THE COMPLETE
TECHNICAL PAPER PROCEEDINGS
FROM:

A COMPARISON OF PON ARCHITECTURES

James O. "Jim" Farmer
Wave7 Optics, Inc.

Abstract

Several Passive Optical Network (PON) standards have been proposed as new architectures for delivering video, voice, and data to homes. PONs are being built in large numbers in Asia, and in increasing numbers in the Americas and Europe. Several cable operators are starting to deploy PONs in selected greenfield applications, typically in situations where required by the developer.

This paper shows the most popular forms of PONs in use today. We compare the performance of the PONs, and talk about how and when one may want to consider PON architectures.

WHAT IS A PON?

PONs, or *passive optical networks*, are just that: fiber optics all the way to the home, with only passive (non power-consuming) devices in the field. With no powered devices in the field, you save on power costs, and maintenance is much lower than with hybrid fiber-coax (HFC). Since the network is all glass (usually called "all dielectric"), you eliminate problems such as sheath current. Lightning issues are generally limited to anything that comes into the home over the power line and, through subscriber equipment, jumps to your equipment.

Figure 1 illustrates the basic PON. A single fiber optic strand extends from the headend to an optical splitter located near a group of homes. Outputs of the splitter supply optical signals to a group of homes. Signals are terminated on each home in a device called an *Optical Network Terminal* (ONT). In many cases the ONT is located on the outside of the home at the utility entrance. Alternate locations include inside the home and in a purpose-built niche in the outside wall.

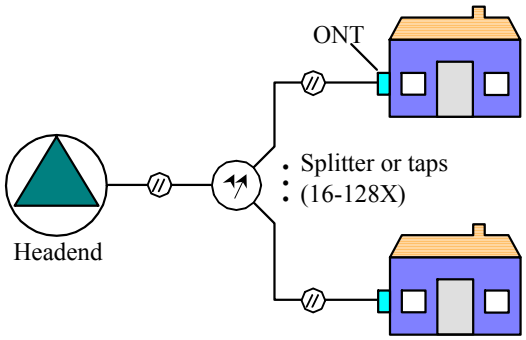


Figure 1. Basic PON

Frequently the splitting is done in a central location as shown. In other cases the splitting may be replaced by a tapped architecture more like that used in HFC architectures. The number of homes served by one PON is limited by the loss budget. While PONs are built with more or fewer subscribers, 32 subscribers is considered the "sweet spot" in PON sizing today. We show up to 128-way splitting, but the optics available today don't support this high a split ratio.

Done correctly, the advantages of PONs include much lower operational expenses, higher quality, elimination of leakage and the resultant measurement requirements, and incredible bandwidth. Data bandwidth of at least 1 Gb/s in each direction, shared over just 32 subscribers is the norm today. This bandwidth is delivered over separate wavelengths from that used for broadcast video, so the entire 54-1,000 MHz RF band is available for video.

TYPES OF PONS

We shall describe several types of PONs in this paper, including BPON (*Broadband Passive Optical Network*, approaching end-of-life), GPON (Gigabit Passive Optical Network), and GE-PON (*Gigabit Ethernet Passive Optical Network*). We shall mention a variant used in some places, called an *active optical network*. We'll also describe an emerging adaptation of an HFC network to extend fiber deeper. It is called RFoG (*Radio Frequency over Glass*), and is an option to consider when a developer requires fiber-to-the-home (FTTH).

GPON and GE-PON systems (and BPON) share a common physical layer architecture, with some differences in optical levels and speeds, so we will cover them together while discussing the physical layer. We'll compare them with the likely RFoG architecture. We say "likely" architecture because work on the RFoG standard has just started this year, and while there are some pre-standard systems entering the market, the standard system has not been defined. Thus, what is described herein is the author's conjecture of what the system may be.

PHYSICAL LEVEL ARCHITECTURE

Figure 2 illustrates the physical layer architecture of PONs. Figure 2a illustrates the BPON/GPON/GE-PON architecture, and Figure 2b illustrates a possible RFoG architecture. In each case, the headend comprises what headends usually comprise in the way of video, voice, and data equipment, except that in the standard PONs of Figure 2a, there is no CMTS – this will be explained later. Downstream RF signals are supplied to a downstream optical

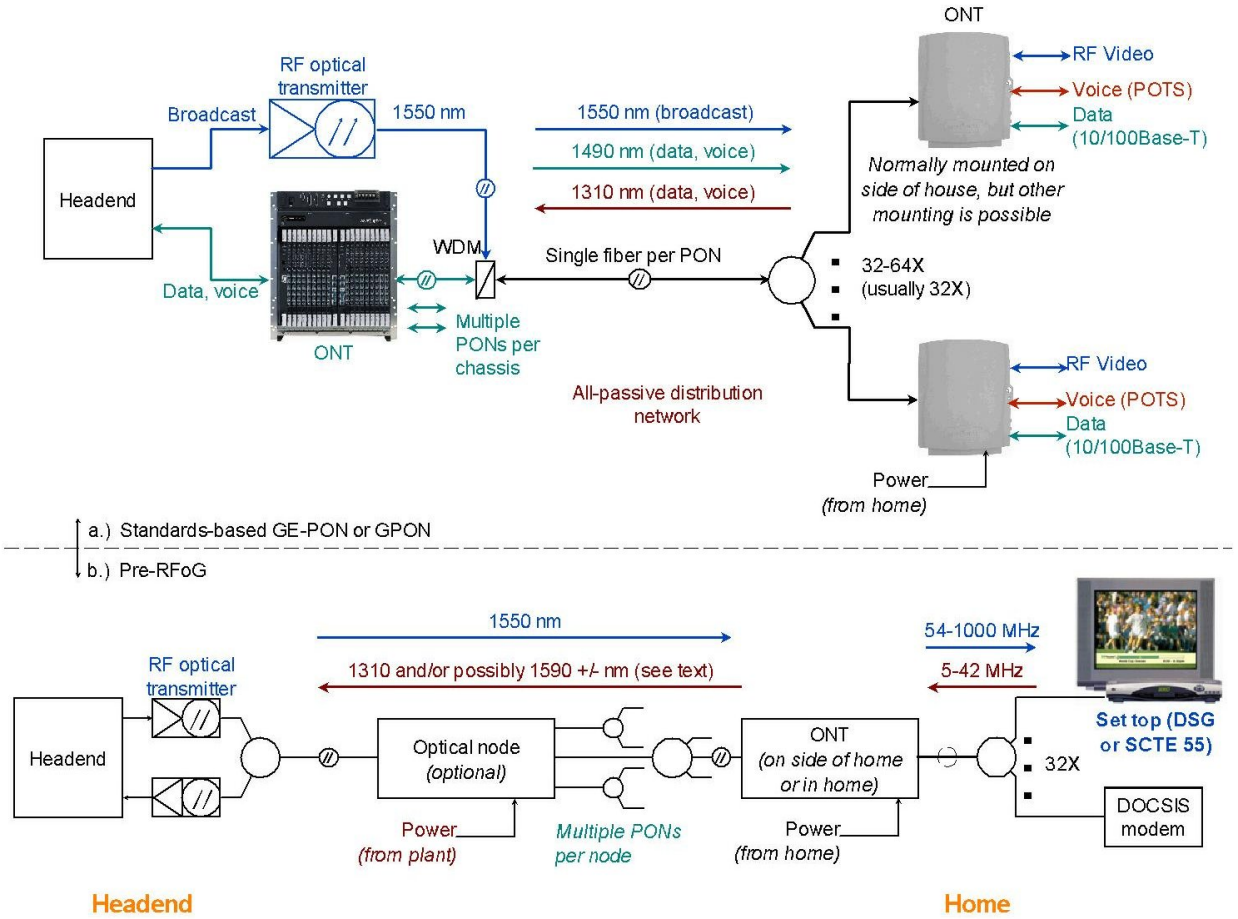


Figure 2. Physical Layer Architectures

transmitter, usually an externally-modulated transmitter and always at 1550 nm because amplification of the optical signal is needed. While you can amplify other wavelengths, amplification at 1550 nm is the most mature and economical process today. All of the standards use 1550 nm for downstream RF broadcast.

GPON/GE-PON

Unlike HFC networks, the network interface for data is not a CMTS – none is used – but rather is an analogous device called an *Optical Line Terminal*, or OLT. It serves the same function as the CMTS in that it converts data (usually delivered as gigabit Ethernet) into the format needed for PON transmission. That conversion includes conversion to the particular PON protocol being used, and conversion to light. The downstream signals are carried at 1490 nm and the upstream at 1310 nm. These are combined with the 1550 nm broadcast signal in a *wave division multiplexer*, or WDM. The WDM operates analogously to a diplex filter in the HFC world.

Typically, the OLT includes many PONs in one chassis, density being very important. There are some cases in which you may need a less-dense solution for outlying pockets of subscribers, and some manufacturers have accommodated this. While we show only one PON, typically many PONs feed into an area and all splitters may be located at a common point called a local convergence cabinet. We can show that this architecture, particularly in greenfields, results in a very economical deployment of equipment.

After splitting, individual fibers supply optical signals to the ONTs at individual homes. An ONT may have one RF output that looks just like the downstream signals from an HFC network, and it may have one or more data connections, usually 10/100Base-T and sometimes 1000Base-T. Also, several analog telephone lines (POTS – plain old telephone service) will be supplied. Other options are shown below.

RFoG

A possible RFoG system is shown in Figure 2b. The headend is identical to that of an HFC system, because RFoG is really an HFC node serving one subscriber. The downstream is again a 1550 nm transmitter, because you will need to amplify the optical signal. The upstream receiver is similar to that used in upstream paths today. The upstream may be analog or it may be digital; this has not been decided in the standardization effort as of this writing.

An optical node in the field is shown as optional. Of course, if used, the network is no longer completely passive. If used, the optical node will likely contain optical amplification in the downstream direction, and combining (in the optical and/or RF domains) in the upstream. Some proposals convert the upstream to digital.

Again, for RFoG we show a 32-way split, though in practice, some may elect to go with different split ratios. Optical budgets will lead to these answers, and as of this writing, optical budgets for RFoG have not been decided.

The RFoG upstream wavelength issue is interesting. One naturally gravitates to 1310 nm as an upstream wavelength, based on widespread availability of low-cost lasers and the zero-dispersion wavelength of standard cable. Since this is the nominal zero-dispersion wavelength of the fiber, it may be possible to use Fabry-Perot lasers, at least for shorter distances. On the other hand, there are applications in which you may want to have some GPON or GE-PON and some RFoG ONTs on the same network. For instance, you may want to serve some businesses with GPON and some nearby residences with RFoG. Or, you may someday

want to upgrade from RFoG to GPON or GE-PON. Since GPON and GE-PON use 1310 nm for upstream data transmission, you cannot put RFoG with a 1310 nm upstream on the same PON.

These considerations would lead to a different wavelength choice for RFoG upstream signaling. 1590 nm is a candidate, but the next generation of GE-PON (and perhaps GPON) has already staked out this wavelength for faster upstream. Any other wavelength that can be passed through the fiber with low attenuation could be used, so the RFoG working group may choose some other wavelength. While the lasers might be more expensive at first, presumably with volume and competition the cost will drop. Of course, it is more likely that DFB lasers will have to be used since we are well away from the zero dispersion wavelength of the fiber.

THE ONT

Figure 3 illustrates the Optical Network Terminal (ONT) at the home. In Figure 3a we illustrate a fully-featured GE-PON or GPON ONT, and in Figure 3b we illustrate a possible RFoG ONT. In Figure 3a we show the optical input to the ONT coming from a 32-way splitter, common practice today. In Figure 3b we are showing a tapped architecture. While people deploying FTTH today tend to favor the splitter architecture, some in the cable TV community are leaning toward a tapped architecture.

Experience has shown that centralizing splitters from a common point within the network and dedicating fiber to each home in a star configuration provides the most cost effective deployment option. An additional benefit centralized splitters provides is the ability to scale OLT ports and splitters in accordance with sub-

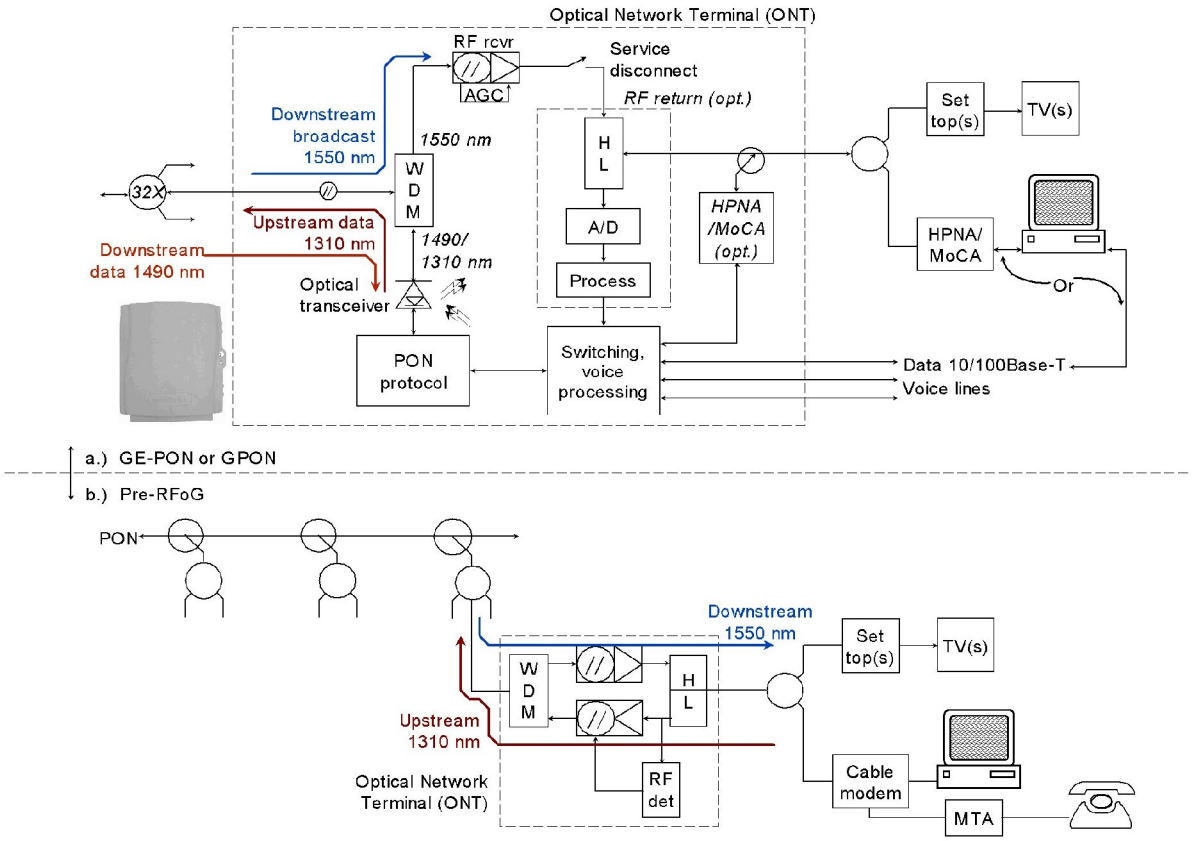


Figure 3. Optical Network Terminations

scriber penetration. In comparison a tapped topology necessitates provisioning the network for 100% of homes passed. A typical serving area for centralized splitters is 250 homes. Either topology will work.

GPON/GE-PON ONT

Figure 3a illustrates a fully-featured GPON or GE-PON ONT. These are three-wavelength systems. The broadcast downstream, 54 – 1,002 MHz, comes in on a 1550 nm carrier. A WDM in the front end of the ONT routes the wavelength to an RF receiver not unlike those in HFC nodes, except that it has been optimized for low cost. Since there are fewer sources of noise and distortion in FTTH plant compared with HFC plant, more contribution can be allocated to the ONT than to an HFC node.

This ONT was described in a paper by this author presented at the 2007 NCTA Convention,[i] so the detailed description will not be repeated here. We shall review enough detail to compare with the RFOG ONT of Figure 3b. The Figure 3a ONT includes a data transceiver interfacing with a PON protocol chip. This is an ASIC (*application-specific integrated circuit*) built by merchant silicon vendors for the appropriate PON standard. It can be thought of as roughly analogous to a DOCSIS modem. Processing on the output of the PON protocol chip converts the data into voice lines and data lines, as well as providing control to the ONT.

Typically, two or more voice lines are provided, with the internal processing supporting any of the common VoIP protocols in use today. Data is usually presented on 10/100Base-T ports, or sometimes on a 1000Base-T port. Many manufactures have a way to put data on coax in order to reduce the amount of wiring that must be done at a home. Two technologies dominate today: HPNA and MoCA. Some manufacturers use an external gateway to provide the data over coax solution, while others use an internal bridge as illustrated. This can be used for delivery of data to a computer or home network, or it can be used for delivery of IPTV (Internet Protocol Television). It can be used for both.

In greenfield applications, it is common practice today to include cat5 data wiring, so for greenfield applications, it may not be necessary to use data over coax at all.

The ONT includes an RF receiver for the 1550 nm broadcast wavelength. As shown, it includes circuitry to convert the upstream RF transmission from set tops to digital for transmission back to the headend. Other systems may use a separate analog transmitter for this function, or it may not be available.

RFoG ONT

Compare Figure 3a, a fully-featured GPON or GE-PON ONT, with Figure 3b, a stripped-down RFoG ONT. Again, we don't know yet what standard RFoG ONTs will have in them, so we start with the simplest possible solution and we'll discuss possible upgrades.

As with the GPON/GE-PON OLT, the fiber is connected to a WDM, which separates the downstream RF on a 1550 nm carrier, from the upstream RF (not data) signal on whatever wavelength is chosen. The downstream receiver could be identical to that in Figure 3a.

A diplexer separates the downstream from the upstream RF signals. Inside the home, RF wiring is exactly as it is for HFC, including the use of a cable modem and, for voice, an MTA, either embedded in the cable modem or separate as shown here.

The RFoG upstream transmitter presents an interesting situation. Analogous to the way upstream RF signals are combined, the upstream

optical signals from many transmitters will be combined before being detected in a common receiver. If we allowed the upstream transmitters to be on all the time, we would have unacceptable interference at the upstream receiver. Thus, each transmitter must be turned on only when something in the house, be it a set top or a cable modem, is transmitting. The RF detector of Figure 3b detects RF signals coming from the house and turns on the upstream transmitter, turning it off when the RF transmission ceases.

A concern is based on the fact that there could be two or more independent systems using the upstream path. The most common situation being a set top upstream transmitter and a DOCSIS upstream transmitter. There is no way to coordinate when the two disparate systems come on, so it is possible to have a set top in one home transmitting at the same time that a DOCSIS modem in another home is transmitting. If the two optical transmitters are close enough in wavelength, it is possible that they will interfere, resulting in neither transmission getting through. Retransmitting routines may mitigate this to an extent, but if a voice packet is affected, there will be a noticeable customer event.

Some people assume that the probability of the above situation is sufficiently small that the industry can live with it if the upstream wavelength utilized is 1310 nm and FP lasers are utilized. Others are not so sure. The assumption is that FP lasers utilize a wide wavelength spectrum with a variance between devices, and with 32 devices being combined statistically this would be ok. The center wavelengths of these devices tend to drift with temperature so determining the statistical frequency in which two or more wavelengths will overlap is rather unscientific. As set tops are used for more applications, it is likely that the percentage of time they transmit will go up, and we know that DOCSIS modems are transmitting a lot. A solution would be to use set tops using DOCSIS set top gateway (DSG), an internal modem, for their upstream. This would work, but restricts you on the set tops you can use. Due to cost, it is not likely that low-end set tops will use DSG.

Of course, the RFoG upstream optical transmitters will need to work with DOCSIS 3.0, which can have multiple upstream data channels in use at the same time. This adds to the performance required of the upstream optical transmitter. DOCSIS 3.0 is likely to work better with RFoG than with HFC because there are fewer sources of distortion, and the RF detector in the ONT will prevent noise funneling.

Since RFoG utilizes optical combining in the upstream direction the architecture will only support one upstream DOCSIS domain per serving group. The upstream bandwidth capacity is now limited by the capacity of a single DOCSIS domain rather than being frequency limited.

It is logical that the RFoG specification, when complete, will have a specification for the RF level threshold at which the transmitter is turned on. This threshold would logically be set as high as possible in order to improve immunity against noise generated in the house. It is desirable to force the highest possible upstream levels, because this puts operation as far above the noise level as possible.

Possible Enhancements to the RFoG ONT

We have shown a basic RFoG ONT in Figure 3b. Some have suggested putting a DOCSIS modem in the ONT. This is possible, but deviates from current cable TV practices. If the market likes the idea of outside ONTs, as are commonly used with GPON and GE-PON now, this would require a wider operating temperature range of the modem, again driving up cost.

An advantage of having some sort of communications in the ONT is that it would allow management of the ONT, something that is

not possible with the simple configuration shown in Figure 3b. A DOCSIS modem in the ONT would allow two-way communication, permitting the ONT to report on it's health and environment, something that is standard with GPON and GE-PON. Lacking two-way communications, a one-way communications path would permit remote disconnect, a standard function of GPON and GE-PON ONTs. Of course, there would be no confirmation, but that may not be seen as too great a price to pay for reducing the cost of the ONT.

ACTIVE ETHERNET

Before we change the subject, we'll mention one non-PON FTTH architecture that is popular in certain places. This is variously called Active Ethernet or Point-to-Point (P2P) FTTH.

In an active Ethernet system, a switch is placed in the field close to a cluster of subscribers. An individual fiber is run from the switch to each home, as shown in Figure 4. The IEEE Ethernet standard has a section that standardizes this configuration. The speed on the fiber to the home can be either 100 Mb/s or 1 Gb/s. However, there is typically no speed advantage with active Ethernet, because the common fiber to the left of the remote Ethernet switch has limited bandwidth, depending on what the operator wants to provide.

Active Ethernet systems are difficult to provision with RF video, because the video would have to be WDM'ed into each individual subscriber's fiber. A few such systems have been built with a second fiber system for video, but for the most part, active Ethernet systems

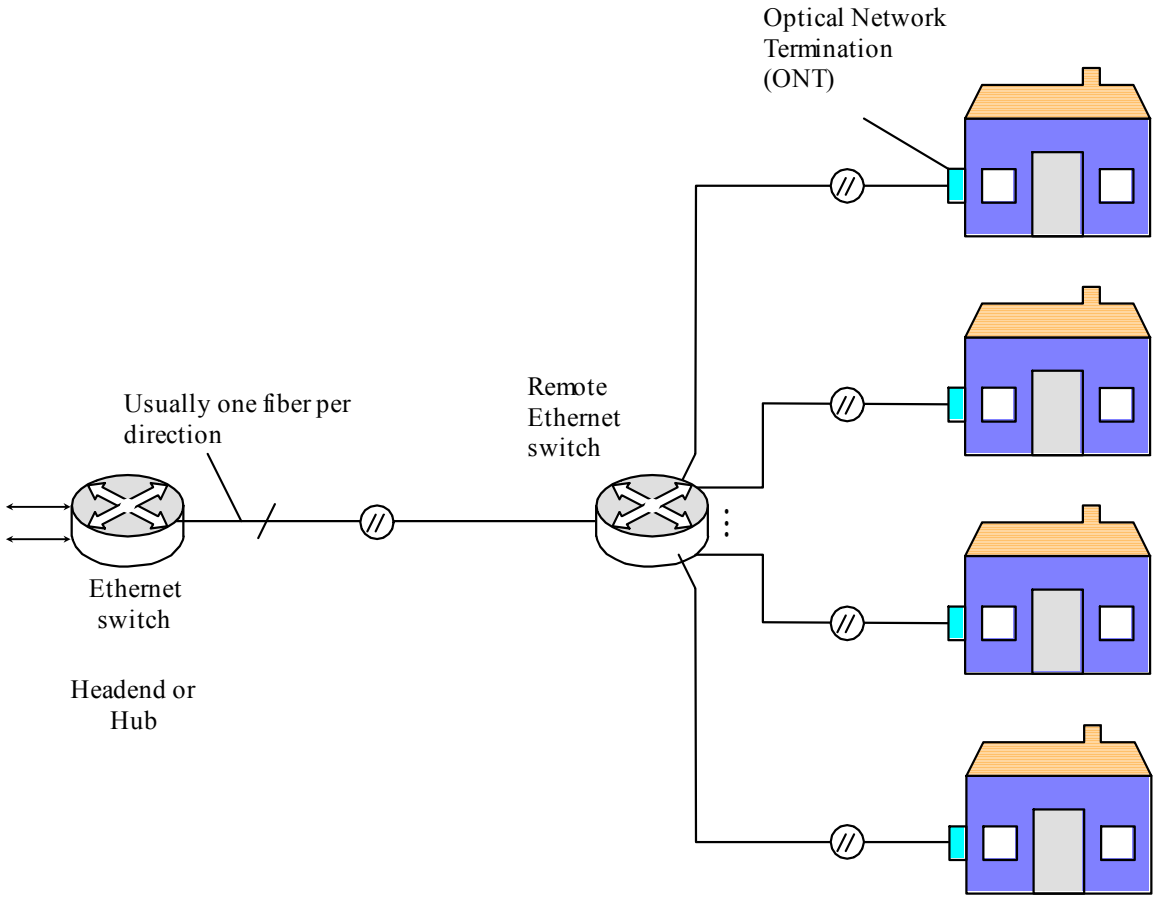


Figure 4. Active Ethernet FTTH System

carry only IPTV or no video at all.

ORGANIZING THE OPTIONS

We've talked a lot about the physical architectures of PONs. Now we need to try to make some sense of the various types of PONs, organizing them so we can understand what each does and where they fit with each other. Figure 5 diagrams the options under discussion. Starting on the right, we have the ongoing development of RFoG. This standardization effort is ongoing within the SCTE, in the fiber optics working group of the Interface Practices Subcommittee. It will be an option for cable operators to consider when required to install FTTH.

In the center of the figure is the IEEE effort, which has been incorporated into the Ethernet specification, managed by the IEEE 802.3 committee. The standard is referred to in this paper as GE-PON, but it is also known as EPON (*Ethernet Passive Optical Network*), 802.3ah (after the IEEE designation of the working group that developed it), or EFM (*Ethernet in the first mile* – someone wanted to emphasize that this applied close to the subscriber, so it was considered to be the first, rather than the last, mile). The active Ethernet architecture of Figure 4 is also a part of this standard, as is a version operating on twisted pair, at much lower data rates.

The specification was approved in 2004, and volume quantities of ASICs became available about 2006. GE-PON is very popular in Asia, which is currently leading the world in FTTH deployment, so most of the PONs in the world are GE-PON. It is also being used in North America and in Europe.

Currently GE-PON operates at 1 Gb/s in both directions. The wire speed, or speed on the fiber, is actually 1.25 Gb/s, but 8b/10b coding[ii]

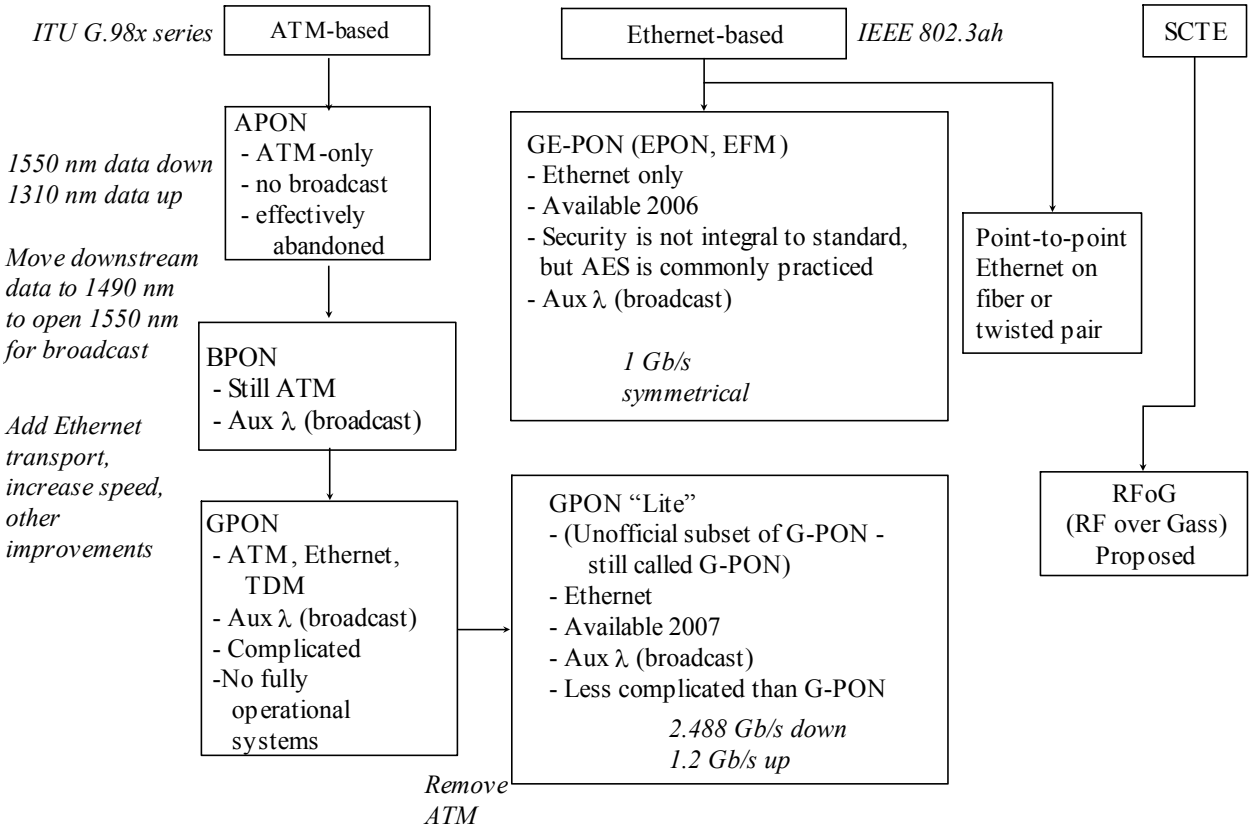


Figure 5. Comparison of PON Types

is used in order to ensure frequent transitions for clock recovery and other purposes, so the net speed is 1 Gb/s. The IEEE is currently working on a new version of the standard that will operate at 10 Gb/s downstream and either 1 Gb/s or 10 Gb/s upstream.

To the left in Figure 5 are the ITU standards. The first ITU standard, ca. 1995, was called APON for *ATM Passive Optical Network*. It used 1550 nm for downstream data and 1310 nm for upstream. It was replaced shortly by BPON (*Broadband PON*), which moved the downstream data to 1490 nm to make room for a broadcast overlay at 1550 nm. This is the version of PON that Verizon is currently deploying, though they have announced an eventual switch to the next standard in the ITU series, GPON (*Gigabit PON*).

GPON, ITU's G.984 series, was approved in parts, in 2003 and 2004. It started as a combined standard that would encompass ATM, Ethernet, and TDM (*time division multiplex*, in this context referring to DS-1 or E-1 transmissions). The standard is written to encompass all three layer 2 technologies. The problem was that implementing the complete standard was exceedingly complex. By the time people started considering implementing G.984, it had become clear that Ethernet was the choice technology for the last mile (or first mile if you use IEEE-speak).

Thus, the real implementation of GPON is based on the Ethernet portion of the standard, with the ATM portion not implemented. The author has called this "GPON Lite," but this is not an official designation – it is still known as GPON. The currently-favored version of GPON has a downstream wire speed of 2.488 Gb/s and an upstream speed of 1.2 Gb/s. It is specified to work with splits to 128 ways, but current optics don't support this many splits over any meaningful distance. The ITU's announced plan for future enhancement has been to use wave division multiplexing, where either each subscriber or a group of subscribers gets a different wavelength. However, this tends to be expensive, and there is some talk in the industry about revisiting the strategy.

Table 1. Comparison of PON Capabilities

Standard:	RFoG	GE-PON	GPON
Year standard available:	Not yet	2004	2004
Year of product general availability	Not yet (pre-standard now)	2006	2008
Field actives?	Optional	Exceptional cases	
Downstream wavelength	1550 nm	1550 nm (broadcast, optional), 1490 nm data	
Upstream wavelength	Probably 1310 nm and one longer wavelength	1310 nm (possibly going to 1590 nm in next generation)	
RF Bandwidth	54 – 1,002 MHz, depending on manufacturer		
Downstream data	DOCSIS	1 Gb/s (after removing 8b/10b)	2.488 Gb/s
Upstream bandwidth	DOCSIS	1 Gb/s (after removing 8b/10b)	1.2 Gb/s
Headend data interface	CMTS	OLT	
IPTV ready?	DOCSIS	Yes	
Service disconnect?	Not decided	Yes (depends on manufacturer)	
ONT management?	Not decided	Yes	
Upstream interference potential?	Maybe	No	

COMPARING THE PONS

Table 1 list comparative features of the PON technologies being discussed. We've listed the year that product started to be generally available to the marketplace, though there could have been limited deployments earlier. Usually, GE-PON and GPON are built strictly as passive networks, with all active equipment being restricted to the headend or hub. However, some manufacturers have made provisions for a smaller field-mounted OLT for several scenarios in which this configuration is optimum.

Everyone carries downstream broadcast on 1550 nm in order to provide for economical optical amplification, and because good optical transmitters are available for that wavelength. This is the only downstream wavelength in RFoG, but the other two standards carry all data (including voice) on a 1490 nm optical carrier. Thus, they don't loose any of the downstream RF band for data – you have up to 158 RF channels exclusively for analog and digital video. If you used them all with 256 QAM, you would have on the order of 6 Gb/s broadcast to all homes.

The upstream wavelength for GE-PON and GPON is currently 1310 nm for economy. There is talk in the industry of using 1590 nm for the next generation of GE-PON (and maybe for GPON, though this is conjecture). RFoG may provide an option of 1310 nm and something else, but this is not decided yet. The trick is to allow interoperability between RFoG and the other standards, while keeping cost low. Interoperability will allow you to deploy RFoG now, and migrate to something else later if you wish. Alternatively, you might deploy RFoG to residences, but need to serve a few businesses from the same PON, using either GE-PON or GPON. Obviously you cannot do this if you are using 1310 nm for the RFoG upstream and the other standard is using it for digital upstream.

Data is where we see the major differentiation between RFoG and the other standards. RFoG data uses DOCSIS for transport and is limited to DOCSIS speeds. At four channel DOCSIS 3.0 bonding, you have the potential for roughly 160 Mb/s of downstream data spread over, using common practice, 32 subscribers. This is an average data rate per subscriber of 5 Mb/s per subscriber, assuming one DOCSIS channel per node. Absent IPTV, this is a lot of data, because of the statistics of data sharing, a subject in which the cable TV industry has developed a lot of expertise. Yet it pales when comparing with the other two standards, which offer, respectively, average data rates per subscriber of 31.25 Mb/s and 77.5 Mb/s.

DOCSIS 3.0 upstream bonding should work better in RFoG than in HFC because of the lack of noise funneling, but the difference in upstream bandwidth is more dramatic than in the downstream direction. Developers demanding FTTH often employ telecommunications consultants who are familiar with GE-PON and GPON, and how they will react to a solution offering less bandwidth is not known yet. We are certainly talking about a lot of bandwidth with any of these PON solutions. Yet the history of data communications is that there has never been enough data bandwidth for long. With all the over-the-top video and peer-to-peer traffic today, it is not clear how long the old bandwidth sharing statistical models will hold true.

IPTV is certainly on everyone's mind today. Both GE-PON and GPON come ready to implement IPTV, and a fair number of users are doing so, some in North America, more overseas. While there are IPTV solutions designed for DOCSIS on the market, the case for putting IPTV over DOCSIS is not as clear as it is with other PON technologies – with DOCSIS/RFoG you still have the broadcast infrastructure, and switched digital video seems to have the potential for doing the same thing as IPTV, using

more mature set top technology and likely minimizing overhead.

More and more subscribers are streaming IPTV from internet web sites as the amount of content available from major networks continues to increase. In essence your subscribers have already launched you into over-the-top IPTV distribution.

SO WHEN DO YOU CONSIDER FIBER?

We have not addressed the question of when a cable operator should build a PON. As we look at the competitive landscape, HFC is in much better shape than is DSL, so the urgency is not what it is for someone with twisted pair plant.

While cable is in better shape than it's competition, bandwidth demands always go up. Your competition is starting to build FTTH. A wise decision today is to build greenfield areas with your choice of fiber technologies, while continuing to operate HFC plant where it exists. Some developers are demanding FTTH because they have learned that it improves the salability of homes.

Conversion of HFC to fiber may make sense when contemplating upgrading old plant to higher bandwidths. This is particularly true when contemplating use of bandwidth above 1 GHz, where massive plant modifications are frequently required. But this conversion can be done only on an as-needed basis, in areas of high demand (and presumably high revenue).

If you start with RFoG and later convert to either GE-PON or GPON, you would need to convert an entire PON (normally 32 or fewer subscribers) at one time. Alternately, if you elected to use a non-interfering upstream wavelength in RFoG, with suitable headend modification and taking loss budgets into account, you could convert one customer at a time. You could also operate in mixed mode for an indefinite time. You will have DOCSIS on the downstream that is not used in the GE-PON or GPON area, but having the signal there will not hurt except for the four RF channels you loose for video (DOCSIS 3.0, four channel bonding).

CONCLUSION

FTTH systems are ready to be deployed now and may make sense for greenfield deployment. The widely-recognized standards are GE-PON and GPON, which are similar in capability from a user perspective, except for speed. If you are not ready to make that leap, you can derive some of the benefits of FTTH by deploying RFoG, though the standard is not complete yet.

Author contact: jim.farmer@w7optics.com

References:

[i] Jim Farmer, *Making FTTH Compatible with HFC*, 2007 NCTA Technical Papers.
[ii] Walter Ciciora et. al., *Modern Cable Television Technology : Video, Voice, and Data Communications*, 2nd ed., San Francisco : Morgan Kaufman, 2004, Chapter 19

ADVANCES IN DWDM ROADM TECHNOLOGY USING PHOTONIC INTEGRATED CIRCUITS

Gaylord Hart
Infinera

Abstract

Modern photonic integrated circuits (PICs) integrate multiple optical subsystems on a single chip, which greatly reduces the traditional cost structure for DWDM ROADMs (reconfigurable optical add/drop multiplexers). This allows ROADMs to be cost-effectively architected for the first time using an Optical-Electrical-Optical (OEO) conversion for every wavelength at every node, thereby allowing the use of digital electronic switches for reconfigurability instead of all-optical, wavelength-only switches. The resulting digital ROADM enables new capabilities and yields significant advantages over analog all-optical ROADMs.

INTRODUCTION

Optical transport systems for delivering digital services have evolved significantly over the last several years. At each stage of this evolution, advances have been driven by economics, as well as the need for greater capacity and scalability. Inflection points in this evolution have typically occurred when technological breakthroughs have enabled a paradigm shift that allowed significant cost reductions or new, advanced capabilities, or both.

The first optical transport systems deployed were point-to-point, single-wavelength systems. To create larger networks, these platforms were often placed back-to-back at a node and electrically interconnected for traffic to transit the node. This is an inefficient and expensive method for building optical networks.

To address these shortcomings, add/drop multiplexers were developed which essentially combined the two back-to-back platforms into the same chassis. In these systems, an OEO conversion is performed by the multiplexer at each node so services can be digitally added, dropped, groomed, or switched. Services merely transiting the node still undergo the OEO conversion, but are digitally directed to the next node. This type of system is typified by SONET multiplexers and offers the benefit that any service can be digitally groomed or reconfigured at any time for add/drop or pass-through.

As the demand for increased transport capacity has grown, these systems have not scaled well because they require a fiber pair for each pair of connected multiplexers, resulting in fiber exhaust in many cases. To address this shortcoming, WDM systems were developed which allow multiple wavelengths to be carried on a single fiber pair, connecting multiple multiplexers at either end.

While these WDM systems relieved the fiber exhaust problem, they do not economically scale well because they utilize an expensive OEO conversion based upon discrete optical components for every wavelength at every node, even for wavelengths which are not adding or dropping services at the node. To squeeze additional costs out of these platforms, fixed optical add/drop multiplexers (FOADMs) were developed so that only those wavelengths adding or dropping traffic at a node undergo an OEO conversion. These systems utilize transponders to add or drop a specific service on a specific wavelength at a specific node. All other wavelengths are passed through the node in the optical domain.

FOADMs relieve fiber exhaust and lower CapEx costs by reducing the number of OEO conversions in a network, but they also created new challenges. Since some wavelengths now pass through nodes in the optical domain, optical engineering for the network is no longer a single-span engineering problem, but a multi-node, network-wide challenge. Additionally, analog optical impairments now accumulate for those wavelengths that must transit through multiple nodes before undergoing an OEO conversion. This can limit the size of networks and may require periodic re-gen of these wavelengths to remove the impairments.

FOADMs also present operational challenges. When a typical FOADM network is initially turned-up, it requires manual power balancing of every wavelength at each node in the network to ensure optimal operation. Moreover, these networks typically require rebalancing whenever services are added or deleted. If the initial network design was not carried out guaranteeing any-to-any connectivity, but only based upon initial services, it is possible the entire network may require a redesign and reconfiguration when services are changed. This is inconvenient and disrupts existing services on the network.

In all but the simplest of networks, FOADMs are time-consuming and complex to engineer and operate. To address some of these limitations, reconfigurable optical add/drop multiplexers (ROADMs) have been developed. Optical ROADMs still use transponders for the OEO conversion for adding and dropping services at a node, and they still pass through a wavelength in the optical domain when no services are being added or dropped locally from that wavelength. But a ROADM offers three capabilities a FOADM does not typically provide: auto power balancing, optical wavelength switching, and a communications control plane to automatically and remotely reconfigure the network when necessary. Some ROADMs also support transponders with tunable lasers to provide enhanced re-configurability.

Optical ROADMs are a great improvement over FOADMs. They accelerate and simplify network turn-up and service changes by automatically balancing optical power throughout the network. They also allow wavelengths to be remotely switched for optical add/drop or pass-through operation at each node. However, they still have many of the limitations of FOADMs, including complex optical layer engineering and a dependence on transponders, which tie services to wavelengths and hence make service layer engineering dependent upon optical layer engineering.

One way to address these optical ROADM limitations is to perform an OEO conversion for every wavelength at every node, then process all services digitally at each node. This would reduce the optical network engineering problem to a series of simple single-span designs, even for mesh networks. Using 3R (regenerate, reshape, retime) OEO conversions at each node, optical impairments would not accumulate, and this would allow networks of essentially any arbitrary size to be built. Finally, digital processing at each node would allow replacing the all-optical wavelength selective switch of the traditional ROADM with a digital switch which can support advanced capabilities not possible with an optical layer switch. Such a digital ROADM would allow for the first time the creation of a flexible digital optical network where the service layer is independent of the optical layer

In the past, digital ROADMs were not economically feasible due to the many discrete optical components required to perform an OEO conversion for every wavelength at every node. However, recent advances in PIC technology allow complete optical subsystems to be economically placed on a pair of chips (TX and RX) less than 5 mm square and supporting 100

Gb/s per chip in a 10λ x 10G DWDM configuration. These PICs integrate over 60 discrete optical components (lasers, detectors, mux/demuxes, etc.) on a single pair of chips, eliminating all the discrete optical packaging and the fiber jumpers formerly required to interconnect the devices (see Figure 1, below). The resulting cost-reduction enables digital ROADMs to be built cost-effectively for the first time, and commercial digital ROADMs built on PIC technology have now been available for three years. PIC technology yields other benefits, as well: reduced power consumption,

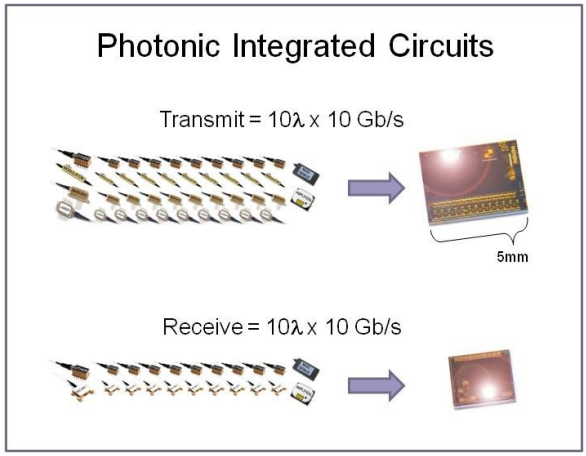


Figure 1 – Large Scale Photonic Integration Places Several Optical Components on a Single IC

smaller footprint, reduced heat generation, fewer modules, fewer fiber jumpers, and higher reliability.

