VPLS, v4, v6, multicast) and firewalling (stateful firewall) features. The virtual machine is able to operate as a regular Juniper device, without the need for hardware to be present.

- Junos Space - a network management application platform that can be used to provision, monitor, and manage Juniper devices
- Centos – a Unix host image for customers to add custom applications or host configurations
- Partner images from leading design and test vendors such as:
    - Cariden Technologies (MATE) [1]
    - Packet Design Insight Manager [2]
    - Spirent Virtual Test Center [3]
    - Mu Dynamics Studio [4]

This paper describes specific experiences, and therefore the images above are restricted to what was available within the existing virtual environment. It is possible that virtual machine image files representing other vendors or technologies could be incorporated into a similar virtual networking environment.

User Interface

The user interacts with the virtual network via a web-based user interface (UI) that lets users access the environment from any browser-equipped laptop or tablet. The UI is an application built as a multi-tenant provisioning tool for account, capacity and library management. It provides the GUI-based control of resources, allowing users to schedule their access times, store their topology files, and build their unique networks on-demand.

IN THE WILD

As previously mentioned, a virtual environment can provide significant value when trying to evaluate new technology and/or test specific large-scale protocol scenarios for a network. A physical lab environment is essential for router/switch hardware testing and validation but in almost all cases cannot provide the topologic resources to determine how a technology or protocol with act on an actual network.

In the next two sub-sections, we will discuss two scenarios where a virtual environment is used to validate network operation in the presence of new technology. For each use-case we will briefly describe the problem and/or challenge followed by a description of how virtual networking resources were used to solve the problem.

Use-case #1: Large Scale Core Network Scaling

In this example, an operator is trying to validate several simultaneous technologies to enable a more efficient method of scaling their core network. This represented a fundamental architectural shift that required a much more detailed test environment than could be provided by a set of off-line modeling tools and a few routers in a lab. The goal was two fold:

- Introduction of a MPLS “optimized” packet forwarding paradigm through the use of BGP labeled-unicast sub-address-family [5]
- Introduction of a multi-plane core architecture and the Aggregation/Edge connectivity

The network and technology migration is illustrated in the figure below.

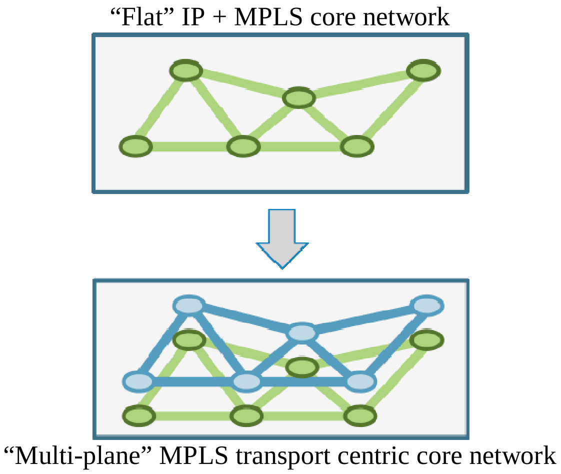

“Flat” IP + MPLS core network
“Multi-plane” MPLS transport centric core network

Figure 2: Network Architecture validation

The challenge the operator faced was how to conceptualize and visualize the target network, test the required protocol modifications, test the introduction of new protocols, and subsequently validate the forwarding properties in the network.

It was essential to be able to validate the changes on a mirror image of the current core network which consisted of a number (10’s) of PoPs geographically dispersed across the U.S. in order to ensure the correct routing policy changes, interaction of additional protocols, and validate the protocol architecture.

In addition to generally validating the modified network architecture, the operator now had a working virtual model of the target network in order to train their operations teams, practice and validate change-order methods and procedures as well a working documented target network.

Use-case #2: Protocol Scaling Characterization

In this use case, an operator wanted to very specifically characterize the memory and forwarding impact on their routing infrastructure if they enabled a new protocol extension. The protocol extension was a Border Gateway Protocol (BGP) extension called Add-path [6]. We will briefly describe BGP Add-path in the next few paragraphs before getting into the specific operator example.

BGP has implicit withdraw semantics on each of its peering sessions, where an advertisement for a given prefix replaces any previously announcement of that prefix. If the prefix completely goes away, then it’s explicitly withdrawn. BGP scaling techniques such as route-reflector and confederations are widely used in networks of all shapes and sizes. These techniques result in information hiding—for example, available backup routes are hidden. This may be good for scaling, but can problematic in other ways. BGP Add-path addresses some of these inefficiencies.

There are a number of reasons to enable BGP Add-path.

- Faster convergence, robustness and graceful shutdown schemes that require backup paths. This is because route reflectors eliminate backup paths.
- Stability and correctness schemes that require additional paths. For example fixes for MED oscillation or MED misrouting
- Multipath schemes that require multiple next hops
- And, implicit withdraw alone is potentially a problem for some types of inter-AS backup schemes

As you can see, much like the previous use-case, the operator was faced with multiple challenges:

- Would BGP Add-path provide the expected functionality?
- How would the additional BGP paths affect the routing resources of their network?

- Do they leverage the current BGP design or could further optimizations be realized?

It was essential for the operator to build a virtual representation of their current International core network to baseline BGP behavior and resource utilization. Another requirement was the need to be able to access and import, as closely as possible, their current peering locations in order to replicate the current BGP table "attributes".

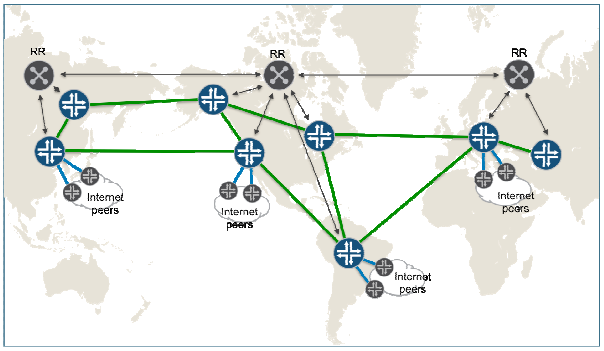


Figure 3: International Core network with regional route-reflectors (RR) for BGP scaling

The resulting virtual network representation allowed the operator to not only characterize their current design, validate BGP add-path and understand specific add-path configuration requirements but also developed multiple future architectural scenarios where indeed BGP Add-path not only delivered the required functionality but could also result in reducing the network resources required to scale BGP.

CONCLUSIONS

Virtual networking environments are a new development that leverage the technologies and concepts popular in cloud computing, and apply them in new ways to solve a fundamental problem for network operators. While virtualized environments will never be a complete replacement for hardware testing, they can provide the resources that allow operators to perform large-scale topology design or testing exercises that would not otherwise be possible. In this paper, we have outlined the technologies behind a specific virtual networking environment implementation, and several use cases, but these technologies and use cases can vary beyond what was discussed within the scope of this paper. In any form, virtual networking environments can be a powerful addition to an operator's design and testing toolkit.

FURTHER READING

QEMU/KVM references/publications

http://www.linux-kvm.org/page/Main_Page
http://wiki.qemu.org/Main_Page

Network virtualization references:

Flexible Cloud Environment for Network Studies:
http://edusigcomm.info.ucl.ac.be/Workshop2011/20110311002

BGP Route Reflection:

http://www.ietf.org/rfc/rfc2796.txt

REFERENCES

[1] http://www.cariden.com/
[2] www.spirent.com
[3] www.packetdesign.com
[4] http://www.mudynamics.com/
[5] http://tools.ietf.org/html/rfc3107
[6] http://datatracker.ietf.org/doc/draft-ietf-idr-add-paths-guidelines/

V-REX – VOICE RELEVANCE ENGINE FOR XFINITY

Stefan Deichmann, Oliver Jojic, Akash Nagle, Scot Zola, Tom Des Jardins, Robert Rubinoff, Amit Bagga
Comcast Labs

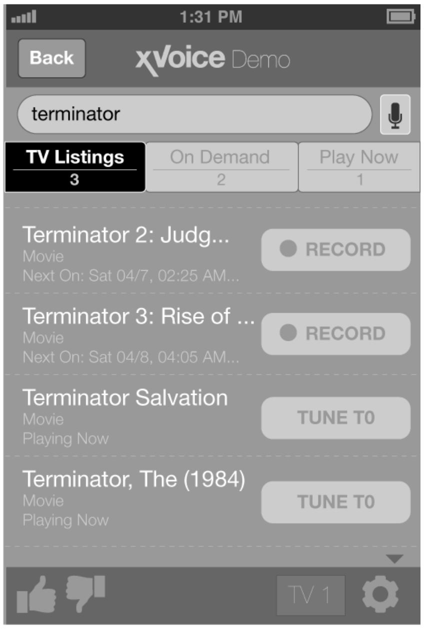
Abstract

V-REX is a new platform Comcast is building to provide speech-based applications for television control and other areas. V-REX applies automated speech recognition, natural language processing, and action resolution modules to interpret the user's request and identify the appropriate response. We describe here how we use V-REX to support an iPhone/Android app that allows users to control their cable set-top boxes by speaking into their phones. The primary focus of the work involves building grammar rules and dictionary entries for the range of requests the app can handle. We use the grammar and dictionary both to guide ASR and to allow NLP to extract the actions and entities in the request. We then convert these results into appropriate database queries that extract the information the user needs.

INTRODUCTION

Using a voice interface provides two advantages over traditional set-top-box remote or web interfaces. First, it eliminates the need for typing or other keyboard-based or remote-based text entry methods. (In the case of TV or cable remotes, this can be extremely tedious.) Furthermore, by allowing the use to directly specify what they want, it eliminates the need to wade through a series of menus or pages to find the desired option.

In order to provide a voice interface, we need to answer three questions: what words did the user speak, what action or information are they requesting, and how do we carry out the action or get the information? Each of these questions is answered by a specific module in V-REX. Automated speech recognition (ASR) identifies the words the user has spoken. Natural language processing (NLP) figures out what action or information has been requested. Action resolution (AR) responds to the request.



Terminator 2: Judg...
Movie
Next On: Sat 04/7, 02:25 AM...

Terminator 3: Rise of ...
Movie
Next On: Sat 04/8, 04:05 AM...

Terminator Salvation
Movie
Playing Now

Terminator, The (1984)
Movie
Playing Now

Figure 1 – Sample Results Display

Our initial V-REX application is an iPhone/Android app for Comcast customers that allows them to look up programs and control their set-top box. The app currently can handle three kinds of requests:

1) "What's on" – list what programs are available on a particular channel and/or at a particular time (including right now). The user can also specifically ask for sports games, or for a particular sport such as baseball or basketball.
2) "Tune" – switch the cable box to a specific channel
3) "Search/Find" – find when and on what channel a particular program is playing; this will also show if it is available in the On Demand library

For "what's on" and "search" requests, the results are displayed on the screen, and the user can select individual programs or channels to get more detail. For example, Figure 1 shows the response when the user asks "What's on CNN tonight"? In subsequent sections, we will describe each step of the process that produces this response.

AUTOMATED SPEECH RECOGNITION

The ASR module is built with CMU's Java-based toolkit, Sphinx4. We used Sphinx because it is a well-developed open-source system that we already had experience with.[1] We replaced Sphinx4's default acoustic model with one from VoxForge, a web site that collects transcribed speech for use with open source speech recognition engines. We built our own application-specific language model, which has two major parts, a pronouncing dictionary and a grammar, incorporating as well a general language model built on the English Gigaword Corpus (www.keithv.com/software/giga).

The dictionary maps words to their possible pronunciations at a phonemic level. The phonemes in our pronouncing dictionary are based on the ARPABet developed for speech understanding systems in the '70s. There are 39 phonemes, not counting variants due to lexical stress. Anything a person can say, including actions like "tune to" and channel names like ESPN, must be stored as a phonetic representation.

The grammar is a textual description of the combinations of words and phrases the system will accept. It is written in Java Speech Grammar Format, an augmented BNF-style format [1]. The grammar contains rules describing how users can ask to change a channel, what exactly counts as a title, and how people can ask about a time of day. Within a limited domain like television, the language model provides a higher level of accuracy in detection than it would normally achieve in an unconstrained system.

The dictionary and grammar are designed around three specific requirements of the application. The first is to handle channel names, many of which are not in the general vocabulary of the basic Sphinx system. For example, "SyFy" and "Tru TV" need to be added. In addition, some channel names may have multiple pronunciations, e.g. "Univision" can be spoken with either English or Spanish pronunciation; and some channel names may contain words that are in the general vocabulary but not commonly used together outside of the domain, e.g. "Fox Business" or "Showtime Family". To handle these cases, we added the names of all of the channels Comcast provides to the dictionary. Including the channel names in the dictionary does more than just improve recognition of these terms; it also lets us use a more precisely tailored language model, improving the overall ASR performance.

The second requirement is to handle a wide range of time and date specifications. While most of these are part of the general language, there are some that are specific to the television domain (e.g. "prime time"). More importantly, we want to directly handle complex phrases such as "next Tuesday evening after 10" and recognize them as indicating the time of a program as early in the processing as possible. To that end, we include in the grammar a large range of ways of referring to dates and times. A portion of this grammar is shown in Figure 2.

Finally, we need to recognize titles of movies and TV programs. This is a particular challenge, as titles can contain deliberate misspellings or ungrammatical phrases that would be rejected by a general language

[1] We are continuing to evaluate other ASR systems, but so far have found Sphinx4's performance and accuracy to meet our needs.

```
<temporalAdverb> = (   <weekday>
                     | on <weekday>
                     | this <weekday>
                     | prime time
                     | [right] now
                     | tomorrow <daytime>
                     | today <dayTime>
                   );

<dayTime> = ( morning | afternoon | evening | night );
<weekday> = (sunday | monday | tuesday | wednesday | thursday | friday |
saturday);
<clockTime> =
      (  at  <hour>  o'clock
      |  at  <hour> [ <minute>  ] [ <amOpm>  ]
      );
```

Figure 2 - A portion of the date/time grammar

model, e.g. "eXistenZ" or "De-Lovely", or subtitles that don't fit into normal sentence structure, e.g. "Dodgeball: A True Underdog Story". Even without these kinds of problems within a title, there is the danger of processing the title as a normal part of the whole sentence. For example "What time is Seven on?" is most likely asking about the movie "Seven", not channel 7 or seven o'clock. In order to handle these problems, we have added movie and TV show titles to our dictionary. This allows us to recognize titles when spoken (in places where titles make sense).

In order to add titles to the dictionary, though, we need to know which titles to add. We can't simply add **all** of the titles that have ever been produced, because that would involve several million titles. This would drastically increase the size of the dictionary, seriously diminishing both speed and accuracy of the ASR system. Furthermore, the vast majority of the titles would be for movies or TV shows that aren't currently available and that the user has almost certainly never heard of. Instead we limit the titles to all shows that are currently available (either on a broadcast or cable channel or on demand). We also include the most popular movies and TV shows, even if they are not available, since there is a good chance the user will ask for them.

The top level of the grammar specifies the range of possible requests the system can handle (and therefore needs to recognize). A simplified version of the grammar is shown in Figure 3. The three possible request types each have their own top-level rule, which is further specified in subsequent rules. The various parts of the grammar are combined to provide a language model that constrains and guides the recognition process.[2]

[2] As the application expands to handle a larger range of requests, we anticipate allowing some of the ASR work to use a more general statistical language model, so that the system can recognize unrestricted language. This will be particularly important when we start handling extended dialog. The grammar will stay play an important role in interpreting the request, though, as described in the section on the NLP module.

```
<whatsOn> = <actionPhrase> [ <modifierPhrase> ];
<tuneTo> = <tuneToPrefix> <channel>;
<search> = (<searchPrefix> <title> [now] | <title>);

<searchPrefix> = ( can i watch | play | search | find );

<actionPhrase> = <whatsOnPrefix> | <whatsOnPrefix> <channel>;

<tuneToPrefix> = ( (tune to | change the channel [to] | change [to] );
```

Figure 3 – a portion of the top level grammar

NATURAL LANGUAGE PROCESSING

Once the ASR module processes the user's request, the output (i.e. the request in text form) is sent on to the NLP module. NLP starts by parsing the text, using the same grammar used by ASR.[3] The text is parsed using the JSGF parsing facility, part of the package used to write the grammar (as described above). The NLP module uses the resulting parse tree to interpret the utterance, inferring the semantics from the rules used and the tags assigned in the parsing process.

For example, consider the request "What's on Disney on Saturday?"; the parse structure for this is shown in Figure 4. Here we can determine the request type from the <whats-on-phrase> node, the requested channel from the <channel> node, and the time constraint from the <temporal-adverb> node. These three different pieces of information are actually obtained in three different ways. The request type is determined to be "what's on" simply from the presence of a <whats-on-phrase> node, which reflects that the request has the structure and content of a "what's on" request. The channel is determined to be "Disney" based on the value of the <channel-name> node, which the grammar indicates is the appropriate value for the text "Disney". (In this case the channel name and the text are the same, but that is not the case for all channels.) The time constraint needs more complex processing, which is provided by special-purpose code in the NLP module that knows the range of possible parse structures and how to extract time and date values from them. Once it has identified all the pieces of the request, the NLP module assembles them into a request structure that is passed on to the AR module.

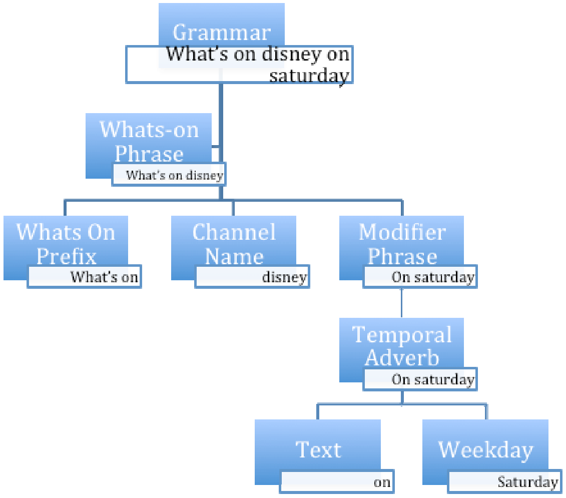


Figure 4 - Parse Structure for "What's on Disney on Saturday?"

[3] The two modules don't have to use the same grammar, although that is the case in the current system. It might be appropriate to use different grammars if, for example, we want to allow incidental comments that are not relevant to the request (e.g. "tune to HBO, please"). In particular, if we switch to using a statistical language model for part or all of ASR rather than a grammar-based one, NLP and ASR will need to use different grammars.

For the current application, we assume that any information not indicated in the request is deliberately left unspecified. For example, if no channel is mentioned, we assume the user wants to know the most popular shows available on any channel in the requested time span; if no time is specified, we assume the request is for programs on during prime time today. Missing information is therefore either left unspecified in the request or filled in with a default value.

In more complex applications, we might need to explicitly mark that the information is missing so that later processing can take necessary action to deal with the situation.

ACTION RECOGNITION

The AR module receives the request structure built by the NLP module and attempts to carry out the request. This involves constructing and sending an appropriate query to Comcast's REX search system. REX is the system we use to index and search through the complete set of programs available on broadcast and cable channels and on demand. The precise form of the query to REX depends on the type of action requested. For "what's on" requests, the query indicates the requested channel (if any) and time span. For "search" requests, the query indicates the title that the user specified. For "tune" requests, the query indicates the name of the requested channel. We need to query REX for tune requests to find the channel number corresponding to the channel name; if the request specified a channel number directly then we can skip the query.

The results returned by REX are packaged up along with an indication of the request type and the output of the ASR module and sent back to the client app. In case of an error, an appropriate error code is returned, along with any relevant information about the error. As mentioned above, a more complex application might require further interaction with the user, either to resolve an error or to get more information needed to carry out the request. The AR module contains a simple text-to-speech component, based on the FreeTTS system [2], to handle such interaction. This capability is not needed currently, though.

THE IPHONE/ANDROID CLIENT APP

The server-side components described above support a client iPhone/Android app that allows the user to speak requests and see and respond to the results. The initial screen for this app is shown in Figure 5; the user can press the microphone and speak a request. A typical response is shown in Figure 6; here the user has asked "what's on CNN tonight?" and the app displays the list of programs returned after processing through the ASR, NLP, and AR modules. The user can select a specific show to get more detail, as shown in Figure 7. Search requests display similar responses, except that the results are organized by relevance to the request rather than by time.

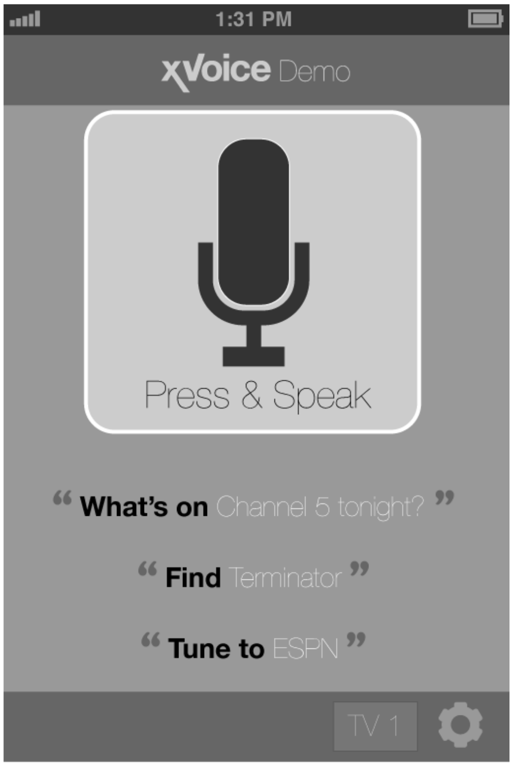
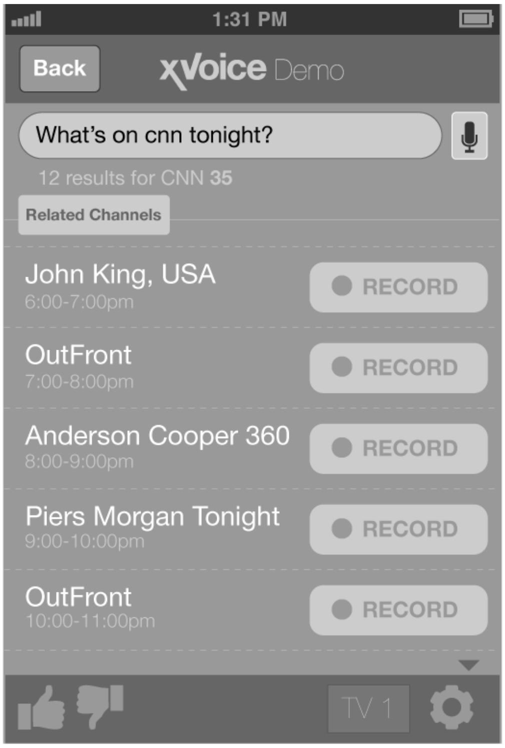


Figure 5 – Initial Client App Screen



What's on cnn tonight?

12 results for CNN 35

Related Channels

John King, USA
6:00-7:00pm

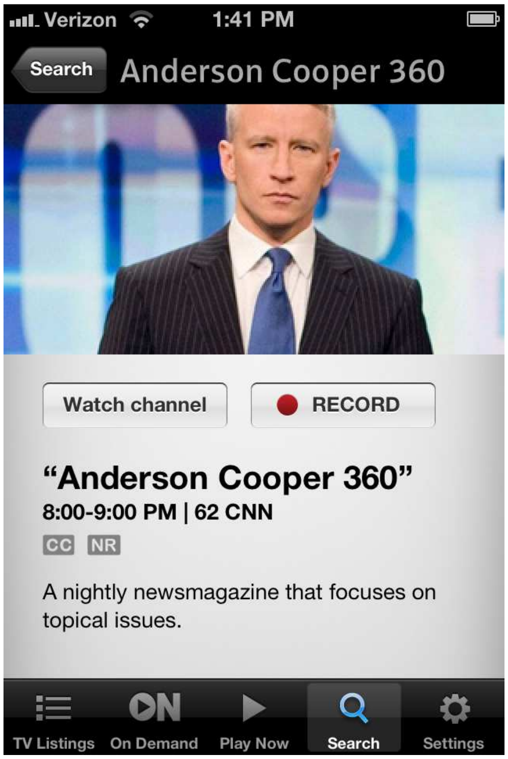
OutFront
7:00-8:00pm

Anderson Cooper 360
8:00-9:00pm

Piers Morgan Tonight
9:00-10:00pm

OutFront
10:00-11:00pm

Figure 6 – Response to "What's on" request



"Anderson Cooper 360"

8:00-9:00 PM | 62 CNN

A nightly newsmagazine that focuses on topical issues.

Figure 7 – Details of a specific program

CONCLUSION AND FURTHER WORK

Speech-based interfaces provide a uniquely simple and direct interface. Users can simply say what they want, without having to type in complex queries or navigate through layers of menus. The V-REX platform combines automated speech recognition, natural language processing, and action resolution to power speech interfaces. Our initial app brings this ability to searching and controlling cable set-top boxes. We are exploring ways to extend V-REX to other applications built on Comcast's cable/internet infrastructure. One possibility is to extend it to our Play Now system for watching programs over the web. A more interesting extension is to apply it to Comcast's home security service, so that people can easily check on the status of their homes while they are away.

REFERENCES

[1] http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/index.html

[2] http://freetts.sourceforge.net/docs/index.php

WHY 4K: VISION & TELEVISION

Mark Schubin
SchubinCafe.com

Abstract

Throughout the history of functional television, there have been moves towards higher definition, countered by the existence of lower-definition standards. Few of the choices of definition have been related to human visual acuity, however, which varies according to many factors.

The latest push for higher definition, to go beyond HDTV, is being driven in part by considerations unrelated to cable television, such as ease of program production and declining movie attendance and TV-set sales. The current era of bit-rate-reduced digital-video transmission, however, might nevertheless be an ideal time to offer consumers what could be the next level of increased picture definition.

THE ORIGIN OF DEFINITION

Language and Vision

Although citations for other senses of the word *definition* in the *Oxford English Dictionary* date back to the 14th century, the earliest citation for the sense relating to a manufactured system's "capacity to render an object or image distinct to the eye" is from 1878 (with two slightly older citations related to visual distinctness as rendered by an artist or in a natural formation).[1] The date might be associated with a different publication.

In 1862, Hermann Snellen published (in German) a book about something that he called *optotypes*.[2] In English, the book might be called *Sample Letters, for determining visual acuity.* The book introduced two concepts that have lasted to the present day: the idea that "normal" vision is 20/20 and the familiar letters on an eye chart, such as the *T* shown in Figure 1 below.

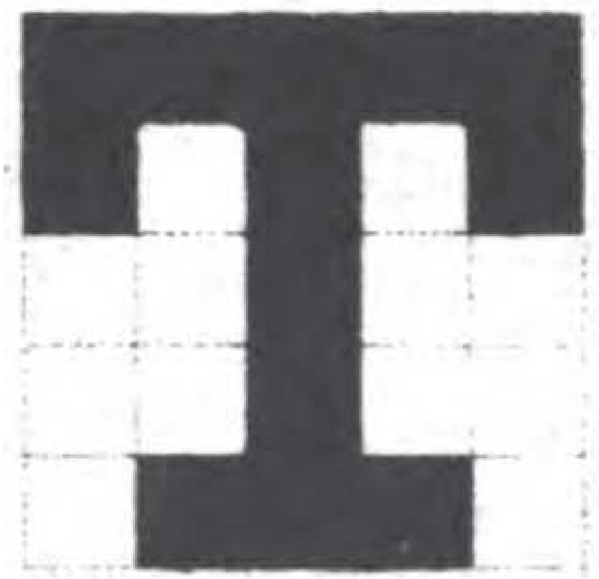


Figure 1: An 1862 Snellen Optotype

As the faint marks behind this optotype taken from his book indicate, the letter occupies a grid five units high by five units wide, and every element of the character, whether black or white, occupies one grid space. Snellen's definition of normal vision involved the ability to resolve features that subtended an angle of one arc minute (one-sixtieth of a degree) on the eye's retina.

If the whole character, therefore, were printed at a particular size and placed at a particular distance so that it would subtend an angle of five arc minutes, it would be able to be read with "normal" vision. The distance chosen was twenty feet so as to avoid visual issues associated with presbyopia (age-related inability to focus at short distances caused by the hardening of the eye's lens), a condition that usually first becomes noticeable around the age of 45.[3] In terms of visual focus, a distance of 20 feet is close to infinite.

Below is the familiar top of an eye chart based on Snellen's optotypes.[4] It was said in 1995 to have had more copies printed and sold in America than any other poster.[5] The top *E* on this chart is labeled "200 ft." on the left and "60 m" on the right.

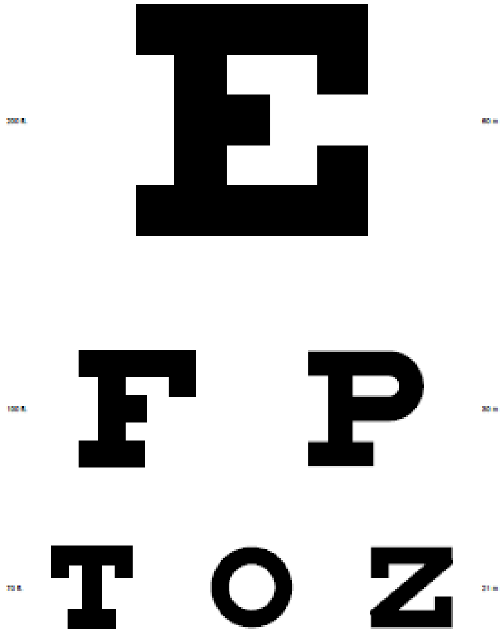


Figure 2: Top of an Eye Chart

The "200 ft." designation means that a person with "normal" visual acuity, as defined by Snellen, can distinguish that *E* at a distance of 200 feet, (or 60 meters). Vision defined as "20/20" (or "6/6") indicates an ability to see at 20 feet (or six meters) what a person with "normal" visual acuity can also see at 20 feet (or six meters).