Just as large-scale integration in digital electronics has enabled radical cost reductions even while increasing processing power, PICs now enable the digital paradigm shift to digital optical networks. Even more exciting, Moore's law indicates significant gains are possible in future generations of PIC technology. Fully functional PICs have now been built in the lab which support 10λ x 40G configurations, and these 400 Gb/s PICs are scheduled to be available in 2009.

ROADM ARCHITECTURE COMPARISONS

SONET has lost favor with most MSOs as they have migrated to Ethernet, but SONET is still widely deployed in the Telco world and in some CATV commercial services environments where some customers require SONET transport. FOADMs are also widely deployed, but are finding more favor in applications where the networks are smaller and less complex. For metro core, regional, and national DWDM networks, most cable operators are now deploying ROADMs to lower total cost of network ownership and to accelerate new service and bandwidth turn-up. In these networks, digital ROADMs now compete with all-optical ROADMs for market share and technical leadership.

Optical and digital ROADMs typically share many common attributes, including the ability to provide dispersion compensation, amplification, and automated optical power balancing at each node. They both typically implement an intelligent control plane (preferably using GMPLS) that allows remote and/or automated configuration as well as other features. Many support automated topology and inventory discovery and optical layer turn-up. Where optical and digital ROADMs primarily differ is in the way they provide core re-configurability: whether they switch in the optical or digital domain. To understand these differences, it is necessary to examine the way optical and digital ROADMs are architected. These differences have implications in the cost, engineering, and operation of a network, and a digital ROADM, as we shall see, provides significant advantages in each of these areas.

Optical ROADM Architecture

Optical ROADMs (just as FOADMs do) only perform an OEO conversion at a node for wavelengths being added or dropped at that node. Wavelengths not being added or dropped

at the node are "expressed" through the node in the optical domain.

Unique to the optical ROADM is a wavelength selective switch (WSS), or equivalent functionality implemented with wavelength blockers or other technology, which allows individual wavelengths received at the node to be switched for local add/drop or for "express" transit through the node. Figure 2, below, shows the high-level architecture of a typical 2-degree optical ROADM.

Looking at signal flow from west to east, the RX optical signals originating from the network's west-facing fiber interface are presented to an optical demux which breaks out the individual wavelengths for processing by a software configurable optical switch. The switch then individually passes each RX wavelength directly through the ROADM and out the reciprocal east-facing TX optical mux, or it drops it to a west transponder for local service handoff. The TX signals originating from these west transponders are usually passively added to the pass-through channels travelling from east to west and sent out the west-facing TX fiber.

The ROADM also provides reciprocal functionality for the TX and RX optical signals originating from the network's east-facing fiber interfaces. For redundancy purposes, this is implemented in the ROADM with separate east and west optical modules with each containing an optical switch and any other required optical muxing, combining, or splitting components. Some ROADMs also provide optical performance monitoring points here for "express" wavelengths, as well.

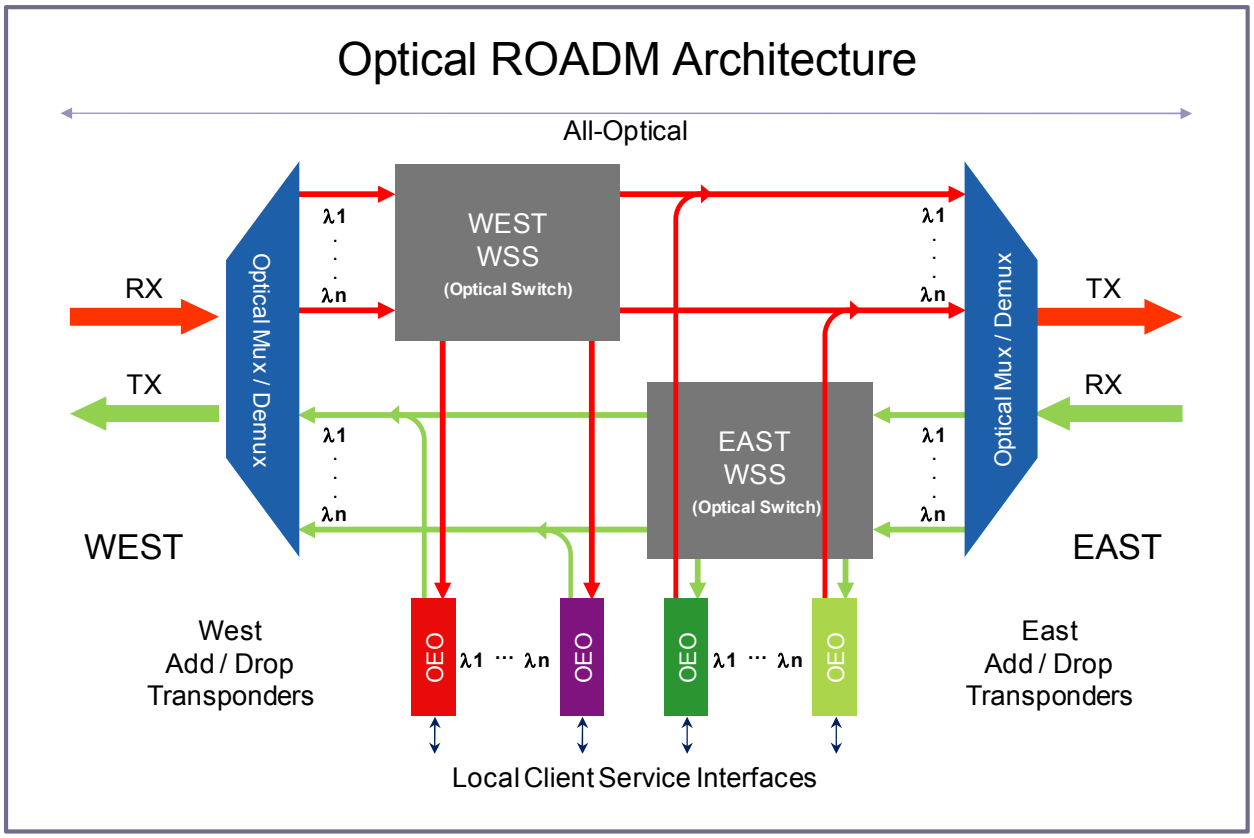


Figure 2 – Typical 2-Degree Optical ROADM Architecture

It should be noted that some optical ROADMs are banded, requiring add/drop or express treatment for a group of wavelengths as a whole (banding is usually implemented in wavelength groups that are an integer multiple of four). Because banded ROADMs must treat all wavelengths in the banded group the same, this can result in stranded bandwidth. For example, if only two wavelengths need to be dropped off at a node, and the banding group size is four, the ROADM will strand two wavelengths at the node.

For the wavelengths being dropped off locally, the transponders perform an OEO conversion and provide client optical interfaces for connection to local equipment. Between the client and line-side optical interfaces, services are processed digitally on the transponder, and it is here that bit-error-rate (BER) evaluation, forward error-correction (FEC), digital wrapper processing (usually G.709), and digital performance monitoring takes place. The transponder also typically provides performance monitoring for any service-specific attributes. Optical performance monitoring points for the add/drop signals are frequently provided here as well.

Each transponder at the node links a particular line wavelength to a particular client service interface (if tunable laser transponders are used, this provides a greater degree of reconfigurability, but at a higher cost than fixed wavelength transponders). Because the same wavelength may arrive from the east and west, and for redundancy purposes, the ROADM supports both east and west-facing transponders, and these typically can only receive wavelengths from the direction of the optical switch module they are attached to. When protected services are required, two transponders must be used, one each for east and west. Service growth is implemented by adding transponders at the service source and destination nodes, and these are typically only installed at a node when services are actually required there.

Digital ROADM Architecture

Digital ROADMs perform an OEO conversion for every wavelength arriving at every node, both east and west. Wavelengths not being added or dropped at the node are still "expressed" through the node, but in the digital domain before the conversion back to optical when exiting the ROADM.

Unique to the digital ROADM is a core digital cross-connect switch connected to every digital signal derived from every optical interface, including the line interfaces and the client interfaces. This switch allows any input to be connected to any output. Because this switch resides in the middle of the OEO conversion (unlike with an optical ROADM, where the OEO takes place in the transponder after optical switching), the line-side optical layer is completely segregated from the client side service layer. Figure 3, below, shows the high-level architecture of a typical 2-degree digital ROADM constructed with PICs.

At each line-side interface, both east and west, an optical band multiplexing module segregates groups of incoming wavelengths into Optical Carrier Groups (typically comprised of 10 wavelengths at 10G each) for handoff to the RX PICs. The same band multiplexing module aggregates the reciprocal group of outgoing wavelengths from the TX PICs for handoff to the line-side fiber. In turn, each PIC simultaneously processes multiple line-side wavelengths in parallel, performing O-E conversions on the RX side and E-O conversions on the TX side. Each PIC processes 10 WDM wavelengths at 10 Gb/s each. A typical initial ROADM deployment starts with a pair of TX/RX PICs on each of two digital line cards facing east and west, respectively. This provides an initial transport capacity of 100Gb/s in both directions. Services

are then added until all 10 wavelengths on the initial PICs are consumed, and then an additional pair of digital line cards is installed to support further service growth.

The electrical interfaces on each PIC are connected to the core digital cross-connect switch, which is also connected to tributary adapters for local service handoff on the client side. This switch is under software control and may be remotely configured. It is implemented in a redundant configuration to eliminate any single points of failure.

Internal to the ROADM, all services are digitally processed. As the PIC's electrical signals interface with the core digital switch, they are processed for BER monitoring, FEC, digital wrapper manipulation, and other performance monitoring. Once inside the core switch, these signals may be directed in any direction, to any interface (line or client), and to any wavelength (east or west, any color). In this manner, express services are switched directly through the ROADM, and add/drop services are directed to local client interfaces.

Services are added or dropped at the node through a tributary adapter. The tributary adapter converts the digital transport stream from the core switch for native service handoff and provides this service on a client optical interface for local use. The tributary adapter also provides performance monitoring for any service-specific attributes and loopback capabilities for test purposes.

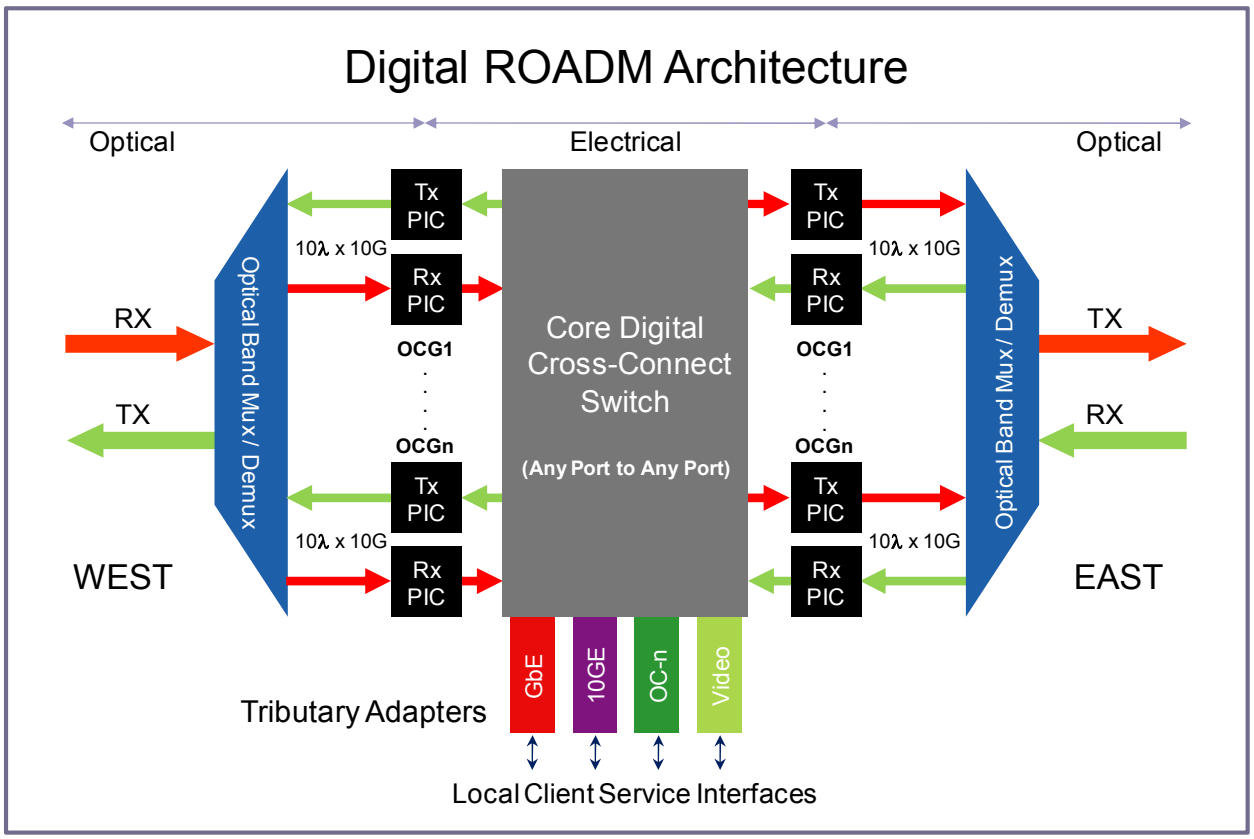


Figure 3 – Typical 2-Degree Digital ROADM Architecture Using Photonic Integrated Circuits

OPTICAL LAYER ENGINEERING

The optical layer engineering process and requirements are very different for optical and digital ROADMs. Optical ROADMs only perform an OEO conversion for wavelengths locally adding or dropping services at a node. All other wavelengths are passed through the node in the optical domain. For these pass-through wavelengths, optical impairments such as dispersion and cascaded filter losses accumulate from span to span. Since the source and destination nodes may be anywhere in the network, each wavelength must be individually engineered for its particular path through the network. But since different wavelengths often share common spans in the network, this is not a simple process. The final design must represent a common denominator that works for all wavelengths in the network, regardless of their paths, and this may result in a less than optimal design that may limit node counts or span distances. In some cases, it may be necessary to perform a 3R re-gen in the network, and this requires back-to-back transponders for every wavelength requiring re-gen.

A digital ROADM performs an OEO conversion for every wavelength at every node. In this case, all wavelengths on a span share the identical path regardless of whether they are being added or dropped at a node, and only individual spans need to be engineered between nodes. Moreover, because a 3R OEO conversion takes place at each node, optical impairments do not accumulate from span to span in the network. This essentially allows networks of any arbitrary size to be built.

Ideally, an optical network's initial design should allow any service at any node to be transported to any other node in the network at any time without requiring any re-engineering or re-configuration of the optical layer to do so. To guarantee this any-to-any connectivity in an optical ROADM network, every relevant analog optical transport parameter (OSNR, dispersion, power levels, etc.) has to be calculated east and west for TX and RX for every wavelength for every possible combination of end-nodes (see Figure 4, below). In a ring network, this requires $2(N^2\text{-}N)WP$ calculations (where N is the number of nodes in a network, W is the number of wavelengths supported by the network, and P is the number of transport parameters to be verified).

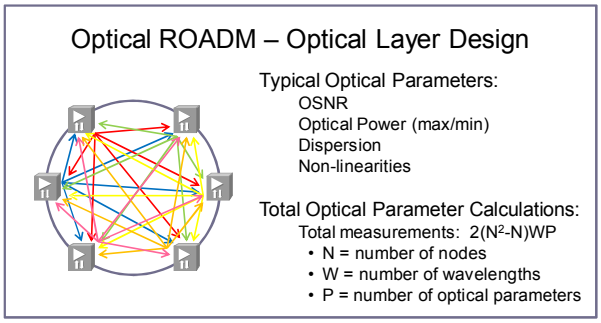


Figure 4 – Optical Layer Design for an Optical ROADM

In a large ring network with a large number of wavelengths, thousands of calculations may be required to verify the optical layer design when optical ROADMs are used. If the initial design assumptions prove unworkable, redesign and recalculation may be iteratively required to find a workable combination of all optical layer parameters. For a mesh network with multi-degree nodes, the engineering problem becomes vastly more complex.

Digital ROADMs limit the optical layer design to a series of independent span designs where the design of one span does not impact the design of any other span (see Figure 5, below). In a ring network, this requires a total of 2NWP calculations (where N is the number of nodes in a network, W is the number of wavelengths supported by the network, and P is the number of transport parameters to be verified). To guarantee any-to-any connectivity in the network, one only has to guarantee that each individual span between a pair of nodes has been

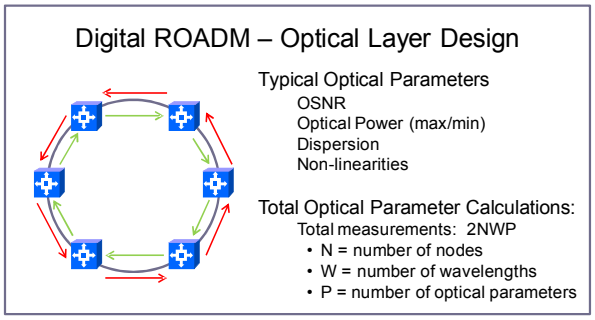


Figure 5 – Optical Layer Design for a Digital ROADM

properly engineered. And this is true even for large, complex mesh networks.

OPTICAL vs. DIGITAL PROCESSING

Aside from the major difference in these ROADMs between switching in the optical or digital domain, another key difference is where the switching is actually performed. An optical ROADM switches above the OEO conversion, and the transponder which actually performs this conversion tightly couples the line-side optical transport layer to the client side service layer. In contrast, a digital ROADM separates the OEO conversion process into distinct line-side optical layer and client-side service layer interfaces, and does its switching in the middle of these.

Having examined the respective architectures for optical and digital ROADMs, we can now examine the specific capabilities supported by each.

Switching

Optical ROADMs, of course, switch entire wavelengths, which are typically transported today at a nominal 2.5 Gb/s or 10 Gb/s line rate. If any of these wavelengths have lower rate services multiplexed up to the line rate by the transponder (muxponder), as is common to maximize the use of transport capacity, these sub-rate services cannot be individually switched, groomed, or added/dropped at intermediate nodes on the network unless back-to-back transponders are used where this functionality is required. This is because the sub-rate services can only be processed digitally, and an optical ROADM only has access to digital signals at the transponders, which are only located at the source and destination points for a service.

While some transponders provide some integrated switching capability, this is typically limited to within the transponder itself or to an adjacent back-to-back transponder for digital add/drop capabilities. It is not uncommon when service grooming or switching is required at a node to accomplish this with manually installed fiber jumpers between the client interfaces on the back-to-back transponders. In these cases, one is often faced with the economic tradeoff between stranding bandwidth or paying for additional transponders.

In WDM optical networks, especially mesh networks, it is quite common to have wavelength contention (or even blocking) between services being carried over the network. This occurs when two independent services using the same wavelength need to travel over the same fiber on at least one span. Of course, one may simply use a different wavelength for one of the services when this occurs, but this may in turn create contention with another service on another span. To fully utilize available network capacity, it may be necessary to perform a wavelength conversion for a service for transport over the contended path. Because the transponders on an optical ROADM tightly couple a service to a wavelength between the transponders, and because an optical ROADM cannot switch services between wavelengths, the only way to perform this wavelength conversion is to use back-to-back transponders using different wavelengths on the contended path. This is an expensive solution, but may be the only option when blocking occurs.

Digital ROADMs provide simple and inexpensive solutions to these switching challenges, and provide additional capabilities as well. As already noted, these ROADMs have digital access to all services and wavelengths at every node and also have an integral digital switch interconnecting all interfaces at every node. This combination yields very powerful capabilities for switching services and maximizing bandwidth usage in the network. An examination of these capabilities follows, but a more detailed discussion of how a digital ROADM processes services is necessary first.

To maximize transport capacity, advanced digital ROADMs utilize PICs providing 10 x 10 Gb/s wavelengths for transport across the network. A 10 Gb/s digital transport frame (DTF) using a G.709 digital wrapper or an enhanced version of it is used on each wavelength to transport native client services end-to-end. The DTF also provides forward error correction (FEC) and performance monitoring, not only between intermediate transit nodes, but between end-to-end service points as well. The DTF in turn has four 2.5 Gb/s digital signals asynchronously multiplexed into it, thus the DTF may transport one 10G or four 2.5G services.

The DTF and its 2.5G sub-rate signals are transparent to the client signals and may carry Ethernet, SONET or other protocols. A 10G DTF can simultaneously support any combination of client signals mapped into its 2.5G signals, up to the full 10G rate. A 2.5G signal may in turn have 2 GbEs mapped into it. Thus a 10G DTF may carry one OC-192, one 10GE, four OC-48s, 8 GbEs, or some combination of these that does not exceed the 10G line rate. Other mapping possibilities exist, and other protocols can be supported as well. The DTF's flexible mapping capabilities allow each wavelength to be used efficiently by multiplexing any combination of sub-rate services into it until the wavelength is fully utilized.

At the switching level, the ROADM's integrated digital switch works at the 2.5G sub-rate granularity, so sub-rate switching is fully supported along with switching at the 10G line rate. This supports add/drop multiplexing of any service at any node at any time. In fact, even the tributary interfaces at a node may be switched to face east or west in the network. Unlike with transponder based optical ROADMs, digital switching makes service transport completely independent of the transport wavelengths: full grooming and switching of every service at every node (including sub-lambda services) is possible to and from any wavelength, and the service bit rate is independent of the transport line bit rate. Figure 6, below, shows several of the processing capabilities provided at every digital ROADM node.

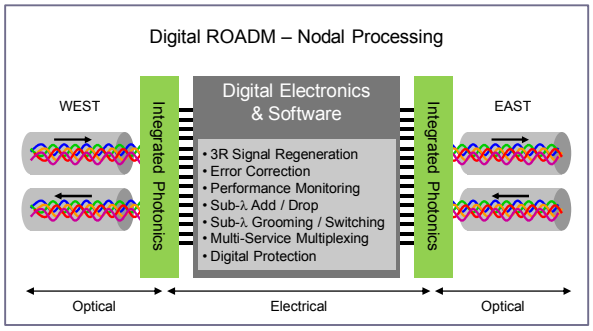


Figure 6 – Digital ROADMs Process All Services Digitally at Each Node

Digital switching is particularly effective in creating multi-degree networks, especially more complex mesh networks, where the complexities of optical layer engineering and wavelength planning become great. With a digital ROADM, the optical layers of each path at a junction node are independent, and traffic is simply switched between them digitally. Services may then be routed end-to-end through any available path in the network, and services may be created or torn down on demand without any optical layer engineering. Mesh networks provide enhanced protection and bandwidth management options

by providing multiple paths between endpoints in the network.

Digital ROADMs are also quite effective in building hierarchical networks, where a single digital ROADM can serve, for example, as a junction site between a regional and metro core network. In this configuration, a digital ROADM can provide north and south interfaces for one network, and east and west for the other. Traffic may then be switched digitally in any direction between or within the networks, and protection is supported for both networks. Because a digital ROADM provides 3R OEO conversions at every node, there are inherently no distance or node count limitations in a digital network. It is actually possible to build a national backbone with integrated regional networks and metro core rings using a scalable digital ROADM.

Bandwidth Virtualization

Whereas transponders in an all-optical ROADM tightly couple services to wavelengths, a digital ROADM for the first time allows the service layer to be fully segregated from the optical layer. This allows wavelengths (and 2.5G sub-lambdas) to be treated as a pool of virtual bandwidth, that is, as an allocatable resource to be assigned to services when and only as needed. As long as sufficient bandwidth exists between any two nodes in a digital ROADM network, any service may be turned up between these nodes at any time. If sufficient bandwidth does not exist, digital line modules may simply be added in those spans where more bandwidth is needed to support the new service. Since PIC based line modules add bandwidth in groups of 100 Gb/s, any unallocated bandwidth on the new line module is added to the pool for future use. In this way, capacity usage is optimized on individual wavelengths and between individual nodes.

Because a digital ROADM provides 3R OEO conversions for all wavelengths at every node and because optical layer engineering is then limited simply to spans between these OEO nodes, service layer engineering is now made fully independent of optical layer engineering. Once a span has been engineered for the first digital line module, additional modules may be installed without the need for any additional optical layer engineering or reconfiguration. This greatly reduces the time and effort required to turn up new bandwidth and services.

Bandwidth virtualization enables flexible reconfiguration of the network at any node at any time without any optical layer re-engineering and provides significant additional capabilities that cannot be provided with an all-optical ROADM. First, unlimited add/drop capabilities are supported for any service at any node at any time, including sub-lambda services. Second, wavelength or sub-lambda services may be groomed or switched from any wavelength to any other wavelength at any node. This enables wavelength conversion for wavelength services and inter-wavelength grooming for sub-lambda services, so a service may actually be transported on several different wavelengths in the network between its source and destination points. This flexibility allows existing wavelengths in the network bandwidth pool to be fully utilized before having to add more capacity and provides a simple solution to wavelength contention or blocking.

Bandwidth virtualization also supports super-lambda services. For example, a 40G service can be delivered using tributary adapters that concatenate 40G service transport over four 10G wavelengths from the available bandwidth pool. In this manner, 40G services may be delivered over a 10G network without any additional optical layer engineering or reconfiguration. Since the client handoffs at either end of the service are standard 40G interfaces, service delivery is fully transparent.

Digital processing enables other features as well, many of which support advanced maintenance capabilities or unique architectures. A digital ROADM supports hair-pinning services, where a service is brought into the ROADM on one client port and is "hair-pinned" out a different client port at the same node. Bridge-and-roll is also supported. Using bridge-and-roll, an alternate path is created for a service between the two endpoints. A bridge is then created to duplicate the service over both paths. Finally, a roll operation is executed which transfers the end-to-end service over to the new path in under 50 ms. The original path is then free for maintenance operations. Non-obtrusive digital test ports may also be created at any node in the network to observe traffic on any service. In this case, a service is digitally replicated at any node in its path, and this copy is then sent to another client port at any node in the network, where it may then be attached to test equipment or otherwise examined. Finally, digital processing enables a wide range of loopback options to be used for circuit verification and fault-isolation purposes, including the ability to perform loopbacks not only at the service endpoints, but at intermediate nodes in the service path as well. All of these maintenance and test mechanisms are remotely provisionable.

Using digital multicast, digital ROADMs also make unidirectional drop-and-continue digital video broadcast architectures simple and inexpensive to implement at Layer 1. In this application, the ROADM's digital switch replicates the broadcast digital video service at any required node, drops it locally, and then passes it on to the next node. At each node, up to three output multicasts may be created, supporting local drop and multicast branching at junction nodes. Since the multicast is handled digitally, no optical splitting or specialized transponders are required, and there is no limit to the number of drop or branching sites. Digital multicast can also be used to support switched digital video when an edge-switched architecture is used for this application. Digital multicast services may be digitally protected for increased reliability.

Performance Monitoring

All ROADMs provide some degree of performance monitoring, and the information derived from this is used for many purposes, including guaranteeing service level agreements (SLAs) with end customers, establishing base network operating parameters so any observed degradation can be used to solve problems pro-actively before an outage occurs, and to diagnose and sectionalize an outage if it does occur. All ROADMs provide performance monitoring points (PMs) for both optical and digital parameters, but significant differences exist between the number of parameters monitored and where they are monitored.

Optical PMs typically consist of optical power levels measured at various points in the ROADM, but may include other parameters as well, such as laser bias current. Power measurements may be aggregate power (combined power of all wavelengths at the PM point) or individual power (power level of an individual wavelength at the PM point). An aggregate power level PM can indicate a problem exists (level too high or low), but it can't tell you which wavelengths are contributing to the problem and therefore is usually less helpful diagnosing and locating a particular problem.

Optical ROADMs typically provide optical PMs at various TX and RX points in the ROADM, but how and where these are measured varies considerably. Optical PMs should be provided for both TX and RX signals on both the east and west line-side interfaces. Ideally, these should provide aggregate and per-lambda data for express wavelengths as well as those originating from local transponders. This provides the greatest amount of useful information for diagnosing and pinpointing any

problems. However, some ROADMs only measure optical power levels on the line-side interfaces in aggregate and only look at individual wavelength levels on the local transponders.

Digital ROADMs, because they perform an OEO conversion for every wavelength at every node, typically provide a full complement of optical PMs, measuring TX and RX power levels for every wavelength on both the east and west line-side interfaces. Aggregate optical power is also usually monitored on the line interfaces. This provides rapid and robust fault diagnosis and location capabilities at the optical layer.

Optical power level monitoring is quite useful, but there are only a limited number of optical layer problems which can be diagnosed with power levels. For example, optical power levels can help diagnose fiber cuts or laser failures, but not dispersion problems. Moreover, optical PMs are of no help in diagnosing digital or service layer problems, and these are usually assessed using digital PMs.

Digital PMs typically provide a large amount of information derived from a large number of monitoring points. Typical information would include loss of frame (LOF), loss of signal (LOS), uncorrected BER, corrected BER, errored seconds, severely errored seconds, and numerous other parameters. This information is usually gathered on the optical transport path for each service and/or wavelength being transported on the network, depending on the particular ROADM. BER data is usually gathered for the optical transport path via a G.709 or enhanced digital wrapper, which provides FEC, as well. Additional service-specific PMs (e.g., Ethernet frame and errored frame counts) are also typically available.

Optical ROADMs, by their nature, only provide digital PMs where signals are processed digitally (i.e., only at the transponders). Since transponders are only used at the service endpoints, digital PMs are only available there, and not at any intermediate nodes (see Figure 7, below). This makes diagnosing and localizing a problem more difficult since intermediate nodes in the path cannot provide PMs which might indicate which section is responsible for the problem.

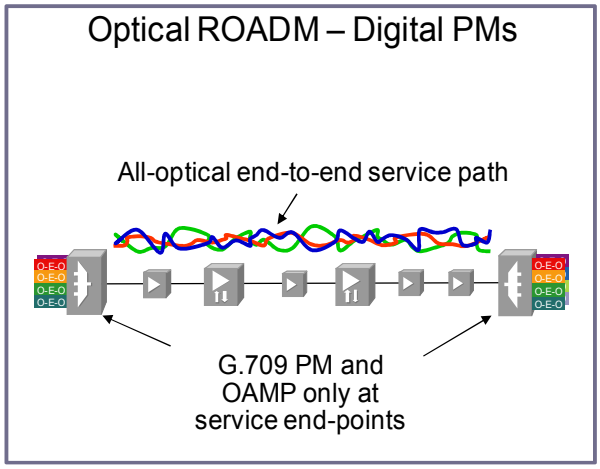


Figure 7 – Digital PMs on an Optical ROADM

Digital ROADMs, on the other hand, perform an OEO operation at every node for every wavelength, so digital PMs are available at every node for every signal that transits the node (see Figure 8, below). Digital ROADMs provide PMs for both the transport section (transport between each node) and the transport path (end-to-end service transport). This allows rapid pinpointing of any problems to an individual link between two nodes.

To facilitate testing, some digital ROADMs also directly incorporate a pseudo-random bit stream (PRBS) generator for direct BER testing between any two nodes without the need for any external test equipment. In PIC based digital ROADMs, a PRBS stream is run continuously on those 10G wavelengths which are installed on operational digital line cards but which are not yet carrying services. This provides the MSO

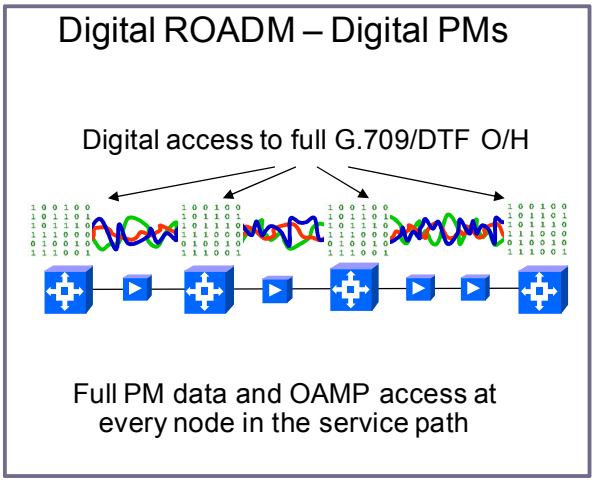


Figure 8 – Digital PMs on a Digital ROADM

with an operational history and track record even for those wavelengths not yet turned up with services and provides a higher degree of confidence in successful service turn-up when the time comes to activate these wavelengths.

Protection

Optical ROADMs typically protect a service by using two route-diverse paths between the endpoints, with a pair of transponders on each path. One transponder pair serves for the working path, one for protection. This configuration provides effective equipment and path protection, but the protection wavelength and its transponders are dedicated and cannot be used for other services. A single client handoff at either end is usually provided by a Y-cable connected to the clients on the working and protect transponders. Some optical ROADMs support a digital backplane interface between the transponder protection pairs, which allows a single client interface on one of the transponders to be used for the protected service without requiring a Y-cable. In this case, the best signal from either of the transponders is presented at the single client. For 1+1 protection, where the local end device connected to the ROADM (typically a switch or router) actually performs the protection switching, the clients on each of the two transponders are connected directly to reciprocal clients on the end device.

Digital ROADMs can also be configured for dedicated protection. In this case, both working and protection paths are pre-defined through the network between the service endpoints. These paths are set up through each node using the ROADM's integrated digital switch and any available wavelengths through the network. Route diversity is used for path protection. At the endpoints, a pair of tributary adapters is used to provide the protected clients. As with optical ROADMs, the paths and tributary adapters are dedicated and cannot be used for other services. A Y-cable may be used to provide a single protected client interface, just as with optical ROADMs, or the ROADM's integrated switch may be used to deliver the best signal from either path to a single tributary adapter client. This second method saves the cost of one tributary adapter and the Y-cable. Dedicated 1+1 protection is handled with two tributary adapters just as with an optical ROADM.

Digital ROADMs, however, support shared protection modes which are not supported by most optical ROADMs. With shared protection, the protection wavelengths are left uncommitted throughout the network, and these remain in the pool of allocatable bandwidth until actually needed. In this way, a small shared bandwidth pool can be used to protect many services across the network, resulting in much lower bandwidth consumption when compared to dedicated protection.

When a failure occurs in a shared protection network, the GMPLS control plane finds a path through the network with sufficient bandwidth to restore service, allocates this bandwidth from end-to-end, then switches the service over to the new path. If sufficient paths and bandwidth are available, protection may be provided over multiple failures at multiple points in the

network. Since multiple failures in different segments of the network are unlikely to occur simultaneously, shared protection provides an effective means of conserving bandwidth while providing a high degree of confidence in protecting against network failures. Shared protection may be used in ring or mesh networks, but mesh networks typically provide more paths through the network between service points and therefore may provide more options for protection routing than would be available in a ring network.

Unlike dedicated protection, shared protection frees up significant bandwidth in the network for other uses, and the network operator has full control over how much spare bandwidth to provide for shared protection. However, there is a tradeoff in using shared protection. A dedicated protection path, because it is always live, provides protection switching in under 50 ms. A shared protection path must be found and routed after a failure, and this can take a few seconds. In a digital ROADM, protection may be provisioned as dedicated or shared on a service by service basis, so this is typically not a problem since any sensitive services can always be configured with dedicated protection.

SUMMARY AND CONCLUSIONS

Photonic integrated circuits represent a major inflection point in optical networking evolution, enabling the digital paradigm shift to digital optical networking and delivering the scalability that will be required for next-generation networks.

Major advances in photonic integrated circuits have resulted in commercial production of inexpensive, highly reliable photonic ICs integrating all the components required to deliver ten 10 Gb/s wavelengths, on a pair of chips (TX and RX) no more than 5 mm square. This allows a low-cost OEO conversion to be used for every wavelength at every optical multiplexer in a network. The PIC cost-savings in turn support integrating a full digital cross-connect switch into the core of the multiplexer, creating for the first time a cost-effective digital ROADM.

The architecture of the digital ROADM (and hence its name) is such that all services are processed digitally, rather than optically. Since a digital ROADM performs an OEO operation on every wavelength at every node, its integrated digital switch has unrestricted access to every service entering or leaving the multiplexer, and therefore has unrestricted ability to groom, switch, or add/drop services at the node. This delivers a much wider range of reconfiguration options than an all-optical ROADM can provide, but it also enables a completely new set of features and capabilities.

Digital ROADMs greatly simplify optical layer engineering, and their 3R OEO architecture supports networks of essentially any size or shape to be built and provisioned easily. Because digital ROADMs segregate the optical layer from the service layer, turning up new services is quick and requires no optical layer engineering whatsoever. Digital ROADMs support wavelength conversion, sub-lambda grooming, inter-wavelength switching, and are in general more bandwidth efficient than their optical counterparts.

Digital ROADMs bring significant benefits to all phases of network ownership, from network engineering, to network turn-up, to service growth, and to network evolution. They offer a lower total cost of network ownership while accelerating network operations, bandwidth expansion, and service delivery.

BUILDING LARGE VOD LIBRARIES WITH NEXT GENERATION ON DEMAND ARCHITECTURE

Weidong Mao
Comcast Fellow
Office of the CTO
Comcast Cable

Abstract

The paper presents an integrated Video On Demand (VOD) content library platform that supports virtually an unlimited amount of media content such as movies, TV shows, Internet video, and user generated content. This approach combines the advantages of the existing Managed Network approach and the emerging Over the Top approach in offering VOD services.

Specifically, this paper describes the overall requirements and architectural evolution of Video On Demand (VOD) infrastructures to support large content libraries. The content libraries will enable a large amount of VOD content (SD and HD) as well as Internet Video content. The solution is based on the Comcast Next Generation On Demand (NGOD) architecture with a key extension of the Content Delivery Network (CDN) that utilizes national and regional IP network and library storage.

Several key architectural building blocks and technology options are include content encoding and transcoding, real time and non real time content ingest, asset metadata and rights management, content library and asset propagation management, VOD backoffice integration, streaming server, as well as shared edge resources.

Finally, this paper also discusses content formats, open interfaces, performance, scalability, reliability, and expandability to future on demand services.

MANAGED NETWORK VS OVER THE TOP

Increasingly cable operators are using Video On Demand (VOD) as a key competitive advantage. Alternative video delivery methods such as movie download or video streaming via the Internet are also becoming more practical and feasible as service providers deploy either DOCSIS 3.0 wideband or Fiber to the Home technologies.

Figure 1 illustrates comparisons between the "Managed Network" and the "Over the Top" approaches for providing on demand video to subscribers. In the existing Managed Network approach that is adopted by Cable and Telco network operators, VOD content is usually encoded in MPEG-2 format and distributed along with metadata via a Satellite or IP backbone to the local VOD systems. The content is usually "pushed" and replicated in every local VOD system. The VOD client on the digital set-top box will be able to setup sessions and perform stream control functions such as Play, Pause, Fast Forward and Rewind.

In contrast, the emerging Over the Top approach uses the broadband Internet as the content distribution and streaming platform. Content aggregators / integrators license and

publish movies and TV shows at the Internet website. PC or CE devices such as TV set-tops are able to access the video content via the Internet using the broadband pipe such as cable modem, DSL, or Fiber to the Home network. Content for the Over the Top services is typically encoded using advanced codec such as H.264 with lower resolutions. Content distributions within the Internet are usually driven by the "pull" requests coming from the subscriber.

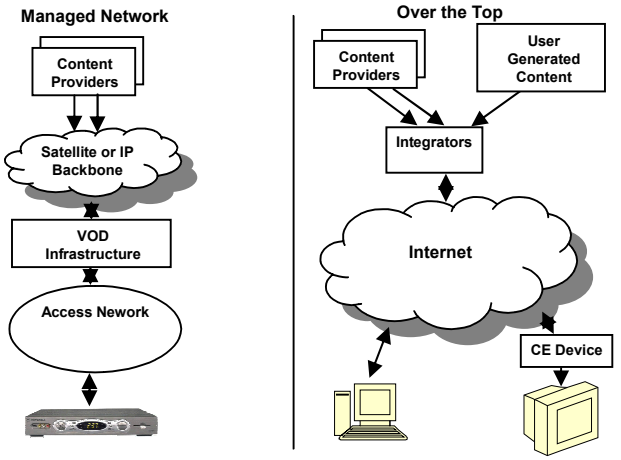


Figure 1 Comparisons of Managed Network and Over the Top VOD Services

The following lists some of the unique features and capabilities of Managed Network and Over the Top VOD services:

Managed Network features and capabilities:

- Extend linear TV programming service with free and premium VOD content
- Utilize managed IP networks for VOD content distribution with better quality of service
- Manage bandwidth expansion at access network to satisfy high concurrency (> 10%) and HD VOD
- Build VOD server streaming infrastructure for streaming capacity
- Leverage existing digital STB at home and digital cable ready TV through tru2way
- Achieve high VOD usage with the large installed subscriber base
- No buffering at the client is required. This enables easy navigation and access of the content.

Over the Top features and capabilities:

- Access to a vast amount of Internet video and user generated video
- Feature rich navigation user interface and search capability using open Web technology
- Benefit from bandwidth competition at access network
- Benefit from advanced codec technology using PC or new CE appliance
- Potential to offer the same VOD service to any device with Internet access

There are several limitations of the Over the Top approach:

- Challenge to achieve high concurrency for HD video streaming
- Utilizes public Internet infrastructure that imposes quality of service constraints (e.g. congestion)
- Lack of end to end network resource management
- Inconsistent premium content offering due to lack of programming agreements with content providers
- Requires subscriber to purchase a separate CE appliance for viewing VOD on TV
- Does not yet have a large subscriber base. Fragmented market with too many players
- Long buffering time at the client may be required

A HYBRID VOD LIBRARY APPROACH

Today's Managed Network VOD system architectures helped network operators bring a compelling product to market. There are significant opportunities for the network operators to expand the current VOD architecture in order to support large VOD content libraries that provide an expansive amount of content including the Internet video. The other opportunity is to provide the VOD offering to devices other than STBs. Most of these are IP enabled devices such as PCs and portable CE devices.

In order to keep and expand the competitive advantages in providing VOD services, network operators are embracing the vision to give customers the ability to watch any movie, television show, user generated content or other video that a content provider wants to make available through Video On Demand. The service would offer the following:

- More HD content:
- More library VOD content
- Time shifted TV (StartOver™)
- Personalized, video rich navigation with better search
- Internet video content
- Cross platform video services (TV, PC, portable devices)
- Extensions for addressable advertising

In order to support these goals of any video content, at any time, to any device, a hybrid approach using VOD content libraries based on the Content Delivery Network (CDN) that will extend the existing Managed Network VOD infrastructure is proposed as shown in Figure 2.

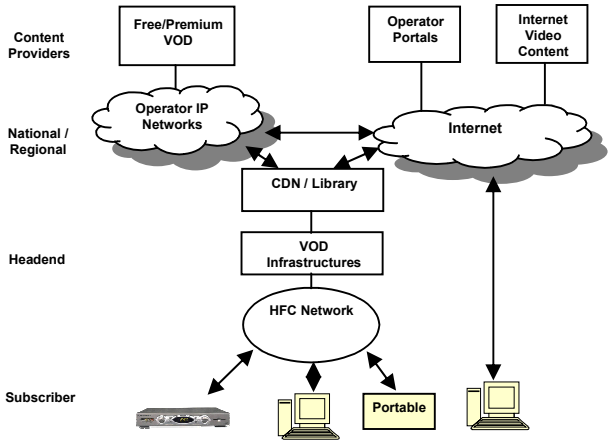


Figure 2 A Hybrid VOD Library Approach

In this approach, the existing Managed Network VOD infrastructure is expanded with the Content Delivery Network (CDN). The large content libraries within CDN are connected via operators' IP backbone and regional networks to both content providers and local headends.

The CDN content library will ingest and store content coming from traditional free or premium VOD sources. The CDN content library will also be able to ingest video content published from an Internet portal and external Internet video content including user generated content.

Typically, all the VOD content is stored in the CDN content library. Popular VOD content can be replicated and propagated ahead of time to the local VOD systems via CDN. Upon a subscriber's request for VOD content, the VOD system can start streaming if the content is already available at the local VOD system. The VOD system will pull the content from CDN content library if it is not available at the local VOD system. The content may be cached at the local VOD system for a period of time to serve other subscribers' requests.

The same CDN content library will serve multiple devices including STBs, PCs, and portable devices. Content may be transcoded to

multiple formats upon ingest to the CDN content library.