Someone who could not read any letter smaller than the top *E* would be said to have "20/200" (or "6/60") visual acuity, the ability to read at 20 feet (or six meters) only what a person with "normal" vision can read at 200 feet (or 60 meters).

Issues Associated with the Definition

In nothing described to this point has anything been said about the illumination of the chart, the perceived contrast of the characters, or the definition of the edges of optotypes printed on the chart. Regarding the last, note, for example, that the edges of the *T* of Figure 1 are not as well defined as those of the *E* of the chart of Figure 2.

Snellen, himself, was aware of other issues associated with visual acuity. Below is a portion of another chart from his 1862 book. It shows not only an inversion of the color of the optotypes and the background but also a variation in contrast between the two lines of optotypes shown here. Snellen was clearly aware that contrast could affect visual acuity.

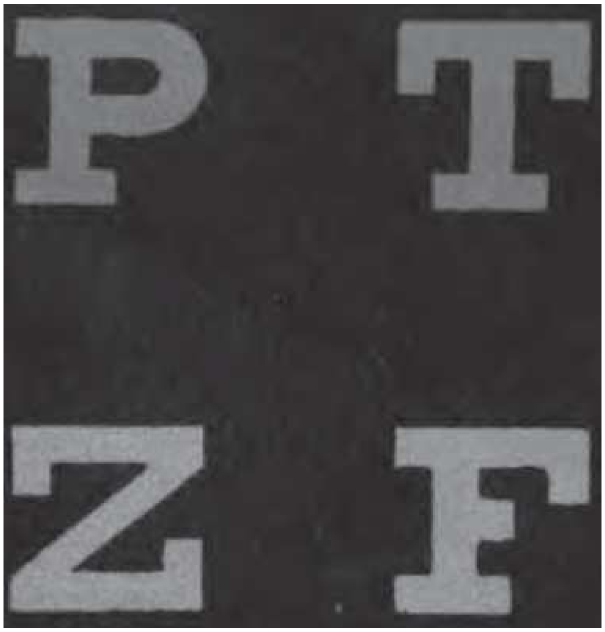


Figure 3: Portion of Snellen 1862 Chart with Color Inversion and Contrast Variation

As for the optotype edges, they contain higher spatial frequencies than the feature sizes would suggest. A pair of lines, one black and one white, suggest a cycle. If each line subtends a retinal angle of one arc minute, there would be 60 such lines in a degree. With half the lines white and half black, the spatial frequency could be said to be 30 cycles per degree (30 cpd).

Unfortunately, at a spatial frequency of 30 cpd, the edges of the optotypes would be soft.

The effect can be observed below. The *E* of Figure 2 was resized several times in image-manipulation software. Is it still readable as an *E*? It should seem clearer farther away.

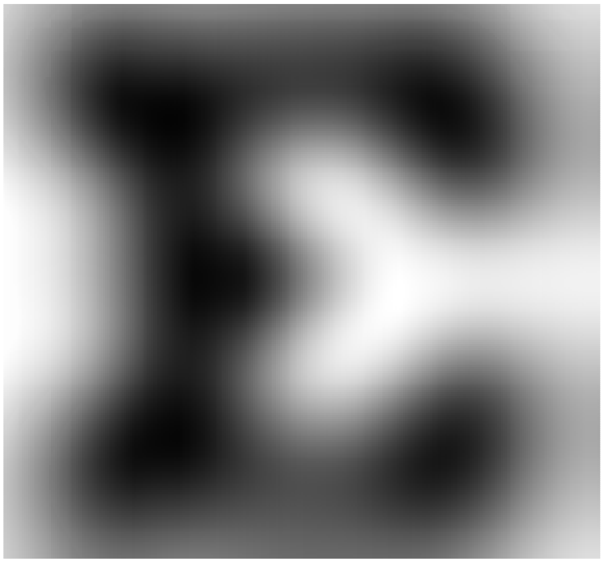


Figure 4: Snellen's *E* Filtered

The images below illustrate how sharp edges require higher spatial frequencies. Snellen's *E* is shown at the upper right. To its left is what its vertical strokes might look like if sinusoidal, and, to the left of that, a graph of the sine function between black and white.

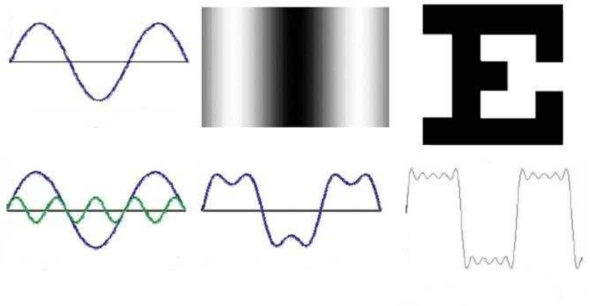


Figure 5: Adding Harmonics for Edges

In the lower row of Figure 5, at left is the same sine wave (the "fundamental") with another sine wave of three times the frequency (the third "harmonic") superimposed on it. To its right is the addition of those two waves. The transitions between dark and light are now shorter and steeper. At the far right is the sum of the fundamental and the third, fifth, seventh, and ninth harmonics. The transitions are shorter and steeper still. A perfect edge would require the sum of the fundamental and all of its odd harmonics, a square wave.

As for contrast, consider the image below. It's called a contrast-sensivity grating. Contrast increases from bottom to top, and picture definition (or spatial resolution) increases from left to right.

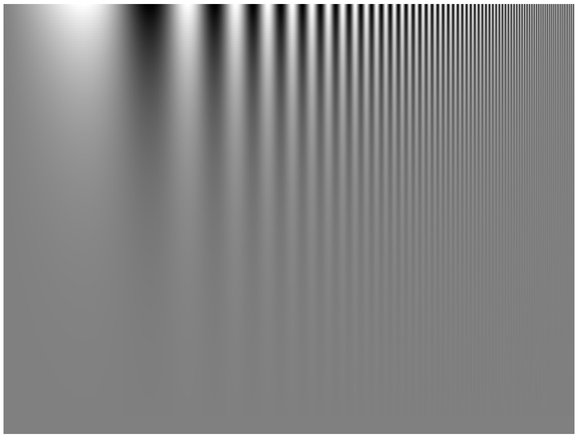


Figure 6: Contrast-Sensitivity Grating

Assuming normal printing or display and relatively normal (or corrected vision), the undifferentiated gray at the bottom of the grating of Figure 6 should appear to have a curve or "V" on top, the left and right edges higher than what is between them. In fact, there is no such curve in the grating. It is being added by the viewer's visual system.

Just as human hearing is most sensitive to middle sound frequencies, so, too, is human vision most sensitive to middle spatial frequencies (varying between about one and eight cycles per degree). Those who are familiar with the Fletcher-Munson curves of loudness sensation might find the visual contrast-sensitivity curve to be similar.[6]

An example of how important the contrast-sensitivity function (CSF) is in human vision may be seen in the pair of composite images of Figure 7 at the top of the next page. They

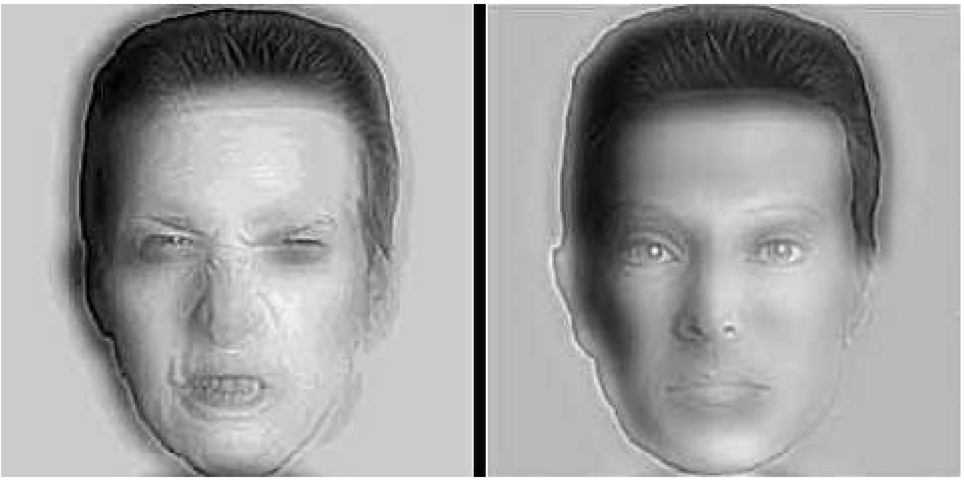


Figure 7: “Angry Man/Neutral Woman,” © 1997 Aude Oliva & Philippe G. Schyns

were created by Aude Oliva of Massachusetts Institute of Technology and Philippe G. Schyns of the University of Glasgow. They are used here with permission.[7]

To a viewer with relatively normal or corrected vision looking at the images from an ordinary reading distance, the image on the left will appear to be that of an angry-looking man, while the one on the right will appear to be that of an emotionally neutral-looking person, perhaps a woman. As the viewer moves farther from the images, however, there will be a distance at which both images appear to be of angry-looking people, followed by a long range of distances at which the angry man appears on the right and the neutral person on the left.

The composite images were created by combining two sets of images. One set, with the angry man on the left, has spatial frequencies intended to be seen near 6 cpd. The other, with the angry many on the right, has spatial frequencies intended to be seen near 2 cpd, in the lower insensitive section of a typical human CSF. As the viewer moves away from the images, both sets of spatial frequencies increase, the lower ones moving into the more-sensitive region of the CSF and the higher ones into the upper insensitive region of the function.

TELEVISION DEFINITION

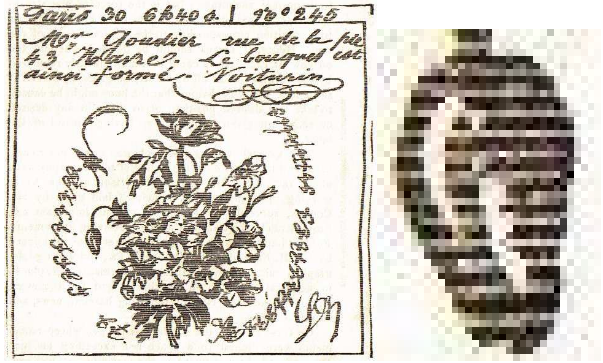
Early History

The Alfred P. Sloan Foundation’s Technology Series includes a book about the invention of television. Its preface has the following: “But who invented television? Nobody knows.”[8]

Nevertheless, as acknowledged by that book and many other sources, the first person to achieve a video image of a recognizable human face seems to have been John Logie Baird. And the first reception apparatus that he used operated with just eight scanning lines at eight frames per second (fps).[9]

That was a drop from the spatial definition of previous image-transmission systems. Although recognizable-face television wasn’t achieved until 1925, television proposals are older and actual, working facsimile-transmission systems older still. British patent 9745 was issued in 1843 to Alexander Bain for a fax system.[10]

A slightly later fax system, Giovanni Caselli’s Pantelegraph (developed in 1856) saw extensive commercial service. Figure 8, on the next page, shows an actual fax page received via Pantelegraph.[11]



Paris 30 6h40 s. | No 245
Mr Goudier rue de la pie
43 Havre. Le bouquet est
ainsi formé. Voiturin

Figure 8: Pantelegraph Fax and Portion

The image at left shows the complete fax page. The image at right shows a magnified portion of just the flower bud on the left side of the arrangement. Fifteen scanning lines can be counted in the bud, alone.

Below is a drawing from German patent 30105, issued in 1885 to Paul Nipkow for an "electric telescope." Television historian Albert Abramson called it "the master television patent" for its video scanning.[10] Each rotation of the scanning disk would produce one video frame. As the scanning disk drawing shows ("D1" through "D24"), Nipkow chose 24 scanning lines per frame.[12]

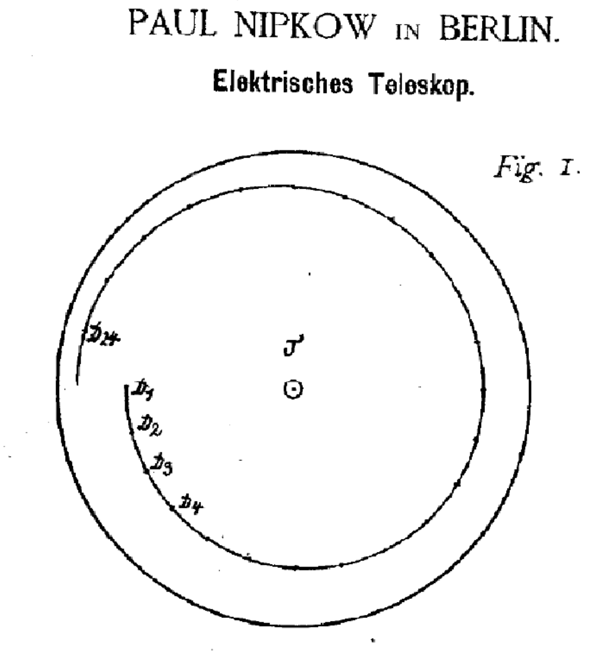


Figure 9: Nipkow's 24-line Scanning Disk

Although Baird and Nipkow might not have conducted studies of optimum image definition, Herbert E. Ives, who headed facsimile and television research at Bell Telephone Laboratories and was also an expert on photography, did. In his introduction to television in *The Bell System Technical Journal* in 1927, he described the definition requirement for what was, at the time, considered primarily an extension of one-to-one telephone service:

"Taking, as a criterion of acceptable quality, reproduction by the halftone engraving process, it is known that the human face can be satisfactorily reproduced by a 50-line screen. Assuming equal definition in both directions, 50 lines means 2500 elementary areas in all."[13]

The 50-line system was soon used, however, to capture larger scenes. In 1928, employees swinging a tennis racquet (as shown below) and a golf club were shown in a video demonstration, and a Bell Laboratories engineer was quoted as saying "We can take this machine to Niagara, to the Polo Grounds, or to the Yale Bowl, and it will pick up the scene for broadcasting."[14]

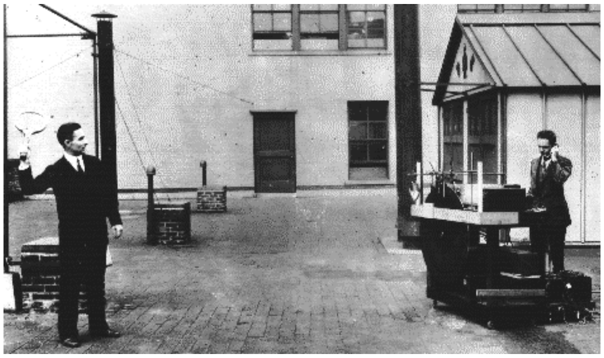


Figure 10: Tennis Swing Televised in 1928 [15]

In fact, the 50-line definition of the Bell Labs system was relatively high compared to that of most of its contemporaries. The second issue of *Television* magazine in the U.S., dated the same month as the Bell Labs demonstration, in its editorial content and advertising listed picture definitions of 24 through 50 lines.[16]

Only August Karolus, in Germany, went to higher definition in 1928. At the 5th German Radio Exhibition that year, he showed images with 96-line definition.[17] They are compared below to 30-line images from Dénes von Mihály at the same event.[18]

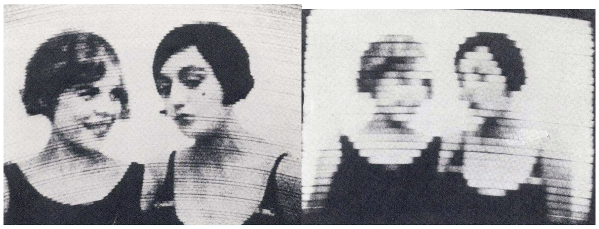


Figure 11: 96- & 30-line TV Pictures in 1928

The First High-Definition Era

Even before television moved from electromechanical scanning to all-electronic systems, there was great interest in higher-definition images. In 1935, a Television Committee, headed by Baron William Lowson Mitchell-Thomson Selsdon, reported to the British Parliament that the government should mandate "high-definition television." It was defined in paragraph 28: "it should be not less than 240 lines per picture...."[19]

Beginning in 1936, British television broadcasts alternated between a 240-line electromechanical system and an all-electronic system with 405 total scanning lines, of which 377 were active (picture carrying). When, in 1952, the Television Society (UK) heard a talk about the events that led to the 405-line broadcast standard, the presentation was called "The Birth of a High Definition Television System."[20]

The use of the term "high-definition television" wasn't restricted to the United Kingdom. Reporting on RCA's 441-line (383 active) television demonstrations at the 1939 New York World's Fair, *Broadcasting* magazine noted,"The exposition's opening on April 30 also marked the advent of this country's first regular schedule of high-definition broadcasts."[21]

When the first National Television System Committee (NTSC) began its work on a U.S. standard in 1940, it surveyed U.S. and non-U.S. proposed and working television systems ranging from 225 to 605 lines. Its last decision (*after* what was supposed to be the committee's final meeting) was a shift from 441 lines to 525 (483 active).[22]

Two other line-number standards saw significant broadcast use after World War II. They were an 819-line standard first broadcast in France in 1949 (with 737 active in the French version and a slightly higher number active in a Belgian version) and a 625-line standard first broadcast in Germany in 1950.[23]

The 625-line (575 active) number was later adopted by most of the world's countries, including France (1963) and the UK (1962).[24] The exceptions were those adopting the U.S.-standardized 525-line system. Although there were many different transmission systems (primarily for the 625-line countries), those two line numbers dominated the standardized, analog, all-electronic television period.[25]

In the previous, largely electromechanical television era, viewers could generally clearly perceive picture improvements with increasing line numbers, as in Figure 11, above left. There were, however, some anomalies even back then.

As early as 1914, Samuel Lavington Hart applied for a patent for interlaced scanning (scanning the image at a lower line number and then re-scanning at a slightly different position to fill-in between the lines), issued the next year.[26] Beginning in 1926, Ulises A. Sanabria applied a three-to-one interlace to electromechanical television, using three, slightly offset scans of 15 lines each to form a complete image of 45 lines.[27] Among advantages reported by an independent observer were a reduction of image flicker and line visibility (called "strip effect),[28] the opposite of complaints about interlaced scanning today.[29]

Studies have shown full, limited, or no effect on perceived definition from interlace over the number of lines in a single scan even in a still image.[30] Increased line visibility and flicker (or "interline twitter") in still images may be seen in video images, which are readily available on the Internet.[31]

Aside from interlace, many other factors could affect image-quality perception. In 1955, a delegation of U.S. engineers went to England to study television there and reported the perceived quality of the 405-line pictures superior to those of America's 525-line pictures. Possible reasons ranged from better operational practices to better allocation of bandwidth to different filtering.[32]

Combining Vision and Television

Ignoring all other television technical characteristics (in scanning system, scene, lens, camera, transmission, reception, display type, and display settings) that could affect image definition, many have reported an optimum viewing distance based only on the number of active scanning lines and Snellen's "normal" visual acuity of one minute of visual arc per scene element. According to that theory, NTSC's approximately 480 active scanning lines, if filling eight degrees of visual arc, would exactly match one arc minute of acuity. That condition occurs when the viewer is slightly farther from the screen than seven times the picture's height.[33]

It is the case that there is an optimum viewing distance. Farther than the optimum distance, the viewer's visual acuity precludes seeing the full resolution being presented. Closer than the optimum distance, elements of the display structure become visible, effectively preventing the viewer from "seeing the forest for the trees."

Consider the image below, an extreme example of this phenomenon. It is possible that you have seen it previously in this paper. It is a lower-case *O* in the Times New Roman typeface used in this text, as it might be depicted on some computer's color LCD screen. As in Figure 7, at a sufficient viewing distance (squinting might help), the image will change, this time from colored blocks to a round, black, lower-case *O*.

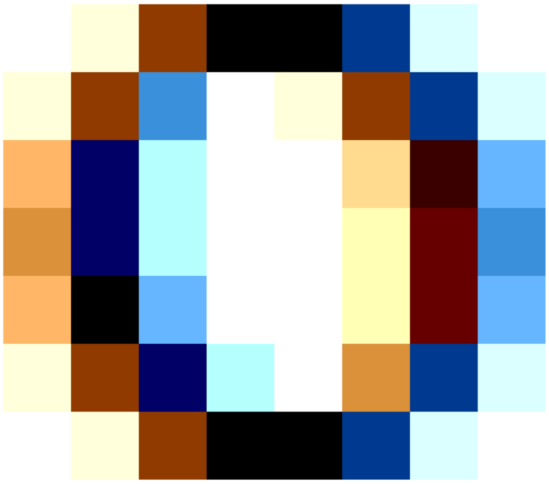


Figure 12: A Fixed-Grid Display *O*

If Snellen were correct about normal visual acuity being 30 cpd, and if active scanning lines (in interlaced television) directly determined perceived resolution, then it *would* be accurate to say that the optimum viewing distance for NTSC video is about seven times the picture height. It would still *not* be accurate, however, to say that viewers typically watched NTSC video at seven times the picture height.

Bernard Lechner, a researcher at RCA Laboratories, conducted a survey of television-viewing distances during the NTSC era. What he found was that those he surveyed watched television from a viewing distance of approximately nine feet, regardless of screen size, a figure that came to be known as the Lechner Distance. Richard Jackson, a researcher a Philips Laboratories in the UK, conducted a similar survey and found a similar three meters, the Jackson Distance.[34]

Assuming, again, that normal visual acuity allows detection of features subtending one minute of visual arc and that active scanning lines determine television resolution, then at the Lechner Distance it would be essentially impossible to see greater than NTSC resolution on a 25-inch four-by-three TV screen. The picture height of a 25-inch screen is 15 inches (1.25 feet); seven times its height is 8.75 feet. The optimum viewing distance for 480 active lines is actually 7.15 times the picture height (1/(2*tan(8 degrees/2))), which means the optimum 25-inch NTSC TV-viewing distance would be 8.94 feet, almost exactly the Lechner Distance.

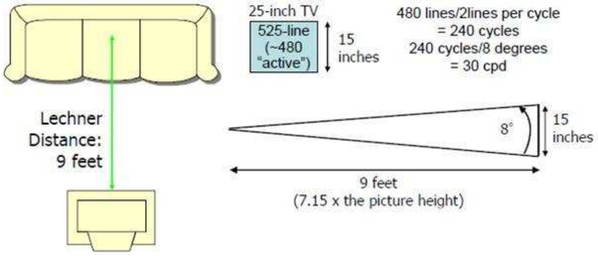

25-inch TV
525-line (~480 "active")
15 inches
480 lines/2lines per cycle = 240 cycles
240 cycles/8 degrees = 30 cpd
Lechner Distance: 9 feet
8°
15 inches
9 feet
(7.15 x the picture height)

Figure 13: U.S. Standard-Definition Viewing

According to figures from the Consumer Electronics Association, the average TV-screen size shipped by factories to U.S. dealers through the year 2000 was under 25 inches.[35] Sales to consumers typically follow a year after factory sales, and TV replacement takes years after that, so, if definition beyond NTSC cannot be perceived on a 25-inch screen, there would seem to have been no incentive for a move to HDTV in the U.S.

In Japan, according to this theory, rooms are smaller, so current HDTV had its origins there. Unfortunately for the theory, TV screen sizes in Japan were also smaller.

Psychophysics

In fact, there *are* reasons why HDTV detail looks better even to American viewers. First, 30 cpd is *not* the limit of human visual perception. In testing conducted in Japan on ultrahigh-definition television (UHDTV), the test subjects were found to have an *average* visual acuity not of 20/20 but of 20/10 (i.e., able to distinguish features twice as small as Snellen's criterion).

That should have meant that their visual acuity was 60 cpd instead of 30. The testing revealed, however, that the subjects were able to distinguish the "realness" of images as high as 156 cpd (the highest spatial frequency that was measured), more than five times supposedly "normal" acuity and more than 2.5 times even 20/10 acuity. "Realness" (degree to which an image was perceived to be comparable to a real object) rose rapidly to beyond 50 cpd and then slowed, but it did increase to 156 cpd (the highest tested), and the data suggest it would continue to increase (slowly) beyond that point.[36]

Realness, like other words ending in the suffix -*ness*, such as *brightness* and *loudness*, is a psychophysical sensation (a psychological response to a physical stimulus). And psychophysical sensations tend to be based on more factors than just the measurable physical phenomenon most closely associated with them. Thus, luminance, alone, does not determine brightness, and sound-pressure level, alone, does not determine loudness.

Similarly, there is a psychophysical sensation associated with picture definition but not determined exclusively by it. That sensation is called *sharpness.*

At the top of the next page is a graph of a typical modulation transfer function (MTF). In the case of spatial definition of the luma (gray-scale) component of images, the modulation is changes between bright & dark.

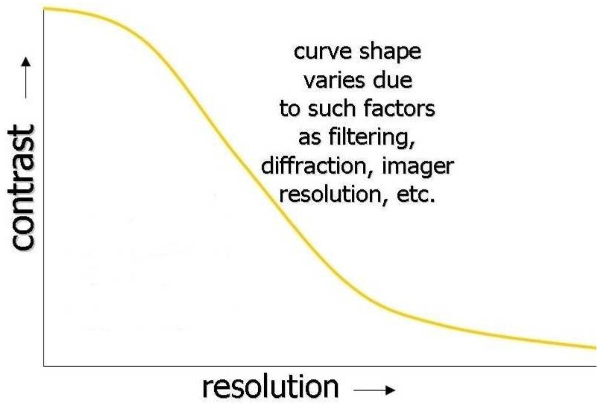


Figure 14: An MTF Curve

The curve of Figure 14 could be that of a lens or a television camera or a complete television system, "from scene to seen." Of most interest, with regard to the sensation of sharpness, is the area under the curve.

There are two significant schools of thought about the relationship of that area to sharpness. One is based on the work of Otto H. Schade, Sr. at RCA Laboratories.[37] The other is based on the work of Erich Heynacher at Zeiss.[38] The former suggests that the sensation of sharpness is proportional to the square of the area under the curve, the latter that it is proportional to the area.

In either case, as shown in Figure 14, image sharpness is most affected by the area under the "shoulder" of the curve (at left) and least by the area under the "toe" of the curve (at right). Sony took advantage of the low contribution of the highest spatial frequencies to sharpness in the design of the HDCAM recording system. It drops 25% of the luma definition at much less loss of sharpness.[39]

One of the factors affecting the shape of the MTF curve is number of digital samples. Digital sampling and reconstruction require filtering. The graph at the top of the next column is a basic filter shape, the so-called SINC or (sin x)/x function. The horizontal scale is arbitrary; on the vertical scale, *1* represents 100% modulation transfer.

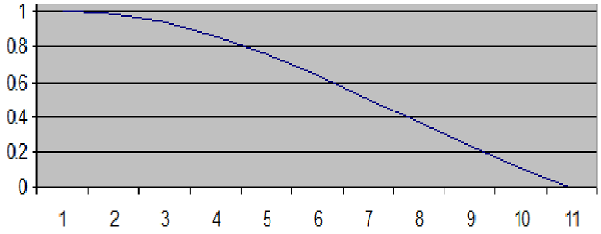


Figure 15: SINC Function Filter

If the vertical axis represents contrast and the horizontal axis represents image definition, then, if number 11 represents, say, 1920 active samples per line, the contrast at 1920 is zero. If, however, number 11 represents 3840 samples, the contrast at 1920 is 64%, a very significant difference.

The next two figures illustrate real-world examples. They are taken from the Bob Atkins Photography web site and are used here with permission.[40]

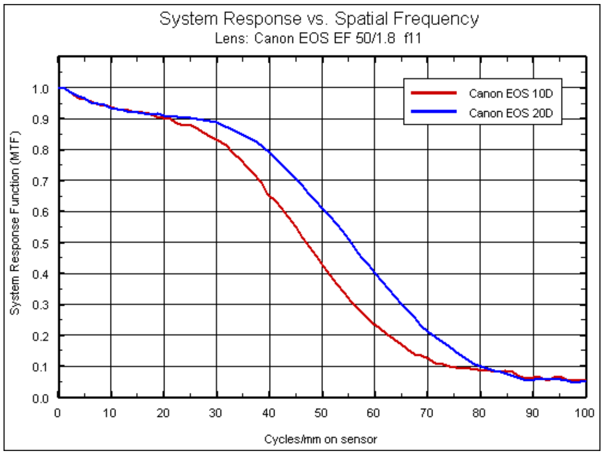


Figure 16: MTF Comparison of Two Cameras
© Bob Atkins

The red (left) curve of Figure 16 is from a Canon EOS 10D still camera, which uses an effective 3072 x 2048 photosite image sensor. The blue (right) curve is from a Canon EOS 20D camera, which uses an effective 3504 x 2336 photosite image sensor.

The horizontal linear increase in definition is just 14%, but significant additional area is created under the MTF curve of Figure 16.

The resulting increase in sharpness can be seen in Figure 17 below.