The proposed hybrid VOD library approach has several significant advantages, compared with the traditional Managed Network and the emerging Over the Top approaches, they include:

- Content:
  - Free or premium VOD content
  - Offer vast amount of movie and television shows, most in HD
  - Enable access to Internet video such as user generated video
- Infrastructure:
  - Utilize a managed IP backbone and regional networks for VOD content distribution with better quality of service
  - Manage bandwidth expansion at access network for high concurrency of HD VOD
  - Expand existing VOD content distribution, management, entitlement, and streaming platform
- Devices:
  - Leverage existing digital STBs and digital cable ready TVs through tru2way
  - Support PCs and other portable devices with Internet access

ARCHITECTURAL BUILDING BLOCKS

Main Challenges

With the advent of new technologies in IP networking and high performance storage and streaming servers, it becomes feasible to evolve the current VOD architecture to support large scale content libraries. However, there are several challenges that need to be addressed:

- Content Library (CDN)
  - Real time and non real time ingest
  - Performance and scalability of streaming from library storage
  - Asset propagation management
- Streaming Capacity and Bandwidth
  - Additional streaming server capacity is required
  - Increased edge QAM and unicast bandwidth is required
- VOD Navigation User Interface
- Internet Video Model
  - Pull versus push
  - Metadata format
  - Transcoding

Next Generation On Demand (NGOD)

Comcast has developed the Next Generation On Demand (NGOD) architecture framework to address both the feature expansion and capacity expansion of the VOD infrastructure to support multiple on demand services (see [1]).

The Next Generation On Demand architecture will continue to be used as a foundation to support the VOD library expansion based on the following principles:

- Open Interfaces: The reference architecture is developed with logical functional components. Standardized open interfaces between the different components in the architecture are also developed. This will enable multiple vendors to innovate in the areas of their expertise and allow seamless integration among the various components. For example, NGOD has specified key components and interfaces for the Session Manager, On Demand Resource Manager, Edge Resource Manager, and Edge QAM.

- Shared Resources for Multiple Services: Today's architectures are typically customized for a limited set of services.

Unfortunately, a significant re-engineering effort is required to support the addition of new services. The NGOD architecture enables the sharing of storage, streaming, network, and edge resources among multiple services. It is an extensible, on-demand platform that allows multiple services to share the same underlying infrastructure. It will create significant cost efficiencies and make it possible to quickly provide new services more quickly and easily.

- High Performance, Scalability, and Reliability: The NGOD architecture is designed to achieve high performance and scalability by allowing each component to be scaled and optimized independently. In addition, redundancy is built in various sources and components to provide high reliability.

Content Delivery Network (CDN)

One of the key enabling technologies for large VOD content libraries is the next generation Content Delivery Network (CDN). The next generation CDN will support VOD content and other media files. The national and regional IP networks connect multiple VOD content libraries in various locations such as national media center, regional centers, and local VOD systems. The CDN will enable operators to provide a large amount of VOD content cost effectively by serving them from the national and regional libraries instead of replicating all content to the local VOD systems. Intelligent caching can be adapted at the CDN and the local VOD system based on the popularity and actual usage of the content to further reduce the network bandwidth usage and enhance the overall performance.

The overall architectural building blocks for VOD libraries based on the NGOD architecture and its extension to next generation CDN are illustrated in Figure 3.

The architecture is partitioned functionally into a number of logical components. Each component is defined in such a way that the interchangeable module implementing the common interfaces can be introduced to work cooperatively with the rest of the system. It is possible that implementations may integrate several components into a single product or solution.

Each logical entity described in the reference architecture may represent one or many physical entities in an actual implementation.

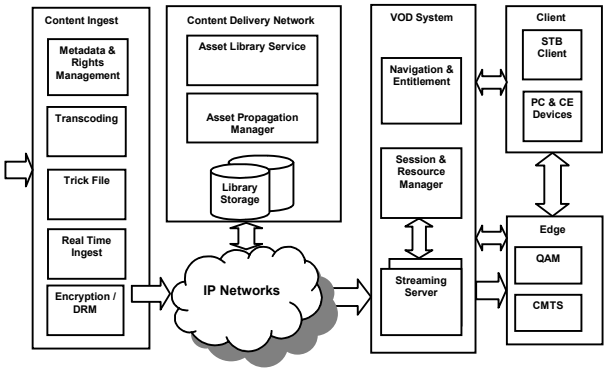


Figure 3 Key Architectural Building Blocks

The key architectural building blocks include:

Metadata & Rights Management – manage the asset metadata and rights for content from various content providers and aggregators, such as licensing windows.

Transcoding – transcode the content into various formats based on codec, resolution, and bitrate.

Trick File – generate fast forward and rewind trick files from the original content.

Real Time Ingest – ingest the real time content streams from content providers and aggregators.

Encryption / DRM – perform encryption and digital rights management packaging on the content.

Asset Library Service – maintain and update a directory of content locations in the CDN / libraries.

Asset Propagation Manager – manage content replication and movement among various library storage locations based on the external business rules and actual content usage.

Library Storage – provide persistent content storage or temporary content caching at various library locations.

IP Networks – provide national and regional IP networks that connect multiple library locations and local VOD systems.

Navigation & Entitlement – provide navigation and entitlement functions to content requested from client devices.

Session & Resource Manager – manage session life cycle and its associated resources for on demand video services requested by subscribers.

Streaming Server – store and outputs content and enables stream control.

Edge QAM – perform re-multiplexing and QAM modulation.

CMTS – Cable Modem Termination System for DOCSIS enabled devices.

STB Client – digital set-top box and its client software that communicate with the VOD system and typically receive video over MPEG-2 transport via an Edge QAM.

PC & CE Devices – PC and CE devices that communicate with the VOD system and typically receive video over IP via CMTS.

OTHER CONSIDERATIONS

Content Formats and Metadata

The content can be realized as various types of media files and real time streams targeting the end devices. Some of the most popular content formats are described in the following Table:

Table 1. List of Content Formats

Format	Resolution / Bit Rate (Typical)	End Device	Delivery	Use Cases
MPEG-2	SD / 3.75 Mbps HD / 15 Mbps	STB	QAM	VOD, StartOver
MPEG-4 / H.264	SD / 2 Mbps HD / 8 Mbps	STB / PC	QAM / IP	VOD, StartOver
VC-1	SD / 2 Mbps HD / 8 Mbps	STB / PC	QAM / IP	VOD, StartOver
Window Media Streaming	800 – 1500 kbps	PC / Portable	IP	Internet
Flash Streaming	400 – 1500 kbps	PC / Portable	IP	Internet
Audio (Dolby AC-3)	192 – 384 kbps	STB	QAM	Music Choice
Audio (Window Media, MP3)	64 – 128 kbps	PC / Portable	IP	Internet
Image (JPEG, PNG, Bitmap)	Various	PC / Portable / STB	IP	Graphics, Photo Sharing
Files	Various	PC / Portable / STB	IP	Application Download

In addition, Advertising content with similar content formats can be delivered via the CDN. SCTE 35 parsing and Ad splicing will be performed at the Streaming Server under the direction of the VOD BackOffice that interfaces with external Ad decision system.

Traditional VOD content is identified using the CableLabs Asset Distribution Interface (ADI) Provider ID and Asset ID. In addition, the metadata for the content is described in the CableLabs ADI 1.1 or ADI 2.0 standard. The content identifier and metadata structure need to be extended to support Internet video that uses Internet media publishing standard such as Really Simple Syndication (RSS).

Content Ingest

The CDN can support ingest of content files from traditional satellite based catchers as well as content files and real time streams via an IP backbone from VOD content providers or Internet Video providers. Content processing may be required upon the content ingest:

- Content may need to be transcoded to a different resolution, bitrate, and codec upon ingest to the CDN
- Trick files (Fast Forward, Rewind) may need to be created upon content ingest into the CDN
- Content with multiple formats (e.g. HD vs. SD) for the same title are treated as different content assets with different content metadata
- The metadata and rights management are performed on the content. This includes content life cycle management such as the licensing window
- Encryption and Digital Rights Management (DRM) are performed on the content upon ingest. It is also possible that tier based or session based encryption can be performed upon streaming

Asset Propagation Management

The CDN contains multiple Content Library nodes connected via national and regional IP networks. The Asset Propagation Manager (APM) is responsible to replicate and/or move the content through the storage nodes of the CDN dynamically based on the content popularity and usage.

The locations for all content within the CDN is maintained and updated by the Asset Library Service (ALS). Upon a session setup request from the subscriber, the VOD session and resource manager will interface with the Streaming Server for the selected content. If the content is already pre-positioned or cached at the Streaming Server, it will stream the content to the subscriber. If the content is not available at the Streaming Server, it will query the ALS for the locations of the requested content within the CDN in order to fetch the content from the content library and stream to the subscriber.

Content Streaming

When the Streaming Server fetches the content from Content Library and streams to the subscriber, it will use the Content Transfer Protocol from Content Library to Streaming Server. The Content Transfer Protocol may be extensions to one of the existing standard protocols such as:

- NFS
- CIFS
- FTP
- HTTP

Open standards, scalability, and performance are some of the key criteria for the selection and design of the Content Transfer Protocol.

Operations and Reporting

An operation model for system monitoring and management is required. Specifically, it will include the following aspects:

- Component level fault monitoring and management
- Content level status monitoring
- Network level monitoring

- Video quality monitoring

In addition, key reporting metrics need to be defined:

- Viewing patterns for content
- Network bandwidth usage
- Storage bandwidth usage
- Peak number of streams and concurrency rate
- Measurement of efficiency of caching algorithm

Scalability, Performance, and Reliability

Scalability, reliability, and performance requirements may include the following:

- Daily hours of content ingest
- Library storage sizing
- Target streaming capacity
- Jitter
- Distribution metrics (delay/latency)
  - Initial Request
  - Peak Utilization
- Caching versus network bandwidth tradeoff
-

VOD for Multiple Devices

It is highly desirable to share the large VOD libraries for multiple end devices such as PCs and other portable media players.

There are several aspects which should be considered when expanding the architecture to support any content to any device.

- Content format transcoding
- Session and resource signaling
- Digital Rights Management (DRM)
- Home networking
- Subscriber and device authentication
- Cross application platform

SUMMARY

This paper describes an architecture framework for large VOD libraries based on the Next Generation On Demand (NGOD) architecture and its extension to the Content Delivery Network (CDN). The architecture combines the benefits of Managed Network and Over the Top approaches. It will enable distribution of any content to any device, at any time. The architectural building blocks are presented and some of the key challenges and design considerations are discussed. On going work includes detailed architecture and interface specifications as well as performance and scalability analysis.

REFERENCES

[1] Evolution of Video On Demand Architectures, Weidong Mao & Kip Compton, May 2004, NCTA Technical Papers

[2] CableLabs Video On Demand Content Specification Version 1.1, August 31, 2006

[3] CableLabs Asset Distribution Interface Specification Version 1.1, May 5, 2006

[4] CableLabs ADI 2.0 Specification Asset Structure, January 5, 2007

[5] ISO/IEC 13818-6: MPEG-2 Digital Storage Media – Command and Control (DSM-CC)

[6] ISO/IEC 13818-1: MPEG-2 System

[7] ISO/IEC 13818-2: MPEG-2 Video

[8] IETF RFC 2326, Real Time Streaming Protocol (RTSP)

[9] ISO/IEC 14496-10, Advanced Video Coding

CABLE'S MOBILE FUTURE: WHICH TECHNOLOGY AND WHY?

Jay Bestermann
Director Product Development, ARRIS
jay.bestermann@arrisi.com

Tarun Chugh
Sr. Software Engineer, ARRIS
tarun.chugh@arrisi.com

Abstract

This paper examines the wireless market and network technology, WiMAX or Long Term Evolution (LTE), cable operators will likely deploy in the not too distant future. Key market trends such as wireless growth, competitive threats and incumbent carrier landline erosion are explored. Technologies are compared based on expected mobile device availability, access technology, core network architecture, and roaming capability. The value of converged services is identified as a cornerstone to cable operators' wireless strategy. Finally a network technology recommendation is made based on the previous market and technology analysis.

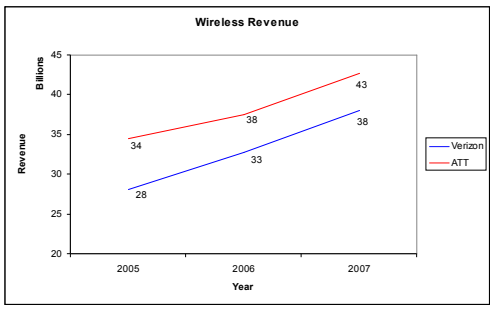
INTRODUCTION

Over the past 3 to 5 years North American cable operators have invested significant resources to add cellular service to their portfolio with varied results. The first major step was the announcement of the SPRINT Joint Venture in November of 2005. This was a very celebrated development, but it hasn't provided an effective Quad Play offering. The next step was when major cable operators in the US, via Spectrum Co, purchased a nationwide footprint of AWS spectrum in September of 2006. This was again a major development, but the companies have yet to officially announce plans for the use of this spectrum. More recently, several major MSOs have entered the January 2008 FCC 700Mhz auctions with Cox, Charter (via Paul Allen / Vulcan Ventures), and Bend Broadband winning additional spectrum. On March 26, 2008, the Wall Street Journal reported Comcast, Time Warner, Brighthouse and others are considering a WiMAX joint venture with SPRINT and Clearwire.

The previous events are noteworthy because they highlight a clear interest from the MSO community in bringing a credible wireless offering to market. This paper examines the prevailing market trends and concludes with a solution recommendation.

WIRELESS GROWTH

According to recent press, and the latest Quarterly reports from AT&T and Verizon, wireless subscriber growth is driving their overall company revenue growth. The graph below shows the revenue growth of the wireless organizations within Verizon and AT&T. Each of these corporations have experienced year over year growth in excess of 11.8% for the past 3 years.



Over the past few years revenue growth within wireless business units has been driven by voice, but as this market matures, revenue growth will be driven by value added data services such as SMS, email, and MMS. For instance, AT&T wireless has experienced exponential growth rates of mobile data usage. Usage of this service on their network has at least quadrupled for each of the last 4 years.

According to CTIA-The Wireless Association®, the average US cellular subscriber spends $50 per month and there were 243.4M wireless subscribers in June of 2007. This is a staggering $146B dollar annual revenue source for carriers. Comparing this to MSOs' most popular and highest ARPU service today, the National Cable Television Association (NCTA) reports US cable operators have 65.1M video households as of September 2007 with an approximate ARPU of $60 for video service (Comcast 2007). This represents a market of $46.9B. The point of this comparison is to illustrate, using current statistics, that offering a cellular service is a potential growth opportunity for MSOs. Cable Operators are already reaping the fruits of their landline efforts, but this opportunity will certainly diminish in the coming years. A viable and even obvious next step is to target a sliver of the wireless market representing a powerful new ARPU growth engine. This growth engine for MSOs will also serve to neutralize the predicted negative growth due to competition from the wide deployment of carrier-based services such as AT&T U-verse[SM] and Verizon FiOS[TM].

In recent years subscribers have been transitioning to a wireless only voice communication paradigm. This transition is fueled by the improved reliability and affordability of wireless communication and subscribers' passion for mobility. As shown in the graph below from OECD, the European average for wireless only households is 22% with Finland leading the charge at an incredible 54%.

Wireless Only Households	
Europe average	22.0%
Italy	38.0%
Austria	39.0%
Finland	54.0%
UK	13.0%
France	18.0%

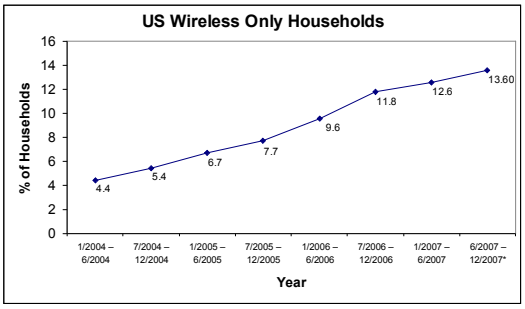
Source: OECD

In recent mobile history, Europe has been a leading indicator for wireless trends in the US for services such as SMS and data. Assuming this applies to users cutting the cord as well, the US is going to follow the European trend that is taking place today. To back this theory the table below from In-Stat[4] shows very strong interest in users migrating to a wireless only paradigm in the US.

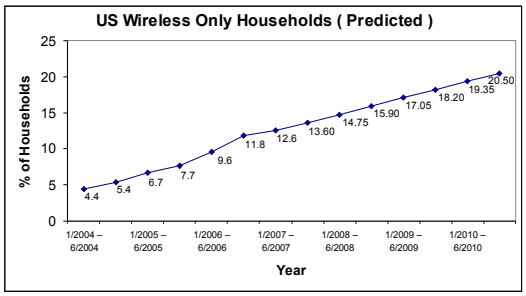
Interest in Going All Wireless	US	France	UK
Extremely/Very Interested	45.0%	44.3%	44.2%
Somewhat Interested	33.5%	31.9%	27.1%
Total WITH Some Interest	**78.5%**	**76.2%**	**71.3%**

Source: In-Stat, 7/07 n=2,049

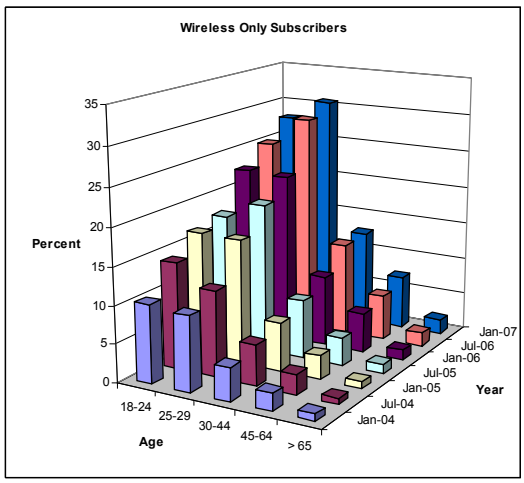
According to NCHS National Health Interview Survey, there were 13.6% wireless only households in the United States in 2007.



Assuming a continued average of 2.25% growth per year until 2010, the US will have approximately 20% wireless only households. This prediction will directly impact the wireline market going forward. Below is a chart of this prediction.

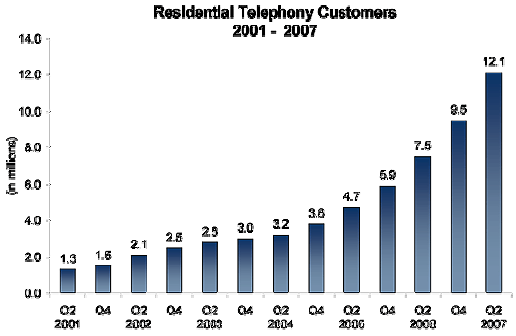


The predictions above are a conservative estimate when the following trend is considered. The younger a subscriber is, the more likely he or she is to be a wireless only consumer. The key message in the next graph is this trend will likely shift to the right as each demographic grows older versus remaining as an age-defined wireless only demographic. The younger generations will be more comfortable with mobile technology and thus more willing to rely on it as their sole communication device, but as they grow older their demand for mobile services will not diminish.

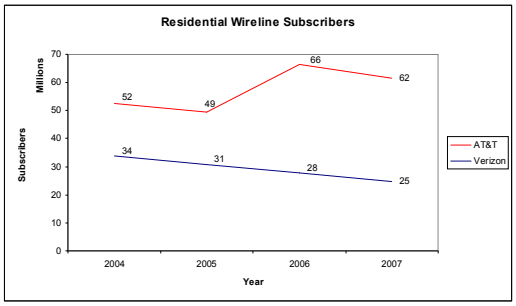


LANDLINE MARKET

In recent years cable operators have benefited significantly from strong subscriber landline growth as shown in the graph below from NCTA.

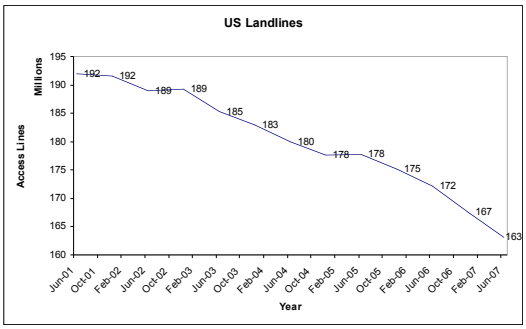


This growth has been the result of the significant landline erosion that has been a market force over the last few years in the traditional fixed line operator space. The traditional landline incumbents have taken the brunt of landline erosion as well as market competition from cable operators and pure play VoIP providers such as Vonage. The graph below shows the landlines lost by the traditional landline carriers over the past 4 years. One detail to notice in the graph is the only growth realized by either of the major carriers in the United States was via acquisition. In particular SBC acquired Bellsouth in 2005 which resulted in significant landline growth for the new corporation, but it is still very clear that the overall trend is landline erosion within the incumbent landline carriers. For instance, Verizon has lost 26.5% of its residential landline subscriber base over the analyzed time frame below.

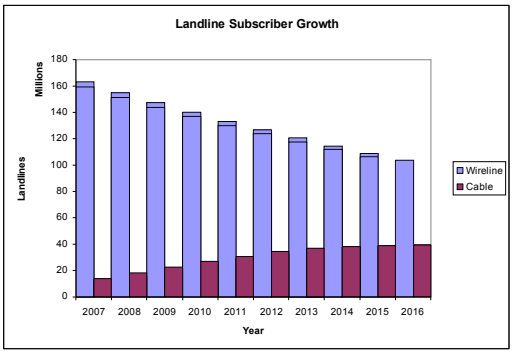


Despite cable operators' significant landline growth in the last couple of years the overriding industry trend is landline erosion as shown in the diagram below. As shown by March 2008 data from the FCC Wireline

Competition Bureau, during the past 5 years, the total landline market compressed by 15.1%. This doesn't highlight other overriding trends such as price erosion, but it certainly shouldn't be ignored.

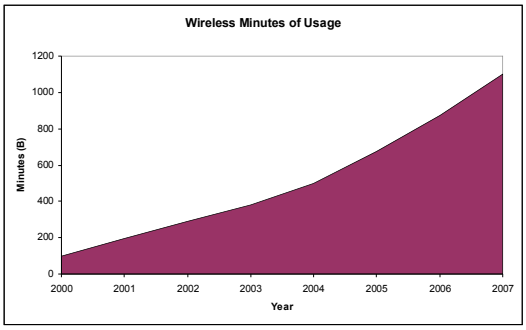


Using the data from the growth plot above and assuming continued -5% annual growth, the projected carrier landline market is shown in the next graph. On this same graph an annual growth rate of 15% for MSO landline growth is assumed for the next 2 years. After this period the growth rate reduces 2% per year until the MSOs reach 40% penetration of their 123M current homes passed. Based on the previous assumptions cable operators will likely experience flat landline growth in the 2012 timeframe.



The key takeaway from this analysis is as follows: Once MSO landline growth stops it is likely that the overriding trend of landline loss will then start to affect cable operators as it has affected the incumbents for the last few years.

These trends are being driven by the consumer's passion for wireless connectivity and continuous connectivity to friends, family and associates. Wireless Minutes of Use actually surpassed residential landline usage in 2003 and the population continues to become more and more reliant on wireless devices. The diagram below shows current and historical and wireless Minutes of Use (MOU). There is certainly a peak to this trend, but for the foreseeable future landline erosion and wireless MOU growth will continue.



To further accelerate the previously discussed trends, incumbent wireless operators are increasing the pressure on landline Minutes of Use and thus the profitability of landline operators' business. Verizon Wireless, AT&T, T-Mobile, and SPRINT all announced unlimited calling plans early in 2008. These unlimited plans will likely drive further growth of wireless MOUs and further commoditize landline voice.

T-Mobile is the most aggressive in its offering of the Hotspot@Home™ and Talk Forever™ services. The T-Mobile Hotspot@Home service is a $9.99 per month service add-on to a minimal wireless plan that provides unlimited voice and data using UMA enabled dual-mode handsets over Wi-Fi. The Talk Forever service is a UMA enabled Analog Terminal Adapter that offers a Vonage-like service to T-Mobile customers. This service is available in limited markets at the time of this writing, but is planned for a nationwide launch at $9.99 as an add-on to an existing wireless plan.

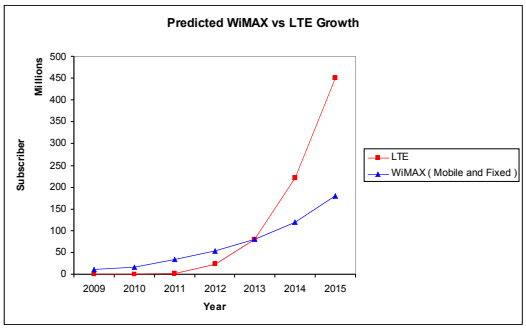
The goal of these new services is best stated by T-Mobile USA CEO, Robert Dotson in a June 27, 2007 release, "More people than ever are looking to drop their home landline phone and pocket the savings. However, they don't want to use all their wireless minutes talking from home. Our new service solves this dilemma once and for all. T-Mobile's HotSpot@Home is a first-of-its-kind service that helps people simplify their lives, save money, and enjoy great call quality on one device — their mobile phone — at home". Joe Sims, VP of T-Mobile Broadband Products was quoted by Wi-Fi Net News stating "T-Mobile is looking to address the remaining reasons people were reluctant to cut the cord."

The expectation is that unlimited calling plans will force the major wireless operators to offer solutions similar to the ones T-Mobile has in its portfolio for various reasons. With unlimited calling plans, subscribers will become more and more reliant on their mobile devices, thus demanding better in-home coverage and bandwidth. Assuming uptake is high, the unlimited services will require operators to add additional network capacity to keep customer satisfaction at high levels. Another key driver of these new converged services is the undeniable cost advantage of landline vs. mobile MOUs.

NETWORK TECHNOLOGY

The key consideration during the selection of a new mobile access technology is the future expectations of access device cost and availability. Both of these factors are typically tied to subscriber bases with the capability to drive the largest handset volumes. Today GSM and CDMA are the prevalent network technologies and handset costs and varieties follow the afore-mentioned trend. According to 3GAmericas.org, as of 3Q2007, there are 2.7B GSM subscribers in 220 countries resulting in 86% of the global wireless market share. In contrast the CDMA Development Group reports 431M CDMA subscribers in 97 countries resulting in approximately 13% of the global wireless market share in 4Q2007.

The above comparison can then be applied to the current handset market for each of these network technologies. The CDMA Development Group reports that 1,950 devices have been introduced in to the market including 509 1xEV-DO Rel. 0 and 48 1xEV-DO Rev A devices. This is a historical figure over the lifetime of CDMA 2000 technology. GSM Arena reports current GSM device availability at 1159 in March 2008. This shows that the current number of commercially available GSM handsets in the market today is over half of the lifetime total number of CDMA2000 devices. This comparison is very useful when thinking about the next technology choices that MSOs will make or are already making. 3GPP Long Term Evolution (LTE) and WiMAX handset availability, variety, and cost will clearly be driven by the addressable market size. Below is a subscriber growth prediction using data from Senza-Fili and Analysys. The Senza-Fili data was extended to project 2014 and 2015 for this paper. The current projection is LTE will quickly overtake WiMAX (including Total Fixed and Mobile) deployments in the beginning of 2013.



Technology decisions will also affect roaming relationships that are possible between networks and operators. LTE has a clear advantage over WiMAX in this area because of the clear evolution path as technologies evolve.

3GPP GSM networks are evolving from GPRS to EDGE to HSPA today and the next evolution to LTE is an extension of that experience. The primary enabler of this evolution is a healthy handset ecosystem that will build handsets that support the upcoming network deployments. A handset that supports EDGE, HSPA, and LTE will be able to get network access virtually anywhere in the world in the coming years. This is a critical capability that must not be ignored as operators build out "greenfield" networks that have limited coverage areas. Technology inclusion and thus roaming relationships for WiMAX devices are very unclear. SPRINT and Clearwire have announced WiMAX network(s) in the US but primarily for laptops and fixed wireless access with a transition to mobile devices. The next step is offering mobile WiMAX in portable devices. Given the limited WiMAX coverage areas, a combination WiMAX/ CDMA2000 handset will likely be required. This combination will have far less volume and a much smaller ecosystem than the 3GPP driven LTE standard. This is driven by the uncertainty or non-linearity in the technology evolution path. Because SPRINT has an existing CDMA2000 network they will require a CDMA2000 /WiMAX handset solution until a nationwide WiMAX network is in place. This is in direct contrast to the recent decision that Verizon has made to utilize LTE technology in its next generation network upgrade. This decision will divide the US CDMA2000 subscriber base between WiMAX/CDMA2000 and LTE/CDMA2000 in the United States. There are many other potential handset technology combinations such as WiMAX/EDGE/HSPA that could be deployed by existing 3GPP or greenfield operators. This decision again would result in a divided market in comparison to the clear 3GPP evolution path and global scale.

The technology decision made by MSOs will also affect the opportunity for inter-carrier roaming. For example, if WiMAX devices are deployed by SPRINT they will be the only established US operator deploying a WiMAX/CDMA2000 technology combination. As previously mentioned, Verizon Wireless has made public commitments to deploying a LTE network in its next round of technology upgrades. This implies SPRINT will not have an established nationwide WiMAX roaming carrier with a large network footprint or the subscriber base to support a nationwide network deployment. SPRINT will have to rely primarily on its network deployments for WiMAX coverage or utilize CDMA2000 based roaming. This also implies SPRINT will receive little WiMAX based roaming revenue from other US based operators except Clearwire. On the other hand, Verizon's LTE approach will immediately expand its roaming partner ecosystem in North America and abroad. Verizon Wireless' 50% owner, Vodaphone, has announced intentions of deploying LTE as have AT&T, China Mobile and NTT DoCoMo. It is likely that many of the over 200 3GPP operators will follow this clear evolution trend. This roaming ecosystem will provide Verizon access to a worldwide roaming based revenue engine as well as offer its customers a much better service availability. Verizon isn't a true greenfield operator as most cable operators are today, but its decision to deploy LTE technology makes it very similar because of the technology discontinuity.

Technically speaking, WiMAX and 3GPP's LTE are very comparable access technologies. Some of the access technology highlights are as follows: OFDM-based, similar modulation techniques, theoretical throughput and capacity. One notable exception to this rule is that WiMAX is typically a TDD solution whereas LTE is typically a FDD solution. WiMAX does have a FDD profile, but this hasn't been deployed. The data in the table below gives a more thorough comparison.

	3GPP LTE	WIMAX
Bit-rate/site(DL)	100Mbps(MIMO 2TX, 2 RX)	75Mbps(MIMO 2TX, 2RX)
Bit-rate/site(UL)	50Mbps	25Mbps
Base Standard	E-UTRAN	IEEE 802.16e
Duplex Method	FDD	TDD(FDD optional)
Downlink	OFDMA	OFDMA
Uplink Multiple Access	SC-FDMA	OFDMA
Channel BW	1.25 - 20Mhz	Scalable:4.375, 5,7,8.75,10 Mhz
Modulation DL	QPSK/16QAM/64QAM	QPSK/16QAM/64QAM
Modulation UL	QPSK/16QAM/(64QAM opt)	QPSK/16QAM
Cell Radius	5km	2-7km
Spectral Efficiency	5[bits/sec/Hz]	3.75[bits/sec/Hz]
Cell Capacity	>200 users @ 5Mhz >400 users for larger BW	100-200 users

From a core network architecture perspective both WiMAX and LTE have been designed to be very flat IP-based solutions that interconnect with an IP Multimedia Subsystem, but there is a key difference. LTE has been architected from the beginning to support seamless handover and global roaming to LTE/2G/3G networks. WiMAX mobility is based on mobile IP and hasn't addressed inter-Radio Access Technology handover or global roaming scenarios.

The current standardization, deployment and mass market timelines show WiMAX reaching market in 2007, which has already happened. WiMAX mass market adoption is to begin in 2009. LTE is trailing WiMAX with expected deployments to start in 2010 with mass market adoption in 2012.

CONVERGENCE OPPORTUNITY

As cable operators become more serious about bringing a wireless offering to market, they must consider the clear advantage of adding a flavor of either device or service convergence to their solution. They must leverage the next generation access and transport networks that have been put in place to serve their other business needs. This investment can be further exploited with an offering that addresses both device and service layer convergence. This need becomes obvious when examining the calling patterns of wireless subscribers. In the US the average consumer makes at least 40% of calls from an indoor home or work environment. This doesn't even consider continuous data usage for services such as Email, IM/presence, Web browsing, etc. This single statistic puts the cable operator with a wireless offering in a very good position to offer compelling new service bundles and capabilities required to compete in a hyper-competitive marketplace. In 2007 there were 123 Million households passed by cable operator networks. If we assume an average of 2 persons per household, MSOs have the ability to offer converged services to 246 Million subscribers. Compelling new services such as these will be required to create an impetus for change and drive the current 50% service penetration enjoyed today even higher.

Device convergence is the concept of embedding both Wide Area Network and Local Area Network technology in mobile devices. The most common example today is Wi-Fi+GSM in a single handset, but in future deployments this will likely become Wi-Fi+WiMAX+CDMA2000 or Wi-Fi+LTE+GSM. Femtocells can also be viewed as device convergence at a slightly different layer. The femtocell combines traditional Wide Area Network technology such as CDMA2000 with high speed local area network backhaul. The purpose of all Device Convergence is to make use of high bandwidth, low cost local area networks when they are available, but make the user experience very simple, cost effective, and truly next generation.

Service Convergence is the idea of blending physically independent device types on one or multiple carrier networks utilizing intelligent core networks. This blending occurs through simultaneous ringing of the independent device types and allows the call to be moved between different terminals very easily. For example, imagine walking into your home on an active AT&T cell call and with a single key press moving that active call from your mobile device to your cable operator managed landline. This would allow the user to select the current device of convenience, best performance, least

cost, or comfort as they prefer. This type of blending would offer a compelling reason for consumers to keep a landline in the home and thus reduce the current trend of landline erosion.

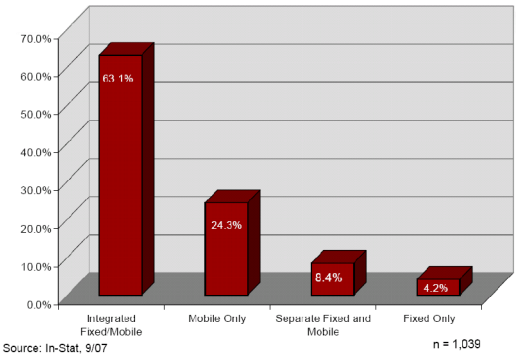
Demand for these services has been shown by recent studies. Of those surveyed, 49% of subscribers rate coverage at home and 42% rate voice plan pricing as the most important factors when selecting a cellular carrier. These are the top two factors which drive a consumer's wireless decision. This clearly shows that a device convergence strategy which attacks both of these issues head on, is a win-win offering for the consumer and operator.

One of the key issues in deciding to deploy a wireless network is developing an offering that is going to be compelling enough to entice a subscriber to switch providers. Based on a recent report, a converged solution is extremely or somewhat interesting to 92.7% of subscribers.

How interested would you be in moving your existing service bundle to another service provider in order to purchase an integrated fixed/mobile voice service?	
Extremely/Very Interested	50.4%
Somewhat Interested	42.3%
Not Very/Not At All Interested	7.3%

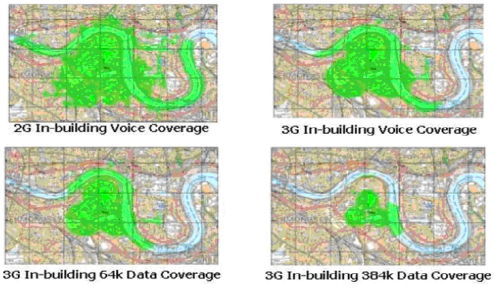
Source: In-Stat, 8/07 n = 1,039

From this same report subscribers are demanding a converged voice experience. Of particular interest to the cable industry is the perception that a fixed-only life is very undesirable to the study group. This indicates a user faced with a forced selection would prefer a mobile-only life rather than a tethered one, but the most desirable solution by far is a mixed offering.



DOCSIS 3.0 has shown the ability to improve download speeds at least 4X (40 Mbps to 160 Mbps) over the current DOCSIS 1.x/2.0 technology that is deployed in today's networks. This is proving to be a competitive advantage because of the relatively low capital investment required to deliver this level of bandwidth to a consumer's door. In the not too distant future converged devices play a significant role to further leverage this investment as part of the cable operator's wireless strategy. Using their HSD infrastructure cable operators can vastly improve the user experience and decrease the cost per bit significantly.

A device convergence strategy addresses an obvious problem MSOs will encounter with indoor coverage as they deploy their new networks. As an illustration of the issue the diagrams below show indoor penetration of 2G Voice, 3G Voice, 3G 64K data, and 3G 384K data by cellular base stations.



Notice the higher the data rate the less effective in-home coverage becomes. This is a well known characteristic of higher bandwidth

and/or higher frequency channels driven by the laws of physics. It may not be intuitive but this phenomenon will also affect outdoor users in terms of effective available bandwidth. The indoor users will consume more of the cell site's resources because adaptive modulation techniques are utilized in mobile networks. The indoor user's device will be using lower modulation schemes and hence require more of the available bandwidth than outdoor users to transmit the same amount of data. These facts show a converged solution will alleviate pressure on the outside network, therefore reducing the network capital investment required to offer best in class coverage and bandwidth to the end user.

CONCLUSIONS

From the market analysis above it is clear that cable operators must enter the wireless space to, at a minimum, prevent contraction of their existing business. Furthermore a much larger opportunity exists to accelerate revenue growth by entering and capturing a small portion of the wireless subscriber base that is searching for a true Quad-Play service offering. Cable operators have a lead in bandwidth to the home and a best in class content offering today, but the carrier community is attacking this safe haven with relentless vigor. With the addition of wireless it is quite clear that the MSOs will continue to be a formidable competitor.

Analyzing market data and comparing technology capabilities leads to a recommendation of LTE as the network technology best fit for MSO deployment. It is evident that the global scale of LTE insures competition as well as innovation in the handset and network equipment space via a vast vendor ecosystem. Its technical capability is second to none with the key element being handover and global roaming capability with not only LTE but legacy 3GPP and even CDMA EV-DO Rev-A networks. This is paramount when it comes to customer satisfaction and revenue generation capability. One potential path to LTE to satisfy operators need for accelerated deployment is a staged approach of deploying 3G HSPA today and upgrading that network to LTE in the coming years. This strategy would allow for early market entry with field proven technology that is comparable in performance to today's WiMAX solutions.

Finally, a convergence strategy is a differentiator that will help insure a successful market entry for MSOs by significantly reducing capital expenditure and providing subscribers with yet another reason to migrate to cable operators' networks.

References:

1. "Statistical Trends in Telephony 2007 Report", FCC
2. "A Comparison of Two Fourth Generation Technologies: WiMAX and 3GPP LTE", http://www.comsysmobile.com/pdf/LTEvsWiMax.pdf
3. "2008 Corporate Brochure", GSMA
4. "HSPA and mobile WiMAX for Mobile Broadband Wireless Access", gsmworld
5. "EDGE,HSPA and LTE The Mobile Broadband Advantage", http://www.rysavy.com/Articles/2007_09_Rysavy_3GAmericas.pdf
6. CTIA mid year surver 2007, CTIA
7."Wireless Quick Facts", CTIA
8."Industry Statistics", NCTA
9."Verizon Communication Inc., Form 10-K"
10."AT&T Inc. Form 10-K"
11."Comcast Corporation. Form 10-K"
12."Time Warner Cable Inc. Form 10-K"
13."Wireless Substitution: Early Release of Estimates From the National Health Interview Survey,January-June 2007, NCHS
14."T-Mobile's VoIP Home Service: Goodbye to the PSTN", In-stat
15."3G Femtocell, Fixed-Mobile Convergence", Light Reading Femtocell Webinar.
16. "Global Mobile Broadband: Market potential for 3G LTE", Analysys

COST EFFECTIVE WATERMARKING IN THE SET TOP BOX

Joseph Oren
Cinea Inc. a Dolby Company

Abstract

Providing Cable consumers with premium (e.g. HD or early window) content via Video on Demand services is projected to become a key revenue source for system operators. Yet the Hollywood studios insist that before this content will be made available, enhanced content protection technologies must be deployed within the content distribution infrastructure [i] *. Specifically, forensic watermarking, defined as the binding of unique traceable information to the video streams, is increasingly mentioned as an essential content protection layer, one that complements existing conditional access and digital rights management solutions.* [ii] *This paper describes how this new business requirement can be technically and economically fulfilled by watermarking technologies now reaching the market. Our focus will be on watermarking technology implemented in the consumer's equipment, commonly called the Set Top Box (STB).*

DISCLAIMER

The author of this paper, Joseph Oren, is employed by Cinea Inc., a Dolby company. Cinea offers commercial products utilizing certain technologies described herein.

INTRODUCTION

DRM and CA technologies have made great strides toward system recognition of the rules agreed upon by content owners and consumers. The available mechanisms to enforce those rules, termed content protection, is, however, limited to encryption during transmission and storage. Once the content is rendered in a consumable form, its digital and analog representations become subject to copying and subsequent unauthorized redistribution (piracy). Figure 1 shows a simplified receiving device, with vulnerabilities identified. A real-world home network may spread these functions over several devices, each with analogous vulnerabilities.

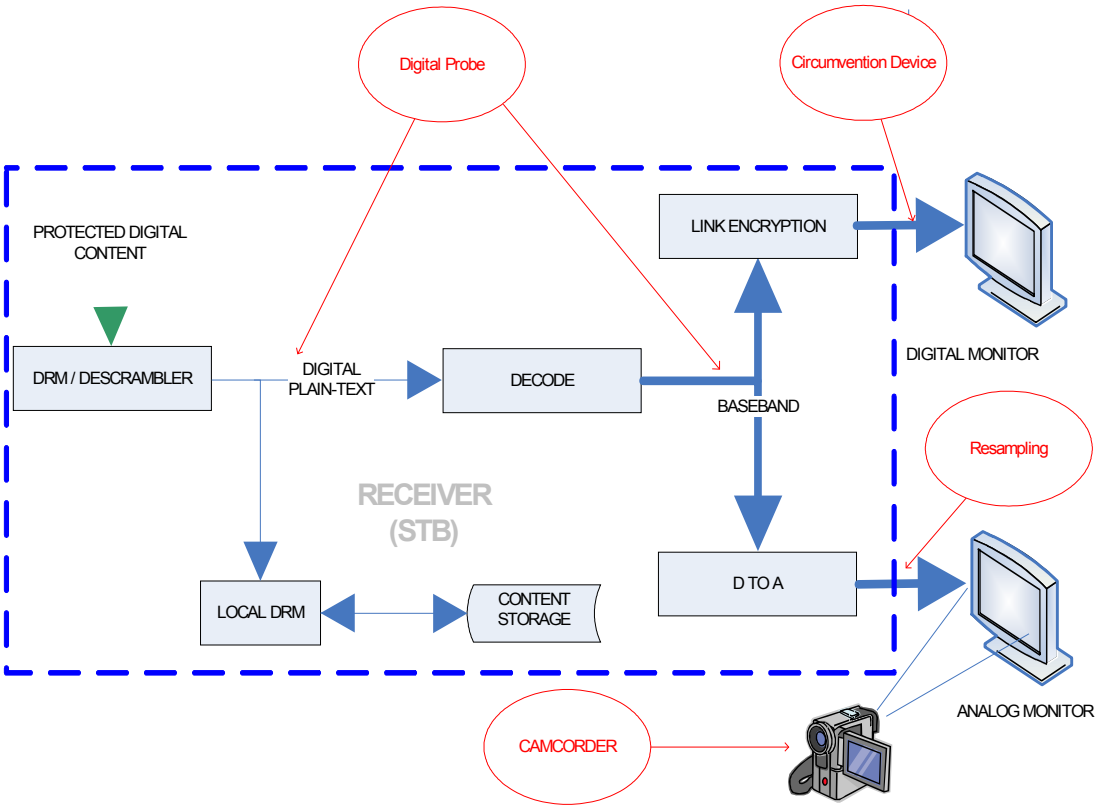


Figure 1 – Receiver with vulnerabilities identified

These vulnerabilities have fed the interest in forensic watermarking technology, whereby each instance of a content item is individuated by information to facilitate the tracing of the content back to its last legitimate holder. Tracing produces valuable evidence in identifying copyright violators. Further, since forensic watermarking is an investigative tool, as opposed to a control tool, it offers the potential to obviate some of the more complex and consumer hostile aspects of strong DRMs. While the DRM acts to constrain the user, actually challenging him to circumvent the technology, forensic watermarking deters piracy by introducing risk of exposure. In the case of large scale re-distribution, forensic watermarking facilitates identification of the point of compromise. With both DRM and watermarking available, a more balanced and appealing approach to content protection becomes possible.