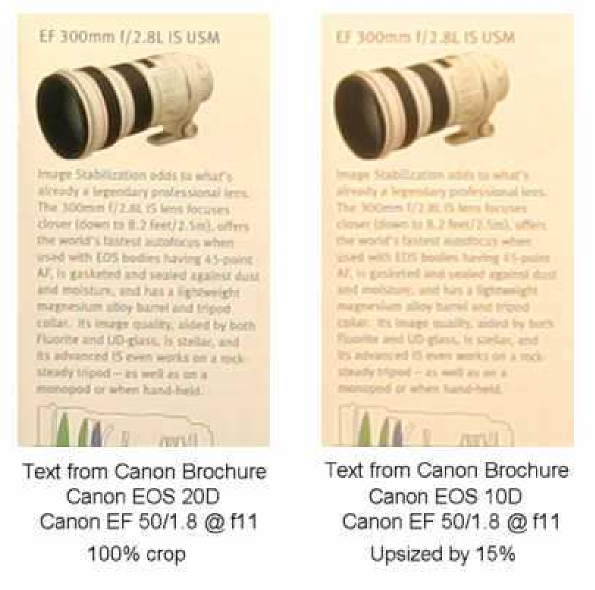

EF 300mm f/2.8L IS USM

Image Stabilization adds to what's already a legendary professional lens. The 300mm f/2.8L IS lens focuses closer (down to 8.2 feet/2.5m), offers the world's fastest autofocus when used with EOS bodies having 45-point AF, is gasketed and sealed against dust and moisture, and has a lightweight magnesium alloy barrel and tripod collar. Its image quality, aided by both Fluorite and UD-glass, is stellar, and its advanced IS even works on a rock-steady tripod – as well as on a monopod or when hand-held.

EF 300mm f/2.8L IS USM

Image Stabilization adds to what's already a legendary professional lens. The 300mm f/2.8L IS lens focuses closer (down to 8.2 feet/2.5m), offers the world's fastest autofocus when used with EOS bodies having 45-point AF, is gasketed and sealed against dust and moisture, and has a lightweight magnesium alloy barrel and tripod collar. Its image quality, aided by both Fluorite and UD-glass, is stellar, and its advanced IS even works on a rock-steady tripod – as well as on a monopod or when hand-held.

Figure 17: Sharpness Difference
© Bob Atkins

The text, photo, and drawing details above indicate very little difference in image definition, as might be expected from the small (14%) linear increase. The additional area under the MTF curve, however, makes the increased sharpness of the left image readily apparent. The vertical definition difference between so-called 1080-line HDTV and NTSC is about 225% (it is a similar 213% from 1080-line HD to so-called "4K").

BEYOND HDTV

Differentiating the Theatrical Experience

Average U.S. weekly movie attendance in every year from 1945 through 1948 was 90 million. By 1950, it was down to 60 million; by 1953, it was just 46 million.[41] The cause of the drop was apparently the rise of home television. The movie industry turned to wider screens, larger (higher-definition) film formats, and such offerings as stereoscopic 3D in order to differentiate the movie-going experience from that of watching television.

In 2011, there were just 24.6 million weekly cinema admissions in the U.S.[42] Only 27% of the number of movie tickets of 1948 were sold in the same year that the population grew to 213% of its 1948 level.[43] So 3D and higher-definition formats are still under consideration in a digital-cinema era.

Today, instead of 70-mm film, the movie industry discusses "4K," images having 4096 active picture elements (pixels) per row (with a number of rows appropriate to the image aspect ratio). The earliest digital-cinema projectors were "1.3K" (comparable to 720p HDTV); many current installations are 2K (comparable to 1080-line HDTV).[44]

Traditional cinema seating arrangements created a wide range of viewing distances for audiences, as shown in Figure 18 below, courtesy of Warner Bros. Technical Operations. Figure 19, courtesy of the same source, shows a typical more-recent auditorium with stadium-style seating. The scales are calibrated in picture heights.

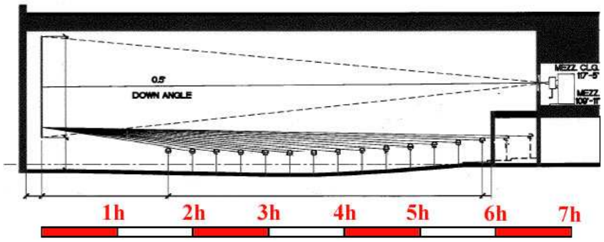


Figure 18: Traditional Cinema Seating

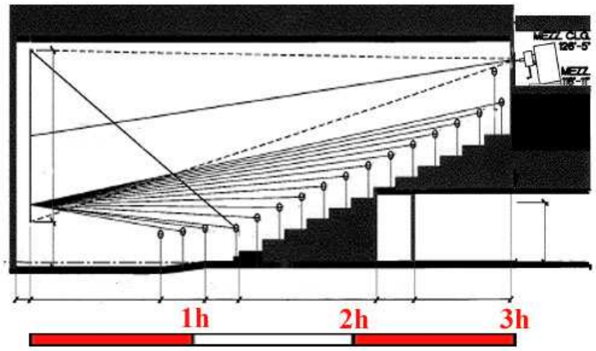


Figure 19: Cinema Stadium-Style Seating

The bulk of the audience is closer to the screen in stadium seating and, therefore, might benefit from additional image definition. Figure 20, below, courtesy of ARRI,[38] shows that definition even beyond 8K (8192 active pixels per row) should be perceptible even at just 20/20 visual acuity in some seats in some cinemas.

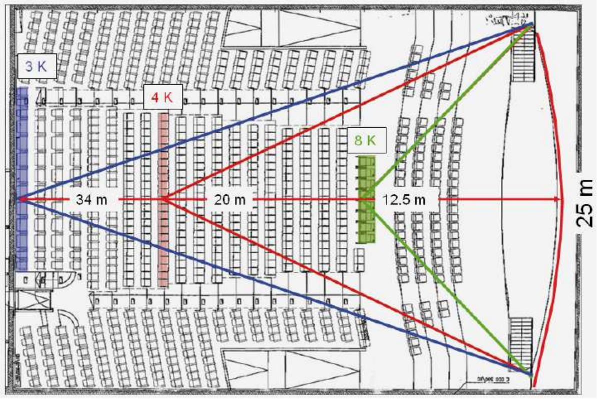


Figure 20: 20/20 Resolvable Definitions

In practice, however, unlike television consumers, who can compare picture definitions side by side in stores (and generally at closer viewing distances than they would experience in homes), cinema attendees cannot easily compare definitions in different auditoriums. Definition of 1.3K was acceptable to audiences when it was used for digital cinema. Perhaps, like "70 mm," "4K" will be a promotional tool.

Production and Post

Even before the modern HDTV era, television-camera manufacturers used oversampling (generally in the horizontal direction) to increase image sharpness. NTSC broadcast video bandwidth restricts horizontal luma definition to approximately 440 pixels; some broadcast-camera image sensors had more than 1100 photosensitive sites per row.

At the beginning of the modern HDTV era, the difficulty of making sensors with even as many as 1920 photosites per line precluded oversampling. The introduction of image sensors into still cameras, mobile telephones, laptop computers, and other devices, however, provided economies of scale allowing multi-"megapixel" sensors to be produced.

Even in the camera-tube era, some cameras used patterned color filters at the faceplate of a single imaging tube to capture color pictures instead of using color-separation prisms with an imaging tube for each of the three primary colors. In the solid-state imaging era, it has become common in still cameras to use a patterned filter over a single image sensor.

In the common Bayer pattern shown below, green-filtered photosites represent half the sites on the sensor. Red- and blue-filtered photosites represent one-quarter of the sites, each. Recreating a full-color image requires a "demosaicking" process to remove the spatial color effects of the filter.

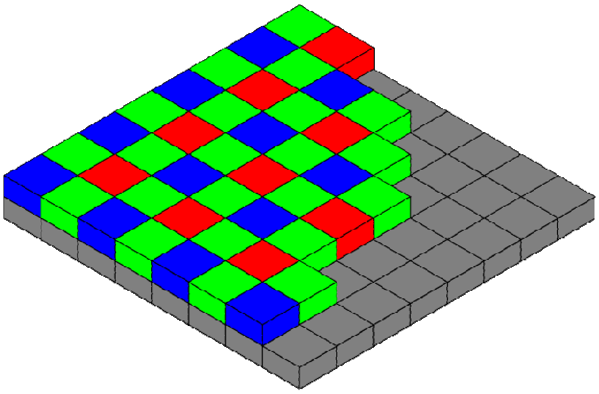


Figure 21: A Bayer-Filter Pattern[45]

There is no consensus about how on-chip color filtering should affect the description of resolution. There is also no consensus about whether 4K requires 4096 samples per line or whether 3840 (twice HDTV's 1920) are sufficient. Thus, one can find labeled 4K, at a single equipment exhibition, cameras with between 8.3 and 20.4 million photosites per image sensor, and with one to three sensors (some older "4K" cameras also used four 2.1-million photosite sensors).[46]

Aside from any advantages in visual definition or sharpness, 4K offers benefits in production and post, as the next three figures illustrate. Figure 22 shows a high-definition video image as a pixel-for-pixel subset of a 4K image, allowing reframing or even zooming after shooting.

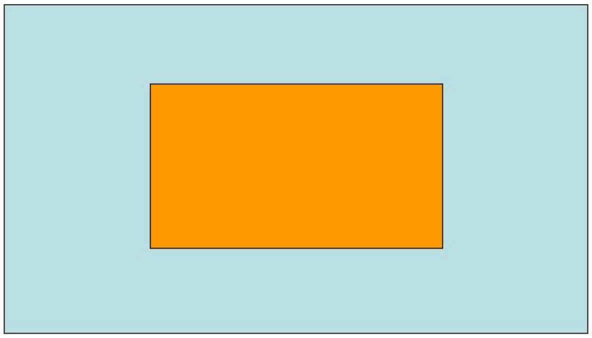


Figure 22: HD as a Subset of 4K

Figure 23, below, shows the effects of image stabilization, which normally causes trimming of the image (illustrated in available short, downloadable video clips[47]). The light inner rectangle (behind the others) is a desired HD image. The skewed rectangle in front of it is the actual image captured by an unstable camera. The smallest rectangle is the trimming that post-production image stabilization would produce. But starting with the outer 4K image allows the full desired HD image to be stabilized.

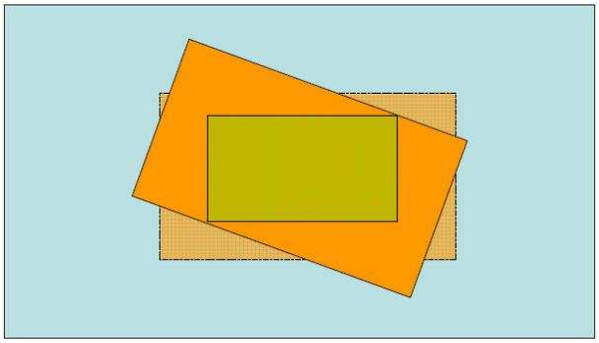


Figure 23: HD Image Stabilization in 4K

Figure 24, at the top of the next column, shows an unusual application of 4K in stereoscopic scene capture. The Zepar stereoscopic lens system attached to a Vision Research Phantom 65 camera, provides side-by-side stereoscopic images on the same image sensor, simplifying processing.[48]

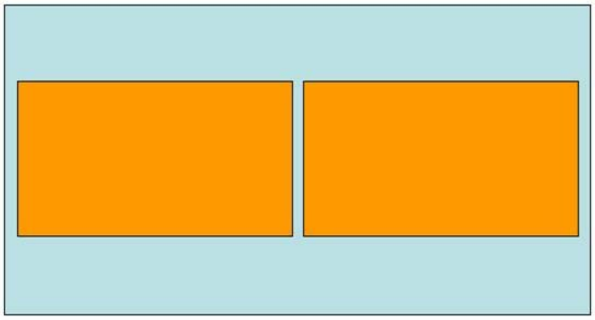


Figure 24: Stereoscopic HD on a 4K Sensor

Very Large TV Screens

When Panasonic introduced its 103-inch plasma TV, at the time the largest consumer flat-panel display, it had the common HDTV definition of 1920 x 1080. As a result, at the Lechner Distance, the image structure could have been perceptible even to viewers with visual acuity somewhat less than 30 cpd.

When the same company later (in 2008) introduced a 150-inch plasma TV, an HDTV image structure would have been even more visible, perceptible even to viewers with impaired vision. The definition of that display, therefore, was 4K (4096 x 2160).

At the top of the next page is an image created by John R. Vollaro for Leon D. Harmon at Bell Labs around 1968, used with permission. Like Figure 12, it doesn’t look like what it is when its edges can be seen.

Like Figure 12, the image will appear to be as it should when viewed from a distance. Unlike the blocks of Figure 12, which were created to help make a color, fixed-grid display present black, round edges, the blocks of Figure 25 were created to obscure a natural photographic image for perceptual study.[49]

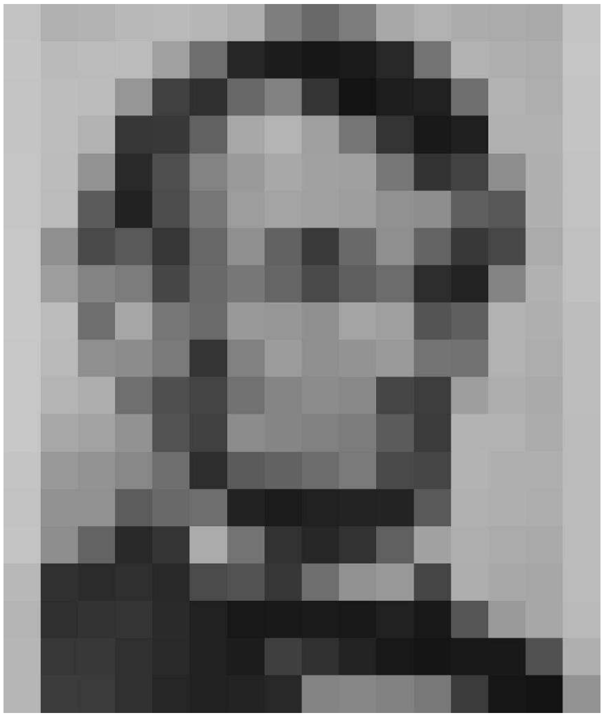


Figure 25: Bell Labs Block Version of Portrait

Surrealist artist Salvador Dali painted a version of the image in 1976, which like Figure 7, changes from one thing to another at a particular retinal angle. The name of the painting is very descriptive: *Gala Contemplating the Mediterranean Sea, Which at Twenty Meters Becomes a Portrait of Abraham Lincoln.*[50]

The effect of Figure 25 occurs because pixels are mathematical points, not little squares, rectangles, or even dots.[51] There are solutions other than increasing display definition, however, such as optical filtering to blur the edges. Holding ground glass or even waxed paper in front of Figure 25, for example, can also reveal its hidden content. Such low-pass optical filtering is commonly used in broadcast television cameras (although it becomes problematic in color-filtered single-sensor cameras: should the filter be appropriate to the luma, the green, or the other colors?).