Figure 2 – Forensic watermarking identifies the source of unauthorized distribution

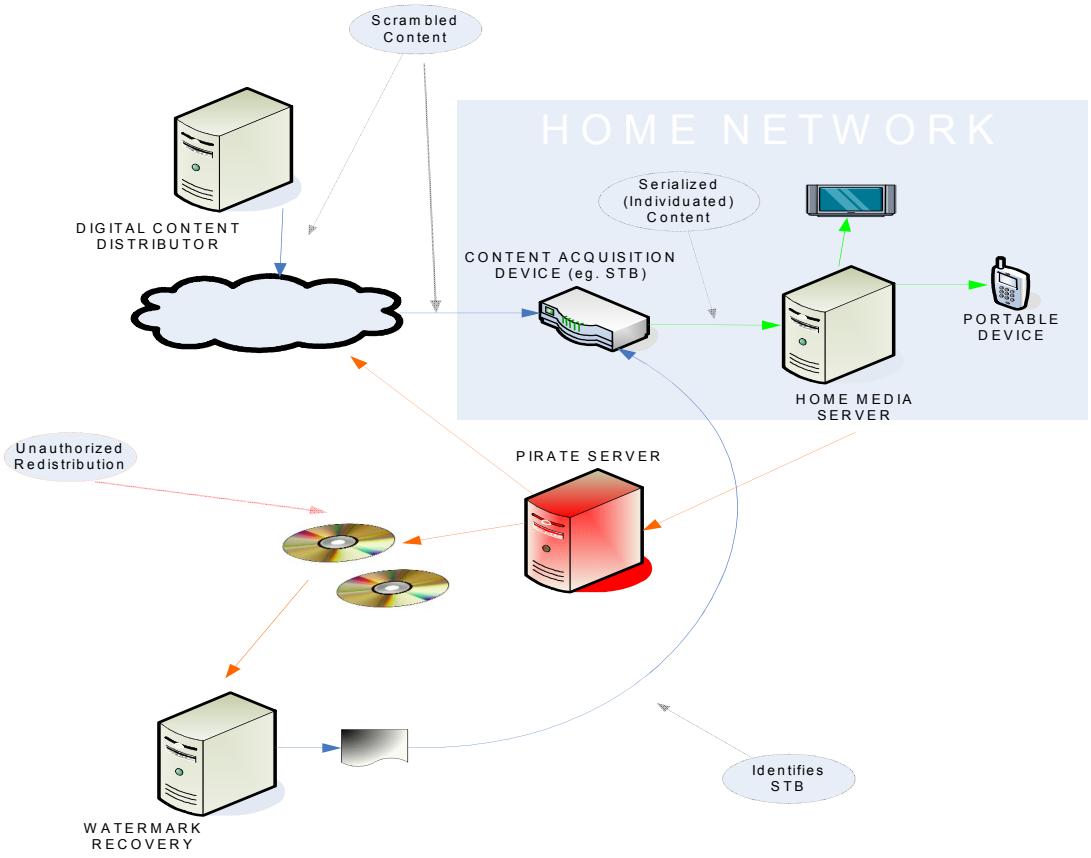


Figure 2 illustrates how forensic watermarking is used to investigate media piracy. In this example, the STB binds an identifying watermark to the content it acquires from the network. The watermark does not interfere with the consumer's enjoyment of the content. But, if the consumer chooses to circumvent the DRM and distribute unauthorized copies, the watermark can be used to determine his identity. Awareness, on the part of the consumer, of the possibility of exposure serves to discourage careless redistribution of the content.

Also called media serialization or content tracing, this paradigm is analogous to the use of serial numbers to track physical machinery or other valuable products. Watermarking technology provides a means of embedding such information into content, through subtle alterations to the content itself.

It is important to differentiate watermarking from simply appending identifying metadata to the content. Identifying metadata can be transparently excised from the content, while erasing a watermark requires specific manipulation of the content itself. Watermarking does not impair the content, but removal of a forensic watermark without impairing content quality is, by design, a very difficult task. We will examine the requirements for an effective forensic watermarking implementation and how those requirements can be addressed within the constraints of the STB.

WATERMARKING CONCEPTS AND TERMINOLOGY

In the broadest sense, forensic watermarking is a steganographic[iii] technique, one that embeds data into an instance of a cover work, in such a way that the data can subsequently be read (recovered) from copies of the watermarked cover work. The cover work may be any communication medium, but we will focus on digital video entertainment content. These principles may, however, be adapted for other media, such as the audio channels.

The process of binding the watermark to a content item is termed *watermark embedding*, and the additional data, in its embedded form, is the watermark itself. The process of reading the watermark from a copy of the content is termed *watermark recovery*.

In a data communications model, the watermark information is data to be communicated, and the cover work is a carrier signal. Indeed, the cover work carrying the watermark is often identified as the *host signal*. In the communications channel, the perceptible features of the content constitute noise that interferes with the watermark's information signal. It is important to recognize that watermark itself is embodied in changes to the cover work features, as opposed to just being ancillary data. Any faithful reproduction of the content (the carrier) will also carry the watermark data. The watermark signal may thus be viewed as modulating a noisy carrier signal, and thus becomes part of the cover work itself.

There are numerous watermarking technologies, with varying degrees of suitability for specific applications. The attributes commonly used to characterize a watermarking technology are as follows:

Perceptibility

Watermarks may either be apparent to the viewer, when the content is rendered, or disguised in such a way that the viewer is unlikely to notice the presence of the watermark. Perceptible watermarks are commonly used to proclaim ownership, exemplified by the visible logo appearing in many network broadcasts. In general, watermarks fall along a continuum of perceptibility, according to the needs of the users and the capabilities of the technology. The field of steganography, the technology of hiding messages in content such that the casual observer is unaware of the message's existence, includes imperceptible watermarking.

Readable vs. Detectable Watermarks

A watermark may carry only a single bit of information, that is, it is significant only in its presence or absence. Such watermarks are classified as detectable. A readable watermark, on the other hand, contains a more complex message, typically many bits of information. Mathematically, a readable watermark with N bits of information could be conceptualized as having been chosen from a set of 2^N detectable watermarks. For a message of useful length, the number of marks in such a set becomes unmanageable, so a practical readable watermark implementation must include a means of decomposing the watermark to reconstruct the message from independent parts.

Forensic watermarks must carry a message: a readable watermark, or a series of detectable watermarks is required to identify the particular source of the content instance. If detectable watermarks are used, the message is treated as a series of independent parts, each of which is represented by a single detectable watermark.

Bandwidth

Bandwidth refers to the of data carrying capacity of the watermark, in proportion to the amount of content carrying the watermark. For multi-media content, bandwidth is commonly expressed in terms of message bits per second. In interpreting bandwidth metrics, however, it is important to distinguish between the original message and an encoded message. Forensic watermarking implementers may apply multiple layers of error control coding (ECC) to the message, to compensate for the "noise" in the channel. Such coding can expand the message several fold, thereby reducing the effective bandwidth of the watermarking technique by the same factor. It is also common to embed several copies of the message into the content. For a robust implementation, the bandwidth required is many times that which is implied by the message length alone.

Robustness

Sometimes termed "survivability", robustness is the degree to which the watermark can withstand the various transformations the content may undergo before reaching the recovery process. An effective forensic watermark can tolerate operations such as rescaling, resampling of analog signals, recompression, cropping, rotation, resolution changes, deinterlacing, gamma changes, and temporal averaging, all of which may occur in the course of pirating the content. Additionally, a pirate may undertake targeted attacks to directly suppress the watermark by filtering, noise addition, collusion or other video processing techniques.

Although no watermarking technique is unconditionally robust, an effective technique requires the adversary either to apply an unreasonable amount of effort to suppress the watermark, or to unacceptably impair the content in the process. In signal processing terms, robustness tends to increase with the amplitude of the watermark signal. Paradoxically, if the signal intensity level reaches the threshold of perceptibility, its nature and location become apparent to the attacker, thereby facilitating the attack. Consequently, the watermark intensity must be carefully calibrated to achieve the required level of robustness.

As mentioned previously, error control coding is an important contributor to robustness. Alterations to the content may erase or distort significant portions of the watermark signal. Effective recovery must include mechanisms to compensate for missing or erroneous signal elements. The watermark system communicates over an extremely noisy channel, requiring aggressive error control.

FORENSIC WATERMARKING REQUIREMENTS

The primary requirement of a Forensic Watermarking application is the placement of the watermark embedder at a point in the distribution network where the legitimate recipient of the content instance is known. In uni-cast or download distribution models, the content instance can be watermarked as it is transmitted to the consumer. The consumer thus receives a unique copy of the content, individuated by the watermark that identifies that consumer's identity, account, or purchase transaction. Any reproduction of the content instance can thereby be traced to the consumer when the Forensic Watermark is recovered.

In broadcast or multi-cast systems, however, each consumer receives an identical copy of the content. It is thus only possible to individuate the content instance in the consumer's content receiving device. Consequently, these distribution models require authentication of the receiver, and sufficient security in the receiver

to ensure that the watermark is correctly embedded.

Figure 3 – Forensic watermarking in a broadcast distribution model

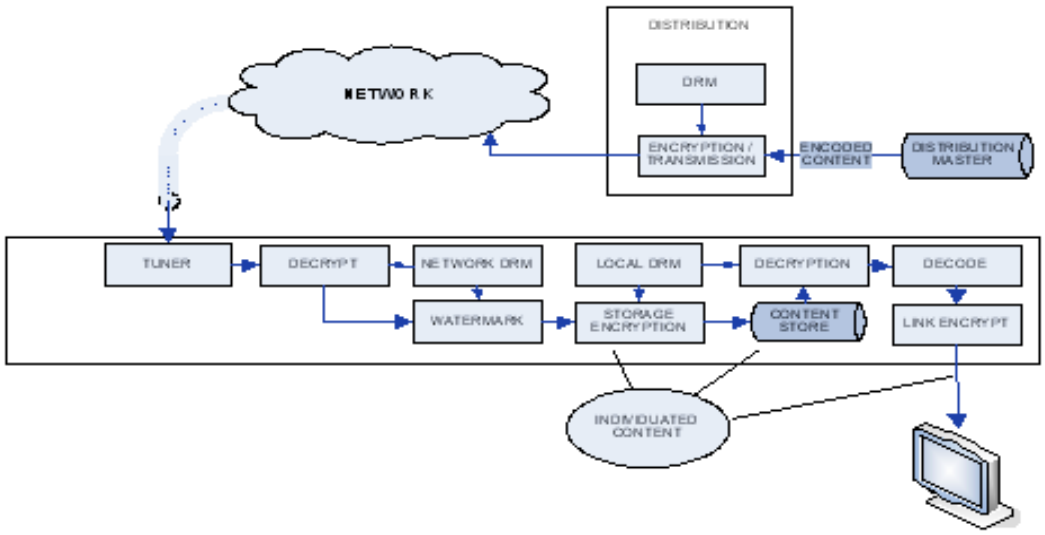


Figure 3 illustrates forensic watermarking in a broadcast distribution model. A watermark is applied following decryption, under the control of the DRM. The content made available to subsequent processes will have been individuated by the forensic watermark.

Watermark Integrity

To prevent compromise, and thus obtain the maximum benefit from forensic watermarking, the watermark should be embedded immediately following decryption. The physical security envelope in the device must enclose all processing inclusive of the DRM and the watermark embedding to prevent interception prior to watermarking or other circumvention. Any accessible data paths carrying unprotected content invite interception and can be targeted by pirates. Once the forensic watermark is embedded, however, the content becomes traceable. Traceable content is less attractive to the pirate, due to the increased risk of exposure, and, consequently, is somewhat less demanding of physical protection. Watermarking early in the content path also ensures that all outputs of the device are protected.

Embedding Performance

A closely related requirement is performance. The STB platform typically lacks substantial spare processing and memory resources. Watermarking in a receiving device must take place at rendering speed for real-time streaming content. Devices supporting background downloads may require watermarking at network speeds exceeding real-time. More sophisticated home network devices may require simultaneous watermarking of multiple content streams. Thus the watermark embedder must only minimally impact the STB computational resources.

Imperceptibility

As noted above forensic watermarks must be imperceptible. The system objective is to preserve the value of the content, so significant quality impacts are unacceptable. Imperceptibility is particularly important for

high definition content, which the consumer expects to be of the highest quality.

Robustness

Robustness is, of course, critical in forensic watermarking. The system is only effective in exposing pirates to the extent that the watermark information can be extracted from the unauthorized copy of the content. Pirated content is often degraded in the capture of the initial copy, as well as trans-coded [iv] for the pirate's distribution channel. The initial capture technique may range from a perfect digital copy, to resampling of analog signals, or even a camcorder directed at a rendered image. Unless the pirate captures a compressed digital signal, recompression - possibly accompanied by cropping, frame rate changes, (de)interlace, and/or resolution changes – will be necessary to re-distribute the content. And finally, the pirate may attack the watermark by injecting noise, dropping frames, filtering, or collusion [v]. It should be made difficult for the pirate to verify that s/he has successfully removed the watermark.

Cost Effectiveness

Economical implementation is of paramount importance in the consumer domain. Security features are of minimal apparent benefit to the consumer, so it is generally not possible to recover a significant cost increment for the material and licensing cost of the watermark embedder in each STB.

Renewability

Another requirement is that of renewability. Content pirates have unfailingly adapted to new media security technology. Watermarking will not be spared. As watermarking technology is deployed, adversaries will build tools to suppress the watermark signal. As such tools are perfected and become widely available, the targeted watermark technology will be rendered ineffective. Renewing the watermark system forces the pirate to analyze a new technique and adapt his countermeasures. Thus the ability to renew watermarking techniques, by varying the watermark signal, is a hallmark of an effective system.

Consistency

Uniform quality is important in an entertainment offering. Similarly, uniform robustness is important in a security system. Both reputation for quality and content security are only as strong as the system's weakest links. Similarly, when an unauthorized content instance is discovered, recovery requires knowledge of the technology used to embed the watermark. If the content has been marked inconsistently, it becomes more difficult to effect recovery, and, if no watermark is detected, very difficult to determine which watermarking technology has failed. An ideal watermark system deployment should, therefore, include a mechanism to ensure that all instances of a given content title are watermarked in a consistent manner.

Flexibility

Analogous to renewability, flexibility describes the ease of adapting watermarking to the needs presented by specific content items. The content universe features broad ranges of exposures to piracy, as well as sensitivity to quality. Content providers are likely to prefer watermark perceptibility-robustness tradeoffs that differ from one content item to another. Ideally, a watermark system should provide a means of control, to conform to the content provider's preferences and policies.

For example, Theatrical content may require very low watermark perceptibility with a corresponding decrease in robustness. Alternatively for the purposes of identifying

service theft, a higher degree of watermark perceptibility may be tolerated in order to achieve an increase in robustness.

Bandwidth

Forensic watermarking makes only modest bandwidth demands: DCI requires only 35 message bits in each 5 minute segment of a motion picture (~.117 bit/sec).[vi] As mentioned above, allowance for error control coding increases the raw bandwidth requirement.

FORENSIC WATERMARKING ENGINEERING CONSIDERATIONS

An effective watermarking system for forensic watermarking (or any watermarking application for that matter) must perform three basic functions. First, it must decide where in the content to place the watermarks. Secondly, it must generate the watermark signal used to modify content features such that the signal can be detected and recovered from a copy of the content. Thirdly, it must convey information in the watermark signal. The choice of a method to perform these functions greatly impacts the performance and cost of the system.

Meeting the application requirements discussed in the previous section, some in direct opposition to one another, is a non-trivial undertaking. It is illuminating to examine the major issues individually:

Perceptibility vs. Robustness

Both perceptibility and robustness are directly related to watermark signal strength. As the signal amplitude increases, other factors held constant, the watermarks become both more robust and more perceptible. A desired level of robustness can thus be achieved by increasing the signal strength, at cost of quality. Conversely, decreasing signal strength to the point of watermark imperceptibility can impact robustness.

Informed Embedding

Certain watermarking techniques favorably shift the perceptibility-robustness tradeoff. Watermark placement and composition can be optimized to take advantage of host signal (content) characteristics[vii]. Numerous studies of human perception have determined that sensitivity to a particular sensory input varies according to context (i.e. background). In the watermarking paradigm, the watermark is the sensory input that should be disguised, and the background context is the content itself.

A particular watermark signal will, thusly, be more or less likely to be perceived over various backgrounds. The effectiveness of a given background in disguising a feature is called its "masking" property. Masking is a function of the characteristics of both the background host signal and the disguised feature. It is thus possible to reduce watermark perceptibility by choosing watermark signal characteristics and placements that leverage the masking properties of the host signal.

Exploiting the host signal masking properties accommodates more watermark signal energy at a given degree of perceptibility, thereby improving the perceptibility-robustness tradeoff. Robustness can also be enhanced by choosing watermark characteristics and placements that optimize recoverability. Watermark robustness depends on the watermark and background image characteristics, as well as the recovery technique being used. Recoverability analysis evaluates interference between the background image and the watermark signal. The technique is analogous to "dirty paper coding" where the signal is positioned to sidestep interference. Optimal watermark composition is often a tradeoff between perceptibility and recoverability, as an image area with a high

level of masking energy may also interfere with the watermark.

Watermark embedding that conforms to the content background characteristics is called "informed embedding"[viii]. Properly employed, informed embedding significantly and favorably shifts the perceptibility-robustness tradeoff. This advantage comes at significant computational cost, however. Analysis of the masking and recoverability properties of motion video signals requires complex algorithms to be applied to several successive frames. The task is particularly challenging at the data rates required to support high definition content in real time.

Sequencing processors or programmable gate arrays capable of this task can add significant per unit costs. ASICs are an option, but only at high volumes, and are difficult to renew.

Receiver Watermarking Security

For broadcast or multi-cast distribution models, as discussed above, forensic watermarks must be applied at the receiver. An optimal security architecture for forensic watermarking in the receiver applies the watermark immediately following content decryption. Both processes should occur within the device's physical security envelope, so that both encoded and baseband plain-text (deciphered) content is protected from eavesdropping prior to forensic watermarking. A serious complication arises, however, due to the requirement of conventional watermarking techniques for access to the uncompressed (baseband) digital video signal.

The baseband digital video is required for informed embedding analysis, and typically for the composition and embedding of the watermark signal. Consequently, a secure architecture in the receiving device requires that the physical security envelope enclose both the decode and watermark processes, in addition to the DRM and decryption blocks. Both the decoding and masking analysis require complex logic, and thus force a potentially costly expansion of the physical security envelope.

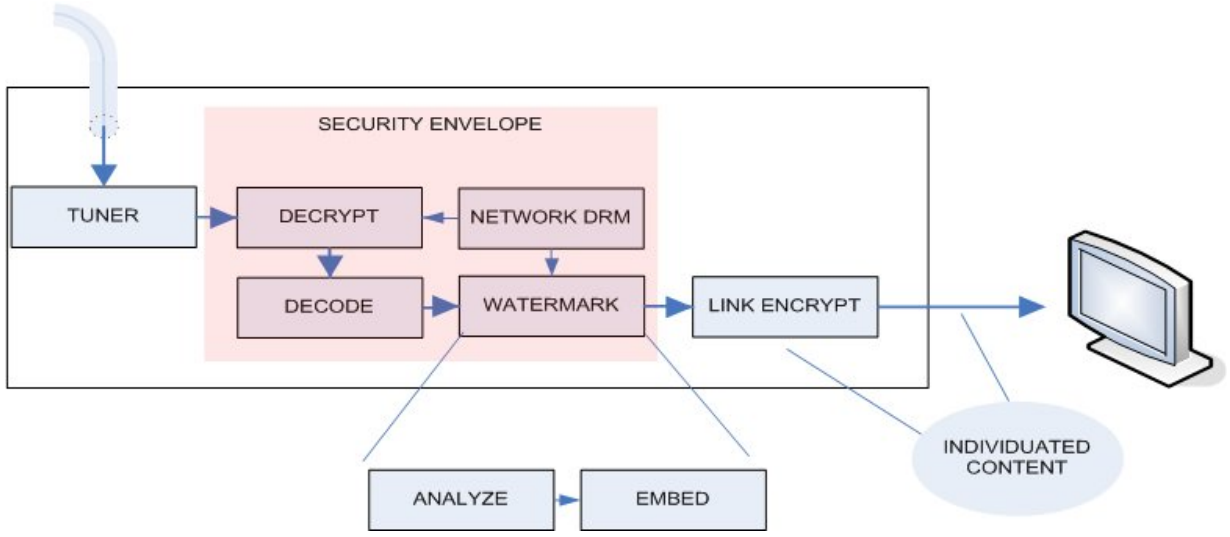


Figure 4 – Forensic watermarking post-decode

Figure 4 diagrams a receiver implementation in which the forensic watermark is applied to the baseband video, subsequent to decode. As shown in the red shaded area, a substantial processing block requires physical security to protect the unmarked data.

Another security issue arises when the receiver imports and stores content, as opposed to rendering in real-time. If the watermark process requires access to baseband content, the receiver must either defer watermarking until the content is decoded and rendered; or decode, watermark, and re-encode prior to storage. The former choice weakens security by distancing watermarking from the initial decryption, in both time and space. Figure 5 illustrates this design, including a very large requirement for the physical security envelope.

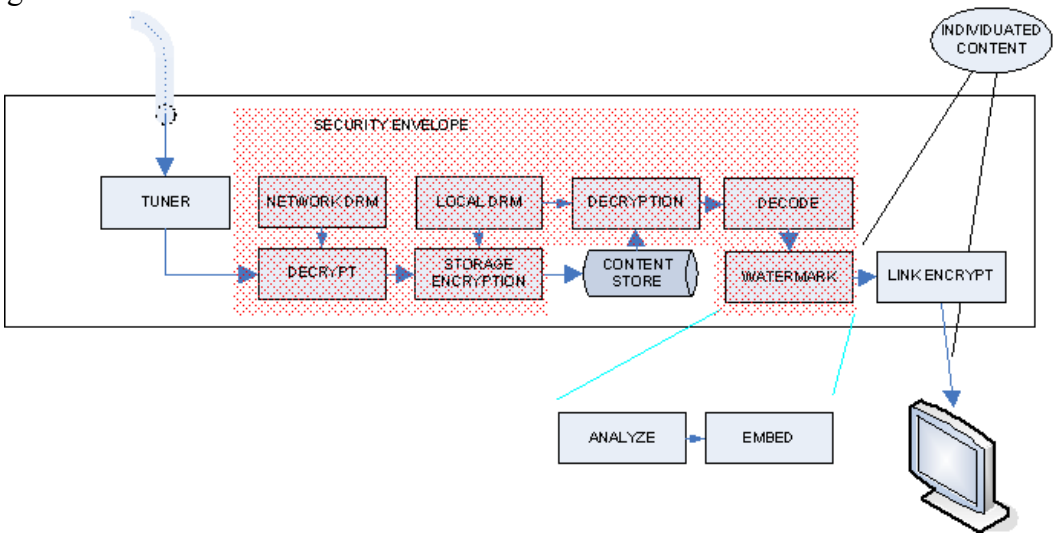


Figure 5 – Late Watermarking Model

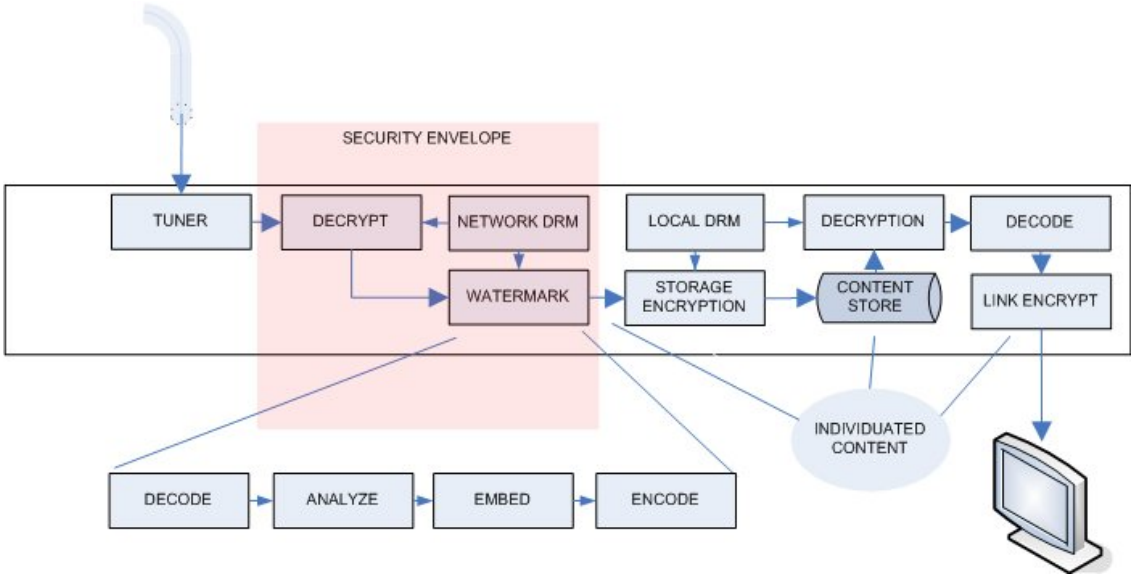


Figure 6 – Watermarking Prior to Storage

The latter choice adds a video encoder to the receiver, at significant component and potential licensing cost, as well as exposure to content degradation due to multiple encodings. This architecture is illustrated in Figure 6. The security envelope appears to have been reduced, but additional decode and encode blocks have appeared as components of the forensic watermarking process.

Integration Issues

A requirement for access to baseband video introduces integration issues for existing equipment designs. It may be difficult to arrange access to the baseband video in a format that is compatible with the watermarking process. Access to several successive frames for sophisticated informed embedding analysis is even more challenging. Additionally, frame latency in the watermarking process could necessitate adjustments to audio synchronization. Large scale integration in media processors can raise formidable barriers to watermark implementation by limiting accessibility to the video frame buffers.

Renewability and Flexibility

As mentioned above, both ease of renewal and flexibility are desirable features in a watermarking system. It is important for the operator to be as nimble as the pirates, so the overhead of a change to the watermarking technique must not constrain his ability to respond to new challenges.

Consistency

The need for consistency in forensic watermarking was discussed above. Consistency is most difficult to achieve when complex algorithms are deployed, particularly in renewal scenarios.

REPLACEMENT WATERMARKING IN THE COMPRESSED DOMAIN

So far, we have discussed forensic watermarking assuming a monolithic implementation – one in which the entire watermarking process takes place in a single device or component, apart from any other device or process. In such architectures, the watermarking process requires access to (at least partially) decoded content. As discussed above, the more sophisticated informed embedding techniques require full access to baseband video. This requirement has caused some system designers to question the feasibility of watermarking in a practical consumer device.

It is clear, on the other hand, that several of the engineering issues could be resolved if the watermark embedder were capable of operating directly on the encoded content. The much lower data rate of encoded content translates to smaller frame buffers and a lesser processing resource requirement. Encoded content is available immediately following decryption, where watermarking could be more securely integrated with the DRM. In use cases where the content is stored or downloaded subsequent to watermarking, the costly decode-encode steps could be avoided. Thus there is ample motivation to develop a technique for watermark embedding in the compressed domain.

Encoded content is, however, highly complex to modify directly. MPEG, the most common video codec family in the consumer domain, contains numerous interdependencies such as inter and intra frame references, differential coding, and entropy coding. Such dependencies make it impossible to interpret a single frame, much less part of a frame, or to make a localized modification. The only obvious approach appears to be the cumbersome decode-watermark-reencode paradigm.

An enhancement to the watermark system architecture can, however, circumvent these problems and actually permit watermark embedding to take place in the compressed domain. Assume that an upstream watermark processing step operates on a single content master instance. This process performs the analysis required for informed embedding, and generates compressed, watermarked video fragments, such that each fragment can be inserted into the encoded content stream, at a specific location, in place of pre-existing video data. Now, package these watermarked fragments with the content, and a downstream watermark embedder need only choose watermarks from this watermark "metadata" to substitute for parts of the encoded content.

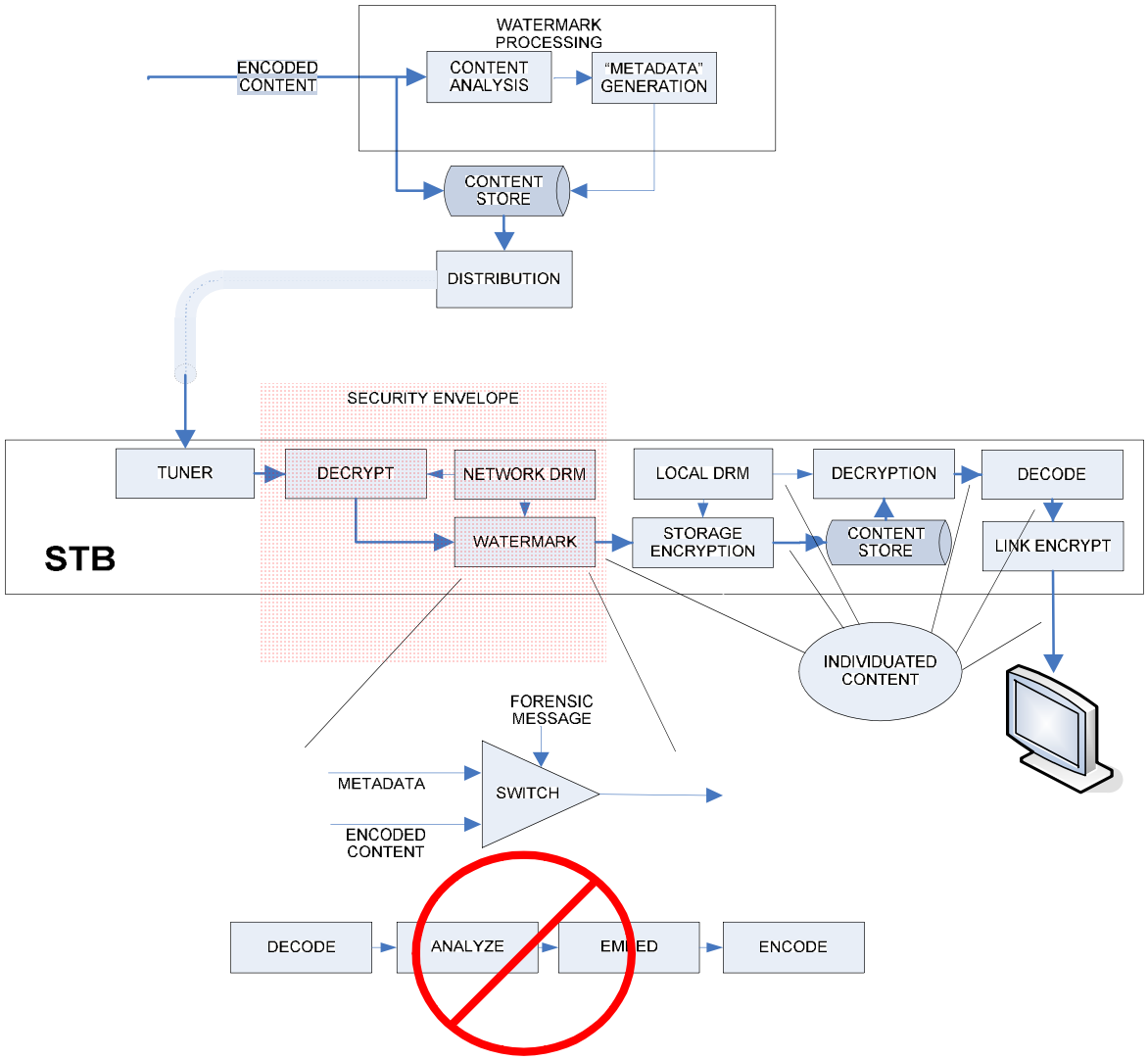


Figure 7 – Replacement Watermarking in the Set Top Box

This architecture, diagrammed in Figure 7, is termed *replacement watermarking*. The embedding process has become a simple switch, capable of selectively replacing segments of the content stream with the watermarked video fragments from the metadata. A substantial efficiency gain is realized by performing complex video analysis and watermark composition only once, where sufficient processing resources can be easily applied, and preserving the results for reuse. Most importantly, decode and encode processing in the watermarking block is completely obviated. Fielded implementations of the replacement embedder require little or no additional processing power than what is commonly available in existing STB designs. The perceptibility-robustness-cost dilemma is thereby alleviated, and the possibility exists that deployed STBs can be upgraded with the light weight embedder.

The STB receiver can utilize replacement watermarking to watermark encoded content immediately following decryption. The very light-weight embedder can be easily integrated with the DRM and decryption processes, and reside within the same physical security perimeter. Whether the resulting content is decoded and rendered immediately, or is stored for later viewing, it will have been individuated by the forensic watermark. The marked content can subsequently be moved throughout a home network, with no requirement for a watermark embedder in each playback device.

Since the embedder is essentially a simple switch, it operates on streams of content data. Frame buffers are not required. Very modest memory demands facilitate integration within the security perimeter inside a consumer electronics device. Integration of the replacement embedder into a STB is straightforward, with reduced impact on time-to-market.

The replacement architecture greatly facilitates renewal. Since all of the watermark placement and composition decisions are made in the upstream watermark processor, these attributes of the watermark system can be altered without affecting the operation of the downstream watermark embedders. Such changes are reflected in the watermark metadata created by the upstream watermark processor, and become effective immediately when the content is processed for replacement watermarking. Significantly, there is no need to update software in widely deployed watermark embedders.

The same techniques used to achieve renewal can be used to increase system flexibility. Watermarking placement and composition are controllable upstream, at the point of watermark metadata creation.

The replacement architecture ensures consistency across the fielded devices, alleviating concerns about heterogeneous fielded watermarking technologies and versions applying watermarks of differing perceptibility and robustness, and requiring diverse recovery techniques. Effectively, the control of watermarking is centralized, and less subject to variations in fielded devices.

REPLACEMENT WATERMARKING ENGINEERING CONSIDERATIONS

Several factors must be considered in the implementation of replacement watermarking. An obvious issue is how to incorporate the watermark metadata into the content package, such that the metadata is available to the watermark embedder. There are several requirements affecting this mechanism.

Foremost, the bandwidth must be available to deliver the metadata to the embedder, properly synchronized with the content. Prior to each frame being processed, any of the watermarked

fragments in the metadata affecting that frame must be available to the embedder. Watermarked metadata volume in existing implementations is minimal, but is subject to a number of factors, principally the density of the marks (e.g. marks per second) and the size of the marks.

To secure the watermark embedding process, particularly when it takes place in the potentially hostile environment of the consumer's receiver, it is necessary to ensure that the watermark metadata is bound to the content such that it cannot be identified and removed, prior to decryption. If an adversary were able to isolate the metadata stream, it would be a simple matter to delete or corrupt the metadata and thus suppress forensic watermarking.

Two techniques for conveying the metadata stream have been explored for commercial implementation. One is to simply utilize the codec "user data" features to carry the metadata within the compressed frame structure to which it refers. This approach has the advantage of transparency in distribution, since the metadata simply appears to be part of the encoded content, secured by the same encryption wrapper.

A second approach is to utilize a side channel. When a side channel is employed, it is necessary to secure the side channel to prevent tampering with the data that might disrupt the watermarking process.

On the upstream process side, the metadata must be created in such a way that a valid encoded content stream results when the watermarked fragments are embedded. The techniques for doing this are largely dependent on the codec in use.

SUMMARY

Forensic watermarking involves mass production of individuated content instances. The process is very challenging to securely implement in a large distribution system, using autonomous watermark embedders. By centralizing computationally intensive tasks, distributed watermark embedding can be accomplished through simple operations in the compressed domain. The perceptibility and robustness advantages of informed embedding can be realized with minimal cost impact at scale, along with improved security, renewability, consistency, and flexibility.

---

[i] Fred Dawson: Studios Eye 1st-Run Service As IPTV Security Advances ScreenPlays January 2008

[ii] Mark Kirstein: MultiMedia Intelligence Identifies Digital Watermarking & Fingerprinting As Key New Opportunity, http://multimediaintelligence.com/index.php?option=com_content&task=view&id=49&Itemid=1

[iii] Steganography is defined as the technology of hiding a secret message inside of a larger cover work, such that the existence and content of the secret message are hidden.

[iv] Media pirates often reencode content at lower data rates or resolutions to suit their preferred distribution channels.

[v] Collusion attacks are those that involve combining content from separately captured instances of the same content in order to dilute the watermark signals.

[vi] Digital Cinema Systems Specification V1.0, July 20, 2005

[vii] Barni, Bartolini, and DeRosa Perceptual Data Hiding in Still Images, Idea Group Publishing 2005

[viii] Barni, Bartolini Watermarking Systems Engineering, Marcel Dekker, Inc. 2004

DYNAMIC INSERTION FOR SHORT-FORM VIDEO ON DEMAND ADVERTISING

John Chandler-Pepelnjak and Brent Roraback
Advertiser and Publisher Solutions, Microsoft

Abstract

Historically Video On Demand (VOD) advertising has required that programming content and advertising content be encoded and delivered together. In 2006, we deployed two field trials of "dynamic insertion," the run-time assembly of advertising and programming content.

This paper details the execution of these trials from an ad-server perspective. We explore the requirements of an ad-serving solution, on both the execution and reporting aspects of a campaign. Sections 1 and 2 provide introduction and background material. Section 3 details the technical requirements for campaign management and execution. Section 4 details the data that is reported on and best practices for its analysis.

INTRODUCTION

What Atlas does

Video On Demand (VOD) advertising has been labeled an "emerging media" channel for far longer than one would think it takes to "emerge." The delay has been due, in part, to the impressively complicated systems that underlie the delivery of cable television. Another part has been the reluctance of cable operators to innovate and thereby risk disturbing service. Finally, VOD advertising has been challenging for the agencies and advertisers that would be likely to use it—markets are small, deployment is tricky and creative requires long lead times.

In this paper we detail a solution that solves many of these agency and advertiser issues through the Atlas On Demand Media Console. Atlas, founded in 2001 and acquired by Microsoft in 2007, makes software for agencies and advertisers to plan, manage, deploy, track, report on, and optimize online marketing campaigns. For the dynamic insertion VOD trials outlined here, we used this technology to facilitate campaigns in two cable markets.

A brief history of VOD advertising

Historically, VOD advertising has been very similar to linear television advertising, at least from the perspective of the advertiser. Typically ad creative is produced, sent to a network where it is encoded with programming and distributed to operator VOD systems. Views of the programming content run via a request from the set-top box (STB) and the collective viewing of the ad asset and programming content is recorded. At the end of the campaign advertisers and agencies receive some data detailing number of views and the reach (the unique count of subscribers or STBs viewing the content). Critically, views of ad content are not separated from programming content views, obscuring the critical information on whether or not the advertisements were fast-forwarded or even viewed at all. Moreover, modifying the scheduled creative mid-campaign is impractical, if not impossible, due to the lead times required to re-encode and distribute the updated content.

For example, movie studios often take their creative material from the finished version of the film. These shots are sometimes ready only one or two weeks before the film is released. With typical VOD lead times of six weeks, VOD advertising becomes untenable for these advertisers. Additionally, it is often advantageous for these advertisers to change creative after the opening weekend, touting reviews from critics or other achievements (e.g., "Number one movie of 2008"). Given that the ad content and programming content are joined, simply swapping creative is impossible. There are many

other industries for whom the ability to change creative depending on external circumstances is valuable.

The final drawback of the current VOD system is perhaps the most damaging in the long-run. The internet's quick ascension in the marketing mix is partially due to the ability to target. Meaning, when an ad impression is called for, the content provider or network can use information about the viewer to determine the most effective creative messaging. In the current incarnation of VOD, this is not possible as all viewers of a piece of programming content will receive the same ads. The ability to make television addressable through dynamic insertion is a critical feature both for advertisers and for the medium itself.

The requirements for a new solution are clear. Advertisers must have the flexibility to insert creative into placements on short notice. This allows creative swaps, different ads targeted to different viewers, and evaluation of creative performance independent of programming performance.

CAMPAIGN MANAGEMENT AND EXECUTION

Background

"Campaign Management" refers to the set of activities concerned with the definition and management of advertising campaigns. Sellers of inventory describe the format and characteristics of media they have for sale. Agencies and advertisers record the inventory they have purchased from various sellers and the terms under which it was purchased. These business terms include information such as the cost method used for describing the unit of media (e.g., CPM or "cost-per-thousand" impressions, time-based costing), the cost per unit (or "cost basis"), target or guaranteed quantity of impressions to be delivered, date ranges, acceptable types of inventory (e.g., length of spots), etc.

Along with information describing the inventory purchased, the second important aspect of campaign management relates to the ads associated with the campaign. Information about the ads (e.g., identification codes such as ISCI or AdID, the names of the assets, the asset durations, etc.) is specified along with the inventory with which they are associated.

Campaign Management Solution

In dynamic advertising scenarios, campaign management becomes an active exercise: users receive frequently updated information regarding the status of their campaigns and have the ability to make changes to the campaign while the campaign is being delivered. In a non-dynamic scenario, such as broadcast television, the same degree of active management and detailed reporting does not exist.

Atlas has created a solution for agencies and advertisers to present instructions to ad execution and management systems directing these systems to display specific ads when a particular piece of content or inventory is delivered. Atlas also collects information about the viewing of advertising content, calculating metrics based on this viewership along with the business terms and goals under which the inventory was purchased.

Our solution, Atlas On Demand, interfaces with the ad execution and management systems over secure connections via APIs defined with our technology partners in this space (e.g., Aaris, SeaChange; Atlas is also a provider of inventory and ad management solutions for sellers of inventory, integrated with other partner systems). The set of services and message structure for managing this communication leverages

emerging industry standards such as SCTE 130 (DVS 629).

Within Atlas, campaigns are created and ads assigned to purchased inventory. Instructions are then published to the ad execution and management systems. When ad-supported content is requested by subscribers the content is assembled based, in part, on these instructions. Ad and program content are "seamlessly spliced" together and streamed down to the subscriber's set-top box (STB).

Campaign Execution

For a variety reasons – disparate, closed network systems; manual or semi-automated processes; emerging standards; pre-existing workflows; etc. – executing dynamic VOD advertising campaigns is still a very complex process. Tight coordination across a range of partners at multiple levels, from senior sales executives and content owners to network engineers and ad operations personnel, is required.

In order for Atlas to be able to communicate with the ad execution and management system deployed at the operator a secure connection must be established, such as through a VPN concentrator or other secure web service connection. The operator's endpoint, transport and protocol are managed through configuration settings in Atlas.

As noted above, program content and ad content have traditionally been encoded together as a single on demand asset. With dynamic VOD advertising, programs and ads are treated as separate assets, "seamlessly" spliced together at runtime and streamed to the viewer. In such a model, content providers and distributors must account for "ad free" versions of the content suitable for dynamic VOD ad campaigns, as well as the ad assets themselves, insuring that the complete package of program and ad assets is distributed appropriately.

To insure the viewing experience in dynamic VOD is of the same quality as other on demand viewing, content encoding standards must be rigorously followed. In our trial campaigns, CableLabs OD encoding standards formed the basis for theses specifications but extra care was taken to insure that bit and frame rates, resolutions, and audio were identical for all assets. Assets were also required to start and end with silent frames of black to ease the transition between assets. The "seamless splicing" of assets, mitigating any remaining discontinuities between MPEG files, was either accomplished in software by the VOD system or in the edge device (i.e., the QAM) level, depending on the VOD system provider's approach.

Aside from system configurations and content preparation, the standard campaign workflow generally begins with media negotiation: sellers of media (content networks, operators) offer packages of inventory to buyers of media (agencies, advertisers). Rates, schedule and other terms are negotiated and agreed to through terms and conditions, insertion orders, etc.

After the contract is finalized, ad campaign information is configured in the respective inventory and campaign management systems of the buyers and sellers (information may already exist in the seller's system, enabling the seller to forecast and book inventory). Ad assets are distributed, generally to the inventory seller but in some cases to the operator directly. Ad and program assets are encoded per the on demand specifications and distributed to the operator and headend systems, usually through existing "pitch – catch" mechanisms but potentially through IP-based distribution.

Once campaigns have been configured and ad and program content distributed, ad

instructions are published by the agency/advertiser to the inventory/ad management system at the operator. Instructions are validated by the operator system and acknowledgements returned to Atlas. If the instructions are valid they will be referenced when ad-supported content is assembled.

Viewers request ad-supported VOD content, initiating sessions with the VOD system. The Atlas instructions are referenced in the assembly of the on demand content, ad and program content is seamlessly spliced together and streamed down to the viewers. Viewer interactions with the content ("trick mode" activity, such as fast-forwards, pauses, rewinds) are recorded by the VOD system. We collect detailed information regarding viewing and playback of ad assets. Data is imported in Atlas' reporting system and metrics are calculated.

Atlas users view statistics related to their campaign's performance through online reports. Statistics are updated several times per day. This granular viewing data allows Atlas to calculate and display multiple metrics describing campaign performance. Impressions (i.e., views irrespective of playback speed), "Brand Exposure Duration" (BXD) (i.e., viewing duration at normal playback speed), completed plays, reach, trick mode counts, and more may used to compare the performance of ads and/or their associated inventory. Campaign performance is assessed by analyzing the metrics corresponding most closely to the advertiser's campaign goals. Our users may then apply this information by "optimizing" their campaign: changing the ads assigned to their campaigns and/or the business rules governing ad rotation to maximize performance. New instructions are then published to the VOD system and enforced during subsequent viewing sessions.

REPORTING AND DATA ANALYSIS

Data description

As mentioned previously, one of the principal benefits of dynamic insertion is the ability to measure ad performance separately from programming performance. In other words, as users fast-forward (FF), rewind (RW) or pause their ad programming, we collect data on each trick-play and can use that to measure viewership.

The data that Atlas collects come in two different styles. The simplest conceptually is what we call "event-level records". These records detail every trick play event and capture the following pieces of data (or meta-data) associated with the event:

- Date
- Time
- Operator
- Headend
- Masked MAC address
- Ad Asset Name
- Ad Asset ID
- Event (Setup, Play, FF, RW, Pause, Teardown)
- Event Speed (1 for play, positive for play or FF, negative for RW)
- Programming Content Name
- Programming Content ID

In data of this format, one row of data represents one event of ad viewing. A viewing session is defined as beginning with a Setup event followed by a Play and ending with a Teardown. There can be any number of interstitial trick play events (FF, RW, Pause). The most complicated data field is masked MAC address. Typically the unique identifier of a STB must be masked for privacy reasons. It is important that the masking algorithm be 1-1 so that no two MAC addresses can be mapped to one masked MAC address so that reach and frequency can be accurately calculated.

Occasionally the detailed data format is not available. In that case, there is an alternative data format ("aggregate-level") where one row of data represents one (potentially partial) view of an ad. In order to use this data, additional fields must be added. These include the following:

- Start Date
- Start Time
- End Date
- End Time
- View time (amount of time asset was viewed in normal playback speed)
- Fast Forward Count
- Rewind Count
- Pause Count

The key metrics that can be derived from these reporting fields are worth pointing out explicitly. All traditional campaign measures—GRPs, impressions, reach, and frequency—are available. Additionally, we can look at viewership patterns by time of day and day of week. Finally, a variety of user-level reporting metrics are available and these will be detailed later in the section.

Measures of Performance

The VOD landscape is crowded with many metrics used to evaluate performance across many dimensions. Ads, placements, asset lengths, and content providers are judged by a variety of yardsticks. Though our dynamic insertion trial, two metrics emerged as the most critical for evaluating these campaign attributes. Because VOD is an accountable media, we take the time to highlight these two different performance metrics and detail their implications.

The first measure has already been mentioned: Brand Exposure Duration (BXD). This is simply the amount of time an asset is viewed in normal playback speed. If you sum up all the BXD values for every impression on a campaign you have "Total BXD". While simple at first glance, Total BXD is a powerful omnibus metric, combining length of asset, asset average view time, and total number of views. Increasing asset length has the effect of decreasing the average view time (jumping from a 15 second to a 30 second spot diminishes viewership) but our research indicates that net viewership typically increases with increasing asset length. We can also look at BXD in several different ways. Average asset BXD can be used to optimize creative—if this method is followed total viewing time will be maximized. Alternatively, if all creative are in rotation in a placement (say, the first commercial break of an ad-supported VOD program) then average BXD will determine the value of the placement and the suitability of that program's audience to the creative message. All else being equal, BXD will tend to choose longer commercial assets. From the data we have seen across VOD on television and video on the web, doubling an asset's length rarely cuts the average percentage of the asset viewed by half. Finally, BXD is a useful cross-platform measurement. BXD can be calculated for video across any screen and is a useful measure of engagement across platforms. Note that the efficacy of *any* video metric diminishes if the ability of users to FF is disabled.

The alternative to BXD that emerged during our trials was "completed plays". This metric can be defined in multiple ways but the simplest is a view of an asset with no FF activity. (An alternative is to look for every second of an asset being viewed at least once in normal play mode, though this is more complicated.) Assets are compared with each other based on the percentage of views resulting in a completed play. Whereas BXD tended to reward longer assets, completed plays unequivocally skew results towards short assets. Completed plays are, however, the only metric that makes sense for certain classes of creative assets. Commercials where the call-to-action or brand message is delivered in the closing seconds require

optimization based either on completed plays or on a "weighted BXD" where certain portions of the asset are worth more than others.

Results from Data Analysis

The two trials detailed in this paper had different compositions. The first had two pre-roll creative assets for one advertiser running over two different time periods across one content provider. This trial allowed us to prove the concept in a simple environment. The second trial was much richer from a data analysis perspective: two advertisers and two content providers running multiple assets of varying lengths in both pre-roll and post-roll positions. In the results that follow we focus on this second trial. It is easier to follow the results when speaking about a specific campaign and the results themselves are much deeper with these data.

As mentioned above, two advertisers took part in the second VOD trial. The first advertiser had several assets in rotation, all 30-second spots, running in both pre-roll and post-roll. The second advertiser had pre-rolls of both 15 seconds and 30 seconds, followed by post-rolls of 60 seconds and 120 seconds. These assets were all in rotation on both content providers.

When reporting on short-form, dynamically-inserted VOD campaigns, there are a number of standard metrics that barely need mentioning in this forum. Although fundamental to campaign evaluation, metrics like reach, frequency, and impressions (divided into various time ranges and publisher and placement groups) are straightforwardly defined elsewhere. Instead we will focus on a series of analyses we performed during the trials that were unique to the dynamic VOD environment.

Initially there were two pieces of "conventional wisdom" that we wished to analyze regarding pre- and post-roll advertising. Are longer pre-rolls fast forwarded more often than shorter pre-rolls? Do viewers stick around to watch post-roll advertising?

It goes without saying that in our data the pre-roll ads received a higher impression count and higher completed plays. This is nearly tautological. Due to the greater length of the post-roll, however, the post-roll commercials resulted in longer BXD. Again, BXD is the average or aggregate duration in minutes that an ad or a brand (if multiple ads) is watched. This mirrors research we typically see with digital video: longer assets perform better from a BXD perspective, shorter assets perform better from a completed play perspective.

Viewership on post-rolls was surprisingly high, including the number of completed plays. This would indicate an undervaluation of post-roll ads given the common assumption of little to no viewership. One factor contributing to the longer BXD was the longer durations of the creative used in the post-roll positions compared to the creative used in the pre-roll positions. On average post-rolls were viewed approximately 36% of the way through (versus 42% for pre-rolls).

Unsurprisingly, our analysis revealed that the 30 second pre-rolls were fast forwarded more than the 15 second pre-rolls. Since one advertiser had only 30 second spots, it is possible there was some burn-out, although this behavior (more FF activity on longer spots) is not atypical.

There was evidence to suggest that viewership of post-roll ads can be augmented by pre-roll ads from the same advertiser. In other words, "bookended" placements (with a single advertiser in both positions) are more valuable than pre- or post-roll placements alone or from distinct advertisers.

From a viewer perspective, post-roll viewership of bookended placements might be perceived in much the same way as mini-long-form advertisements. Indeed, our research indicates that, if maximizing BXD is the goal, advertisers would do well to treat post-roll viewership as a "conversion" and focus on pre-rolls that are most likely to achieve viewership of the post-roll. Given the viewers discretion to watch or FF in this context post-roll viewership starts to look like a click or performance-type media on the web. Our research also indicates that longer pre-rolls (:30s over :15s) make viewers more tolerant of longer post-rolls.

Another question we asked of the data was, Is there an interaction effect between advertiser and content provider? A display of the data appears in Figure 1.

This figure shows the interaction between the advertiser's commercial content and the programmer's content adjacent to which the commercials run. On the y-axis we see BXD expressed as a percentage of asset length—higher numbers indicate more of an asset was watched. As we can see, Programmer 2 performed better overall, but there is an interaction between the advertiser's content and the programmer's content (indicated by the crossing of the lines). This is important: not all programming is optimal for all advertisers and advertisers may reasonably value different pieces of VOD inventory differently. Currently many

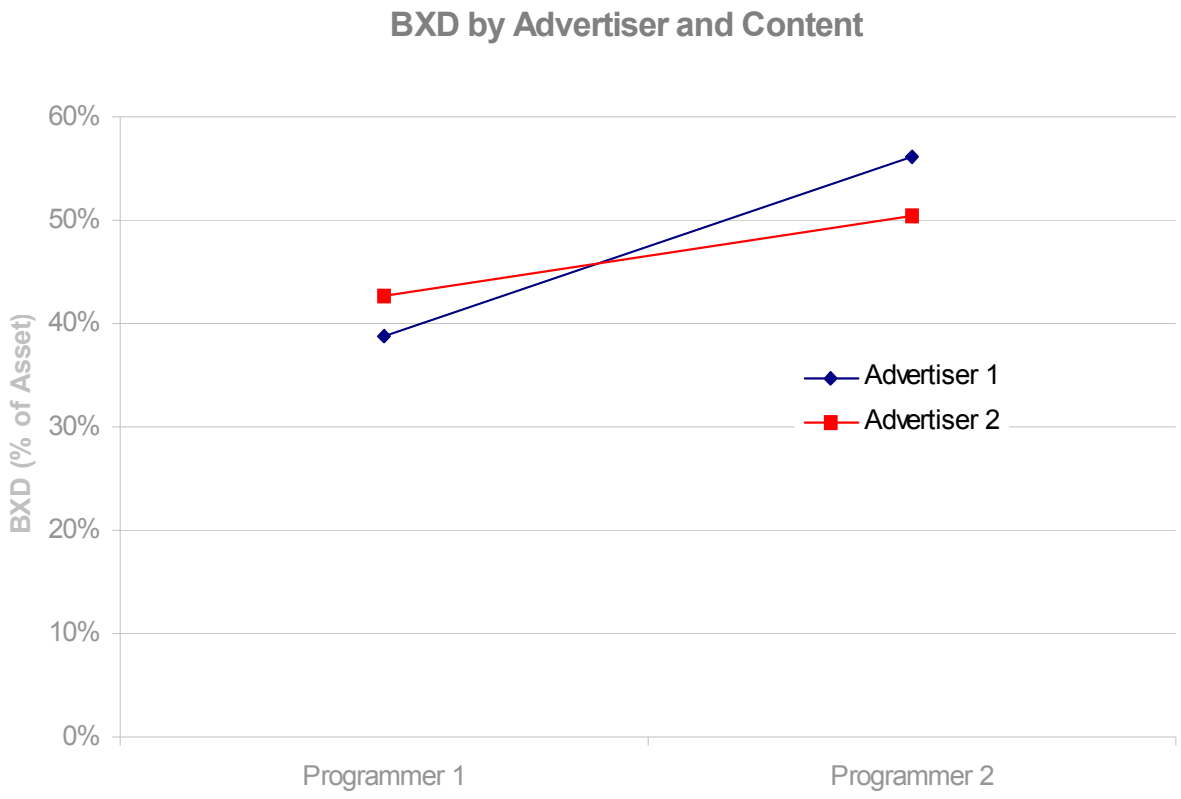


Figure 1: In this figure we see the interaction between the advertiser's commercial content and the programmer's content adjacent to which the commercials run. On the y-axis we see Brand Exposure Duration (BXD) expressed as a percentage of asset length—higher numbers indicate more of an asset was watched. As we can see, Programmer 2 performed better overall, but there is an interaction between the advertiser's content and the programmer's content (indicated by the crossing of the lines).

marketers believe that pre-roll video viewing is dependent entirely upon the content following the spot. This research contradicts that view.

There was, however, a tradeoff to consider between reach and impressions on the one hand and increased duration on the other.

As for asset lengths, longer advertising assets generally resulted in longer viewing durations and more total minutes viewed, shorter commercials enjoyed more completed plays and were watched in their entirely a higher percentage of time.

Although you might assume that the average BXD percentage might decrease as the length of the commercial increases, it was found that the percentage of :120 spot viewership was higher than that of the :60s.

Post-rolls were watched approximately the same percentage of the time regardless of asset length (roughly 35%). This confirms previous research conducted by Atlas on long-form VOD advertising. One interesting effect of post-roll viewership was noted. For one advertiser, viewership of :120s in the post-roll position increased by 20% if the pre-roll was a :30 instead of a :15.

The :30 commercials of one participating advertiser performed better in the pre-roll than in the post-roll in terms of both BXD average minutes and as a percent. Some of this effect could be due to the repetition of the same ads given that if the same spot appeared in the pre-roll position as in the post-roll position a viewer would be predisposed to fast forwarding through the second appearance.

CONCLUSION

There are two critical components in the deployment of a short-form VOD campaign. The first is the ability to package ad content separately from programming content. This is fundamental to many needs of advertisers including creative management, creative swapping, decreasing creative lead times, accurate reporting. The second is detailed measurement of individual ad viewing duration. The first requirement is a near-prerequisite for the second, but it is only with this measurement that the power of the addressable television medium is achieved.

ECONOMIC IMPACTS OF VIDEO FOR THE BROADBAND WIRELESS OPERATOR

Paul Steinberg, Mark Shaughnessy, Lloyd Johnson, Philip Fleming
Motorola, Inc.

Abstract

The ability to offer compelling wireless video services over a wide area has been a long awaited goal for consumers and operators alike. Finally, 4G wireless technologies such as WiMAX and Long Term Evolution (LTE) offer adequate performance characteristics to support IP based video services to a large number of consumers simultaneously with a quality that most will find attractive. However, there are a number of economic and technical factors that the operator will need to consider when it comes to actually implementing video services in a wireless network.

In this paper we begin by summarizing the history of broadband wireless technologies and their shortcomings relative to the desired consumer experience and operator economics. Next we review the key attributes and performance implications of 4G broadband wireless technologies and relate those to the capabilities necessary for a service provider to deliver a viable mobile video service offering.

The paper then describes an overall network architecture and essential elements to deliver an end-to-end video solution. The mobile wireless environment enables the operator to tailor the offered content, services, and advertising, dependent on user location and context. Operators that offer both wireless and wireline access networks have the opportunity to integrate them under a common IPTV video headend which provides not only consistent consumer experience and content access, but also the ability to provide mobility between these networks. Video streams in progress can be moved among wired and wireless devices, and content can be made available to any device, anywhere. We conclude with a brief look at some anticipated future trends and applications of mobile video.

INTRODUCTION

Our world is increasingly mobile, and this is driving the demand for easier access to content and services from any location at any time. Existing *wireline* networks have made universal access to the Internet possible, but only for those that have a physical cable or short-range WiFi connection to the network.

The underlying demand is for untethered wide-area mobility, i.e. accessing content and services without wires. To partially satisfy this demand, a vast market has grown in support of a "cache and carry" model. Since digital technology has enabled the efficient storage of vast amounts of content in small portable devices, content can now be downloaded from a multitude of public and personal sources and stored (or "cached") in a portable device that is carried wherever a person goes.

However, "cache and carry" requires the step of pre-loading, which doesn't solve the problem of access from any location whenever the need arises. Nor does it support streaming real-time content or social interaction services like real-time communications. Fortunately, *wireless* mobile data networks, coupled with a converged content and delivery network, are a solution to these mobility aspects. However, only recently have wireless technologies become fast enough to support delivery of bandwidth heavy content, in particular the type of content that is the focus of this paper: video.

EARLY WIRELESS DATA TECHNOLOGIES

Since they were first deployed, there have been attempts to use wide-area wireless mobile data networks to provide some level of video service. Early examples of wireless mobile data networks were in the military and public safety realms [1]. Early public mobile data networks, such as Cellular Digital Packet Data (CDPD), Mobitex, and DataTac offered data rates of 19.2 Kbps or less, and suffered from relatively low subscriber interest. The first public wireless data networks to gain significant subscriber uptake were extensions to the 2G cellular telephony networks, primarily GSM (3GPP) and CDMA IS95A (3GPP2). Early technology demonstrations of video services on these networks centered on video telephony and the exchange of video clips via the multimedia message service (MMS).

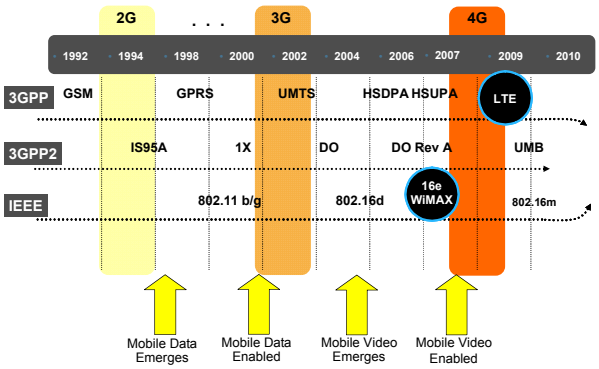


Figure 1. Timeline for the Introduction of Mobile Data Functionality

Figure 1 summarizes the evolution of the commercial wireless broadband standards [2]. With the advent of 2.5G technologies – such as GSM GPRS and CDMA 1xRTT, basic data services such as email and web browsing were offered commercially and were somewhat successful. There was much discussion of extending the cellular data offerings to include some form of video service, with attention focused on video telephony and multimedia messaging. Technical feasibility was demonstrated for these services but commercial versions of the services never materialized. While there were undoubtedly multiple reasons for these commercial failures, the three most significant were high cost of data delivery, low bit rate performance, and limited mobile device functionality.

The device limitations in particular, such as very little memory, slow processors, and small low resolution displays, were the pivotal factors that drove the data services offered for 2/2.5G cellular phones to be primarily limited to browsing of cellular-aware web sites and specialized email programs; and even these were painfully slow. There have been some 2/2.5G data commercial successes in controlled system environments. A notable example was the i-Mode service offered by NTT DoCoMo in Japan [5]. Success was achieved through creation of unique applications and experiences that were optimized for low bit rates, by a large development community.

A wider range of services could be offered using PCs with 2/2.5G access cards. In theory, subscribers could even attempt video downloads using these cards, assuming they had enough time and money. The predominant business model employed by operators was to charge for usage (bytes transferred). Download of even a small 5MB highly compressed video file at low resolution with a typical data rate 50 Kbps or less would have taken over 13 minutes; and given the tariffs of the time (often as much as $1 per 100 KB for GPRS), cost as much as $50.

The introduction of 3G and 3G+ technologies – UMTS, HSDPA and EV-DO – during the middle of the current decade provided a marked increase in wireless data capabilities, with typical data rates on the order of 0.5 to 1 Mbps achievable for HSDPA and EV-DO. These new technologies also provided a dramatic decrease in the cost of delivering data services, as is shown in Table 1.

	Peak Speed	Realistic Speed	Cost to Deliver MB ($)	Cost to Deliver 5GB/Month ($)
HSDPA	14 Mbps	900 Kbps	0.021	$105
1xEV-DOr0	2.5 Mbps	300-500 Kbps	0.022	$110
UMTS	2.0 Mbps	150-200 Kbps	0.069	$345
1xRTT	625 Kbps	60 Kbps	0.059	$295
EDGE	384 Kbps	30 Kbps	0.138	$690
GPRS	120 Kbps	12 Kbps	0.415	$2075

Source: CSFB, HSDPA Networks Group

Table 1. Data Delivery Costs for 2G/3G Wireless Technologies

"Unlimited" data services (where unlimited is defined as a few GB per month) were now being offered by several operators at less than $100 per month. During this same time period, dramatic advances were made in the quality and cost of the cameras and displays that can be incorporated into a mobile phone. With the combination of the new functionality and reduced cost, non-text services finally began seeing substantial growth. Multimedia messaging incorporating both still and moving images is now in wide use; and, as is demonstrated almost daily on the evening news, the uploading of embarrassing film clips from mobile phones to YouTube has become a world-wide hobby.

BROADBAND WIRELESS TECHNOLOGIES BECOMING AVAILABLE NOW

4G wireless broadband in the form of WiMAX and 3GPP's Long Term Evolution (LTE)[1] (see [7], [8], [10]) provide a further significant step forward in consumer experience and network capacity, as well as network price performance.

Before diving into the specifics of 4G wireless performance for video, it's useful to compare to wireline performance since that's the more traditional reference for carrying video.

[1] 3GPP2 is also defining a 4G air-interface standard called Ultra-Mobile Broadband (UMB), which is technically very similar to LTE.

Figure 2 compares fixed/wireline and wireless broadband technologies in terms of throughput capability over time [3]. In general, we see that wireline bandwidths fairly consistently exceed those of the contemporary wireless technologies by approximately two orders of magnitude. The dotted line approximates our view of expectations for "good enough" mobile wireless performance. We see that it's not until 4G that wireless throughput reaches a level that is satisfactory.

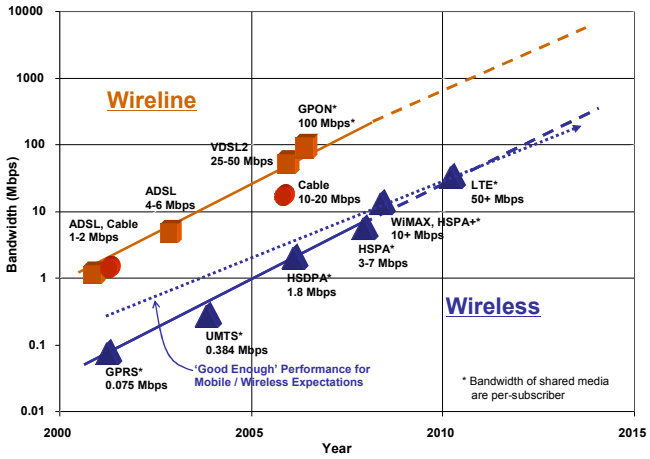


Figure 2. Technology Bandwidth Comparison

Mobile Video Needs

Table 2 illustrates typical bandwidth demands for video services using the current generation of mobile terminals and video displays. It is reasonable to assume that these values are the minimum acceptable; going forward, higher resolutions and faster frame rates will become the norm since consumers will expect the better picture quality that they experience at home to be duplicated in the mobile environment.

Resolution Name	Picture size (& frame rate)	H.264 (MPEG-4 part 10) Bit Rate
QCIF	176 x 144 (10 to 15 fps)	64 to 80 kbps
CIF	352 x 288 (7.5 fps)	192 to 240 kbps
QVGA	320 x 240 (10 fps)	192 to 240 kbps

Table 2. Example Data Rates for Mobile Video

4G Air Interface Advantages

Table 2 shows that QVGA resolution video (currently used by YouTube) at 10 frames per second can be supported with a data throughput of 240 Kbps (max). Using the performance projections in Table 3 and assuming a 10 MHz carrier deployment, we estimate that WiMAX can deliver a sustained downlink streaming video session with a throughput of 240 kbps to about 60 randomly scattered video users simultaneously in an area of roughly 3 km^2. This results in the capability of supporting 20 video sessions per square kilometer with a good (i.e. QVGA) user experience. LTE, using the same spectrum allocation and cell size, can support slightly more than 1.5 times as many video sessions. Furthermore, in the near future, LTE and WiMAX, may exploit increased bandwidth allocations up to 40 MHz. Support of 100 or more simultaneous streaming video users per square kilometer will then be within reach.

The uplink is typically the weaker link, supporting 20% to 50% of the data throughput of the downlink. For uplink streaming applications such as See-What-I-See that require about 120 kbps we can expect today's WiMAX to support about 10 such sessions per square km and LTE and future WiMAX to support up to 50.

Parameter	HSPA 5+5 MHz FDD	WiMAX 10 MHz TDD	LTE 10+10 MHz FDD
DL Peak Rate (Mbps)	14	32	60
DL Peak SE (bps/Hz/Sector)	2.8	6.3	6
DL Sector Throughput (Mbps)	3.3	7.9	16.7
DL 5%ile User Throughput (Kbps)	120	210	450
DL SE (bps/Hz/Sector)	0.66	1.30	1.67
UL Peak Rate (Mbps)	5.8	5.0	20.0
UL Peak SE (bps/Hz/Sector)	1.15	1.0	2.0
UL Sector Throughput (Mbps)	1.5	1.4	7.6
UL 5% User Throughput (Kbps)	43	52	192
UL SE (bps/Hz/Sector)	0.31	0.37	0.76

Table 3. Wireless Broadband Performance Comparison

The estimates in Table 3 are based on technology simulations performed by Motorola and other major suppliers of wireless broadband access points and network equipment [11], [12]. An important metric used in these comparisons is Spectral Efficiency (SE). SE is calculated by taking a ratio of the throughput as measured in bits per second (bps) and the amount of radio spectrum allocated (Hz). So the units are typically given in bps/Hz (or equivalently, Mbps/MHz). Higher spectral efficiency translates to lower cost per subscriber for the operator. 4G wireless technologies are able to achieve higher spectral efficiency through the use of advanced transmitter and receiver designs, multi-antenna arrays, adaptive coding and modulation techniques, and smart packet scheduling methods, all of which take advantage of the changes in the radio environment during a data session. While earlier technologies possessed some of these methods they were not able to utilize them all in concert over a wider spectrum allocation to achieve the high data rates of WiMAX and LTE. In addition, the improvement in uplink spectral efficiency possible with LTE is primarily enabled by the use of an advanced coherent detection scheme [13].

As the relative total cost of ownership curves presented Figure 3 show, LTE and WiMAX, which are based on OFDMA (Orthogonal Frequency Division Multiple Access) radio technology, provide a substantial improvement in network price performance (Total Cost of Ownership, or TCO), especially as the amount of data consumption increases per device. Based on the above QVGA estimation, a heavy mobile video user (say one who averages an hour of viewing per day) would consume on the order of 7 GBytes per month of data service. Examination of Figure 3 indicates that it would cost 2 to 4 times as much to operate an HSDPA or EVDO network when providing this level of per user capacity, as it would to operate an equivalent LTE or WiMAX network.

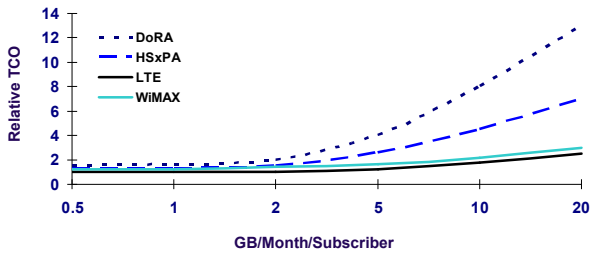


Figure 3. Relative Total Cost of Ownership for Broadband Data Delivery[2]

4G Network Technologies & Topologies

Prior to 4G, macro-area wireless networks were first and foremost cellular voice systems optimized for carrying narrow circuit voice style traffic. These systems had packet data facilities "glued" on almost as an afterthought. WiMAX and LTE are pure, broadband data systems that have no specific circuit voice provisions (other than as a constant bit rate QoS class) – they are IP packet based access technologies. This results in simpler and more decentralized network architectures relative to earlier networks and eliminates the need for complex protocol interworking between the wireless network and the operator's IP backbone.

The architectures of the 4G WiMAX and LTE networks are contrasted with that of the 3G UMTS network in Figure 4.

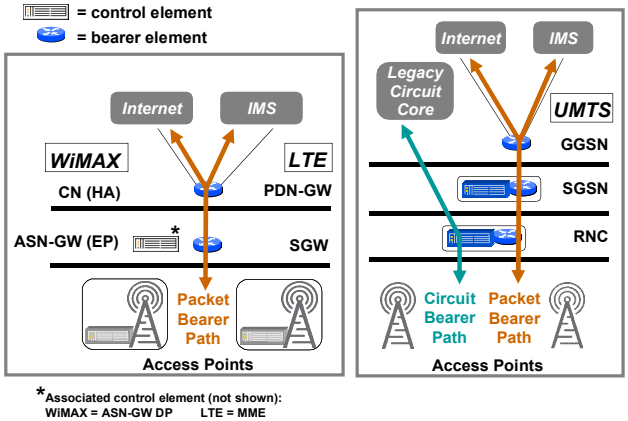


Figure 4. Flatter, Data-Only Architecture for LTE and WiMAX

The WiMAX and LTE network architectures were specifically created to support packet data services and are optimized for those services. Voice services for WiMAX and LTE are based on VoIP technology and are treated just like other data applications. Both WiMAX and LTE make extensive use of Internet concepts and protocols. This allows them to limit the amount of domain-specific equipment that must be used and leverage the volume production of components designed for Internet use. In contrast, the 3G UMTS network has both a data architecture and a legacy circuit architecture. Since legacy circuit and packet data are dramatically different concepts, this combination architecture gives rise to complex, domain-specific components that limit network performance and drive up infrastructure costs.

The WiMAX and LTE architectures are flatter (i.e. more distributed, with fewer layers of system elements) than the UMTS data architecture. In WiMAX and LTE there are two levels of components in the bearer path between the BTS / AP and the application core network,

[2] For this illustration: a population density of 1000/km^2 is assumed with a 15% subscriber penetration rate. The spectrum usage is normalized across the technologies.

while in UMTS there are three. In networks designed to support legacy voice services this additional level was useful for a variety of reasons including aggregation of low speed circuit traffic, scheduling of timeslots on circuit transport facilities, and setup/teardown of transport bearer circuits. In a high speed data network however, these circuit-related functions are not needed, and this additional layer accomplishes nothing other than driving up cost and slowing performance. The flatter architecture of LTE and WiMAX also aids in simplifying network operations, as there are fewer different types of components to be managed.

One final characteristic of the high-level architectures of WiMAX and LTE that contributes to their cost advantage over 3G networks is better separation of wireless control functions from data plane functions. By definition, wireless control functions are unique to the wireless domain. Separating the unique wireless control functions from the data plane functions makes it possible to leverage some network elements that can be produced in higher volumes with lower costs. This separation also simplifies network sizing and expansion. For example, wireless control functions are influenced by the frequency with which subscribers enter and leave the network and the speed with which they move around in the network. And, the data plane processing load is determined by the number and size of data packets that are sent between terminal devices and applications servers. The ability to scale wireless control network elements and data plane elements separately in LTE and WiMAX enables operators to size their networks to their exact needs and to focus capacity increases on those functions that are under stress.

WiMAX and LTE will also include specific provisions in support of video broadcast. For example, 3GPP is in the process of defining the Multicast Broadcast Multimedia Service (MBMS) for LTE. In principle, MBMS provides facilities in the network that define sets of base stations over which a given service or media stream should be broadcast. There are provisions to control when a particular service is broadcast including facilities to dynamically enable broadcast based upon user demand in a particular location. IP Multicast will be the network level distribution method to direct content from the content source via a gateway function (MBMS-GW) to specific base stations as required.

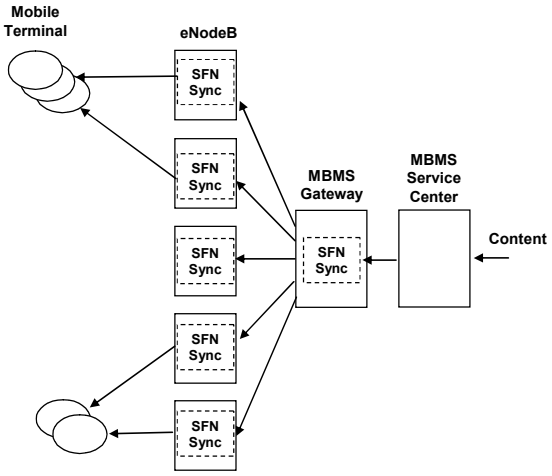


Figure 5. LTE Multicast Broadcast Multimedia Service (MBMS) Operation

MBSFN (Multicast Broadcast Single Frequency Network) further optimizes broadcast operations of the network by synchronizing the delivery of specific content across multiple base stations so that the media can be transmitted simultaneously by each station, as shown in Figure 5. This provides a much improved RF environment and corresponding signal/noise ratio by allowing the mobile terminal to combine reinforcing signals from multiple adjacent base stations.

Mobile Device Advances

Device technologies will continue to improve with the same technology advances that drive the desktop computing environment. Moore's law coupled with substantial improvements in power management and memory density advanced the

functionality available on the handheld computing platforms. Just as the laptop computer became a staple among business professionals and college students, now the handheld mobile device is as well.

The potential for new and expanded video services is also being impacted dramatically by technological advances in the mobile device domain. Advances in materials and Micro Electro-Mechanical Systems (MEMS) technology are spawning a new generation of miniature advanced antennas systems that enable devices to fully exploit the capabilities of LTE and WiMAX. MEMS is also a key to new designs that leverage the DLP™ technology developed for HDTV to provide high-quality, reflective-light displays that are well suited for mobile applications. These new displays promise to provide much improved visibility is high ambient light environments and extremely low power consumption, ameliorating two of the most serious challenges to supporting video applications in a truly mobile environment. 12 GB flash memory cards will be available for mobile devices this year. One card has enough storage to hold 1500 songs, 3600 photos and over 24 hours of video at the same time[3]. One manufacturer has even announced a 120 fpm video capability in one of its mobiles to support slow motion video functionality comparable to today's in-home DVD players. By early 2009, wireless devices are expected to have built-in projectors. These devices have the potential to eliminate the most frequent cited inhibitor to video on mobile devices – their small screen. Finally, advances in lens, flash lighting, memory cards, and other technologies are making certain that the explosion in bandwidth utilization for video will not be a one way street.

[3] Approximation based on 4 minute songs using 128 kbps MP3, pictures taken with 2Mpixel camera and MPEG-4 video at 384 kbps. Pictures and video assume typical compression and resolution.

A FULL SERVICE SOLUTION IS NEEDED

A high-performance wireless access network is important, and generates revenue on its own. But access networks are really just an enabler for providing content and services to consumers. The Internet model has taught consumers to expect universal access to hundreds, if not thousands of valuable applications and content sources. For a network operator to maximize revenue and profitability, it's desirable to participate in every way possible in the business of delivery of content and applications. Two key ways to do this include supporting an open environment in client devices, and establishing unique value added applications.

Open Client Environment

The open client environment is the first tier in the application value chain. This allows 3rd party application providers to create and deploy applications quickly, constrained only by competitive capitalism. Even if these applications are hosted outside of the operator's network, the open client environment benefits the network operator by increasing 'stickiness' since consumers know (and expect) that they can enhance what their device can do, freely and at their own discretion – said another way, they don't need to move to a competitor's network to gain access to new services.

Operator Provided Applications

The 2nd tier up the application value chain is when the network operator provides some applications themselves. Again, these are supported within the same open client environment, and may be developed by 3rd parties or by the operator, but they are hosted within the operator's own network. Now the operator gains revenue not only from the use of the access networks, but also by charging for the services themselves. Of course the goal for an

operator is to identify as many compelling applications as possible and bring them into an integrated environment in their network, rather than passively watching as others gain revenue for those applications.

Over time, consumers have come expect to access applications and content from the Internet as well as from the operator. However, there are still many ways that operator hosted applications can differentiate and more easily provide capabilities that Internet hosted applications can't. Here are some examples:

- Integrating multiple applications by linking data between them
- Sharing user preferences
- Sharing user identity information
- Improving performance through application linkage to QoS enforcement in the access network
- Tailoring content and functionality based on client geographic location *(note that Internet hosted applications can do this too – if the client knows its location.)*

Video (streaming, VOD, and linear) will be a major component of the future service offerings for the mobile and converged service provider. Table 4 shows the results of a survey done by *M:Metrics* showing the percent of users of a given device type who performed a number of popular data activities in the month of January 2008 [6]. There is a clear trend towards significantly increased video usage by consumers who have an easier to use device (and typically a flat-rate data service plan).

Activity	iPhone	Smart-phone*	Market
Any news or info via browser	**84.80%**	**58.20%**	**13.10%**
Accessed web search	**58.60%**	**37.00%**	**6.10%**
Watched mobile TV and/or video	**30.90%**	**14.20%**	**4.60%**
Watched on-demand video or TV programming	**20.90%**	**7.00%**	**1.40%**
Accessed Social Networking Site or Blog	**49.70%**	**19.40%**	**4.20%**
Listened to music on mobile phone	**74.10%**	**27.90%**	**6.70%**

Table 4. Mobile Content Consumption: iPhone, Smartphone and Total Market, January 2008

Video content and applications are most cost effectively delivered by a comprehensive end-to-end network architecture. This architecture should leverage the all-IP nature of video and voice services to provide a converged set of functionality which can support service delivery across both wireline and wireless networks.

CONVERGED VIDEO DELIVERY ARCHITECTURE

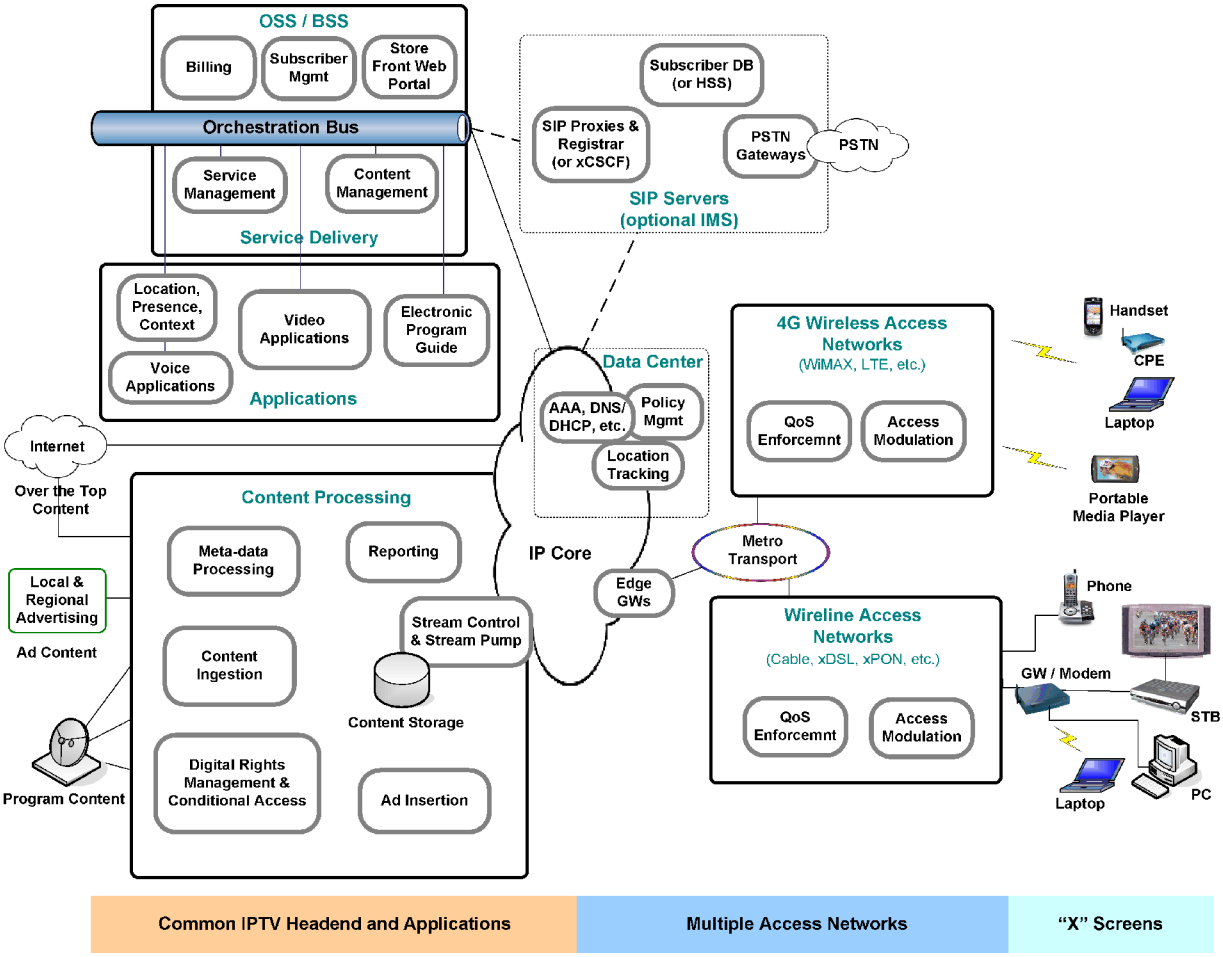


Figure 6. Functional Blocks of the Converged Video Architecture

The primary functional blocks of the converged video architecture are shown in Figure 6.

Content Processing – Includes direct video manipulation such as content ingestion, storage, and stream playout, ad insertion, logo insertion, linear (real time) and off-line encoding / transcoding, and off-line production tools.

Most video content has associated informational data files that are delivered to the system either coincident with or at some time before the content is available. This supporting information is referred to as meta-data. Some examples of information contained in the meta-data include program title and synopsis, length, encoding, points where ads can be inserted, etc.

Linear (real-time) content is ingested, processed, and streamed out immediately. Video on Demand content processing includes delivery of the VOD program (asset) to the system, ingestion of the asset meta-data into the VOD catalog, and storage of the content itself into the VOD stream pump.

Digital Rights Management (DRM/IPRM) and Conditional Access functions include linear and non real-time (offline) encryption, and Key / Certificate Management.

Middleware (not shown) – Middleware typically describes a software layer in the client and network that enables and supports video applications. The open client environment described earlier is embodied as part of the middleware of the solution. Middleware typically supports many or all of the following functions:

- Retrieval of the meta-data information described above from the Electronic Program Guide or store front portal server, and formatting that for display. This includes a list of what video programs are available for viewing, what's being played now, what can be purchased on demand, and what has been stored locally.
- Actions related to viewing, including Picture in Picture, control of video display (play, pause, rewind, fast forward)
- DVR functions such as recording video for later viewing, or starting playback of content already recorded.
- Emergency Alerts and display of desired informational streams
- Content Advisory / Parental Control

Motorola's preferred implementation for middleware is to use a 'thin-client' strategy, such that the centrally managed servers deliver "display ready" user interfaces to the clients, from which the viewer makes a selection. This allows having the same user experience across multiple devices, which will be described in more detail in the next section entitled *Media Mobility*.

Applications – Includes the multitude of application servers that provide operator managed video related services to client devices (as well as voice services, if offered). Some examples of these will be given shortly. Supporting applications, such as Presence, Location/Mapping, and user context services are also ideally integrated here with the other application servers in an orchestration environment. This approach allows for sharing of data and linking of multiple applications together to create more feature rich offerings.

SIP Servers – The converged architecture takes advantage of the capabilities defined within a SIP environment to provide a consistent mechanism for managing control of applications that are naturally session based such as Video on Demand or Voice over IP. Operator implementations will likely rely on SIP for providing converged session control of voice, multi-media, and video streaming services. This convergence allows for rapid deployment of compelling combined services, such as concurrent voice and video (talk to your friend while you both watch the same content, integration of presence for dynamic personalization of content, and so on. Many carrier class operators are expected to use variations of 3GPP's IP Multimedia Subsystem (IMS) architecture, which is SIP based [9], to enable this convergence.

Service Delivery – Includes Content Management functions such as VOD asset management, and the Electronic Program Guide / store front portal servers, as well as Service Management functions such as service bundling and merchandising, access rule management, and mediation functions.

Policy Management & Quality of Service - Includes Quality of Service (QoS) mechanisms which coordinate the assignment of capacity to individual clients based on session needs. The converged architecture proposed here uses information within the Session Description Protocol (SDP) structures of SIP messaging to determine the session needs, thus the SDP contains information on flow rate and the end points which are used to establish a connection

from video source to client. The underlying system is then expected to enforce the requested connection during the session, or to work with the Policy Management and Application servers to adapt the video stream to match revised system capabilities.

Access Networks – In previous sections, this paper has focused on wireless broadband access networks such as LTE or WiMAX. However, the real power of this converged architecture is the applicability of the same video application and content delivery environment to devices in any broadband IP access network, whether wireline or wireless.

Media Mobility™

An important class of applications that this converged architecture enables is what Motorola calls Media Mobility™.

Media Mobility applications enable a commonality of service between all of the content access and display devices used by consumers. These devices include PCs, televisions (and set-top boxes), and mobiles, including phones and PDAs. Drawing on the lessons that content providers have learned through the growth of the web, Media Mobility promises to provide consumers with entirely new forms of video entertainment that combine traditional program styles with emerging social network-driven entertainment modes across a wide range of consumer devices.

Benefits to the Consumer

The converged SIP-based solution offers consumers access to advanced applications that provide a personal video experience. For example, the ability to pause a video stream on one device and pick it up on another allows consumers to seamlessly carry their video with them wherever they choose to view it. As discussed earlier, the simple cache and carry approach limits video mobility to content that has been previously stored, whereas the Motorola Media Mobility solution applies to live broadcast (linear) content as well as personal video sharing among consumers.

The flexibility of the converged applications architecture means that the underlying mechanisms that have been developed to allow video transfer among devices can be readily applied to other applications as well. Presence notifications, news feeds, home caller-id, and many others are all examples of information that could easily have their target context changed from one device to another, as a user wishes.