Aside from their pixel definition and resulting sharpness, very large television displays also stimulate more of the visual field at any given distance. Research into UHDTV (which can be a form of either 4K or 8K) has shown that the increased visual angle not only increases “presence” sensation but also increases dynamic visual acuity (the ability to perceive fine detail moving relative to the eye’s retina).[36] Of course, as shown in Figure 22, a very large 4K screen can also be seen as providing an HDTV image in each quadrant.

<u>Distributing 4K</u>

Very large television displays are, and should continue to be, rare in homes. Again, the average screen size of TV sets shipped to the U.S. through 2000 was less than 25 inches. Screen-size increases continued slowly through 2005, followed by a spurt in 2006 as inexpensive widescreen HDTV sets became available.[52]

Size growth then slowed again. According to Display Search, in 2010, the average TV screen size for shipments to North America (which has the world’s highest average) was 36 inches. In 2012, it is expected to be 37.8 inches. In 2014, it is expected to be 39.2 inches. Globally, TV screen sizes of as little as 50-inch and above accounted for only 5.3% of shipments in 2010 and are expected to account for only 6.3% in 2014.[53]

Figure 26, below, shows actual and estimated average screen sizes for global TV shipments. Not only is the average below 36 inches, but it also appears to be leveling off.[53]

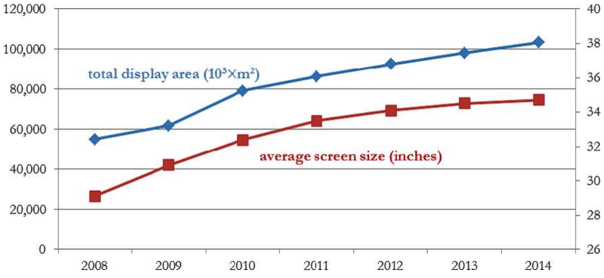


Figure 26: World TV Shipment Average Sizes

Given those screen sizes, it is unlikely that, at the Lechner Distance, average viewers will notice the pixel-grid structure of their television displays. They will also not be close enough to their displays to appreciate the high-definition detail offered by a quarter of a 4K display (or a sixteenth of an 8K display). They *should* be able to appreciate the additional sharpness that 4K image capture offers, but, as Sony's HDCAM filtering indicates, they will see the bulk of that sharpness even on ordinary HDTVs.

It would seem, therefore, that there is little for an average TV viewer to gain from 4K display resolution. The TV-set industry, however, is in a quandary. It was described this way in *The New York Times* in 2011:

"By now, most Americans have taken the leap and tossed out their old boxy televisions in favor of sleek flat-panel displays. Now manufacturers want to convince those people that their once-futuristic sets are already obsolete.

"After a period of strong growth, sales of televisions are slowing. To counter this, TV makers are trying to persuade consumers to buy new sets by promoting new technologies."[54]

That article, which appeared at the time of the 2011 Consumer Electronics Show (CES), indicated that such features as stereoscopic 3D and Internet connections "have not generated much excitement so far." At the 2012 CES, therefore, a new feature being promoted was 4K (and even 8K) definition, with demonstrations from such major manufacturers as AMD, JVC, LG, Panasonic, Sharp, Sony, and Toshiba.[55-58] *Consumer Reports* called 4K "one of the most talked about innovations" at the show.[59]

Aside from promotion by TV-set manufacturers, 4K programming is just starting to become available, primarily in the form of movies, based in part on the production and sharpness advantages of 4K and in part on the use of 4K to differentiate digital cinema from home theater. The late-2011 American version of *The Girl with the Dragon Tattoo,* for example, has been called "the first large-scale end-to-end 4K digital cinema release."[60]

Portions of the 2012 Olympic Games are also expected to be shot and shown in beyond-HDTV resolutions.[61] Thus, cable-television operators might wish to take advantage of all of the promotion by providing a 4K offering. Fortunately, it need not require a large amount of data capacity.

All else being equal, a 4K image sensor has more than four times as many photosites as a 1080-line HDTV image sensor. A 4K display similarly must deal with more than four times as many picture elements.

As might be expected, an uncompressed, 8K (7680 pixels per line, or four times HD's 1920, rather than 8192) "Super Hi-Vision" link would require 16 high-definition serial digital interface (HD-SDI) connections.[62] It is not clear, however, that 4K or 8K require multiples of high-definition data rates in the compressed domain.

There are three main reasons. One may be seen in the previous Figures 14 and 16. As detail gets finer, the energy in the signal is reduced, so there is less to compress.

Another reason relates to motion estimation in bit-rate reduction (BRR) systems that take advantage of temporal redundancy as well as spatial redundancy. The better defined a point is, the more accurately its motion can be estimated and, therefore, the lower the bit rate at which errors will be imperceptible.

Both of those factors suggest that, in an ideal BRR system, the "overhead" to increase from HDTV (or 2K) to 4K will be very much less than an additional three times the original signal value. The third factor is that, absent the very large retinal angle of a cinema screen or very large TV display, viewers are less sensitive to image defects, or, as one BRR-comparison paper put it, "the quality requirements are more stringent when the viewer is in a cinema."[63]

Figure 27, below, is based on a graph in another paper comparing BRR systems for digital cinema. The full graph compares seven BRR systems out to data rates of 260 Mbps for the test sequence called *CrowdRun*. At those high data speeds, the 2K systems all outperform the 4K systems in PSNR.[64]

The small section of the graph shown below, however, is restricted to such low data rates as might be used for delivery of HDTV on a cable-television system. As shown by the five identified data points (all JPEG2000), at those low data rates (starting at 14 Mbps), 4K actually outperformed 2K. It was only beyond roughly 26 Mbps (extrapolated) that 2K outperformed 4K.

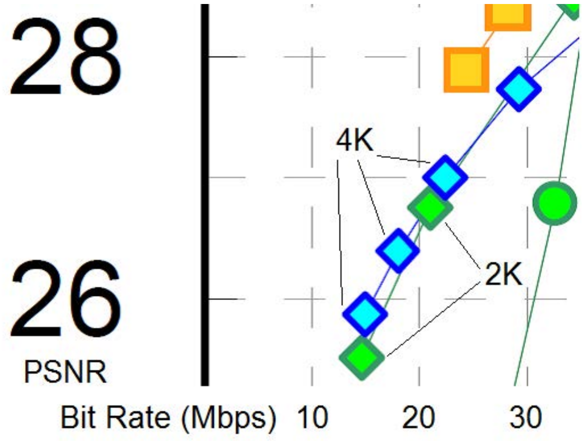


Figure 27: 2K vs. 4K BRR Comparison[64]

BRR quality results can vary according to many factors, but Figure 27 shows that 4K can be transmitted at rates comparable to HDTV with comparable results. It is certainly conceivable that a layered transmission system can also be used, adding only 4K's additional information to that already carried for HDTV. It is not clear, however, what the efficiencies of such layered transmission would be in comparison to the use of a single signal for 4K distribution.

CONCLUSIONS

In program production, 4K is well established and growing. Cameras are available from ARRI, Astro, JVC, Red, Sony, and Vision Research and have also been shown by Canon, Dalsa, Hitachi, Ikegami, Lockheed-Martin, Meduza, NHK, and Olympus.[65]

Though the *4K* designation of some of these cameras can be questioned (largely due to the use of color-filtered single image sensors), all are intended to capture definitions beyond those of HDTV. There is even a technique to extract 4K resolution from masked HDTV image sensors, so as to reduce uncompressed data rates.[66]

In post production, 4K is also well established and growing. *The Girl with the Dragon Tattoo* might be "the first large-scale end-to-end 4K digital cinema release," but all of the individual processes used have been available for some time.

In cinema, 4K is also established and growing. NHK's Super Hi-Vision has been used in cinema-like applications (community viewing of a single, giant screen in a dark room), intended to be viewed at just 0.75 times the picture height.[67]

Super Hi-Vision is also intended to provide a home-viewing experience. Although displays with 4K and even 8K resolutions have been shown, it is not clear at this time either that sufficiently large displays will be purchased by consumers or that consumers will move sufficiently close to smaller displays to give them the “presence” and “realness” intended for Super Hi-Vision.

Figure 28, below, shows a 152-inch plasma TV. Simply getting it into a room in a home is cause for concern. Even a 152-inch size is too small for 0.75-height viewing at the Lechner distance; that would require a 294-inch screen, with a 12-foot-high image (not counting its frame). Clearly, as-intended 8K Super Hi-Vision viewing in the home will require a change in viewing-distance habits.

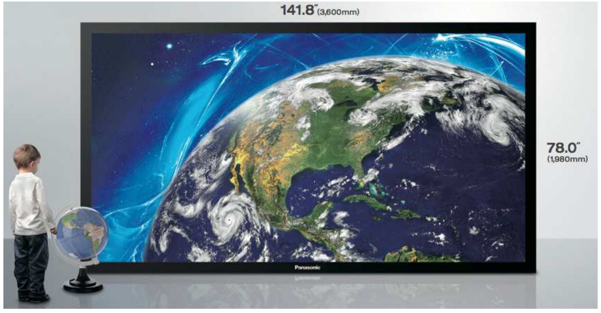


Figure 28: Panasonic 152-inch Plasma TV

The increased sharpness of beyond-HD-resolution imaging is largely available to viewers using existing TV sets. The extraordinary images of 4K and 8K television displays have been reported by observers who could view them at closer than home-viewing distances (e.g., the Lechner Distance).

Cable-television operators can nevertheless take advantage of the promotional aspects of moves to 4K resolution by offering 4K distribution. It is not clear whether *any* increase in bit rate is required, but, due to the low energy of the highest octave of spatial frequencies in a typical, real-world 4K-captured image, improved motion estimation provided by better-defined pixels, and relative insensitivity to compression artifacts at typical TV viewing distances and display sizes, any such increase should be minimal.

REFERENCES

1. *The Compact Edition of the Oxford English Dictionary*, Oxford University Press, 1971

2. Snellen, Dr. H., *Probebuchstaben, zur Bestimmung der Sehscharfe* [Sample Letters, for determining visual acuity], P. W. van de Weijer, 1862 http://archive.org/details/probebuchstaben01snelgoog

3. *A.D.A.M. Medical Encyclopedia*, National Center for Biotechnology Information, U.S. National Library of Medicine, reviewed May 24, 2010 http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002021/

4. Schneider, Joel, “Block Letter Eye Chart,” created May 2002 http://www.i-see.org/block_letter_eye_chart.pdf

5. Bordsen, John, “Eye Chart Still the Standard for Vision,” *The Seattle Times*, August 9, 1995 http://community.seattletimes.nwsource.com/archive/?date=19950809&slug=2135585

6. Fletcher, H., and Munson, W.A., “Loudness, its definition, measurement and calculation,” *Journal of the Acoustic Society of America*, vol. 5, 82-108 (1933) http://www.sfu.ca/media-lab/archive/2011/386/readings/Misc.%20Readings/Loudness,%20Its%20Definition,%20Measurement%20and%20Calculation%20.pdf

7. Oliva, Aude, and Schyns, Philippe G., “Dr. Angry and Mr. Smile: a series,” *Hybrid Images*, Computational Visual Cognition Laboratory, Massachusetts Institute of Technology, 2006 http://cvcl.mit.edu/hybrid_gallery/smile_angry.html

8. Fisher, David E., and Fisher, Marshall Jon, *Tube: the invention of television*, Counterpoint, 1996 http://books.google.com/books?id=eApTAAAAMAAJ

9. Burns, Russell W., *John Logie Baird: Television pioneer*, IET, 2000 http://books.google.com/books?id=5y09hpR0UY0C

10. Abramson, Albert, *The History of Television, 1880-1941*, McFarland, 1987 http://lccn.loc.gov/86043091

11. Prescott, George Bartlett, *Electricity and the Electric Telegraph*, volume 2, D. Appleton, 1892 http://books.google.com/books?id=9_5KAAAAYAAJ

12. Nipkow, Paul, German patent 30101 - 1885-01-15, link to the UK Intellectual Property Office copy: http://bit.ly/H4f5Iu

13. Ives, Herbert E., “Television,” *The Bell System Technical Journal*, volume 6, October 1927 http://www.alcatel-lucent.com/bstj/vol06-1927/articles/bstj6-4-551.pdf

14. “Television Shows Panoramic Scene Carried by Sunlight,” *The New York Times*, July 13, 1928 http://select.nytimes.com/gst/abstract.html?res=F50D13F7395C177A93C1A8178CD85F4C8285F9

15. Schubin, Mark “The First Sports Video,” SchubinCafe.com, July 10, 2009, http://www.schubincafe.com/2009/07/10/the-first-sports-video/

16. *Television* magazine, volume 1, number 2, Experimenter Publishing, New York, July 1928

17. Goebel, Gerhart, “From the history of television - The first fifty years,” *Bosch Technische Berichte*, volume 6 (1979), number 5/6; note: much of the information is available on the web from the Deutsches Fernsehmuseum Wiesbaden (see next reference)

18. “Kapitel 6 (ab 1923)” [Chapter 6 (from 1923)], Deutsches Fernsehmuseum Wiesbaden http://www.fernsehmuseum.info/fernsehgeschichte06.html

19. *Report of the Television Committee*, His Majesty’s Stationery Office, 1935 http://www.thevalvepage.com/tvyears/articals/comrep/comrep.htm

20. Preston, S. J., “The Birth of a High Definition Television System,” *Journal of the Television Society*, volume 7, number 3, 1953

21. *Broadcasting* magazine, May 1, 1939

22. National Television System Committee, *Proceedings*, 1940-1941 http://lccn.loc.gov/45051235

23. Pemberton, Alan, “Line Standards,” *World Analogue Television Standards and Waveforms*, Sheffield, England, 2010 http://www.pembers.freeserve.co.uk/World-TV-Standards/Line-Standards.html

24. Pemberton, Alan, “Timeline” from “Overview,” *World Analogue Television Standards and Waveforms*, Sheffield, England, 2010 http://www.pembers.freeserve.co.uk/World-TV-Standards/index.html#Timeline

25. Schubin, Mark, “Special Report: TV Around the World,” *Videography*, March 1979

26. Hart, Samuel Lavington, “Improvements in Apparatus for Transmitting Pictures of Moving Objects and the like to a distance Electrically,” British patent 15,270, published 25th June, 1915, link to the UK Intellectual Property Office copy: http://bit.ly/H7Gqvx

27. Yanczer, Peter, “Ulises Armand Sanabria,” in “Mechanical Television,” Early Television Museum, http://www.earlytelevision.org/u_a_sanabria.html

28. Dinsdale, A., “Television in America To-day,” *Journal of the Television Society*, volume 1, 1932 http://books.google.com/books?id=O2cPAAAAIAAJ

29. Watkinson, John, *The Art of Digital Video*, Focal Press, 2008 http://books.google.com/books?id=8uLEXlN9ouAC

30. Hsu, Stephen C., “The Kell Factor: Past and Present,” *Journal of the Society of Motion-Picture and Television Engineers*, volume 95, no. 2, February 1986 http://journal.smpte.org/content/95/2/206.abstract