Benefits to the Operator

A comprehensive network architecture which allows an operator to ubiquitously deliver video and multimedia content to their subscribers offers significant business advantages, such as:

Network efficiencies from a common video headend and converged session control - Operators with fixed and wireless networks can especially benefit from a converged architecture. It is a single video headend delivering content independent of access technology. Significant savings in CAPEX and OPEX are realized by this approach, as opposed to deploying individual independent video delivery solutions for each.

The SIP mechanisms used to move the content from one device to another also enable the network to be aware of what is being viewed, by whom, and on what device. The Motorola solution that supports the gathering of these kinds of information, including service brokering and service orchestration, is being designed as a modular architecture that allows for easy integration into existing IMS environments.

Ubiquitous access and meeting increased expectations - By having ‘always on’ access regardless of location, there are more

opportunities created for pay-per content access, or advertising 'eyeballs' reached. Additional ARPU should be possible simply by providing consistent service across multiple environments. For example, consumers clearly find YouTube to be an interesting way of consuming video at their PC and this is already becoming desirable in a mobile environment (it is reported that in 2007, YouTube video generated more Internet traffic in the United States than all of the Internet traffic combined, worldwide, in the year 2000). As consumer expectations grow, it becomes a competitive imperative for a carrier to meet the common denominator.

Taking advantage of location – Targeted advertising is possible based upon an individual's past behaviors as well as the content that is currently available (e.g., when a particular piece of media is about to be distributed by a MBMS broadcast, the user can be alerted). With feedback of current subscriber location information back into the video headend, it is possible to customize individual streams (VOD for example) based upon a particular user's location. This enables location based advertising with a fine degree of granularity (e.g., for individual shopping malls, restaurants, events, etc.).

Additional public services – Enhanced emergency communications services are also possible such as broadcast alerts and/or updates on surrounding context (e.g., traffic, etc.) based upon location. Minimally, this becomes a crucial part of a user's bundle and dependency on the mobile media environment. Many of these information services can be subscription based. It is also possible that public carriers may be able to "wholesale" selected broadband capabilities to public safety agencies.[4]

[4] Note that the FCC's recent attempt to formalize this with the D-block spectrum auction failed in that no carrier was willing to accept the business terms associated with operating in this spectrum.

Stickiness – Providing consistent access, management of preferences, as well as look and feel (e.g., EPG, storefront) to the user in a mobile environment as well as the home/fixed locations increases the stickiness of the subscriber to the carrier.

THE FUTURE – EMERGING SERVICE TRENDS

Many services and capabilities that appear in wireline networks (i.e. the Internet) tend to make their way to wireless, gated by the ability of the wireless access technologies to support them. 4G wireless technologies are clearly able to support most activities people do with the Internet. Also, in the past there tended to be a slower pace of client development for wireless. However, given the inevitable trend toward web based applications and open wireless client environments discussed earlier, as well as the convergence of networks, this development gap is also expected to disappear.

A few examples of new and emerging video related applications that are likely to be popular among wireless consumers are:

Social TV – This is a range of enhancements to today's one-way video viewing experience which will provide people with the ability to share their viewing experience, regardless of location. It is envisioned to include things like integration of presence so you see on your screen not only which of your "buddies" are online, but also what they're watching, and give an easy ability to switch to that content yourself. Unlike today's passive experience, the ability to have a voice or text conversation with your buddies while viewing the content, or sending your rating on whether you like it or not, is supported by the device and 4G wireless network being inherently bi-directional.

See-What-I-See – This is a real-time linking of people's visual experience. One-to-one video telephony is a basic form of this. However, it can be extended to share video among a small group, such as sharing what I see with my fellow construction or firefighting team. Or it might be offered as a fee based service for enhancing the viewing of public events, such as broadcasting point-of-view video from a sports star to fans who have subscribed to it.

User Generated Content – Revenues for the total user generated content market have been estimated to grow between 66% and 99% per year on average over the next 5 years [4]. The success of social networking and blogging sites reflects the desire for people to make, publish, and view their own video content. The ability to do this from a wireless device will accelerate the quantity of content produced. Video blogging will become an extension to today's simple video file upload, giving consumers the ability to easily create multi-media narratives for public or private viewing. Revenue opportunities exist all along the user generated content value chain, from the creation of easy to use clients for creating the content, to the application servers which store catalog, and transmit it.

Access to Personal Content – Mobile wireless enables access to personal content libraries from anywhere. A potential future revenue opportunity for the operator is to provide an easy to use hosting service for storing and managing personal content (perhaps for a flat fee or ad based), as well as integration of a network based Personal Video Recorder (PVR) capability.

Peer-to-Peer Sharing – Like in the wired Internet, peer-to-peer sharing of content among wireless users will likely grow to dominate traffic on the 4G network.

When examining these new and emerging applications, we find an interesting trend that will have significant impact on wireless networks: most of them require vastly more uplink bandwidth (from client to network, or from client to network to client) than has been needed in the past. Earlier we discussed the improved uplink capacity of 4G wireless technologies over previous generations. This will help to ameliorate the network cost impacts of widespread uptake of these new services. However, since the uplink is the limiting direction for wireless, its utilization will need to remain a parameter that is closely watched.

CONCLUSION

With the advent of 4G wireless technologies, operators will have the flexibility and network robustness to economically deliver a truly mobile video and multimedia experience. WiMAX and LTE offer significant performance and economic advances over their 2G and 3G predecessors. These networks are based on OFDMA radio technology and use a flat, IP-based network architecture. WiMAX deployments are well underway and their pace continues to increase. LTE will follow, and bring with it even more advances and capabilities that are applicable to video delivery (e.g., MBMS). Mobile device technology has advanced commensurately to provide a reasonable platform for mobile multimedia.

To best exploit the opportunities for revenue that these Radio Access Network technologies enable, a comprehensive end-to-end service and content delivery architecture is essential, and is of particular benefit in a converged wireline/wireless environment. In addition to the radio access infrastructure, the architecture needs to include the following:

- An open client environment enabling consumer installation of new services and applications, which are deployed either by

the network operator, or by 3rd parties in the Internet.

- Content Processing: Encoding and Device Adaptation, Ingest and Storage, Metadata processing, DRM.
- Middleware: Consistent implementation among all clients and network servers of key supporting capabilities such as EPG, content merchandising mechanisms, and interactive features.
- SIP Session Management: Consistent orchestration of multimedia sessions.
- Service Delivery: Management of application introduction, deployment, and charging for services.
- Policy and QoS: Determination of a session's QoS needs and entitlements, and orchestration of this across the end-to-end network via content selection/adaptation, bearer flow establishment, etc.

A unified mobile media solution offers many carrier revenue generation opportunities including ubiquitous always-on access to existing services, location based derivatives such as location based advertising, public services, and increased dependency (stickiness) on existing operator services and content.

REFERENCES

[1] Robert E. Kahn, Steven A Gronemeyer, Jerry Burchfiel and Ronald C. Kunzelman, Advances in Packet Radio Technology, Proceedings of the IEEE Vol 66, No 11 November 1978.

[2] Motorola Strategy and Business Development, internal analysis.

[3] Bandwidth evolution prediction by Motorola, extrapolated from publications and presentations from Arther D. Little, Motorola, Ericsson, and Alcatel.

[4] User Generated Content--More than Just Watching the YouTube and Hangin' in MySpace, In-Stat, http://www.instat.com/Abstract.asp?ID=212&SKU=IN0602976CM, September 2006.

[5] I-Mode Overview. MobileInfo.com, http://www.mobileinfo.com/imode/index.htm.

[6] M:Metrics: iPhone Hype Holds Up, http://www.mmetrics.com/press/PressRelease.aspx?article=20080318-iphonehype, March 2008.

[7] 3GPP TS 23.401 [3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access (Release 8)].

[8] 3GPP TS 23.402 [3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Architecture Enhancements for non-3GPP Access (Release 8)].

[9] 3GPP TS 23.228 [3rd Generation Pertnership Project; Technical Specification Group Services and System Aspects; IP Multimedia Subsystem (IMS); Stage 2].

[10] Long Term Evolution (LTE): A Technical Overview, http://www.motorola.com/mot/doc/6/6834_MotDoc.pdf, Motorola, 2007.

[11] F. Wang, A. Ghosh, C. Sankaran and P. Fleming WiMAX Overview and System Performance, *IEEE Vehicular Technology Conference (VTC)*, Sept. 2006.

[12] F. Wang, A. Ghosh, C. Sankaran and S. Benes, "WiMAX System Performance with Multiple Transmit and Multiple Receive Antennas", *IEEE Vehicular Technology Conference (VTC)*, April. 2007.

[13] Y. Sun, W. Xiao, R. Love, K. Stewart, A. Ghosh, R. Ratasuk, B. Classon, Multi-user Scheduling for OFDM Downlink with Limited Feedback for Evolved UTRA, IEEE 64th VTC Conf. Fall 2006.

ACKNOWLEDGMENTS

The authors wish to thank the following people for their valuable input to this paper:

Kishore Albal
Pete Armbruster
Samantha Buechele
Herb Calhoun
Amitava Ghosh
John Harris
Marie-José Montpetit
Mike Needham

FEMTOCELLS—THE GATEWAY TO THE HOME

Sheriff Popoola
Senior Manager, Product Line Management
Motorola Connected Home Solutions

Abstract

Femtocells are small, low-cost cellular base stations optimised for use in the home and small businesses. This paper discusses this exciting new market and concludes that femtocell and Wi-Fi technology will be co-existing, rather than competing, to deliver a comprehensive digital home experience.

It describes how femtocells will enhance the delivery of telecommunications services in the home and the new possibilities arising from the integration of femtocells with home gateways and set-tops. It will also point out technical challenges cable operators must assess and outline the opportunities for cable operators complementing cable access infrastructure with femtocells to enhance market share and customer retention through enhanced triple play and quad play services.

A FEMTOCELL OVERVIEW

Femtocells—miniature cellular base stations that connect via cable infrastructure to provide enhanced 3G signal within the home – represent arguably the most exciting development in home networking since the arrival of Wi-Fi®.

Interest in femtocell technology is reflected by growing activity among telecom operators and hardware vendors alike. Research firm IDC, predicts that spending on femtocell-enabled services will grow to $900 million by 2011.

The unique demands of a high-performing femtocell ecosystem demands competence in several key areas; a combination of expertise not previously required from a single cellular infrastructure product.

- It must perform in a hostile RF environment.
- It must meet the high expectations of a mature cellular subscriber base.
- It must integrate seamlessly with existing HFC access networks.
- It must be capable of being deployed and supported in high volume.
- It must extract maximum performance from HFC backhaul and it must be capable of being remotely managed without excessive operator effort.

These *"must haves"* attributes demand a skillset that cable operators can rely on for delivering end-to-end femtocell solutions that increase Average Revenue Per User (ARPU), grow market share, and enable innovative partnerships with wireless service providers that increase brand value and allow cable operators to develop innovative triple play and quad play services.

STANDARDS INTEGRATION

Femtocell technology is very new and developments are fast moving and exciting. As a consequence, product development is far ahead of standardization. Cable operators require that new femtocell deployments rely on industry standards and enable smooth integration and the

ability to deliver seamless mobility of voice and data services.

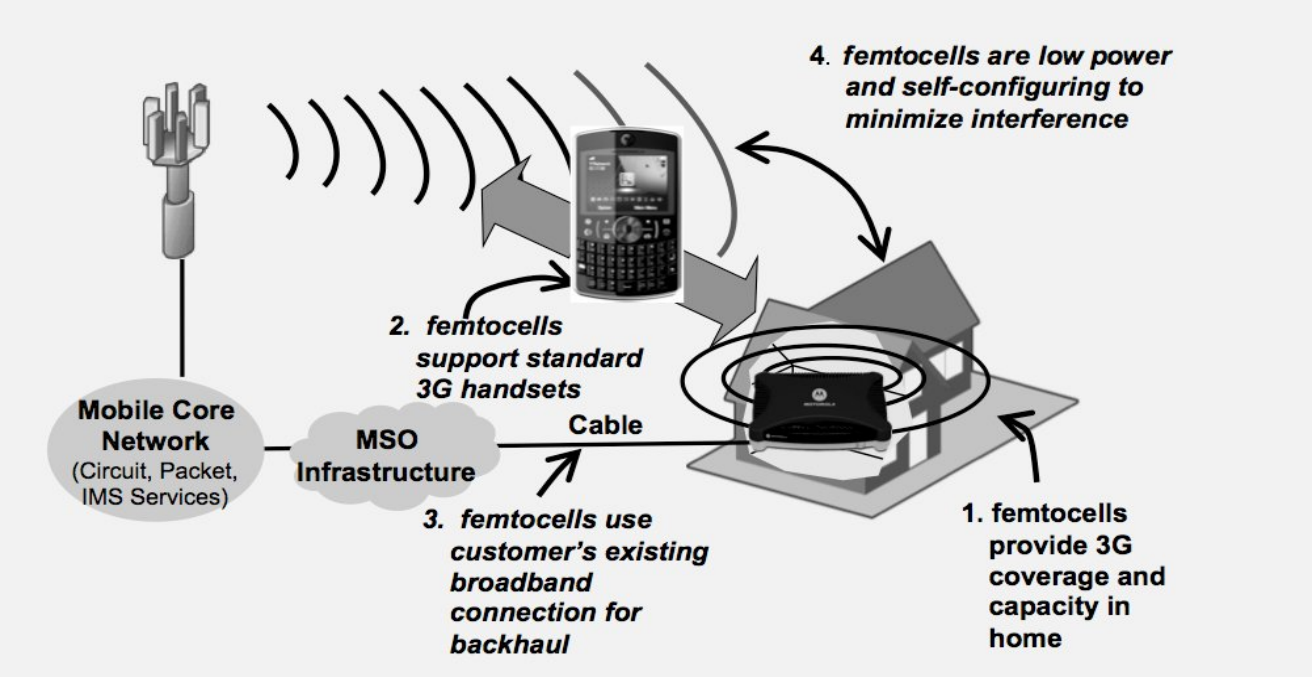
If the femtocell market is to grow as expected, the evolution of common standards is the best way to achieve this. The interface between the femtocell Customer Premises Equipment (CPE) and the femtocell aggregator (sitting in the operator's headend) is key to interoperability.

This is also the interface that's most open to interpretation, so it is important to monitor the evolution of industry standards. Motorola is very active in the standards bodies to move this debate forward but in the meantime is supporting two subtly different versions of the CPE-aggregator interface within the company's femtocell product portfolio. Femtocell trials are being conducted this year, with commercial launches expected to begin in late 2008 or 2009.

THE NEED FOR FEMTOCELL DEPLOYMENTS

Femtocells will deliver home broadband communication like never before by enabling personal devices to perform seamlessly in and out of the home. Connection to the mobile network via a gateway over an existing HFC connection, femtocell solutions make indoor coverage for mobile communications truly pervasive while delivering additional benefits to both the cable operator and the end-user.

A femtocell is a small, low-power, self-installed cellular base station optimized to deliver cost-effective coverage in the home and small office environment. Once installed in a customer's home, it enables the operator to provide higher-quality and higher-performance wireless voice and real-time data services to their customers inside their homes. Today, 3G is the focus for femtocell technology. The full 3G service set can be delivered in the home from a small stylish device, which is connected to the mobile operator's core network using open 3GPP based standards via the consumer's HFC connection



Femtocell products require extensive experience in collapsed architectures, RF techniques, home gateways, CPE management, and fixed-mobile solutions to enable cable operators to launch new services with fast time-to market and low risk. Femtocells are aimed at providing high-performance 3G voice and data communications in and around the immediate home environment. Connected to the operator's mobile network over existing broadband connections in the home, femtocells have the potential to make indoor coverage for mobile communications truly pervasive while delivering additional benefits to both the operator and end-user.

The femtocell network architecture provides operators with a complete indoor coverage solution, all in a small, low-cost, low-power, easy-to-install base station that can be seamlessly integrated into existing mobile networks and provisioned for service within minutes of switching it on. Such platforms enable a host of new applications and revenue opportunities, and provide cable operators with the means to prevent the loss of subscribers to carriers offering bundles of wired and wireless access services.

Femtocell allows cable operators to aggressively enter the fixed/mobile convergence market. Femtocell deployments will address the driving need for seamless mobility of voice and data service.

ADDRESSING SUBSCRIBER PAIN POINTS

Every consumer of wireless voice services has had the experience of being on an important wireless phone call and losing their cell signal. In light of that consumer pain, femtocell technology is arguably one of the most exciting developments in home networking since the arrival of Wi-Fi – both are enabling operators to better meet consumer demands for seamless connectivity. The low-power, wireless femtocell access points are optimized for use in the home and small businesses, connecting via broadband to provide an enhanced 3G signal within the home.

Femtocells—which may look like a stand-alone consumer device sitting on a kitchen counter—actually function as part of the provider's network infrastructure. Consumers primarily use their mobile phones at home, even when they have a fixed-line telephone. People—especially younger consumers—have become accustomed to the mobility *convenience*s of a single communication device.

Additionally, more people would prefer one number and one device to handle all their communications needs, whether they are in the home or at work or play. An October 2007 survey commissioned by mobile content backup services provider, FusionOne, Inc., found that more than half of respondents indicated that their social lives would "suffer" if their mobile phone were to go missing.

Most wireless operators agree that a significant proportion of all calls made from mobile phones are initiated indoors, so it becomes understandable why providing good indoor coverage is essential to provisioning cost-effective, high-quality and higher-performance wireless voice and data services to consumers.

Allowing subscribers at home to connect to the wireless operator's mobile network over existing HFC infrastructure allows cable operators to make mobile communications truly pervasive, creating long-term bonds with subscribers that minimize churn and enabling new revenue opportunities from bundles mobile and HFC service packages.

FEMTOCELL BENEFITS FOR THE CABLE OPERATOR

Femtocells enable cable operators to provide higher-quality and higher-performance wireless voice and real-time data services to their residential and small home office customers.

They will be able to offer subscribers high-quality 3G services at lower costs while they are in their homes.

In addition, they enable a lower cost of delivery of wireless traffic in comparison to the macro cell network. Femtocells can be used as part of integrated triple or quad play services, which meet consumer communication needs—increasing the competitiveness and customer retention for the cable operator.

The integration of cable and femtocell technologies will allow cable operators to fend off attacks from carriers, create longer-lasting relationships with subscribers, and drive new revenue growth by offering attractive seamless mobility services.

Femtocells have an important role to play in driving premium mobile service adoption, finally turning the 3G service vision into a reality by encouraging a culture of usage through low-cost high-performance mobile data services.

FEMTOCELL BENEFITS FOR THE CONSUMER

For consumers, the benefits of femtocells include:

- A seamless communication experience as they roam from inside to outside their homes.
- Greater convenience via effective fixed-mobile substitution by removing the need for users to have separate home phones and offering the flexibility for consumers to rely on a single phone for access on the road or at home.
- Reduced in-home call charges.
- Excellent indoor coverage.
- Lower-cost voice calls from within the home.
- Consolidated billing for voice and data services.
- The convenience of using a mobile handset with its personal phonebook and other cool handset features, without the concerns of poor call quality or additional cost.

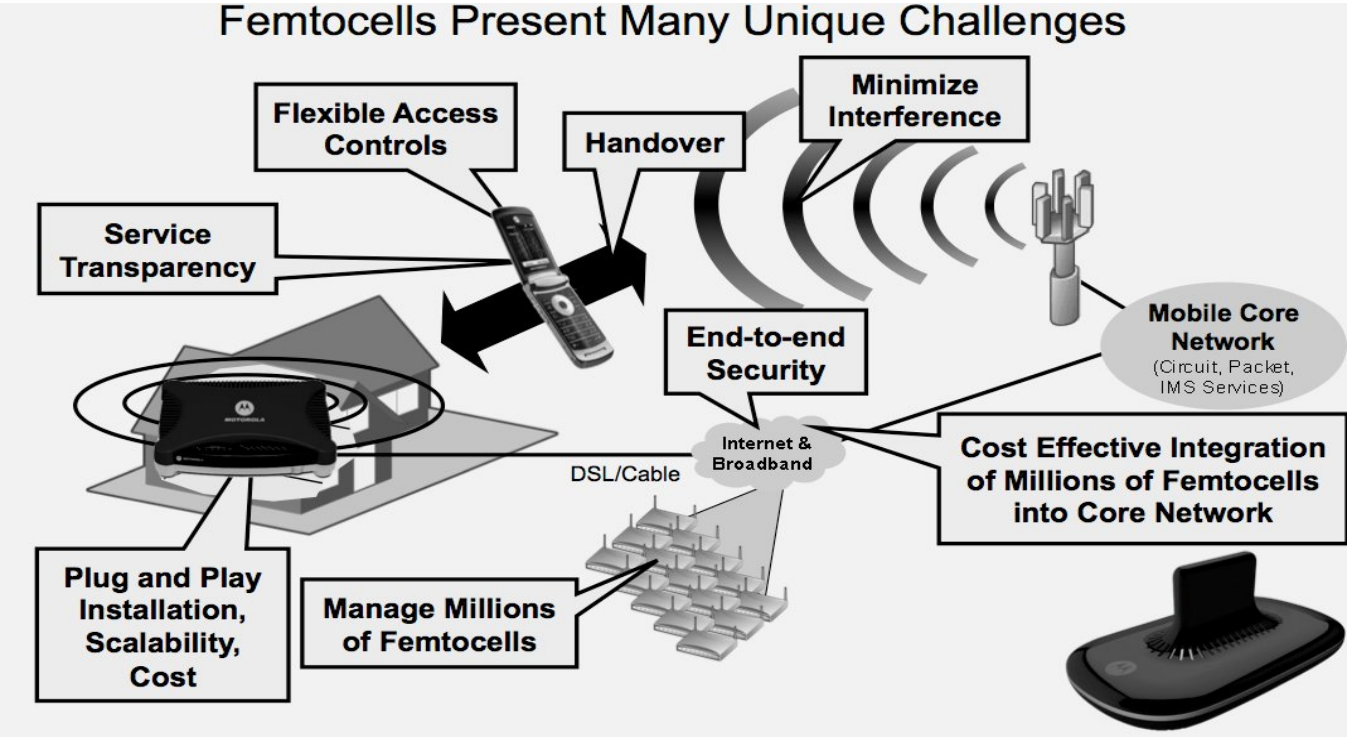
ADDRESSING TECHNOLOGY CHALLENGES

The successful deployment of femtocell technology can bring great rewards, but require that cable operators overcome diverse technology challenges.

Operators face significant challenges with the deployment of new technology, new applications and the ever-increasing usage demands placed upon mobile networks. At the forefront of these challenges is making 3G coverage as *"near ubiquitous"* as practically possible, both outdoors and indoors.

The traditional 3GPP 3G network architecture, made up of numerous macro base-stations, with its centralized RNC function and ATM backhaul was designed to provide wide-area coverage. It was not however designed to scale, physically or economically, to provide effective coverage for individual indoor/residential situations.

Cable operators can make a compelling value proposition for partnering with mobile operators. It is common knowledge within the mobile telecommunications industry that the use of outdoor macro-cells to provide indoor/residential coverage quite simply does not provide an effective solution, from both a coverage and economic perspective. It also impacts practical matters, such as site acquisition, which is becoming increasingly problematic.



Not only is site acquisition costly, forming a major proportion of network build-out costs, it is also very time consuming with many local authorities closely regulating the sale and usage of potential cell-sites. Assuming suitable sites can be acquired, increasing cell-site density through the use of smaller cells, may not overcome all coverage issues but will lead to increased backhaul costs and other practicality issues.

People are becoming increasing reliant on their handset device to the extent that it forms "part of their identity"; similarly more and more people would prefer one number and one device to handle all their communication needs be it in the home or elsewhere. Many end-users prefer to use their mobile phone when in the home, even where a fixed-line telephone is available. People have become accustomed to and take for granted the convenience that the mobile phone provides in terms of mobility and in having a single device to communicate that includes their contacts and even takes and stores their messages in a variety of formats.

3G signals, operating at very high frequencies and high bandwidths have a poor ability to penetrate through structures. This often leads to service quality and service experiences that do not meet end-user expectations and can lead to dissatisfaction, reduced minutes of use, increased customer churn and ultimately, lost revenues.

Most end-users of 3G services invariably have to settle for the coverage provided by the macro base-station serving their location at that point in time, whether stationary out in the world, in a building or while on the move. The issues associated with providing coverage for indoor situations from macro base-stations are well known; 3G and buildings, or to be precise their fabric, are inherently not a good mix.

Since late 2006, interest in femtocell solutions has increased to the extent that most industry analysts suggest femtocell deployment will become widespread in the coming years.

FEMTOCELL CONSIDERATIONS AND CHARACTERISTICS

Having considered the major drivers for femtocell deployment, this section looks at some of their practical aspects. Femtocells overcome the issue of providing effective indoor coverage from the 3G-macro layer by their placement in the end-users' homes.

Once installed in an end-user's home a femtocell will enable the cable operator to provide higher-quality and higher-performance wireless voice and 3G data services in and around the immediate vicinity of the home environment.

Femtocell products are in many ways similar to Wi-Fi access points in that they enable access through an unobtrusive device; however femtocells enable full 3G service delivery in the home. Similar in size to a cable modem, a femtocell is a low-capacity base-station, radiating only sufficient power to cover the area of a home environment. The femtocell connects to the operator's core network using open 3GPP based standards through the end-user's household broadband Internet connection rather than traditional cellular backhaul methods. Accordingly femtocells must also fulfill a number of other criteria:

Low-impact—Space may be limited for some households. As a result femtocells must be physically small, aesthetically pleasing and easy to position. Furthermore, they should also be silent in operation, generate low levels of heat output, and be inexpensive to run in terms of on-going electricity costs.

Low RF power—The transmit RF power output of femtocells is low, typically less than 10 mW. Put in perspective, this is a lower power level than many Wi-Fi access points, which transmit at 100 mW of output power. Additionally, by being close to the femtocell the 3G handset itself is able to transmit at lower power levels than it might otherwise have to when on the macro network.

Capacity—Femtocells are aimed at delivering dedicated 3G coverage to a household and in doing so can provide a very good end-user experience within the home environment. As a result, femtocells have a design *"capacity"* of up to 20 registered users and 4 simultaneously active calls.

Low-cost—There is significant competition for access solutions in the home space. Wi-Fi is commonplace, and easy to install/configure. Femtocell platforms in the home should compare favorably with Wi-Fi base stations in cost and performance.

Low-power consumption—Clearly if the end-user is to foot the bill for the electrical energy consumed by the femtocell base-station then this figure must be low enough not to raise concerns as to its impact on the fuel bill.

Easy end-user installation—Like cable modems and DSL routers, femtocells will be installed by consumers and activated through service providers. This means that the cable operator will not have to employ installation teams or have a truck-roll every time a new femtocell is deployed. From the end-user perspective the unit must be a simple *"plug and play"* installation with a minimal amount of intervention required.

Interference—The use of femtocells in spectrum also currently used by the macro layer may, if not managed correctly, give rise to issues with interference between cells; macro with femtocell and in the instance of close proximity of two or more units, femtocell with femtocell. Operators will likely want to launch femtocells on the same channel as macro cell networks for capacity reasons.

Handovers—Current macro RF planning techniques are inappropriate for femtocells

because of the sheer potential numbers of femtocells. Also the potential to "ping-pong" between layers, especially as an end-user moves around the home and enters into areas where the signal strength from the macro-cell is greater than that of the femtocell, must be considered very carefully to ensure that the networks provide the best overall coverage without issue. Femtocells introduce new complexities in macro to Femtocell, Femtocell to macro, and Femtocell to Femtocell handover scenarios

Security—Given the requirements for low-cost and easy installation, the use of the broadband Internet as the network interface becomes very easy to understand. However this raises security risks in that broadband Internet has open access. There are various approaches to address this issue including the embedding of the interface within the IP signaling itself while network security is managed by the IP security (IPSec) protocol.

Worldwide cellular network standards–Understandably femtocell products are likely to appeal to many end-users around the world. As a result differing models will be developed and offered to satisfy the various needs from the different regions. Products should offer support for their respective and existing (3GPP) UMTS and (3GPP2) CDMA standards, as well as emerging standards such as Imax, UMB and LTE.

Support for existing 3G handsets and devices—Support for existing handsets and devices is a very important consideration for the end-user and operator alike. In each technology market, femtocells will support existing handsets and devices, further helping to drive uptake of 3G services and femtocells in particular.

Operator control—Femtocells operate in licensed spectrum and as such cable operators must ensure that they comply with regulatory requirements. Femtocells need to feature client software that enables remote configuration and monitoring via a centralized Operations, Administration, Management, and Provisioning (OAM&P) system.

New services and applications—Femtocells are likely to become an integral part of managing all communications in and out of the home environment. They will enable cable operators to cost-effectively offer in-home pricing and integrate mobile services into triple-play / quad-play service offerings. Femtocell architectures will need to include provisioning for a complex service environment on which applications may be added, thereby facilitating new revenue opportunities.

Service Assurance—Remote management is needed to enable an operator to provide the end-user quality of service needed at the edge of the network.

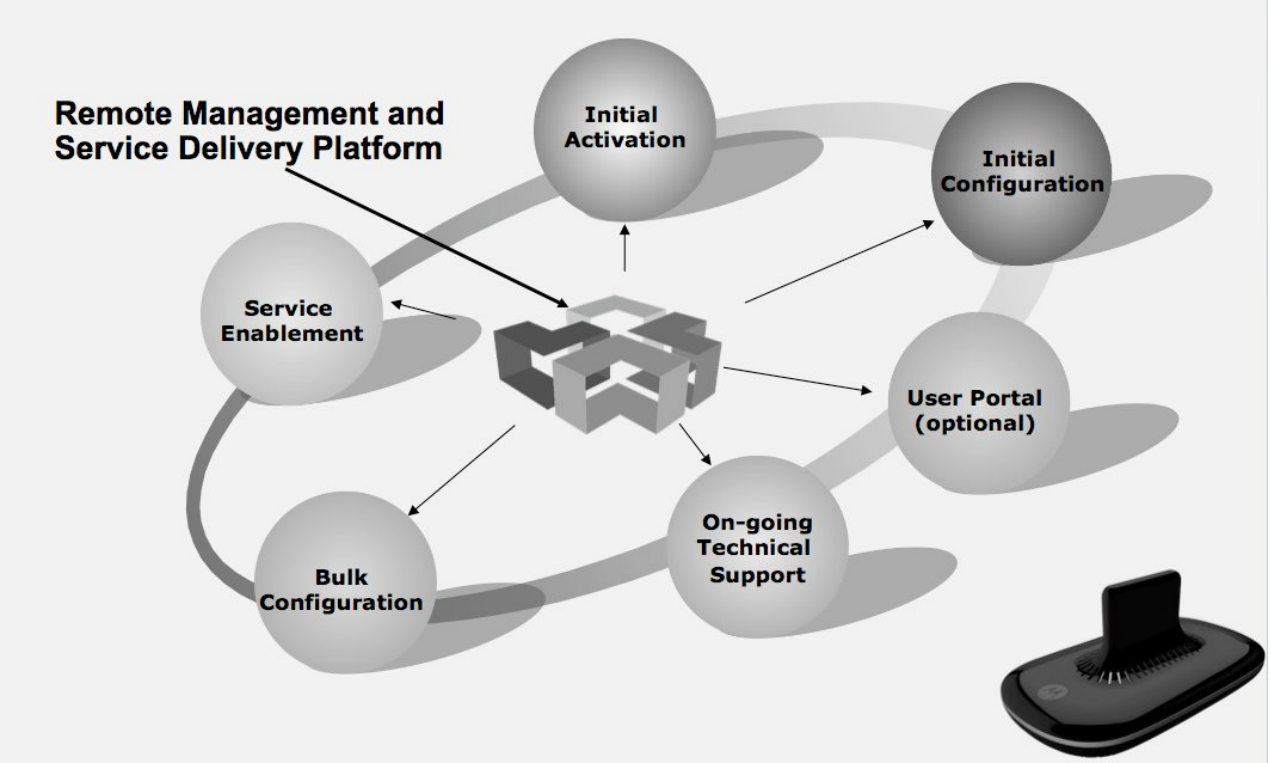
INDUSTRY TRIALS

Cable operators can turn to the following URLs for up-to-date information on femtocell technology:

- www.femtoforum.org
- www.motorola.com/femtocell

While the previously mentioned criteria and challenges are being addressed throughout the industry, femtocell testing and deployment continue to happen around the world. Selecting equipment vendors with experience in femtocell trials is essential, and trials are now underway primarily in Europe and North America.

These developments and the industry's early groundwork are leading toward the realization that as technical and commercial challenges are resolved, a femtocell in every home could become a reality.



FEMTOCELL MANAGEMENT

Cable operators need software platforms that will allow them to remotely access, configure, and troubleshoot the full portfolio of consumer devices – including mobile phones, CPE and femtocells. This helps to lower operators' costs by reducing truck rolls and lowering operational expenses. It also helps to increase revenues by accelerating new service introduction.

Scalable, carrier-grade systems are needed that can manages devices, home networks, and services. MSOs deploying femtocell solutions will need to an excellent end-user experience, and they need centralized OAM&P systems that enable the efficient provisioning, management and operations of femtocell solutions. Centralized control is essential so that cable operators can deliver an excellent user experience while minimizing support costs and swiftly generating revenues from mobile services.

MSOs need the ability to centrally automate service provisioning, upgrade femtocell platforms at customer locations and activate and support residential subscribers without bearing the burden of unnecessary truck rolls.

END-TO-END FEMTOCELL SOLUTIONS

Cable operators implementing femtocell connected home solutions can deliver highly differentiated services that drive revenue growth and lead to longer-term relationships with subscribers. Considering that the majority of mobile calls originate in the home and end-users prefer to use a single handset, cable operators can now turn to emerging solutions available to them that overcome the issues of poor in-building coverage.

Until now, providing good mobile coverage for homes has largely been overlooked. That is changing. Femtocells will provide a one-box solution: a small, low-cost, low power unit that can be self-installed to provide mobile 3G coverage to the home.

For the end-user femtocell solutions will provide dedicated and reliable mobile 3G

coverage in the home with opportunities for preferential tariffs. For the cable operator, femtocells deliver cost-effective coverage and new revenue and customer satisfaction opportunities.

Once in the home, the femtocell is likely to encourage end-users to use their mobile as their single communications device irrespective of their location. Femtocell solutions are also likely to increase minutes of use and ARPU and also open up brand-new revenue streams for cable operators through the integration of mobile services into triple-play and quad-play service offerings. Cable operators need to minimize the risk of deploying new technologies by relying on vendors that already have proven expertise in the many technology areas required for successful femtocell deployment.

Solutions however are not simply about only technology; they are about capabilities and delivery. That's why cable operators need access to expert professional services to deploy femtocell solutions that leverage existing wireless standards.

End-to-end femtocell solutions will allow cable operators to differentiate their service offerings by providing seamless mobility and allowing subscribers to access voice and data services as they move throughout the connected home.

While femtocell technology offers great promise, selecting equipment from the right vendor is key to successfully launching new services. Cable operators should rely on fully integrated and tested end-to-end solutions based on open standards, which includes:

- A range of low-cost, easy to deploy CPE.
- A core network concentrator.
- A centralized management and provisioning system.

By relying on proven technologies already tested in femtocell trials, cable operators can safely explore the value of seamless mobility by offering integrated voice and data services available throughout the home over wireless cellphones and PDAs.

Connected home solutions that leverage emerging femtocell technology will enhance the user experience, allowing the operator to increase ARPU through better home coverage and new multimedia and location-based applications. Selecting vendors experienced in delivering high-volume CPE for both mobile and fixed networks is crucial, and the ability to develop and manage end-to-end femtocell solutions will allow cable operators to prosper by enabling the future of the connected home.

ABOUT THE AUTHOR

Sheriff Popoola is Senior Manager, Product Line Management with product responsibility for Motorola's Femtocell CPE offering covering UMTS, GSM, xDSL and cable technologies. He has been with Motorola for 13 years, and his experience includes Networks in Public Safety, the Japan CDMA group, Advanced Radio Technology group, iDEN Dispatch and Packet Data services group, and the Aspira Network Products group, in functional areas spanning Systems Engineering, RF Development Engineering, Digital Hardware Development Engineering (where he received a patent in CDMA), and Product Line Management.

Prior to joining Motorola Popoola worked for Cellutech Communications as the Cellular Service Manager running the Chicago-North facility. He has a BSEEE degree from Obafemi Awolowo University in Nigeria, a MSEECS degree from University of Illinois, and an MBA degree from Northwestern University's Kellogg Graduate School of Management. He can be reached at Sheriff.Popoola@motorola.com.

FUTURE DIRECTIONS IN CABLE BROADBAND BANDWIDTH CAPACITY

John M. Ulm
Motorola, Home & Networks Mobility Technology Office

Abstract

This paper takes a long term look at cable's broadband bandwidth needs over a 10+ year horizon. It discusses several drivers that forecast bandwidth needs down the road. We compare HFC potential against FTTP architectures and shows that there is plenty of capacity left in coax to compete with GPON or even co-exist with it. HFC also has the significant advantage in that it can incrementally expand bandwidth through this entire process without requiring a massive infrastructure overhaul like pure FTTP providers.

How far can the MSOs take coax looking way out on the 10+ year horizon? With the recent evaluations of the RF overlay systems, we take a closer look at the theoretical maximum capacity of coax. The paper discusses how HFC can offer 1+ Gbps services over coax and what is needed beyond our current DOCSIS systems.

Introduction

HFC has long enjoyed the position as the leader in providing broadband content to the home for both video & high speed data. Now it's being besieged by FTTP and satellite HD technologies and many have claimed its days are numbered. Our analysis shows that there is plenty of capacity left in coax to compete with GPON and even 10G PON technologies, given the appropriate investments. HFC also has the significant advantage in that it can incrementally expand bandwidth through this entire process without requiring a massive infrastructure overhaul like FTTP.

DRIVING BANDWIDTH NEEDS

This section discusses several drivers to forecast bandwidth needs down the road. Some of these may appear to conflict with each other, but each gives a unique perspective on bandwidth needs.

Moore's Law continues

Over the years, Moore's Law has driven data bandwidth growth and it looks to continue for the foreseeable future. Simply stated, Moore's Law is the doubling of transistors per device every 18-24 months. This increased technology capacity has resulted in a corresponding improved performance, density and power. Some studies have shown that high speed data broadband service offerings have closely tracked Moore's Law over the last 8-10 years. The implication that it will continue to track means that we'll need to offer 100 times today's bandwidth in another 10-15 years. With broadband providers offering data services around 10Mbps today, this means that the broadband providers should be prepared to make 1 Gbps services generally available in 10-15 years.

There are several aspects of data bandwidth growth that need to be examined. In Andrew Odlyzko's paper "Internet traffic growth: Sources and implications," he contends that the data traffic growth will continue to follow Moore's Law and is primarily driven by file transfers as opposed to streaming traffic like video. As bandwidth growth is modeled, it may be necessary to separate traffic that is streamed and requires constant bit rate service from the more

general, best effort high speed data traffic that is following the Moore's Law growth.

But there are some questions with this argument. Will Moore's Law continue forever? Will technology improvements continue in the density and power areas but less so in the performance area? We may not know the answer for another decade, but must prepare in case it does continue to track.

Supply Side Economics

Is the right question to ask: "What is the demand for bandwidth?" Some insist it is not and instead we must look at the bandwidth supply rather than the demand. Internet demand is (nearly) infinite and users have been shown to try and consume everything given to them. So the operators should assume that any high speed data offering could probably be 100% utilized.

Instead of looking at demand, the right question should be "How much bandwidth should I supply?" And the answer to that is based primarily on competition. In the early days, cable modems were competing with dial-up services and only needed to offer 1Mbps service, even though the channel supported almost 40 Mbps. As DSL penetration increased, cable ratcheted up its data rates to maintain a sufficient performance edge. As FTTP penetrations like Verizon's FIOS increase, competitors will be able to flip the tables on cable and it will be up to cable to race and keep up with FTTP.

The difference between FTTP and HFC is most notable in the upstream direction. Since the HFC upstream capacity is significantly less than the downstream, competitors could cause an immediate disruption in cable operator's business by emphasizing services that use significantly more upstream capacity. We are just starting to see this in current marketing efforts.

Video dominates & HD becomes Mainstream

Among Video, Voice and Data services, by far the one with the largest bandwidth requirements is video. Not only does video require a high data rate of multiple Mbps, it streams for extremely long periods of time, maybe even several hours. We are also reaching the point where High Definition (HD) is becoming main stream. HD has a significant impact on bandwidth requirements, increasing the video bandwidth by a factor of 2-4X over Standard Definition (SD) video.

Many indications over the last year have shown that HD adoption is accelerating. As HD content rolls out, it will cause a significant impact on short term bandwidth needs. But what are the longer term impacts once the majority of video content is delivered as HD? To deliver one or two narrowcast (i.e. personalized) HD streams along with a couple of SD streams to each subscriber would require ~20Mbps per home using MPEG-4.

However, this may actually be understating the bandwidth needs. As HD penetration becomes truly pervasive, then we may need to support 3-4 HDTV's per home with a couple extra HD streams going to a DVR for recording. On top of this, some of these streams may also be high quality 1080p HD content requiring additional bandwidth. This longer term scenario would require closer to 40-50Mbps per home to support streaming HD video. The bottom line is that Service Providers will need to continue to offer more bandwidth capacity to the home as the percentage of HD content increases along with the percentage of HDTV penetration.

Burst speeds vs. Sustained rates

Throughout computer history, data networks have been useful because of the nature of statistical multiplexing. By offering a shared resource with high burst rates, users get the impression that they have high data

rate services. In reality, data usage is very bursty and average data rates are significantly lower than the burst rate.

An example broadband system might be engineered today to provide 1% concurrency for a 10Mbps data service. This means that during peak busy hour each user will get an average of 100Kbps. As network speeds increase, the statistical gain also increases. This implies that the concurrency can be reduced as burst speed increases. So in this example, the operator could decide to offer a 100Mbps service with 0.25% concurrency. The result is that each user would get an average of 250Kbps during peak busy hour.

The significance of this is that the data service rate increased by 10-fold while the actual bandwidth provided by the operator only had to increase 2.5x. This factor will become especially important as broadband providers look to start offering Gbps services. Note, this analysis applies only to bursty data applications and does not apply to streaming voice or video. Care must be taken with concurrency if the bulk of internet traffic becomes streaming video.

Not All Subscribers are created equal

Another important aspect of understanding bandwidth needs is to look at the usage across the many different subscribers. Many operators have noted that a relatively few users consume a proportionately large amount of the bandwidth. One piece of data indicated that 5% of the users consume two thirds of the total bandwidth and that 25% of the users consume 95% of the bandwidth.

This has a major implication in how an operator rolls out its bandwidth capacity increases. If it can increase bandwidth incrementally to a small number of power users, then it can avoid costly upgrades that go across the board to all subscribers.

Peer-to-Peer (P2P)

As discussed previously, Internet demands appear to be virtually unlimited and one of the driving applications behind this is Peer-to-Peer (P2P) applications. From its start in music file sharing, P2P is branching out into many more main stream applications. As P2P video file sharing becomes prevalent, it will significantly increase the bandwidth demands on the system, especially in the upstream.

Home Generated content

Consumer devices and in particular mobile devices have made tremendous progress in recent years and look to continue in coming years. They have all become rich multimedia devices. Mobile devices some day will include 10 Megapixel digital cameras that can stream video clips. Digital camcorders are becoming HD capable and inexpensive. All of this combined with the ease of use being introduced by UPnP/DLNA will create a tremendous amount of user generated content that will be shared over the Internet.

In addition to these consumer devices, the introduction of low cost video cameras (a.k.a. webcams) will increase the number of applications like video surveillance or “nanny cams” that can be shared over the Internet. Video telephony will also increase the amount of traffic coming from the home, and as previously mentioned, these applications will have a much larger and earlier impact on the upstream bandwidth needs.

Wireless Backhaul

Mobile Devices are evolving from their roots as a telephony device to a full multimedia device supporting streaming video. This evolution is going to create a tremendous demand for 4G capable wireless networks like WiMax and LTE as well as technologies like Metro-Wi-Fi. As the number of mobile devices continues to rise along with

improved video quality and screen sizes, the bandwidth needs will continue to escalate. Cell sizes will need to continually shrink to accommodate the wireless bandwidth density. Cable operators have a further incentive for creating wireless networks to maintain a competitive position with Telcos who already own spectrum and provide cellular services.

As 4G cell sizes are reduced to several hundred meters, the number of cell sites increases significantly and the site become more geographically dispersed. This creates a great opportunity for the Broadband Service Providers to supply the backhaul over their existing infrastructure. This may put tremendous additional bandwidth demands on their access networks.