31. Yerrick, Damian, “Demonstration of interlace and so-called ‘interline-twitter,’ based on part of an RCA Indian Head Test Card, ca. 1940,” http://en.wikipedia.org/wiki/File:Indian_Head_interlace.gif

32. *Television Digest*, volume 11, number 36, 1955

33. Taylor, Jim, Johnson, Mark R., and Crawford, Charles G., *DVD Demystified, Third Edition*, McGraw-Hill, 2006 http://books.google.com/books?id=ikxuL2aX9cAC

34. Poynton, Charles, *Digital Video and HDTV: Algorithms and Interfaces*, Morgan Kaufman, 2003 http://books.google.com/books?id=ra1lcAwgvq4C

35. Wargo, Sean, Consumer Electronics Association press presentation, New York, November 2006

36. Sugawara, Masayuki, et al., “Research on Human Factors in Ultrahigh-Definition Television (UHDTV) to Determine Its Specifications,” *SMPTE Motion Imaging Journal*, vol. 117, no. 3, April 2008 http://www2.tech.purdue.edu/Cgt/courses/cgt512/discussion/Chastain_Human%20Factors%20in%20UDTV.pdf

37. Schade, Otto H., *Image Quality : a comparison of photographic and television systems*, RCA Laboratories, 1975, republished in the *SMPTE Journal*, volume 96, number 6, June 1987, http://journal.smpte.org/content/96/6/567, described more recently here, http://www.panavision.com/sites/default/files/24P%20Technical%20Seminar%202.pdf

38. Heynacher, Erich, “Ein Bildgütemaß auf der Grundlage der Übertragungstheorie mit subjektiver Bewertungsskale” [Objective Image Quality Criteria, based on transformation theory with a subjective scale], *Zeiss Mitteilungen*, volume 3, number 1, 1963, described in the *ARRI 4K+ Systems* brochure http://www.scribd.com/doc/52408729/4K-Systems-Arri

39. Thorpe, Laurence J., Nagumo, Fumio, and Ike, Kazuo, “The HDTV Camcorder and the March to Marketplace Reality,” *SMPTE Journal*, volume 107, number 3, March 1998, http://www.smpte-pda.org/resources/The+HDTV+CamorderThorpeMar1998.pdf

40. Atkins, Bob, “Canon EOS 20D DSLR Review,” *Bob Atkins Photography* http://www.bobatkins.com/photography/digital/eos20d.html

41. Vogel, Harold L., *Entertainment Industry Economics: A guide for financial analysis*,, Cambridge University Press, 1986 (with data from *Reel Facts*) http://books.google.com/books?id=3TwrQgAACAAJ

42. “Yearly Box Office,” *Box Office Mojo*, http://boxofficemojo.com/yearly/

43. “Population Estimates,” Historical Data, United States Census Bureau, http://www.census.gov/popest/data/historical/index.html

44. “Help Documents,” *The Big Screen Cinema Guide*, http://www.bigscreen.com/about/help.php?id=36

45. Burnett, Colin M. L., “A bayer pattern on a sensor in isometric perspective/projection,” 28

December 2006, http://en.wikipedia.org/wiki/File:Bayer_pattern_on_sensor.svg

46. Schubin, Mark, “Fun Out of the Sun in Las Vegas - 2011: a different kind of NAB show,” 2011 May 19, http://www.schubincafe.com/2011/06/01/nab-2011-wrapup-washington-dc-smpte-section-may-19-2011/

47. Schubin, Mark, “Things You Can or Can’t Fix in Post: Video Acquisition,” San Francisco Public Television Quality Group, 2010 June 8, http://www.schubincafe.com/2010/06/15/things-you-can-or-can%E2%80%99t-fix-in-post-video-acquisition/

48. “Phantom 65-Z3D System,” Abel Cine, http://about.abelcine.com/wp-content/imported/images/pdf/phantom_65-z3d.pdf

49. Vollaro, John R. “Commentary on the History of ‘Photomosaic’ Images,” March 2006, http://vollaro.com/WebScrapbook/docs/Clinton/NudeStory.html

50. “Gala Contemplating the Mediterranean Sea which at Twenty Meters becomes a Portrait of Abraham Lincoln,” Authentic Society, http://www.authenticsociety.com/about/GalaMediterraneanLincoln_Dali

51. Smith, Alvy Ray, “A Pixel Is *Not* A Little Square, A Pixel Is *Not* A Little Square, A Pixel Is *Not* A Little Square!,” Microsoft Computer Graphics, Technical Memo 6, July 17, 1995, http://alvyray.com/Memos/CG/Microsoft/6_pixel.pdf

52. Wargo, Sean, Consumer Electronics Association, e-mail communication to the author, January 31, 2007

53. Park, Won Young, et al., *TV Energy Consumption Trends and Energy Efficiency Improvement Options*, Environmental Energy Technologies Division, International Energy Studies Group, Ernest Orlando Lawrence Berkeley National Laboratory, July 1, 2011 http://www.superefficient.org/~/media/Files/SEAD%20Televisions%20Technical%20Analysis.pdf

54. Grobart, Sam, “A Bonanza in TV Sales Fades Away,” *The New York Times*, January 5, 2011 http://www.nytimes.com/2011/01/06/technology/06sets.html

55. Pogue, David, “Sampling the Future of Gadgetry,” *The New York Times*, January 11, 2012 http://www.nytimes.com/2012/01/12/technology/personaltech/in-las-vegas-its-the-future-of-high-tech-state-of-the-art.html

56. Putman, Peter, “HDTV Expert - CES 2012: Another Opening, Another Show,” *HDTV Magazine*, January 18, 2012, http://www.hdtvmagazine.com/columns/2012/01/hdtv-expert-ces-2012-another-opening-another-show.php

57. Walton, Jerry, “The Best of CES 2012,” *AnandTech*, January 17, 2012, http://www.anandtech.com/show/5437/the-best-of-ces-2012

58. Healey, Jon, “CES 2012: 4K TV sets make their debut, minus the hoopla,” *The Los Angeles Times*, January 11, 2012, http://latimesblogs.latimes.com/technology/2012/01/ces-4k-tv-sets-make-their-debut-minus-the-hoopla.html

59. “CES 2012 Video: Could 4k TV technology bring better 3D TV?” Consumer News, *ConsumerReports.org*, http://news.consumerreports.org/electronics/2012/01/ces-2012-video-what-is-sonys-4k-tv-technology.html

60. Koo, Ryan, “Fincher Reframes in Post! The 4K Release of ‘The Girl with the Dragon Tattoo,” *NoFilmSchool*, December 28, 2011, http://nofilmschool.com/2011/12/fincher-reframes-post-4k-release-the/

61. Carter, Jamie, “BBC Talks Super Hi-Vision Plans for London 2012,” TechRadar.TVs, *TechRadar*, http://www.techradar.com/news/television/bbc-talks-super-hi-vision-plans-for-london-2012-1068914

62. “World’s First Live Relay Experiment of Super Hi-Vision,” *Broadcast Technology*, number 25, Winter 2006, NHK STRL, http://www.nhk.or.jp/strl/publica/bt/en/to0025.pdf

63. Shi, Boxin, Liu, Lin, and Xu, Chao, “Comparison between JPEG2000 and H.264 for Digital Cinema,” *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2008 http://www.cvl.iis.u-tokyo.ac.jp/~shi/files/Shi_ICME08.pdf

64. Baruffa, Giuseppe, Micanti, Paolo, Frescura, Fabrizio, “Performance Assessment of JPEG2000 Based MCTF and H.264 FRExt for Digital Cinema Compression,” *Proceedings of the 16th International Conference on Digital Signal Processing*, July 2009 http://dsplab.diei.unipg.it/files/baruffa_DSP2009.pdf

65. Schubin, Mark, “Beyond-HD-Resolution Cameras and their Workflows,” 18th HPA Tech Retreat, Hollywood Post Alliance, Indian Wells, California, 2012 February 15, http://www.schubincafe.com/2012/03/11/4k-hpa-tech-retreat-2012/

66. Schöberl, Michael, et al., “Increasing Image Resolution by Covering Your Sensor,” 18th HPA Tech Retreat, Hollywood Post Alliance, Indian Wells, California, 2012 February 17, http://data.memberclicks.com/site/hopa/2012_TR_Pres_SFoessel.pdf

67. “8K Television System ‘Super Hi-vision’ is the TV technology of our dreams,” NHK, http://www.nhk.or.jp/digital/en/superhivision/

WI-FI NETWORKS IN HFC CABLE NETWORKS - HOW TO UNDERSTAND AND MEASURE USER QUALITY PROBLEMS IN A WIRELESS ENVIRONMENT

HUGO RAMOS, NET Servicos SA

An initial effort to try to bring the quality of service of our carrier grade high speed data node based service from cable network, to a cell wireless network, presenting the concept of WUEQi (Wireless User Experience Quality index).

ABSTRACT

In this paper, is presented the concept of WUEQi, an index like a MOS index, to give ability for cable operators that are deploying Wi-Fi networks to have better control of this technology to deliver services. Looking basically to give to the users a better user experience and to the operator a good way to plan the expansion of the Wi-Fi network.

Keywords

Wi-Fi, Wireless networks, CATV, Strand mount AP, average power, EIRP, antennas, user experience, DOCSIS.

1. INTRODUCTION

Today with the dissemination of a new generation of very powerful mobile devices such as smartphones and tablets the way of our society communicate and consume information is changing, creating an exponential rise in the importance of mobility services, with the anything, anytime, anywhere concept. With this in mind a lot of cable MSO are dealing with a new and different ways to deliver mobile services to their customers with the deployment of wireless networks over the cable HFC plant.

With the proliferation of mobile devices the bandwidth requirements in the traditional mobile networks has also change pushing these mobile operators to find new ways of deliver these services with adoption of new technologies and that search can be a very good opportunity for the cable industry as well, since the deployment of this networks over the cable HFC plant can be very fast and efficient.

One technology that is getting a very big attention of the market right now is Wi-Fi. The Wi-Fi standard today is undoubtedly the most heavily used wireless technology in terms of number of devices that is capable of use the technology and as the amount of data traffic transmitted over networks using this wireless technology. The technology is very mature and in any report that you have access to, it is set that around the world, millions and millions of devices such as smartphones, tablets, netbooks, notebooks, TVs, STBs, home gateways, even cars and refrigerators has embedded and certified chipsets with this technology and this number is growing, actually skyrocketing each day. That is why deploy a Wi-Fi network is a very attractive option to any convergent operators to give fixed broadband customers the mobility that they look forward.

But unlike from the common sense thinking that says the Wi-Fi technology is a very mature technology, this standard is not mature yet to deploy a service provider type of network. This is very easily pointed by the fact that to gain the scale of number of devices, this standard utilizes non license frequencies and since it is inherent to a wireless network (different to a cable HFC that is confined cable network) this type of solution deals with devices that are moving, unstable customer demands and different device power transmission that changes the way of get to know, understand and solve the problems that affect the most important thing at the end that is the user experience.

Since motivators of this deployments, in cable could be retention and churn reduction, 3G offload to traditional mobile operators and location based services, user quality experience is became very important to the SP to provide the service and the use of analytics, throughout the WUEQi, could be the answer to try to have a little control to the chaos of delivering Wi-Fi mobile services to a customer that is used to have the quality of service of our carrier grade high speed data node based service from cable network.

This paper is organized as follows. In the next section we present the advantages of deploy a Wi-Fi network over cable HFC plant. In section 3.Important concepts about Wi-Fi networks deployment (Link budget calculation). Section 4. Show the challenges to deliver good and control user experience to the chaos of a non-license frequency wireless network, the problems about different link budgets with different devices and the concept of WUEQi.

2. ADVANTAGES OF WI-FI NETWORK IN A CABLE ENVIRONMENT

There are three big obstacles that service providers which intend to deploy a Wi-Fi network will probably face, named: reliable powering the access point units, make the backhaul link with Internet traffic to the APs and acquiring mounting sites. These three obstacles can make a huge difference in deploying big networks more efficient. HFC cable networks are already design and build with reliable power supply that can feed the Wi-Fi access point, of course if the access point is capable of handle the power from HFC plant and there is some units at the market that can not only handle but also being mounted in strand, solving the second issue. Also in a cable network the IP connectivity can be delivered in the same cable that is connected to feed the AP throughout the high-speed data DOCSIS network that is already in place. With this issues solved in HFC networks, cable operators will can quickly and easily deploy this wireless networks in the existing plant.

So as shown above, the implementation of a Wi-Fi network in an HFC plant is faster, easier and more efficient but something that we are not used to work and it is a big problem to control is the wireless access part mainly in deal with average transmitted power (TX power from a device) that is going to be translate to coverage and differ from the DOCSIS very much since the cable modem is easily controlled and does not move around. This problem is going to reflect in a bad user experience

3. IMPORTANT CONCEPTS ABOUT WI-FI NETWORKS DEPLOYMENT

3.1 Link Budget

A lot of questions are important when designing a Wi-Fi network such as:

1. What type of devices I want to provide the service?
2. What type of services I want to deliver?
3. What type of environment I am going to deploy the network?

So, with this in mind you can find out how to calculate the link budgets to provide services to the devices that you chose taking in consideration the environment and the services that you want to provide.

In our case we are going to consider 3 devices to provide the services, a typical notebook, a typical table and a typical Smartphone. Of course in a real scenario we are not going to see typical devices but real information that can be get from devices providers, FCC test and in our case ANATEL(Brazil Telecom Agency). Our environment will be a very dense and populated metropolis with a lot of buildings and reflections.

To determine the link budgets we should consider the usage of some models. For a free space, line of sight environment, the propagation law would result in square law decay. i.e. for each doubling of distance the power drops by 6dB. The free space model is only an idealized model and the real world is not line of sight. There are objects in the environment such as trees, buildings, poles that signals reflect from and with this, the signal can add and subtract to produce a signal that decays faster than the square law.

Below we show a typical configuration of some device that we are going to consider calculating the link budgets.

	AP	**Typical Notebook**	**Typical Tablet**	**Typical Smartphone**
Average Power (A) dBm	21	16	17.5	16.0
Antenna Gain (B) dBi	5	6	2.0	0.0
EIRP (A+B) dBm	26	22	19.5	16.0

Table 1. Typical power of devices

Using some models we can show below the link budget in an urban and very dense city environment.

Link Budget between an AP and a typical notebook:

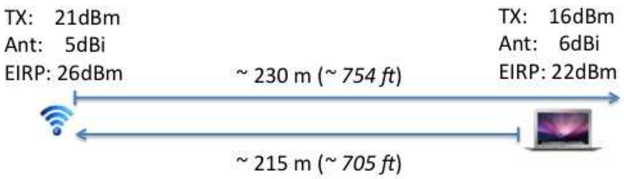


figure 1. Link budget – AP and notebook

So in this case the link budget is limited in 215 meters from the notebook to the AP.

Link Budget between an AP and a typical tablet:

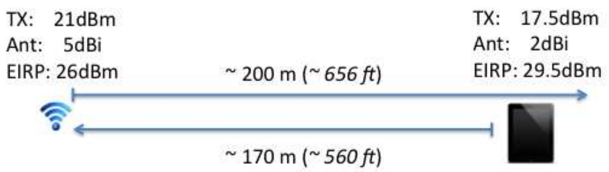


figure 2. Link budget – AP and tablet

So in this case the link budget is limited in 170 meters from the tablet to the AP.

Link Budget between an AP and a typical smartphone:

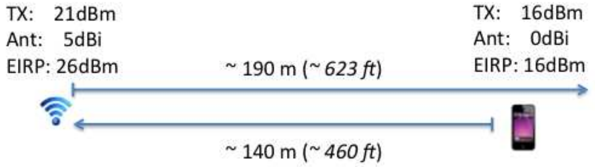


figure 3. Link budget – AP and smartphone

So in this case the link budget is limited in 140 meters from the smartphone to the AP.

This will show to us that the limit of the link budget will be defined by the transmission from device to the AP in upstream direction rather than in downstream direction since the average power transmitted in a downstream direction is bigger than the upstream direction.

3.2 Relation between type of services and SNR

Since the Wi-Fi is a system that uses adaptive modulation, the modulation nominal rate is dependent of the SNR of the link budget that is the difference from the noise floor (typically -90dBm in our city) to the signal that are receive by the device and the access point. Of course this SNR will make our user experience better or worse so we have to consider this in our index.

Below we show a table of typical SNR versus the type of service that is possible to use.

Service	SNR
Video 99% sucess	>= 35 dB
Voice 99% success	>= 25 dB
Web surfing 99% success Voice 90% sucess	>= 20 dB
Email – Web surfing 90% success	>= 7 dB

Table 2. Tested Services vs. SNR

	Rate (Mbps)	SNR (dB)	Signal Level (dBm)
802.11b (DSSS)	1	4	-81
	2	6	-79
	5,5	8	-77
	11	10	-75
802.11g Data Rate (OFDM)	6	4	-81
	9	5	-80
	12	7	-78
	18	9	-76
	24	12	-73
	36	16	-69
	48	20	-65
	54	21	-64

Table 3. Theoretical table data rate vs. SNR

4. THE CHALLENGES TO DELIVER GOOD USER EXPERIENCE

Since the limit of the link budget will be defined mainly by the transmission from the device to the AP some situations can occurs like a smartphone see the network and could not associate (getting a fail association status a the system) or have a small SNR, having a bad user experience, in a place where a notebook can associate.

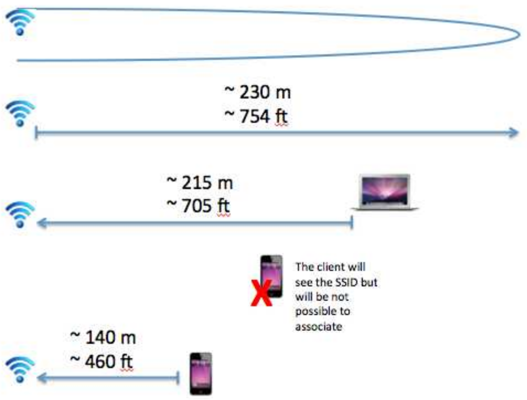
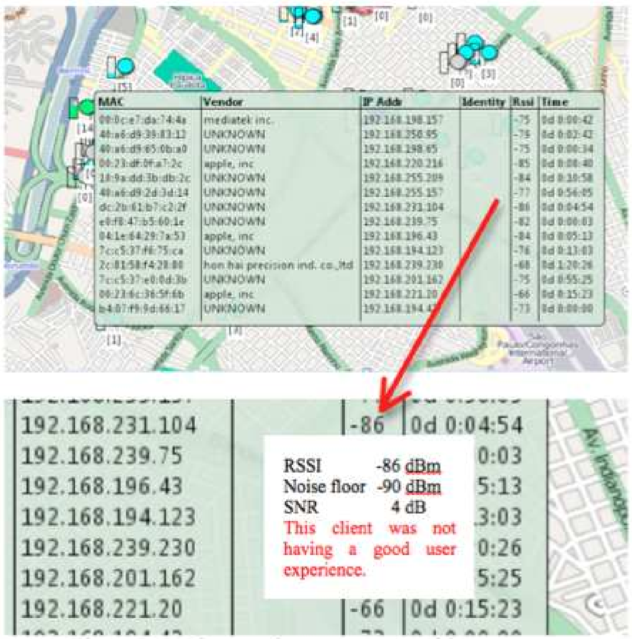

~ 230 m
~ 754 ft
~ 215 m
~ 705 ft
The client will see the SSID but will be not possible to associate
~ 140 m
~ 460 ft

figure 4. Association problem



MAC	Vendor	IP Addr	Identity	Rssi	Time
00:0c:e7:da:74:4a	mediatek inc.	192.168.198.157		-75	0d 0:00:42
40:a6:d9:39:83:12	UNKNOWN	192.168.250.95		-75	0d 0:02:42
40:a6:d9:65:0b:a0	UNKNOWN	192.168.198.65		-75	0d 0:00:34
00:23:df:0f:a7:2c	apple, inc	192.168.220.216		-85	0d 0:08:40
18:9a:dd:3b:db:2c	UNKNOWN	192.168.255.209		-84	0d 0:10:58
40:a6:d9:2d:3d:14	UNKNOWN	192.168.255.157		-77	0d 0:56:05
dc:2b:61:b7:c2:2f	UNKNOWN	192.168.231.104		-86	0d 0:04:54
e0:f8:47:b5:60:1e	UNKNOWN	192.168.239.75		-82	0d 0:00:03
04:1e:64:29:7a:53	apple, inc	192.168.196.43		-84	0d 0:05:13
7c:c5:37:f6:75:ca	UNKNOWN	192.168.194.123		-76	0d 0:13:03
2c:81:58:f4:28:80	hon hai precision ind. co.,ltd	192.168.239.230		-68	0d 1:20:26
7c:c5:37:e0:0d:3b	UNKNOWN	192.168.201.162		-75	0d 0:55:25
00:23:6c:36:5f:6b	apple, inc	192.168.221.20		-66	0d 0:15:23
b4:07:f9:9d:66:17	UNKNOWN	192.168.194.43		-73	0d 0:00:00

RSSI -86 dBm
Noise floor -90 dBm
SNR 4 dB
This client was not having a good user experience.

figure 5 – Real network management with customers

The figure 5 above show a particularly customer that was not having a good user experience, what we notice was that this customer could associate, was able to get an IP address (192.168.231.104) but he was not having a good throughput. In this example, we noticed that in most cases the algorithm of the device shown at the screen two of three bar of signal giving the erroneous impression to the user that it has sufficient signal strength to associate or have a good user experience but unfortunately this is not true.

In other cases we also notice another customers with less SNR was not be able to get an IP address because the throughput was too low and another that could not accomplish to associate and generate a trap registering that he had a fail association.

Because of this less control of the CPE and the type of the business models delivered where the product is offered free of charge to the subscriber as an extension of their fixed broadband service, there is a natural tendency to think that if the client can not connect to the Internet in a place where the Wi-Fi network is broadcast, there is no problem because the customer will get another form of connection to solve their connection issue. This will cause a bad user experience to the client and will surely threatened one of the most important asset that we have that is our brand and also cause a money issue in the project since one of the reason of deploying this network is to retain the customer.

Clearly we need to seek technical "thermometers" so we can monitor the user experience in the wireless network service, not only depending from customer surveys to measure the quality of the service, and try to deliver a wireless services at the same level of quality that our industry is providing carrier-grade services to our customers. That is the difference of being called a carrier grade Wi-Fi service provide

4.1 Wireless User Experience Quality index

Since this problems of low SNR and fail association could occurs, we need to find a way of get, register and understand this data to solve problems like user experience, find places where the signal coverage is not good enough, control the network and know if the network is getting better or worse in time and also with this understanding make the upgrade and make the coverage bigger with a more cost effective way because you with those information can understand better where your customers need coverage. That why we present the concept of WUEQi, that is a index that is a scale of 1 (bad) to 5 (excellent) user quality of experience, similar to a MOS type index.

Variable #1 – SNR of the connection – S
The SNR variable has a weight of 2 in the total.

Margins	S
< 7 dB	1.0
7 dB < S < 15 dB	4.5
15 dB < S < 25 dB	4.9
> 25 dB	5.0

Table 4. “S” variable

Variable #2 – Fail Association - FA
The Fail Association variable has a weight of 3 in the total.

Margins	FA
< 3 fail association in a AP per hour	5.0
3 < Fail Association per hour < 10	2.5
10 < Fail Association per hour < 20	1.0
> 20 Fail Association per hour	0.5

Table 5. “FA” variable

Variable #3 – Get an IP address - I
The fail getting IP Address per hour variable has a weight of 3 in the total.

Margins	I
< 3 fail getting IP address in a AP per hour	5.0
3 < fail getting IP address in a AP per hour < 10	2.5
10 < fail getting IP address in a AP per hour < 20	1.0
> 20 fail getting IP address in a AP per hour	0.5

Table 6. “I” variable

Variable #4 – Medium throughput of the connection - M
The medium throughput of the connection has a weight of 2 in the total.

Margins	M
M < 6 Mpbs	1.0
6 Mbps < M < 9 Mbps	2.5
9 Mbps < M < 12 Mbps	4.75
12 Mbps < M < 18 Mbps	4.8
18 Mbps < M < 25 Mbps	4.85
25 Mbps < M < 50 Mbps	4.9
50 Mbps < M < 75 Mbps	4.95
M > 75 Mbps	5.0

Table 7. “M” variable

Using the formulas show below we can find the WUEQi.

$$MT = \frac{\sum M}{\sum users(1h)}$$

$$ST = \frac{\sum S}{\sum users(1h)}$$

$$WUEWQi = \frac{2xMT + 2xST + 3xI + 3xFA}{10}$$

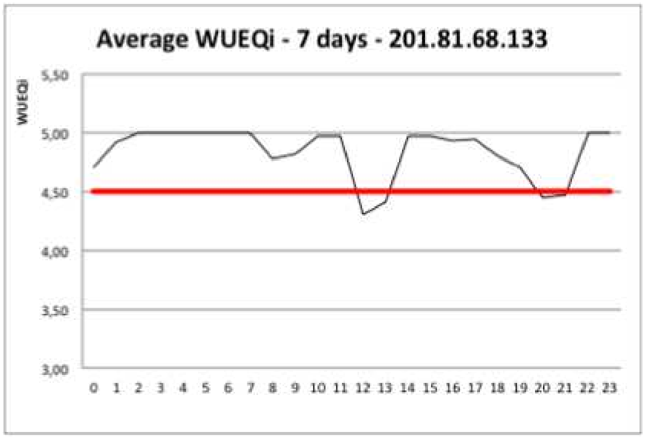
We can calculate from the WUEQi of an AP to the WUEQi of the network using a simple mean formula:

$$WUEWQi(network) = \frac{\sum WUEQi(allAP)}{\# AP}$$

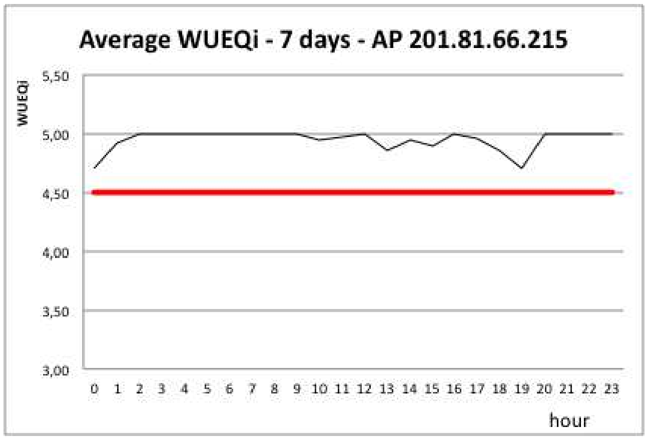
4.2 Using the index

With this index we can identify which access points are facing problems with bad user experience and make site surveys to implement more coverage in the places that the customer are demanding. With this data you can make better planning for the growth of the network, more efficient and cost effective. This index can also keep the track of changes of the network keeping history of the network grade.

The case showed below is an AP that was facing problems with a bad user experience and has to have a survey and make the coverage better.



The case showed below is an AP that has a good comportment



5. CONCLUSION

This work does not intend to be the final word in measuring the quality of service of a Wi-Fi network but a starting point of a more controlled network and a point of discussion. As a service provider, we must have the concern with the quality of any service that our companies are delivering. When we were deploying this network we faced with a lot of our users (testers) saying that in some places they could look the network and connect well, but in other places they found the network associate and could not use the Internet or found and could not associate, so with this concepts we could not only measure the network quality but more important get to know where to focus our effort to make expansion to grow our coverage and serve our customer better.

REFERENCES

[1] Kemisola Ogunjemilua, John N. Davies, Vic Grout and Rich Picking, “An Investigation into Signal Strength of 802.11n WLAN” Centre for Applied Internet Research (CAIR) Glyndŵr University, University of Wales, Wrexham, UK.

[2] Friis H.T.,(1946) Proc. IRE, vol. 34, p.254. 1946

[3] Garg, V.K., (2007). Wireless Communications and Networking, Morgan Kaufmann PublishersGast M., (2002), 802.11

Wireless Networks: The Definitive Guide, O’Reilly Media Incorporated.

[4] BELAIR Presentation - BelAir SNMP Overview – January 2011

[5] BELAIR Link Budget Presentation - 2011

[6] CISCO - Cisco ClientLink: Optimized Device Performance with 802.11n

[7] CISCO Datasheet - Cisco Aironet 1550 Series Outdoor Access Point

ISBN 0-940272-01-6; 0-940272-08-3; 0-940272-10-5; 0-940272-11-3; 0-940272-12-1; 0-940272-14-8; 0-940272-15-6; 0-940272-16-4; 0-940272-18-0; 0-940272-19-9; 0-940272-20-2; 0-940272-21-0; 0-940272-22-9; 0-940272-23-7; 0-940272-24-5; 0-940272-25-3; 0-940272-26-1; 0-940272-27-X; 0-940272-28-8; 0-940272-29-6; 0-940272-32-6; 0-940272-33-4; 0-940272-34-2; 0-940272-35-0; 0-940272-36-9; 0-940272-37-7; 0-940272-38-5; 0-940272-39-3; 0-940272-40-7; 0-940272-41-5; 0-940272-42-3; 0-940272-43-1; 0-940272-44-X; 0-940272-45-8; 0-940272-46-6; 0-940272-47-4; 0-940272-48-2; 0-940272-49-0; 0-940272-50-4; 0-940272-51-2; 0-940272-52-0; 0-940272-53-9; 0-940272-54-7