3-D and Multi-view technologies

Since video is the main bandwidth hog today, it is important to understand future variations that will be hitting the market and potentially creating the next big impact to the industry. At consumer shows like CES, we are starting to see High Definition technology that supports both 3-D and multi-view technologies. The initial thrust for this technology will be the gaming world, studios starting to create 3-D movies as well as sports casting. Eventually, this technology may become main stream just like HD. The 3-D technology may cause a 50-100% increase in the bandwidth required to deliver the content. The multi-view causes an even larger increase, as it needs to deliver separate HD streams for each view provided to the user.

Future User Experiences

As we gazed into a very foggy crystal ball, we tried to imagine what applications of the future might drive a new paradigm in bandwidth requirements. There may be other technologies in addition to 3-D and multi-view just discussed that will impact the user experience. An ultra-HD technology may be developed that offers even higher video resolutions, with corresponding higher bandwidth demands.

As we think about enhancing the user video experience, the simple 2-way video call may be improved through the use of tools such as avatars and facial recognition. Beyond this, we may get into other enhancements to the user video experience with things like visualization and holography. At this point, there is no hard data on the bandwidth capacity impact, but we should continue to monitor and keep an eye for killer bandwidth applications. The only thing we do know is that bandwidth needs have always increased.

EXISTING HFC CAPACITY UPGRADES

Current techniques for expanding HFC bandwidth are well known and include: RF Upgrades (1GHz), node splitting and deep fiber expansion, Switched Digital Video (SDV), reclamation of analog channels, MPEG-4 encoding, DOCSIS 3.0 & M-CMTS.

Node Splits and Deep Fiber

In today's HFC plants, there are many that still have 500, 750 or even >1000 Households Passed (HHP) per fiber node. These plants were designed and optimized for a broadcast system. As we evolve to a completely on-demand system, one of the most effective means of increasing bandwidth capacity is to reduce the node size by a factor of 2, 4, 8 or even more.

Typical Fiber Nodes have 2 to 4 coax outputs. This allows the operator to split a node 2-way to 4-way by replacing electronics inside the existing housing without pulling any additional fiber. The operator also has the option of splitting the upstream independent of the downstream. Splitting the downstream does require additional narrowcast wavelengths to be sent to the node.

To reduce node size further, the operator can push the fiber deeper into the HFC plant. This can often be done cost effectively using smaller satellite nodes. While many of today's HFC systems support a six amplifier cascade (N+6), newer deep fiber systems may eliminate (N+0) or have a single (N+1) amplifier. A deep fiber system may reduce node size to 125 HHP or even less.

A key issue with node splitting and pushing fiber deeper is available fiber count. The number of fibers in an HFC plant can vary dramatically from plant to plant. For those plants with low fiber count, Wave Division Multiplexing (WDM) becomes a critical technology in providing additional narrowcast wavelengths and hence additional bandwidth capacity. A good example of this technology is Motorola's Enhanced Coarse WDM (E-CWDM) system that was announced at the 2007 SCTE Cable-TEC show last June.

A key advantage for HFC systems is that they only need to split nodes that need the extra capacity. For example, if an entire community is configured with 750 HHP, but only one neighborhood has exhausted its on-demand bandwidth, the operator only needs to split that one node to expand bandwidth capacity to meet demand.

SDV

Switched Digital Video (SDV) is different than the other bandwidth approaches. Rather than increasing bandwidth capacity, SDV provides a mechanism to better utilize existing capacity. SDV allows the operator to convert a fixed number of broadcast channels into a potentially unlimited number of video channels within existing spectrum. This becomes even more important as the amount of HD content being offered increases.

In addition to increasing the number of offered channels, SDV is an important mechanism to allow the transition to new technologies, such as migrating to 1 GHz tuners, or from SD to HD as well as from MPEG-2 to MPEG-4. Over time, once the new technology becomes the dominant installed based, then any remaining broadcast technology can also be converted to the newer technology while the older technology will be completely switched using SDV.

Analog Reclamation

Analog TV channels consume 50-75% of the spectrum in today's typical HFC system. Analog TV channels are also extremely inefficient with the use of spectrum. A single 6-MHz TV channel can be replaced by a digital QAM channel delivering 10-15 SD programs. Digital QAM channels are also very versatile and can be dynamically assigned to VOD, SDV or high speed data.

So why aren't cable operators just dumping all of their analog channels and going all digital? Supporting analog TV channels has become a competitive advantage for cable operators over satellite providers. This will become even more critical after 2009 when the Over-the-Air analog channels are no longer broadcast. Many homes have a 2nd, 3rd and even 4th TV in the house. Satellite providers must add a new STB for every additional TV in the house. Once the Over-the-Air analog channels are removed, cable service will be the only way for consumers to get basic local services to these other TVs.

Over time, those operators that want to continue to offer an analog service can still reclaim a large portion of the analog channels to get a significant increase in bandwidth capacity while still offering consumers a reasonable analog service. For example, a HFC system with 125 6-MHz channels could reduce the number of analog channels from 75 to 25. This would double the bandwidth capacity available for digital QAMs, while

still offering a basic analog service with all local and major network TV channels.

DOCSIS 3.0 and M-CMTS

DOCSIS 2.0 systems are currently in their prime, but will be running out of steam over the next several years as it tries to compete with FTTP. The DOCSIS 2.0 cable modem can support up to 38 Mbps downstream and 30 Mbps upstream before it hits a brick wall. DOCSIS 3.0 will be coming on line shortly and its channel bonding feature will allow significantly improved data rates to the subscriber. A DOCSIS 3.0 cable modem with 8 downstreams and 4 upstreams could enable a 300 Mbps downstream and 100 Mbps upstream service.

While DOCSIS 3.0 enables a bigger IP pipe to the subscriber, it does not impact the additional headend costs relative to video bandwidth associated with today's integrated-CMTS. A new Modular-CMTS (M-CMTS) architecture has been defined by DOCSIS to address this. It decouples the upstream from downstream and separates the RF technology from the CMTS core. This allows commodity driven Universal Edge QAM modulators to be shared between VOD, SDV and CMTS resources. This is the first major step in reducing the cost of delivering IP packets over DOCSIS and should reduce its cost relative to delivering video from 10-20X in the early days down to ~2X. Over time, the gap should continue to shrink as improvements are made to the all-digital CMTS core.

Upstream splitting and stacking

The current HFC bandwidth capacity is extremely asymmetric, with the downstream spectrum occupying 54 MHz to 1GHz while the upstream spectrum is limited to 5-42 MHz in North America. The upstream bandwidth capacity is further hampered by operating in a much noisier environment. To get the most out of an existing upstream requires the use of advanced technologies like SCDMA and improved ingress cancellers. SCDMA allows the operator to re-coup data bandwidth from the lower 5-15 MHz spectrum. All told, an operator can get a total of 140 Mbps of DOCSIS upstream bandwidth with appropriate improvements.

After improving the existing upstream, another common method of increasing upstream capacity is to split it into smaller node sizes. This can be done independent of splitting downstream node sizes and can also reduce the noise per upstream. Most fiber nodes support two to four coax legs. Each coax leg is potentially a separate upstream return spectrum.

HFC VS. GPON COMPARISON

But how does HFC today stack up against GPON in raw bandwidth capacity? It turns out quite well . . . in the downstream. Up until now, we've discussed many mechanisms that HFC may use to incrementally increase bandwidth. At this point, we will compare an HFC system using these available enhancements to a GPON system. The HFC system under consideration has 1GHz RF with deep fiber nodes [125 House Holds Passed (HHP), 100 subs]. We also assume about half of the analog channels have been reclaimed, so the system reserves ~40 analog channels and 8 digital simulcast QAM channels (i.e. 80-120 digital video broadcast streams). This means about 100 QAM channels are available for switched VOD/SDV/Data. This is approximately 1 QAM per sub or slightly less than 40Mbps downstream capacity per sub.

This is almost identical switched downstream capacity to a GPON system with 64 subscribers (i.e. 2.4 Gbps / 64 subs = 37.5 Mbps per sub). The implication here for the HFC system is a significant increase (almost 100-fold) in the number of QAM modulators

to achieve this. Edge QAM devices will need to continue to make significant improvements in cost, density and power to achieve this. However, the undesirable alternative for MSOs is to pull fiber to all 100 homes. It should be noted that some FTTP vendors are deploying GPON with 32 subscribers per PON. The HFC can continue to match this by pushing the fiber even deeper (e.g. N+0 with 50-65 HHP).

The GPON system still has an edge in RF bandwidth, burst speed and upstream capacity. The GPON system has an optional 750MHz RF carrier while the previous HFC system had set aside about 350MHz for analog and digital simulcast channels. For certain applications like large file transfers, the burst speed of the network is critical and GPON still has a large edge in this category. For upstream bandwidth, GPON supports ~1Gbps while the DOCSIS 3.0 system only supports about 100Mbps which it will still fall short of GPON's upstream capacity.

UPCOMING HFC BANDWIDTH CAPACITY UPGRADES

Modulation improvements

As HFC systems continue to improve, they will eventually be able to support higher order modulations like 1024 QAM. Increasing the QAM constellation density increases the MPEG-2 transport bit rate almost 28% to almost 50 Mbps. However, this higher throughput comes at a cost: the required threshold signal-to-noise ratio (SNR) to achieve the same reliability (bit error rate) is at least 6 dB higher using conventional J83.B FEC coding. Using 1024 QAM with advanced FEC can yield up to 23% higher throughput with a moderate increase of ~3dB in threshold SNR.

RFoG & CablePON – FTTP for cable

MSOs may have certain targeted areas where it makes economic sense to install a complete fiber based solution. This may include support for an industrial park or FTTP in a new housing development. If the existing bandwidth requirements allow, the cable operator may initially support existing cable services over the FTTP, while having the fiber in place for future expansion. This approach is called RF over Glass (RFoG) and is currently being standardized by the SCTE.

There may be some applications where the current HFC bandwidth solutions are inadequate. Some of these include Gbps Commercial Services or 4G Cellular & other wireless backhaul. For these, Motorola's CablePON solutions allow operators to offer a GPON solution where needed but within their existing legacy cable equipment infrastructure. This solution is different than a traditional GPON deployment in that the MSO leverages the fiber portion of its HFC to transport the services. This is also different from traditional GPON in that the devices fit into the MSOs back office infrastructure. Another important aspect is that this solution fits within the cable operators existing router and M-CMTS infrastructure without requiring expensive B-RAS equipment.

RF Overlays

Other technologies like RF Overlay systems offer the prospect of increased RF bandwidth capacity. Some recent technologies offer a 2-3 GHz system adding additional downstream and upstream capacity. These systems are being evaluated from both a technology and business perspective. While it may not make business sense to apply this technology to an entire plant, it may prove useful to bring additional bandwidth to a particular site (i.e. a surgical strike) where it is not feasible to extend the fiber portion of the plant. RF Overlay technology will most likely

be used in conjunction with other plant upgrade techniques.

DOCSIS 3.0 Mid-Split Systems

In addition to channel bonding, DOCSIS 3.0 also supports an option for a mid-split HFC system with additional upstream capacity. The upstream spectrum is increased from 5-42MHz to a 10-85MHz range. The added frequency range is also significantly less affected by impulse and ingress noise, so the net effect is that upstream bandwidth capacity may almost quadruple (e.g. from 27MHz usable at 16-QAM average to 70MHz usable at 64-QAM average). The mid-split upgrade needs to be done in conjunction with reclaiming the lower analog TV channels (channels 2-6) and a plant upgrade that replaces/eliminates the old diplex filters.

DIBA – CMTS By-Pass technology

As the world migrates to an all IP environment for the delivery of video, the additional cost of delivery over a DOCSIS network becomes more important, even with an M-CMTS approach. Motorola has pioneered a concept called DOCSIS IPTV Bypass Architecture (DIBA) that allows session oriented IP traffic to bypass the CMTS core and go directly from its server to the Edge QAM device. This greatly reduces the needed CMTS core capacity which in turn reduces the relative cost of delivering IP packets over DOCSIS by more than two thirds.

Over the long term, DIBA holds the promise on economically converting cable systems to an all IP infrastructure all the way to the home. Motorola has published several white papers on this topic, including a paper at the 2007 SCTE Emerging Technology conference. These papers provide a detailed description of DIBA.

Hybrid PON Coax (HPC) systems

As bandwidth needs for residential users continue to increase, it will eventually make sense to take the small percentage of users who consume a significant portion of the total available bandwidth on existing HFC systems and offer them Gbps services thru an FTTP solution. In this scenario, it would be highly desirable that an MSO have the capability to drop fiber to any existing individual subscriber.

One method of accomplishing this is with a concept called Hybrid PON Coax (HPC). The HPC system is basically a CablePON/GPON overlay on top of existing HFC plant. HPC allows the MSO to install pockets of FTTP within its existing HFC infrastructure. Having the ability to drop a fiber connection to a single home without upgrading its entire plant can give MSOs marketing leverage to combat its telco rivals.

Statistics have shown that a small percentage of users consume a large portion of the data bandwidth. The MSO can potentially move 5-10% of its subscribers to FTTP and free up two thirds of its high speed HFC data bandwidth. This means that with a relatively small fiber plant investment, the MSO can significantly extend the life of its HFC. The HPC also allows the MSO to combat the telco marketing in being able to offer FTTP to anyone, but with the huge advantage compared to them that it only needs to pull fiber to the select few that need the Gbps service.

Headend Impacts from Bandwidth Increases

Increasing HFC bandwidth capacity is more than just upgrading cable plants, it is providing equipment at the access edge to deliver this capacity. For HFC to compete with GPON, it will require a significant investment in additional Edge QAMs and CMTS core capabilities. Other video

capabilities like transcoding and encryption will need to be scaled as well. As technology progresses these functions will get pushed from the core further out to the access edge.

In addition to the video components just described, there will need to be new advances in the high speed data transport beyond the current M-CMTS developments to economically scale data traffic for tens or hundreds of Gbps data rates. Current CMTS architectures provide high touch, per subscriber services similar to B-RAS equipment. In order to economically get a hundred fold increase in data bandwidth, we will need to start migrating to Class based services and other technologies that can leverage standard Ethernet switching equipment.

There will also be the need to converge future generations of DOCSIS CMTS and GPON OLT products. As just discussed cable operators may need to start deploying HPC systems. With M-CMTS decoupling the PHY layer in DOCSIS, a similar approach can be taken with GPON to create a common packet processing core for both technologies going forward. Both share similar capabilities in routing, traffic shaping and policing as well as seamless mobility support. This platform should also be extended to support other access technologies such as WiMax, LTE and metro-Wi-Fi.

NEXT GEN COAX – FUTURE OF HFC

RF Upgrades – 3GHz & beyond

With the recent evaluations of RF Overlay equipment, we have taken a closer look at the theoretical maximum capacity of coax. It turns out that the typical hardline coax being used today has a limit of ~5GHz before waveguide effects take over. There may be other effects that limit the total coax capacity, but we should continue investigation in this area to understand completely how far we can push cable.

For RF designs, the complexity is often related to the number of octaves (i.e. doubling of frequency) that the design must cover. Current 1GHz systems must cover more than 4 octaves since they start at 50MHz. Going from 1 to 4 GHz adds two more octaves. As investigation continues on multiple GHz systems, the lower couple octaves (e.g. 50-200MHz or -400MHz) should be considered dropped for new systems to help minimize cost and complexity.

As deep fiber architectures eliminate other active components, we will reach the day where the coax is the final limitation on bandwidth capacity to the home. If the cable hardline is ultimately replaced by fiber, the coax drop line has a much higher RF frequency limit due to its much smaller diameter. This could allow cable to the home to even exceed 5 GHz. Other areas of investigation could look at ways of increasing the theoretical limit of the cable as well as using different waveguide mode(s) along with the lower order mode to increase the cable's capacity.

Next Gen Coax System

How far can the MSOs take coax looking out on the 10+ year horizon before it hits the brick wall? From above, the RF hardline cable might be capable of supporting systems in the 3-5GHz range. However, no standard exists yet for devices above 1GHz. With the increased bandwidth needs, the old 6MHz channel size no longer makes sense in this range. This gives us the opportunity to define a new Next Gen Coax system above 1GHz.

To be competitive with FTTP, Next Gen Coax must support extremely wideband channels (e.g. >100MHz wide) with dense advanced modulations (e.g. 1024/4096 QAM or equivalent) that are capable of delivering

greater than 1 Gbps symmetrical bandwidth to individual subscribers in a single or a small number of bonded channels. The new standard should also eliminate MPEG-2 transport and DOCSIS layers and define a simple all IP infrastructure that unifies all devices, including cable modems and STBs with end-to-end IP connections. An all IP infrastructure will also help with integrating the Next Gen Coax system with other FTTP and wireless access technologies.

These wideband channels would be of great value in existing HFC as well. The ability of offering symmetric Gbps services within the existing 1 GHz spectrum would significantly level the playing field with FTTP competitors. It may make sense to roll the wideband channel support out initially within existing frequency ranges and then allow for 1-5 GHz operation in future releases.

Next Gen Coax vs. 10G PON

Given the expected timeframe, a Next Gen Coax system would need to compete with a 10G PON system. How does this compare? Assuming the 1-5GHz range is split 2:1 in favor of downstream traffic with a 1024 QAM modulation or equivalent, the Next Gen Coax system could theoretically support a total of 20Gbps downstream and 10Gbps upstream bandwidth capacity. This is roughly equivalent to a pair of 10G PON systems. This means that MSOs could keep their node sizes roughly twice the size of the PON group (e.g. 125 HHP) and still provide the equivalent bandwidth per subscriber as a 10G PON.

If it turns out that there are other limitations in reaching a 5GHz system, the MSO can still provide a 3GHz system that would provide an additional 10Gbps downstream and 5Gbps upstream capacity. This along with smaller node sizes (e.g. 50 HHP) can still keep coax competitive with 10G PON systems.

Next Gen Coax development should start now so it can keep pace with 10G PON development. As MSOs upgrade to N+0 and N+1 cable architectures, we should consider enabling at least 3GHz RF, if not the full 5GHz as feasible. When you combine these changes with the Hybrid PON Coax architecture that allows select users to migrate to FTTP if needed, then the HFC system appears to have a very long life ahead of it. And most important, all bandwidth increases are incremental and invested as needed.

HOME NETWORKING IMPACTS

The Broadband pipe into the home is becoming a fire hose. The prospect of offering 100's Mbps or even >1Gbps to a home only makes sense if the home network can handle that in addition to all of its local LAN traffic. This will place a burden on existing multimedia home networks. To meet this need will require multiple wired and wireless home networks to be interconnected.

For home networking over cable, next generation MoCA will be needed to scale to several hundred megabits per second to match or exceed DOCSIS 3.0 speeds. This next generation MoCA will become critical for cable operators to deliver its DOCSIS 3.0 bandwidth throughout the home and compete with FTTP.

Another important piece of the future home networking scene will be 802.11n. This next generation Wi-Fi network supports more than 100 Mbps and provides a number of features that will improve robustness and range. We also expect to see smart antenna developments coupled with 802.11n to improve performance. Work on the following generation to 802.11n has started and may be a step towards approaching Gbps wireless rates in the home.

Looking even further out there is the possibility of 60 GHz technology evolving to provide higher bandwidth home networking. The major issue with 60GHz technology will be the ability to propagate through walls to provide whole home coverage. It is possible that 5 GHz technology may be the limit in terms of whole home wireless coverage.

Next Gen Coax as Home Network technology

As the Next Gen Coax technology is developed as a multi-gigabit access technology, it should also be extended within the home as the first multi-gigabit in-home network. No other home networking technology seems poised to address this challenge. While GPON currently provides 2.4 Gbps to the side of the house, actual burst data rates to consumer devices will be limited to the home networking technology, which today is on the order of 100 Mbps or less.

If cable operators can develop multi-gigabit technology delivered all the way to consumer devices throughout the home over coax, it can once again leapfrog its Telco rivals. Another advantage with this approach is the economies of scale from the consumer devices sharing the same technology as the access devices. We've seen the benefits of this in the Wi-Fi world.

CONCLUSION

We have seen that the bandwidth needs will continue to increase for the foreseeable future and that cable operators will need to extensively support Gbps services within 10-15 years. The cable operators will also get severe competitive pressure from FTTP providers like Verizon FIOS further accelerating the need to increase bandwidth.

With existing upgrades and continued technological developments in devices like Edge QAMs, the cable operator is in excellent position to compete with GPON with respect to downstream traffic. This bodes well for offering personalized HD video services. As new applications start to drive upstream capacity, this will expose the upstream as the cable operator's Achilles heel. Increasing upstream capacity is an area that needs continued research and development.

In the near term, there may be scenarios such as offering Gbps Commercial Services, wireless backhaul, or residential green field builds where the cable operator needs an FTTP solution. For these, Motorola's CablePON solutions allows MSOs to offer FTTP where needed while operating completely within a legacy cable equipment environment. As the bandwidth race continues, it may become necessary for MSOs to be able to offer Gbps service to select power users on existing HFC thru a Hybrid PON Coax (HPC) system. Moving the heavy users off the HFC network extends the HFC useful life and having the ability to drop a fiber connection to any home without upgrading its entire plant can give MSOs marketing leverage to combat its telco rivals.

Looking far out on the 10+ year horizon, there is a possibility of a new Next Gen Coax system. Capable of operating up to 3-5 GHz with very wideband channels, the cable operator can offer symmetric Gbps services to the user while matching the overall bandwidth of a 10G PON system while maintaining 50-125 HHP node architectures. This Next Gen Coax architecture may also be a catalyst to enable Gbps services throughout the home network as well.

With all told, the future does not look bleak for HFC, but it looks to have a long and healthy life. And the most important piece of this is that the HFC can grow incrementally as needed without the need for a forklift upgrade like its competitors.

HOW TO MONETIZE OVER-THE-TOP (OTT) VIDEO

Author Name: Eitan Efron
VP Marketing & Business Development, Oversi

Abstract

The MSO role in the new media ecosystem is under scrutiny. How can cable operators financially benefit from the demand for over-the-top (OTT) video services and become key players in the content delivery supply chain? At present, most operators are mere bandwidth conduits which do not receive any financial gain from the increasing amounts of Internet video flowing through their broadband networks.

This paper will outline how edge content distribution networks (CDNs) deployed at the cable operator core network can accelerate Internet video delivery and improve quality of service through content distribution platforms. This opens up new monetization opportunities for cable operators with both content owners that are willing to pay for guaranteed service levels, and customers who can choose from a variety of tiered service packages.

MARKET OUTLOOK

Growth in OTT Traffic

Content is traveling through the Internet in ever growing quantities. The most ubiquitous segment is OTT video, which is increasing by volume and quality. The growth of video on the Internet is so marked it is predicted that very soon it will account for the vast majority of all traffic.

A recent white paper from Cisco showed that Internet video sites, such as YouTube, Xbox 360 movies, and MySpace, already generate more traffic than the entire US backbone in the year 2000.[i]

In a joint declaration by the Broadband Services Forum in January 2008, it stated that "current projections on the growing popularity of OTT video predict that service provider networks are going to be operating at near or complete capacity by 2010."

Even with service providers spending billions of dollars to build better, faster and more reliable pipes, OTT video could bring many networks to a grinding halt in the near future.[ii]

The changes noted in Internet traffic patterns and OTT video reflect the customer's desire for convergence. Customers want access to endless varieties of content on multiple devices, or "anything, anytime, anywhere." However, due to high levels of congestion, service providers can't keep up, and customers are experiencing ever lower levels of quality of experience (QoE), leading to customer dissatisfaction and increased churn.

OTT Delivery

Because of the importance of the user's video experience, QoE is becoming a number one priority for content providers. Content providers have traditionally been using the services of CDNs (Content Distribution Networks), such as Akamai and Limelight, in order to expedite the delivery of content and ensure high levels of QoE.

In recent months, OTT video providers, such as the BBC and NBC, are using peer-to-peer

(P2P) or hybrid P2P/CDN technology to distribute high quality content to their customers. P2P is helping them to simultaneously reduce their distribution costs and scale their services to serve more customers with higher quality video content (flash crowds).

However, while CDNs are paid for content delivery, MSOs continue to deliver content across their networks without receiving compensation from content owners or CDN operators.

It is clear that MSOs must find new ways to manage video traffic of all kinds through their networks.

Pain with No Gain

To meet the rising demand of OTT video, infrastructure costs for the MSO are steadily increasing, while the revenues that MSOs can enjoy from their subscribers as a result of this investment is partial at best.

Because of this changing landscape, monetization of OTT video has become a must for MSOs.

THE OPPORTUNITY

From Pain to Gain

In order to close the gap between rising expenses and decreasing revenues, and to capitalize on market opportunities, MSOs can adopt a combination of the following business models:

1. Manage the surge of OTT video on their networks, without hurting the user experience and creating net neutrality issues, by implementing caching platforms which support OTT video.

2. Charge content providers, including OTT providers, for the delivery of content with assured QoE. In this context, MSOs are complementing CDNs and providing service to the last and most important segment in the content delivery value chain: the end user.

3. Establish mega video portals, thereby becoming part of the video distribution value chain and leveraging their existing relationship with their customers. Under this model, the MSO joins the OTT value chain through shared revenues with the content provider.

4. Generate advertising revenue sharing with content providers for ad-supported content. As cited in a recent report, the largest revenue opportunity for online video will come from advertising, which could reach $4.4 billion in the US by 2011.[iii] Under this model, the consumer receives the content at no charge, but advertising has been added to the content by the MSO. Advertising can be general or personalized to specific user groups.

5. Introduce a host of services that can increase the average revenue per user (ARPU) based on subscriptions or advertising. These new services could include nPVR, catch-up TV, and so on.

6. Sell tiered services to users, which also increases the ARPU. MSOs can offer enhanced delivery over specific time periods for special deals. For example, a customer can purchase a movie for download for $3 and pay $1 extra for a faster download, or access their favorite OTT video with accelerated service for an extra $10 per month. This service improves the customer QoE and reduces churn.

A New Content Delivery Infrastructure

To actualize these opportunities, MSOs need to implement an infrastructure that will:

- Cost-effectively manage and deliver their OTT video over the network while maintaining QoE, including high definition TV (HDTV).

- Enable users to view content on all devices, whether on their TV, home computer or portable.

- Support monetization schemes for content delivery, such as local advertisement injection.

In essence, MSOs need their own internal content distribution platforms (CDPs), housed within their existing network. Today's CDNs are deployed outside the MSO network and can only deliver content in the open Internet environment, where they cannot guarantee quality of service inside the MSO network.

MSOs need a solution that can deliver content in both the open and the private network. While open Internet traffic travels on a best efforts basis, using a private network, the MSO can provide content providers/OTT partners with quality of service guarantees and enhance the services it offers its customers.

THE CDP SOLUTION

The content distribution platform (CDP) solution is designed to meet the needs and requirements of MSOs in the delivery of rich media content through their networks and enable monetization opportunities.

The CDP is an edge content delivery platform deployed within the MSO's network.

The following figure illustrates the overall CDP concept:

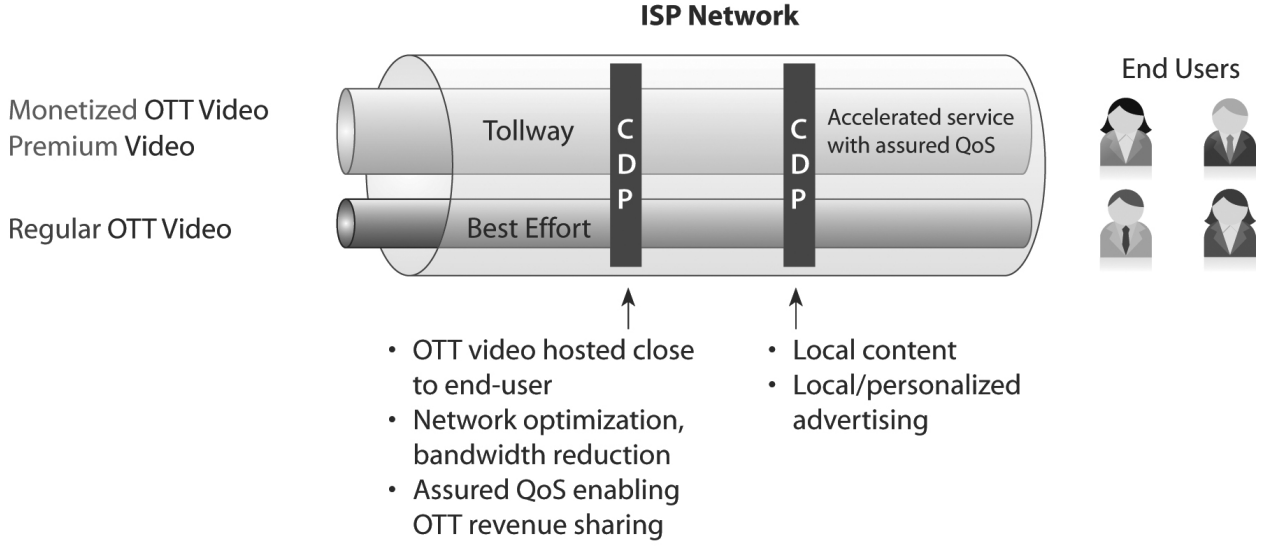


Figure 1: CDP Solution

The CDP solution consists of a two-tiered content delivery network:

1. A best effort freeway, which is today's open broadband connection for P2P, user-generated content (UGC), and other OTT traffic, that provides equal (net neutral) service on a best efforts basis to all incoming traffic.

2. A QoE tollway, which is designed as a gated garden that allows for service level agreement (SLA) based delivery of premium content for users and OTT content providers.

Both tiers are supported by the CDP infrastructure at different service levels, which is the key to both a rational network/traffic management model and an effective content monetization solution.

CDP Conceptual Design

The CDP is based on the smart deployment of a robust and scalable caching and acceleration system. A proven way to ensure the delivery of popular, high quality content is to cache it close to the user. The caching of content avoids randomly created network bottlenecks, saves on bandwidth and ensures prompt content delivery upon request.

The CDP is a multi-protocol caching platform that provides all of the necessary functionality from content collection and smart caching to the delivery of huge quantities of the cached content. The CDP supports all of the major and most popular protocols that are being used for content delivery today and anticipated in the future, including: P2P, HTTP, RTSP, etc.

The following diagram illustrates the CDP overall architecture:

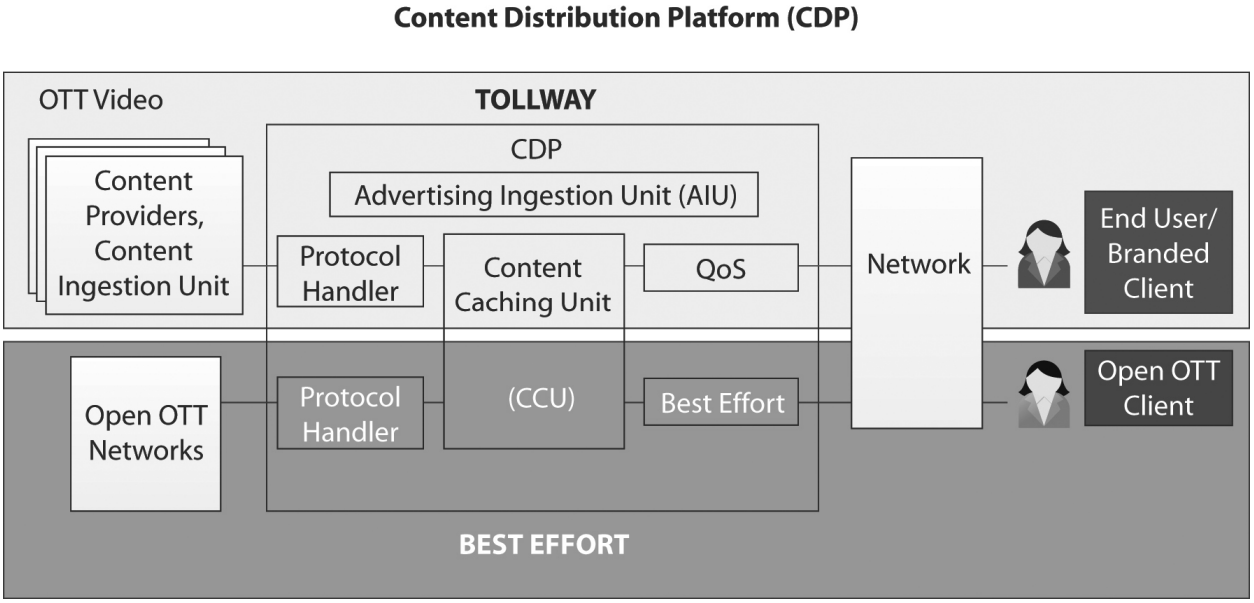


Figure 2: CDP Conceptual Design

A CDP typically consists of the following main units in its architecture:

1. **Content Ingestion Unit (CIU)**: Located within the MSO or at the Content Provider (CP) facility, the CIU enables content providers to upload content and define content-related

policies. Uploaded content has business data attached to it, which describes the business model between the MSO and the CP.

2. **Content Caching Unit (CCU)**: A caching unit that is characterized by strong bandwidth generation capacity (it can deliver the same content to many concurrent users simultaneously). Once content is ready for caching, it is uploaded into the CCUs. CCUs are also typically installed in the access network since they need to support high demand delivery, such as streaming content. The subsystem keeps track of all of the content handled by the system using smart caching technology and is responsible for sourcing and delivering content.

3. **Advertising Ingestion Unit (AIU)**:
The AIU enables the CDP to deliver content and advertisements in accordance with policies established in the external system. The unit interfaces with external advertisement control systems or ad networks. The CDP's ad-insertion system is located close to the end user.

4. **End Unit Client (EUC)**: This optional MSO branded client is installed at the customer's premises, providing branding opportunities and advertising functionalities and enables quality of service guarantees.

Functional CDP Requirements

The CDP solution meets the following key functional requirements:

- Works in managed and unmanaged environments, and gated garden (B2B) modes.

- Scalable: Supports millions of users and assets.

- Personal: Per-subscriber SLA and accounting/charging.

- Modular architecture that supports multiple access protocols, including HTTP, RTSP, P2P, etc.

- Modular solution: Enables gradual transition from single multi-server node to managed network of multi-server nodes.

- Virtual CDP architecture enables providers to offer managed CDN services to third parties.

SUMMARY

To avoid becoming dumb pipe operators, and carrying the costs of delivering OTT video, MSOs must adopt strategies that enable them to become part of the OTT video distribution value chain.

A CDP deployment enables the cable operator to monetize OTT video by:

1. Generating revenue from video traffic passing through its network. MSOs can charge content providers for hosting and delivering their content.

2. Delivering virtual video portals for their customers, using a pay-per-view model or subscriptions.

3. Participating in advertising revenue sharing with content providers for ad-supported content directed at an MSO's existing customer base.

4. Selling tiered services and different advertising/subscription packages to customers, thereby increasing the ARPU.

5. Taking advantage of personalized advertising opportunities through the MSO's own content portal.

Additional MSO benefits of a CDP deployment include:

- Full control over video content.
- Reduced MSO bandwidth costs.
- Improved QoE for end users generated through accelerated content delivery and a better overall experience of all online services, which will ultimately increase the MSO customer base and reduce churn.

---

[i] Statistics from Figure 4 from The Exabyte Era White Paper (based on the paper: Traffic Forecast and Methodology 2006-2011), Cisco Systems, 2007.

[ii] Joint Declaration of the Broadband Services Forum, January 2008.

[iii] Report: IPTV Competitors are Over-the-Top (Quoting James Crawshaw's report: Internet TV, Over-the-Top Video & the Future of IPTV Services, Heavy Reading), Ryan Lawler, Light Reading, June 28, 2007.

IMPLEMENTING ADDRESSABLE ADVERTISING IN LINEAR NETWORKS

Steve Riedl, Principal Architect, Time Warner Cable
Doug Jones, Chief Architect, BigBand Networks

Abstract

Although the potential for addressable advertising to increase revenues is widely accepted, the infrastructural changes needed to support it are still being examined. Existing network functionality needs to be enriched to allow the following abilities to be implemented:

- *Replication of program streams such that more than one copy can be simultaneously supported on the network – since bandwidth-intensive HDTV is increasingly being offered to subscribers the ability to place its replication closer to the edge of the HFC network could conserve IP distribution network capacity;*
- *Managing the multiple copies within the current systems without causing issues with the large increase in sources;*
- *Implementing IGMPv3 to insure only viewed networks are distributed on the IP sections of the operator network;*
- *Association of a particular program stream with viewers of like demographics to enable relevant ads to be delivered to those viewers – grouping these viewers can be accomplished either when theviewer initiates a channel change or by force-tuning that viewer before the ads are to be played;*
- *Addressing of program streams to a single user in order to deliver a particularly relevant ad to them;*
- *Selective insertion of advertising into a program stream to match viewers' demographics – ad insertion can occur either through seamless splicing or a playlist;*
- *Interoperability between the HFC resource management system and the advertising decision manager to ensure that the available capacity of the HFC network is taken into consideration;*
- *Instantiation of an overall control mechanism to coordinate these activities – the SCTE DPI committee has begun the process of developing standardized interfaces to meet this need.*

The delivery of advertisements should include extensions of switched digital video known as microcast and unicast, and set-top based ad insertion mechanisms. These methods are complimentary because content can be delivered either over the network or from a local hard disk drive in the set-top box. This paper unifies these topics, describes an architecture designed to satisfy the requirements listed above, and explains its operation. The intent of the authors is to prove a technology framework for converged resource management and addressable advertising for emerging marketing services.

The paper also provides relevant context on the addressable advertising service including an architecture for a next-generation advertising system. The authors show how ads can be stored and how the developing standard interfaces will enable next-generation ad delivery systems. Additionally, it identifies a variety of parameters that can be used to size the advertising system by leveraging projections of initial deployments.

INTRODUCTION OF ADDRESSABLE ADVERTISING

Addressable advertising is the selective insertion of advertising into a program to match viewers' interests, thereby making that advertising more relevant to the viewer. Ad relevancy is important, because if presented with a choice many viewers would likely opt-in to

minimize or eliminate certain types of advertising that does not match their interest; for example, parents without kids might prefer not to see another baby diaper ad.

With addressable advertising, sets of viewers can be targeted at home watching a program like ESPN's SportsCenter and delivered different ads more relevant to there interest at precisely the same time. For example, one set of viewers who prefer full-size high-performance vehicles can be shown a commercial for the new Chevy Blazer and another set of viewers who prefer mid-size fuel-efficient vehicles can be shown a commercial for the all-new Chevy Equinox. This type of viewer ad targeting was pre-planned by the advertiser and tools are becoming available in the network to make it a reality based on aggregate geographic, demographic, psychographic or other characteristics of the consumers residing within specific areas.

Making advertising more relevant to viewers can have several benefits. First, viewers will come to see advertising as part of the programming, and will most likely be happier and more interested in viewing it. Additionally the relevancy will become more valuable to the advertisers making cable programming a better place to spend their advertising dollars. Finally, intelligently choosing advertising benefits the advertiser by minimizing the placement of irrelevant advertising in front of viewers.

While there are many methods to address advertising to viewers, this section will introduce two widely available methods; ZIP+4 and Prizm Codes. Both methods are based on the premise that "birds of a feather flock together," and that people with similar demographic traits tend to behave in the same way in the marketplace.

ZIP+4 codes are specific from a single to a few dozen households, which clearly is more specific than the average cable service group which can pass several hundred to a thousand homes. Commercial databases are available which will provide specific demographics for the households within all ZIP+4 areas and advertisers already use these for mass mailings. This advertising principle can easily be carried over to cable if the proper advertising infrastructure were in place to provide advertising to just a few dozen homes, or the viewers associated with those homes.

Prizm Codes are associated with several demographic parameters. There are 66 defined Prizm Codes and one or more can be assigned to each cable customer. In fact, one of the 66 Prizm Cluster Codes is already assigned to every address in the U.S. Prizm Codes are based on one of four urbanicity categories which is determined by the population density of an area and it's neighboring areas. Within each urbanicity category, segments are further sorted into groups based on affluence, another powerful demographic predictor of consumer behavior.

CABLE PRIVACY ACT

A discussion of addressable advertising requires an up front disclosure on the implications on customer privacy. Addressable advertising implies a knowledge about a customers demographics and with all the information available about consumers, a major issue in addressable advertising is that cable operators must comply with some very strict privacy guidelines, including the Federal Cable Privacy Act of 1984 and other specific laws passed by individual states. Nothing in this paper should be taken as legal advice. Always check with your company privacy attorneys before implementing any system.

TODAY'S CABLE ADVERTISING LANDSCAPE

This paper primarily focuses on Linear programming but many of the concepts will also apply to On Demand programming. Linear programming is traditionally a broadcast medium. On a cable plant, broadcast bandwidth is a very precious resource. Narrowcast bandwidth can be created by reducing node size and/or adding additional QAM bandwidth to a node by removing analog networks. Once the node size reaches about 500 tuners, then it is usually more economical to expand the number of QAMs per node.

Since unicast services such as VOD and multicast services such as SDV share this narrowcast bandwidth, an Edge Resource Manager needs to be put in place to allocate these resources. This paper show how this service plays an important role in utilizing this bandwidth for advertising purposes.

Not all channels are insertable; it generally depends on the programming contract. There are about 60 ad-insertable networks today. Most operators insert on 40 channels, four 30 second spots per hour is normal – and only 12 hours of the day matter. So while there are about 4,000 insertion opportunities each day, only about 2,000 are useable for any significant revenue. The top ten networks provide about 80% of the ad revenue and the top 20 networks provide over 90% of the revenue.

Given that cable local / spot advertising is estimated to be a $5 billion business in 2008 (NCTA), the last 10% is not small change at $500 million using current methods. By making that smaller number of eyeballs addressable, measurable and interactive, this 50% of the avail space could bring in well over $5 billion.

Some programs are addressed by their very nature. The typical ESPN viewer is different than the typical Lifetime Movie Network viewer. Others networks like CNN attract a wider demographic of viewers. Even a network that attracts a niche demographic will have some variation and if third-party data can be used, like if the car lease is expiring, then targeting on those networks has real value.

Today the most sophistication associated with advertising is geographic zoning, where customers in one part of a city will see different ads than customers in another part of the city. The degree of geographic zone subdividing within the cable system would generally be measured in tens of square miles associated with different regions of a city. Figure 1 shows an example of carving a city into five geographic zones.

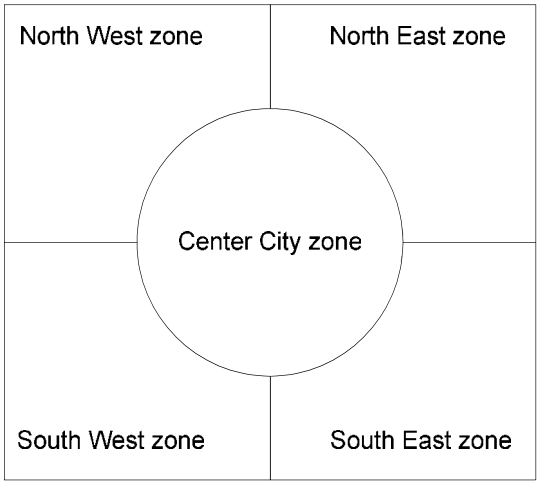


Figure 1 – Geographic Zoning

To deliver this type of zoning, the ad-insertable programming is replicated, generally in the headend, and each copy of the programming is sent through equipment which inserts ads for a particular geographic region. Then that programming is distributed to the appropriate geographic zones and is delivered to the viewers.

Migrating to more personalized addressability requires pushing the functions of program

replication and ad insertion further out to the edge of the network, and coupling these functions to a system which can make decisions about which streams to which viewers to personalize and how.

ARCHITECTURES FOR NETWORK-BASED ADDRESSABLE ADVERTISING

This section introduces both a method to do addressable advertising with current network infrastructure as well as an emerging architecture to deliver addressable advertising.

The architecture being developed with today's network components for linear addressable advertising is based on modifications to the On Demand system. Based on how VOD is deployed today, most of the resources are centralized at one or more super headends while other processing can be moved out to the hub to manage the metro network bandwidth utilization. If the transport capabilities are inexpensive enough and can handle the required streams, then centralizing the resources makes more sense. If the transport is more expensive or just not possible, then functions can be pushed to the edge, which is a key component of the emerging system to be described following the rest of this example.

Figure 2 shows one logical flow through this architecture. The network signals are received off the satellite which then goes to a stat-mux/splicer to be rate-shaped either as a feed for SDV, network PVR or as part of a multi-program multiplex. The splicer portion is used to insert advertisements at every available opportunity.

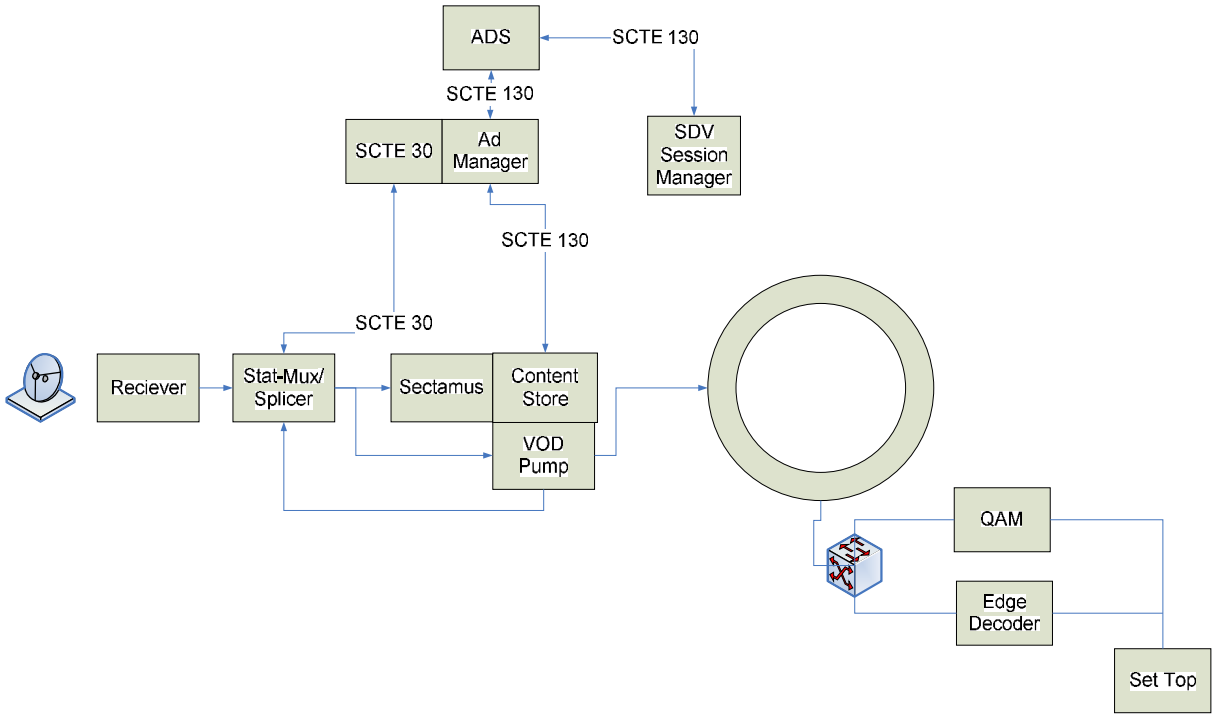


Figure 2 – Linear Addressable Advertising Based on VOD Modifications

These will typically be the advertisements for the largest geographic ad zone. By performing this initial splice, the network stream is prepared for easier content segmentation. This feed is then distributed onto the metro ring to the hubs to provide a constant reliable network feed source. This processed feed also then goes into the On Demand server complex for real-time acquisition. If additional ad zones are required, real-time feeds can be set up out of the On Demand server and new ads can be inserted by using playlist techniques from the On Demand server.

Figure 3 shows the emerging method to implement an addressable advertising system which is not based on the On Demand system. With this new method, the equipment is located closer to the viewers such as in a Distribution Hub. In this example, a new category of network component, the media services platform, is capable of the traditional stream replication and ad insertion, but also more advanced forms of personalization including managing bound applications (Enhanced Television, or ETV, and OCAP bound applications) as well as the possibility for creation of personal mosaics.

The media services platform allows the operator to grow the new addressable advertising service on purpose-built equipment without impacting existing services such as VOD. In this example, the media services platform ingests program streams, national or local feeds which come with default ads, and has the capability to personalize these streams by replicating them and inserting advertising specific for the viewers.

The media services platform interfaces with a personalization engine which is the decision maker that instructs the media services platform when and how to personalize streams. The personalization can include inserting addressed advertising, or a particular enhanced programming or to create a mosaic. The personalization engine can make decisions on what streams to personalize based on the household viewing of that content, their geographic location, assigned Prizm Codes or any number of other factors known about that household including specific opt-in opportunities.

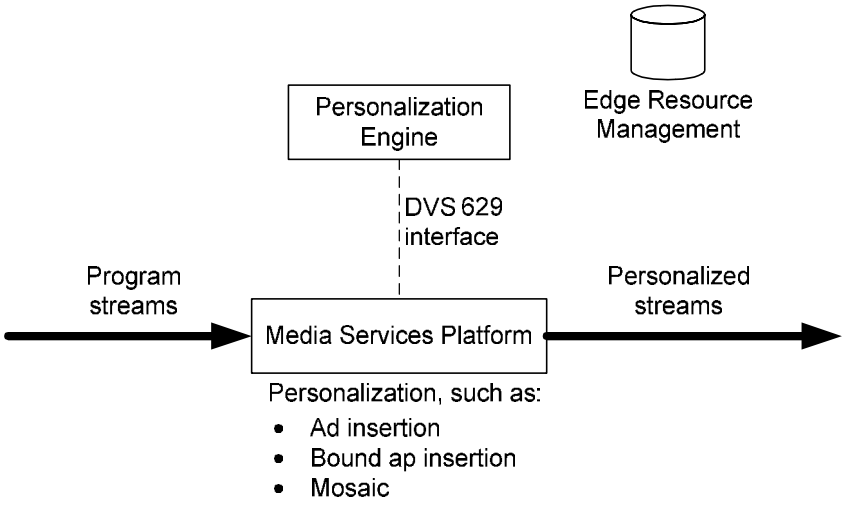


Figure 3 – Personalization System

When either a customer changes channels or an advertising avail comes up in a program stream, the personalization engine needs to make a real-time decision to decide if:

a) The program is ad-insertable;
b) There is an available campaign for this subscriber on this network;
c) Is there enough QAM capacity on the service group to do something like support replicating the stream for just that one viewer.

Separating the decision making from the actual delivery system allows for open systems. The SCTE (Society of Cable Telecommunications Engineers) DVS (Digital Video Subcommittee) is developing an interface standard, SCTE 130, which supports this separation of functions and will allow innovation to occur within both the personalization engine and media services platform while allowing operators to choose the components independent of each other.

Finally as shown in Figure 3, the addressable advertising system has to interface with an ERM (Edge Resource Management) system which is used to manage the digital QAM bandwidth to subscribers. Because the personalization system can cause program streams to be replicated on a service group, it uses edge QAM capacity more so than a geographic zoned ad system. As such, the personalization system needs feedback from the ERM system as to how much QAM capacity is available for personalization. During most parts of the day, the QAM service group is underutilized and there is excess QAM capacity available for personalization. It is only at peak viewing times that the QAM service group can become full and if the operator wants to continue high levels of personalization during those times, additional QAM capacity should be considered.

With a system as show in Figure 3, the options for delivering addressable advertising in linear programs are expanded from just geographic zoning to include three basic types of addressability which should cover the broad spectrum of advertising models to be developed in the foreseeable future; broadcast, groupcast and unicast linear ad insertion models.

Network-based addressable advertising uses SDV (Switched Digital Video) as an enabling technology for several reasons. First, SDV has every channel change request go through a server which allows subscribers to receive the appropriate personalized programming each time they change channels. Additionally, tuning information from set-top boxes, including both switched and broadcast networks, can be collected and used to understand how to apply demographics to viewers.

Broadcast is the current ad environment and is addressed through the specific network, the show, time of day, etc. As shown in Figure 4, a single version of the program stream is delivered to all viewers who all see the same ads. This solution is the basis for today's advertising business and cable is continually adding better audience measurement capabilities that will allow us to better utilize this segment.

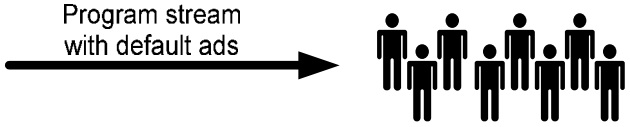


Figure 4 – Broadcast Advertising

Groupcast is a solution where viewers of like demographic are "grouped" onto a common copy of a program which is then enhanced for them. Take a program such as CNN which can have a wide variety of viewers, the personalization system would create several copies of CNN, one for each demographic which an advertiser is trying to reach and each copy of CNN would then be personalized with ads for that particular demographic.

An example of groupcast is shown in Figure 5 where there are three copies of the program stream created in the media services platform and viewers of particular demographic are grouped onto those copies which each carry ads specific for that demographic. In this case there are copies of the program generated for demographics A and B and a third copy of the program is available for viewers which do not fit these demographics.

Groupcast is very effective at demographic targeting-based techniques such as ZIP code or Prizm code. A household is either in a ZIP or Prizm cluster and doesn't move (at least during the duration of the program). Groupcast is not as effective as unicast at hyper-targeting households for certain ads based on a specific household (such as a specific credit card holder or not).

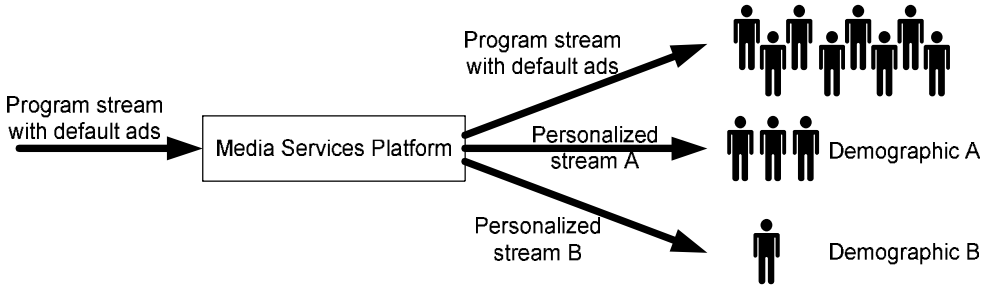


Figure 5 – Groupcast Advertising

Before a new groupcast feed is created for a particular demographic, bandwidth availability can be determined by polling the ERM. If bandwidth is not available then the additional program feed is not created which can result in a slightly lower yield, but requires no additional edge bandwidth spending. Reporting can indicate if this is a frequent occurrence and the operator might consider adding QAM capacity to those particular service groups.

Unicast is a technology where a single program stream is customized for a single viewer, as show in Figure 6.

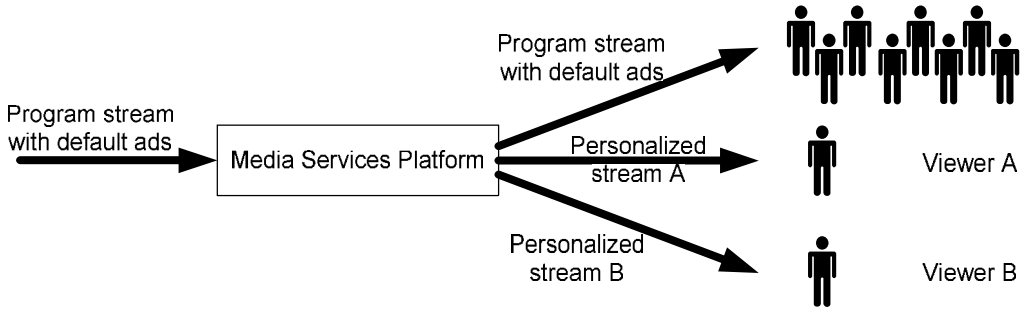


Figure 6 – Unicast Advertising

With unicast, whenever a viewer changes channels, the personalization system has to determine if a new unique copy of that program should be created just for the viewer. This solution is narrowcast, where there is a stream for each viewer and can be used to fill available capacity on a QAM service group.

While unicast means that every subscriber could have a dedicated stream, it does not necessarily mean that they will have a dedicated stream, just that if their demographics were such that they matched a specific set of available ads to play and bandwidth is available on the service group, a custom stream could be created for that single viewer.

Groupcast is a good example where the system can dynamically deliver advertising using both broadcast and unicast, as well as the specific case of groupcast. When used fully as a bandwidth saving multicast service, groupcast resembles broadcast. When QAM capacity is available, the demographic parameters used to replicate programs can approach unicast mode where it may be possible to create completely customized programs for each viewer.

The migration to unicast needs to be managed by taking advantage of available service group capacity, as more QAM capacity is added over time. Unicast does not have to imply that a separate stream is available for every viewer. Unicast may mean that separate streams are available for only some viewers. For example the QAM capacity of a given city may support 80,000 narrowcast SDTV slots, but the peak TV viewing population of that city may surpass that at certain times of the day. When this happens, it is simply not possible to deliver all unicast streams and some blend of broadcast, groupcast and unicast will be most efficient. It is the personalization engine which makes these decisions to best optimize the delivery of advertising and personalization while ensuring customers can watch as much programming as desired.

The goal is to add addressable streams based on bandwidth, demographics and ad availability. At startup, the system sets up an initial stream for every insertable network and assigns a base demographic to that stream. The system could also create a "spare" stream of every ad insertable network with no assigned demographic. This spare stream could be switched onto the service group if the personalization system determines it is appropriate to customize it for a new viewer (at which time a new "spare" stream would be created). If on the other hand there is no bandwidth available (or above a limit set for ad purposes) the new viewer gets tuned to an existing groupcast which best fits their demographics.

SDV ADVERTISING PHASED IMPLEMENTATION APPROACH

The following is a proposed phased approach to addressable advertising based on groupcast using switched digital video. The first three phases can be done with existing equipment and small modifications to the SDV session manager. Note that in this example transition the edge QAM resources will be managed through the SDV system to ensure the additional copies of programs do not overflow the QAM resources available on the service group. A more detailed discussion on QAM resource management is presented following the phased implementation approach.

Step 1 – Study the current ad insertable networks to determine which demographics are needed. Also learn which demographic groups watch the networks and during what times. For example if the goal is to personalize programming for Senior Citizens, don't advertise on Nickelodeon.

Step 2 – Use the data from step 1 to pick a small number of networks that typically have four demographics watching them. For a deployment of this scope, the SDV server can perform the demographic selection based on a look-up table with set-top box IDs, networks and the demographic groups. Ad selection in this case can be done with current traffic and billing systems by generating four schedules or, the system could choose to use an SCTE 130-based system. Alternatively in Step 2 one could use fixed demographics such as ZIP+4 or political party affiliation (democrat, republican, libertarian, unknown) as a stable demographic selector.

Step 3 – Allow the demographics on any given network to change during the day to better represent both the viewing patterns and the available ad inventory. This additional degree of flexibility will allow better addressability, albeit the combinations can become more specific and migrating to a true personalization engine may be prudent at this time.

Step 4 – Extend this groupcast example to all ad-insertable networks. This will maximize the revenue by bringing addressability to all the networks where the rights are available to insert ads.

Step 5 – Extend groupcast to unicast when there is an available ad campaign. Unicast advertising most closely resembles the Internet advertising model and promises the greatest revenue per ad.

With respect to QAM service group bandwidth, the following group of figures show the utilization of the QAMs in a typical SDV service group over a week, note these figures do not yet include addressable advertising but they can still be instructive to understand how viewers are using the linear lineup. While only about 50 networks are ad inserted, typically 150 or more networks are carried and only a percentage of viewers are watching ad inserted networks that would use the additional bandwidth if available. The non ad inserted networks would always use multicast and only one version per service group.

Figure 7 shows how the amount of traffic on the switched tier is allocated between unicast traffic (a single unique viewer on a program) and multicast traffic (2 or more viewers on a program) and indicates that there is a significant percentage of time when there is but a single viewer on a program which could provide an opportunity for unicast advertising without having to replicate an additional version of the program on the service group.

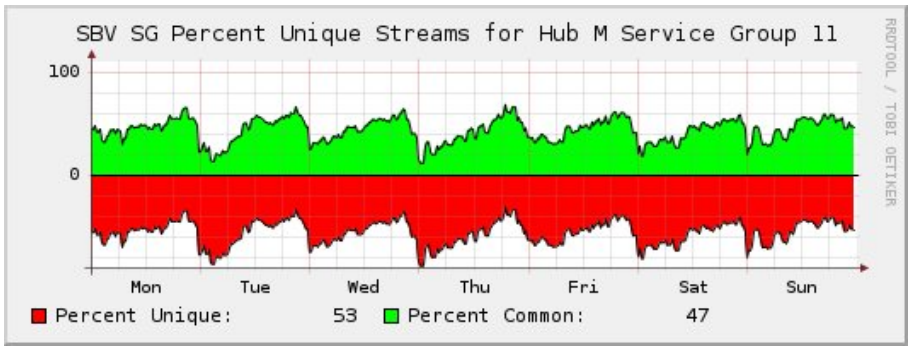


Figure 7 – Percent Unique Streams per Service Group

Figure 8 shows the number of active viewers on a service group. Note the periodicity of the graph with peaks building up through during prime-time each day.

Viewers can be reached at any time during the day with addressable advertising; however, the largest numbers of viewers are present during prime-time.

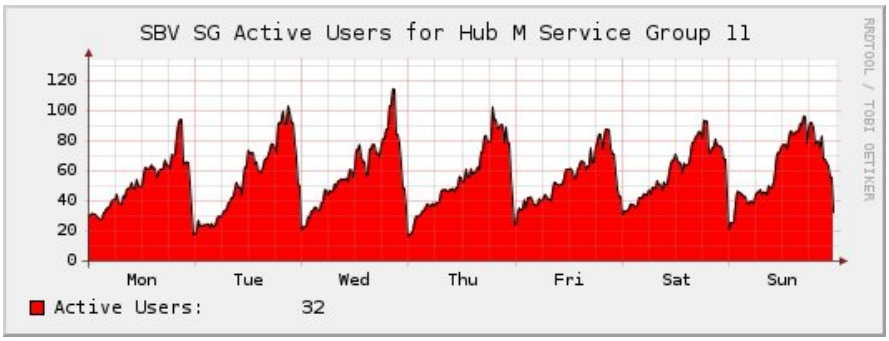


Figure 8 – Active Viewers per Service Group

Figure 9 shows the percentage of bit-rate used on a service group during the day. Again note the periodicity of the graph with peaks building up through during prime-time each day. As noted earlier in the paper, the stream personalization needed for addressable advertising uses more capacity in the service group then otherwise switched digital video would. Hence, overlaying addressable advertising onto this service group would require more bandwidth throughout the day

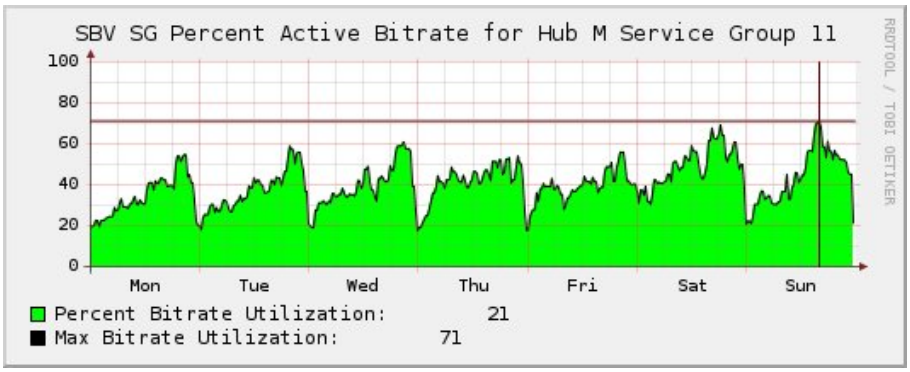


Figure 9 – Active Bit Rate per Service Group

And when coupled with the information from Figure 8, that more viewers are on the system during prime-time, it should be clear that to reach the most viewers with addressable advertising that additional service group capacity will be needed. However, because there is extra service group capacity available during off-peak times, it is possible to get started with addressable advertising with a switched digital system as engineered today. The addressable advertising system can make more efficient usage of service group capacity by filling up an otherwise empty service group with revenue generating addressable advertising during the times of the day when the QAMs might otherwise sit empty and un-used.

To more fully reap the benefits of addressable advertising as the system begins to pay for itself with additional ad revenues, the operator can grow additional QAM capacity into the service group to deliver additional addressable advertising during the times of peak service group usage. And once those QAMs are available, they can be used for other types of personalization as well including personalized guides and mosaics.

In general, a working philosophy for edge QAM utilization is to pick a value that if the QAM group is above the threshold, additional streams will not be used for advertising (unless the required network is not already in the Service Group). The current number is believed to be between 80 and 90% utilization. The following examples illustrate how the operator can get started with addressable advertising with the currently available QAM service group capacity.

Examples

Scenario 1: A viewer tunes to a program and there is already a version of that program on that service group assigned to demographic 3, but the addressable advertising system decides that this viewer should get demographic 7 in order to play different ads for additional revenue. The service group is currently only 20% full, so the SDV manager allows the creation of a new feed of the program assigned to demographic 7, places that program onto the service group and feeds the tuning info to that viewer.

Scenario 2: Same as scenario 1, except that the service group is now 93% full meaning there is no additional capacity to create a new feed of the program. In this case, the SDV server will join the viewer to a program that already exists and most closely matches their demographic.

Scenario 3: The viewer selects a program which is not currently switched onto the service group; however, the service group is 93% full. This viewer should get the demographic 7 version of the program. Since this is a tuning request for a program which is not already on the service group, the SDV session manager will add that program with a demographic 7 version and then tune the viewer to that program.

Because the addressable advertising system interfaces with the edge resource management system, the net effect of addressable advertising on edge bandwidth therefore should be effectively zero. By monitoring service group usage, if over-time there is not enough edge QAM bandwidth to run the scheduled ad campaigns, then the operator should consider a bandwidth expansion on specific service groups where capacity is an issue. But at this point in time, there should be definite revenue numbers associated with the addressable advertising which will offset the bandwidth expansion.

SUMMARY

Cable local and spot advertising is currently a $5 billion business, and it is widely accepted that adding addressability will grow that number. This paper presents an architecture for linear addressable advertising which builds upon existing linear splicing by adding Switched Digital Video as a means of personalization. Using SDV allows the operator to offer addressable advertising in a number of contexts including both switched groupcast and switched unicast.

The paper identifies several areas where planning may be necessary for addressable advertising including reviewing transport network capacity and switched service group usage. Planning can help ensure that the service meets the needs to better target advertising while giving the customer a better viewing experience, more HDTV and a more user friendly guide experience.

INFRASTRUCTURE CAPABILTIES SUPPORTING CABLE'S NATIONAL PLATFORM

James Mumma, Sr. Director of Video Product Development, Comcast Cable
Doug Jones, Chief Architect, BigBand Networks

Abstract

One of the major initiatives for the cable industry is the introduction of functionality giving subscribers opportunities to interact with applications and services through their televisions. Doing so will enhance viewing experiences, usher in new revenue opportunities and provide competitive differentiation to satellite broadcasters and the telephone companies.

The ETV and the tru2way family of specifications available at CableLabs describe how applications can be bound to programming allowing cable to deliver a national platform for advertising and other services. While there are industry specifications for delivering bound applications to a set-top box, there are no specifications defined on the infrastructure capabilities needed to manage these bound applications. This paper proposes a technical architecture and capabilities that can be used to manage and deliver bound applications (in both ETV and OCAP formats) capable of providing operators with a flexible platform for advanced services delivery.

ETV and OCAP applications are bound to individual programs by carrying those applications on MPEG-2 PIDs (Program Identifiers) that are included along with the programming. There can be multiple PIDs associated with a bound program and the paper proposes a flexible architecture to manage them. These include:

- *Passing bound applications, which include extra PIDs, through headend equipment;*
- *Capability to dynamically add or drop individual PIDs associated with bound programs;*
- *Protocol interfaces to manage the manipulations of identifiers associated with bound programs;*
- *Interoperability between the HFC resource management system and the PID insertion function to account for the additional bandwidth used on a QAM as bound applications are managed;*
- *An overall control mechanism to coordinate the management of bound applications with programmers, both national and local.*

With a proper management framework bound applications will provide both a platform for national services as well as personalized services. The ETV and OCAP toolset provides for a plethora of services, but the management and control architecture needs to be designed in order to achieve the full potential for innovation of which it is capable. The authors examine the requirements associated with management and control and explain how present capabilities can evolve to satisfy them.

THE NATIONAL PLATFORM

Until recently, implementing interactive subscriber services on a national basis was not feasible due to the lack of implementation standards in the cable industry. The splintered approach of proprietary technologies was cost prohibitive for content providers and distributors. Today, with new and emerging specifications such as ETV, which includes the Enhanced TV Binary Interchange Format (EBIF) and the tru2way, which includes the Open Cable Application Platform (OCAP),

specifications developed by CableLabs, the proliferation of interactive TV in North America is closer to becoming reality. However deployment challenges still remain for cable operators.

The benefits of interactive TV to subscribers, programmers, advertisers and cable operators is mutual, as figure 1 shows. The shared benefits that interactive TV offers each of these key stakeholders provides fertile ground for a nascent ecosystem, with the potential to improve the viewing experience, while driving new corporate revenue streams.

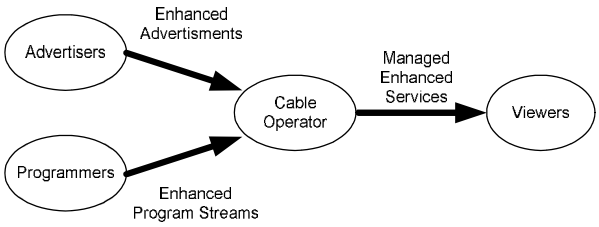


Figure 1 – Enhanced Programming Flow

Interactive TV application will come in two forms:

- Synchronous, or bound to the programming;
- Non-synchronous, or unbound to programming.

Bound applications are associated with specific programming. While a subscriber is watching a program, the ETV application will run, resulting in an enhanced viewing experience. A bound application provides the ability to interact with a program or with advertising. For example, a subscriber would be enabled to vote during a reality TV programming directly through the use of the remote control rather than through a secondary device such as a phone or PC. In another scenario a subscriber can request more information about a product or service promoted during an advertisement. While clearly enhancing the subscriber experience, bound applications garner the most interest from content providers because of the economic potential. Currently bound applications can be supported in linear and VOD content and will be supported in time-shifted content in the near future.

Unbound applications are not associated with specific programming and are not implemented as part of a programming service; they are, instead, resident in the set-top box and can be run regardless of the programming being watched. Examples of unbound applications include the standard guide or CallerID to the TV, both of which can be rendered on the TV regardless of the currently tuned linear broadcast.

The delivery of applications, specifically bound applications, remains in its infancy. The CableLabs' specifications define how the bound applications should be interpreted at the set-top box but stops short of governing supporting infrastructure components; justifiably so, since this area needs to evolve to ensure the most efficient and innovative ways to manage bound applications.

The focus of this paper is a suite of capabilities that will support the national deployment of interactive TV by streamlining specific components that enable the localization and personalization of bound interactive TV applications.

BUSINESS DRIVERS OF INTERACTIVE TV

Interactive TV offers advantages to key stakeholders including subscribers, content providers, programmers, advertisers, and cable operators through interaction, personalization and localization of content.

Subscribers

One does not need to look far to find evidence of subscribers' interest in

participating, shaping and enhancing content. Personalized websites, video outlets such as youtube.com and facebook, and the popularity of reality TV in which subscribers can determine the outcome of the programming align with the principles that interactive TV will enable. The adoption of these, and similar, phenomena indicate that subscribers should quickly embrace interactive TV as well.

Programmers

With increasing alternatives to broadcast TV, programmers continue to vie for 'eyeballs' of live programming. Interactive TV differentiates their programming and strengthens their brand with enhancements. Straightforward opportunities to enhance the programming with bound applications include the capability for voting and trivia questions while providing near real-time feedback about how other subscribers responded at either the national or local level, thereby creating a sense of community.

Additionally, programmers can use interactive TV to keep subscribers 'on brand' before tuning away to alternative networks by offerings opportunities to view their VOD content, perhaps at the conclusion of program, or even offer content provided by sister networks (i.e., NBC, Bravo, USA). The programming community is well-positioned to provider subscribers with personalized and localized experiences that are compelling and difficult for the competitors to mimic.

Advertisers

As broadcast and cable advertising continue to be threatened by internet or mobile device alternatives, interactive TV provides advertisers with an effective response. Advertisers can enhance ads to fulfill requests for information about products or by using interactivity to telescope directly to VOD clips about their products to provide additional information to the subscriber. Consider the local car dealership that can promote its latest campaign during a national advertisement for the car chain by 'piggy-backing' onto that national advertisement. Interactive TV also provides incentive to subscribers to watch time-shifted advertisements once this functionality is supported.

As the CableLabs specifications are adopted and implemented among MSOs, an unrivaled national platform will emerge, providing advertisers and cable operators alike with a robust opportunity.

Cable Operators

Since subscribers, programmers and advertisers can all be counted as customers or partners, cable operators will benefit as the enabler of interactive TV. They can capitalize on their scale versus that of DBS and/or telco providers as well as their established relationships with programmers and advertisers. In addition, operators can use enhanced advertisements to promote their own offers and services. For example, a cable operator could enhance a linear promotion to telescope to a VOD clips to learn more about the available On Demand services, how to interpret their cable bill, or even sign up for a service offering directly from their TV.

Each of these key stakeholders shares a common ecosystem that powers the television business today. The benefits afforded by interactive TV across these key stakeholders provide a recipe for the broad adoption and consequent success of interactive TV that will introduce a new era in the TV viewing experience.

UNDERSTANDING THE SPECIFICATIONS

Within a proper framework, bound applications can provide both a platform for

national services as well as individually addressed services. But in order to architect the right framework one must first fully understand the applications, the available specifications and how they affect the existing architectures.

The industry specifications for ETV and tru2way are developed by CableLabs. This set of specifications provides a basis for product interoperability. The specifications were designed to be non-proprietary and open in order to support the national reach for the platform. The guiding principle varies little from the 'write once, run everywhere' model common in computer programming today.

Enhanced TV

ETV provides a way in which interactive TV applications may be deployed to legacy set-top-boxes (STBs), such as the Motorola DCT-2000 and Scientific-Atlanta Explorer 2000. Since it is estimated there are millions of deployed legacy STBs, ETV was created to allow operators to deploy interactive applications across this large footprint of STBs. ETV applications will also run on tru2way host devices.

It is important to note that ETV is supplemental to tru2way. In fact, an ETV User Agent could be implemented as a tru2way application to support ETV applications on tru2way hosts.

ETV applications are set-out in an EBIF, for use in decoding and rendering ETV constructs on the TV screen. Applications consist of a collection of one or more partitions containing resources and programmatic data. ETV applications are interpreted by a User Agent resident in the set-top box. On a tru2way host, the ETV User Agent is a bound application.

In terms of transmission from the headend, ETV requires a number of additions to a standard MPEG-2 transport stream. These include two new descriptors in the MPEG-2 Program Map Table (PMT), an EISS Table (ETV Integrated Signaling Stream) containing applications signaling and timing information and a Data Carousel for carriage of the application itself. When the receiver tunes to the transport stream that contains the ETV application, the receiver reads the PMT and determines there is an ETV application present and alerts the ETV User Agent to run the application.

ETV is well beyond lab testing and has entered field trials that are important to validate the technologies in use. ETV is generally recognized as a "fast track" item, garnering deployment priority as soon as feasibly possible.

tru2way

tru2way has several components, including a host specification, CableCard interface specification and the OCAP middleware specification. Middleware is software that provides an interface between applications and whatever system software a manufacturer chooses for a host device. The middleware is based on the widely accepted Java™ technology. By abstracting away the various consumer electronics device operating systems to a common set of middleware APIs, application developers can write an application only once and it will run on all models of tru2way devices. Cable subscribers with tru2way-enabled digital televisions, retail set-top boxes, and other interactive digital cable products will be able to receive all of the cable operator's services just as if the subscriber was leasing a comparable set-top box from the operator.

In terms of transmission from the headend, tru2way bound applications require similar additions to a standard MPEG-2 transport stream as does ETV. The application and the data files that it accesses are packaged into an

Object Carousel (OC) format, which is an extension of the MPEG-2 transport environment that exposes a file system to the device at the other end of the network. In addition, an Application Information Table (AIT) is required to tell the tru2way host both that there is an application present and where to find it. This collection of files is then multiplexed into the MPEG program stream. When the receiver tunes to the transport stream that contains the application, the tru2way system reads the AIT and launches the application; if the receiver tunes away from the service, the application is terminated.

OVERVIEW OF THE NATIONAL PLATFORM

While specifications exist for the client (i.e., how the set-top box is supposed to receive and handle bound applications), the server-side infrastructure requirements are more loosely defined. This is analogous to the VOD infrastructure in which there are numerous variations on how the service can be deployed and managed. Like VOD, this scenario presents opportunities and challenges to the broad deployment of a national interactive TV platform. The goal is to create an infrastructure framework that supports the national platform using industry specifications for defined interfaces.

In creating this framework, there are at least five different technical components of the National Platform that must be considered, as shown in Figure 2.

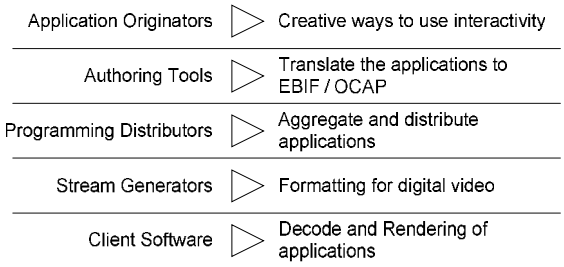


Figure 2 – Enhanced Programming Ecosystem

The applications originators are the programmers and advertisers who choose to enhance their programming. These groups may create the application in-house or outsource the application development to third-party developers.

Third-party developers often provide authoring software tools to create the enhancements for the bound applications. For the purposes of this paper, the word "enhancement" refers to an MPEG-2 program that contains an ETV or OCAP application. The authoring tools generally include easy-to-use interfaces and simulation tools to aid in the design and development process. Ultimately, these tools are used to put the applications in the correct format to be considered compliant to the industry specifications.

With bound applications there will be variations on how the enhancements will be passed from a programmer to the headend and on down to the market level. In addition to enhancements being originated directly from the programming studio, the ecosystem includes programming distributors such as the Comcast Media Center (CMC), TVN Entertainment, Headend In The Sky (HITS) and others. Additionally, there is always the possibility to swap enhancements at the local level either in the headend or deeper in the network, closer to the consumer.

Another component of the National Platform is the stream generators which put the applications into the proper format to be included with a digital program stream and the devices which actually place that enhancement into the program stream so it can be delivered to a digital set-top box. These stream generators will need to develop the capabilities to add and drop different enhancements based on the needs for localization. These stream generators need to adhere to a number of industry specifications (e.g., SCTE-130) to ensure the digital

program streams are of the proper type and format to be decoded by a set-top box.

Finally there are the digital set-top boxes which need User Agent to read the interactive signals in the broadcast stream. The User Agent is the software programs in the set-top box that interprets the interactive applications.

CHALLENGES OF THE NATIONAL PLATFORM

Broad deployment of the National Platform faces many challenges surrounding the management and control of interactive TV applications. Some of the issues to be considered are listed below.

Data PID Integrity

Data PID integrity includes successfully passing and maintaining bound applications through the National Platform without adverse impacts. The bound applications create additions to the PMT associated with that program which needs to be handled by a number of pieces of equipment in the infrastructure.

Data PID Control

The capability to dynamically manage individual or multiple PIDs, including the ability to administer and manipulate identifiers associated with the bound application, is key. If there is no business agreement between the programmer and the operator, the infrastructure needs the capability to recognize and remove the enhanced PIDs from the programming. The infrastructure needs the ability to insert non-enhanced ads into enhanced programming to support regular advertising capabilities.

Localized Operations

To support enhanced advertising, the infrastructure needs the capability to insert enhanced ads into either enhanced or non-enhanced programming. In addition, the coordination of the control of bound applications and local enhancements needs to be coordinated between the programmer / advertiser and the infrastructure. These additional control interfaces need to be developed to support more sophisticated use of the tools. This includes a general operational readiness from a business and technical level between the operator and programmer or advertiser.

Bandwidth Management

Adding enhancements to the programming has the effect of increasing the amount of bandwidth needed for that program. Since those extra PIDs associated with the PMT carry data, those extra bits and bytes need to be accounted for by the infrastructure to ensure the complete bit rate through a QAM modulator does not exceed the capacity.

Return Path Capacity

Since the programming enhancements rely on interactivity, the capacity of the return path needs to be managed. If the enhanced program is widely viewed there can be bursts of activity when all those subscribers respond to an enhancement. The goal here is to not overwhelm the return path. Legacy boxes will use the existing back-channel which is relatively low capacity compared to newer technologies such as DOCSIS®/DSG. Since there are different tiers of set-top boxes, the enhancements can be different and look better with more advanced boxes with a higher capacity back-channel.

Data Collection & Reporting

Since interactivity is managed across a national platform, there has to be a method for that interactivity to be aggregated on a massive scale. For example, some popular programs can garner tens of millions of

concurrent subscribers and if a significant portion of the subscribers vote simultaneously, the ensuing avalanche of data needs to be handled, aggregated and acted upon in a scaleable and quick fashion.

A PROPOSED ARCHITECTURE

Bound applications can be inserted into programming at either the national or local level, including personalizing the bound application at the local level. Therefore, a dynamic and flexible architecture is required to manage, control and deliver these services while accounting for national and local footprints.

Figure 3 shows an architecture which supports the National Platform including personalization at a local level. National programming originates on the left side of the figure, including the insertion of national advertising. At this point both the programming and the advertising can include enhancements. When the programming reaches the local cable operator (on the right side of the diagram), local personalization systems can further modify the enhancements and direct the programming to groups of subscribers and potentially even individual subscribers. The local personalization is implemented in conjunction with the programmer or advertiser who wants to craft a custom experience for their subscribers which otherwise would not be possible with just nationally originated programming.

Figure 3 proposes several new components to the local cable infrastructure including a new category of edge, video processing platform, known as a media services platform. This platform interfaces to the personalization engine.

Media service platforms are responsible for personalizing streams for subscribers and can selectively insert specific ads into the programming, and specific bound applications into both the programming and advertising. The media services platform will offer the data PID control interfaces as well as ensure data PID integrity for the program streams. Since personalization is done based on specific subscribers of the programming, it is best done toward the edge of the local cable network, as close to those subscribers as possible.

The bandwidth management function will be handled by the "last mile" network. The assumption for cable is that this last mile is an HFC network hence the Edge QAM plays a significant role in bandwidth management, ensuring that the enhanced programming does not overrun the capacity of the QAM channel. There are other last miles networks, including wireless, where the wireless access point would have the responsibility to ensure the wireless channel is not overrun.

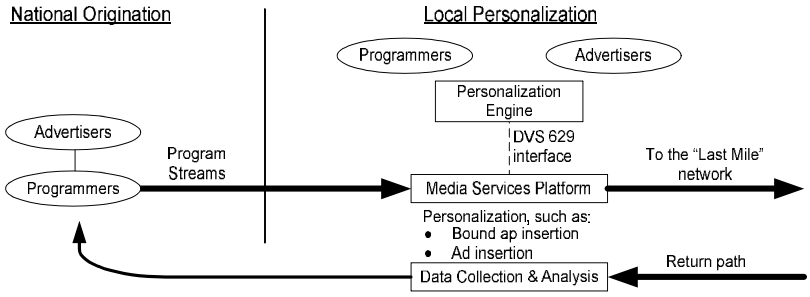


Figure 3 – Proposed Architecture

The personalization engine makes decisions about what personalizations should occur in the programming. The personalization engine has separate interfaces to both the programmers and advertisers to make close to real-time decisions about how to personalize programming based on available campaigns, operator business agreements and viewership.

These personalization decisions are passed to the media service platform over the DVS 629 interface. DVS 629 has recently been developed by the Society of Cable Telecommunications Engineers Digital Video Subcommittee (DVS) specifically for the purpose of personalizing programming. It is expected that DVS 629 will be ratified into an SCTE standard later this year and then be known as SCTE 130.

Finally, the figure shows the data collection and analysis function. Since the programming is interactive, the user responses need to be collected, aggregated and passed back to the programmers and advertisers. The analysis can be either real-time, such as voting which can be provided as feedback during the programming, or non real-time if there is no impact on the current programming.

The cable industry has been working toward personalization and the DVS 629 interface allows separation of the personalization engine and the media services platform to allow innovation to occur around that interface. Figure 4 shows additional detail around the insertion of enhancements to programming.

This figure represents how a program delivered to a customer can have both local and national enhancements for both the program itself and advertisements associated with that program. Note that the enhancements associated with the program are available during the program, but not during the advertisement where different enhancements can be available. The media services platform must not only be able to insert the correct enhancements, but must enforce the boundaries between the program and advertisements to ensure that the proper enhancement is presented to the proper subscriber, with the proper enhancement at the proper time during the program or advertisement.

Considerations

Bound applications are actual software programs and data associated with a TV program or advertisement. The bound application is inserted either at the national origination of locally, and that application is run in the set-top box. Technically the bound applications are inserted into the MPEG-2 program stream which represents that program. The bound applications are inserted as additional data on specific PIDS associated with the digital programming.

Since these bound applications represent real data, the media services platform must account for them as they are multiplexed onto the last-mile network, which for this paper could be considered a 256 QAM modulator able to carry approximately 38.8 mbps of data. Traditionally the programming is statistically multiplexed to best fit within the

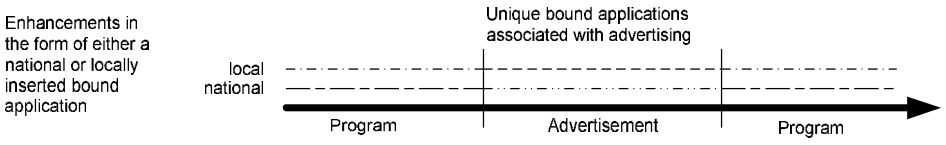


Figure 4 – Enhancements to Programming

38.8 mbps; the bound applications, however, represent additional bit rate overhead which must also be carried by that QAM but which cannot be statistically multiplexed. Clearly inserting bound applications will impact the amount of bits through the QAM and this information will need to be signaled back to the edge resource manager associated with the QAM resources in order to accurately account for QAM usage. This information will allow the edge resource manager to efficiently utilize the resources in that QAM.

Additionally, media services platforms should be able to manipulate the Program Map Table associated with the program such that the additional PIDs associated with the enhancements (both national and local) are accurately represented such that the set-top box is presented with a valid MPEG-2 bit stream to decode. The integrity of the PMT and the MPEG-2 bit stream has to be maintained even though locally the enhancements can be added or dropped in near real-time based on instructions from the personalization engine.

The personalization engine will require interfaces back to the programmers and advertisers in order to manage the personalization. These interfaces are yet to be defined but could be considered a next phase to the work being done to create DVS 629.

CONCLUSION

Interactive television offers the cable industry opportunities to improve subscribers' viewing experiences, reduce churn and enhance advertising revenues. By providing distinctive benefits to subscribers and content partners, each stakeholder will gain a significant advantage as the 'handshake' among them is redefined.

Cable operators can address the infrastructure challenges described in this paper by leveraging the architecture proposed by the authors. This can also provide the potential to launch a migration towards a more flexible national platform for personalized services and advertising.

The necessary standard interfaces are becoming available, as is the equipment needed to implement the required services and functionality. At that point, the programming experience will increasingly become limited only by the creativity of the application developers. Such an evolution will usher-in new viewing experience for cable subscribers.

IS IMS THE ANSWER?

Bruce P. McLeod
Cox Communications, Inc

Abstract

Is IMS the answer? The 3GPP IP Multimedia Subsystem is a topic of hot debate among technologists in the MSO community as to its validity as a service integration platform and core technology. It is regarded by many as a solution looking for a problem and by others as a panacea for simplifying the rapid introduction of new service types that have voice as a key component. This paper discusses the real world learning garnered by Cox Communications during our technology research and prototype development beginning in 2006 and throughout 2007. Specifically addressed will be the strengths and weaknesses of Session Initiation Protocol as an enabling integration technology and the challenges of providing next generation voice services in a world where the rules of the Public Switched Telephony Network still define much of what can (and can't) be done with new voice services.

Communications Technology has been on an evolutionary path to convergence since the need to transmit computer data from one computer to another arose in the mid twentieth century. The telephone network was adapted to support transmission of data. At the same time pure data network technology evolved and over time. With the advent of internet technology it was recognized that voice service could be considered as just another data type and in many ways could be transmitted within a data network as effectively as any other data type and Voice over IP was born.

Voice over IP technology has begun to replace traditional Public Switched Telephone Network elements across the globe. In most successful cases this transformation has gone un-noticed by the end user. This apparently seamless evolution has transpired because the design of VoIP technology has followed a path of replication of telephone service to a handset. In the future, the traditional telephone handset will remain part of the Voice service network but it will not be the only interface as it has been for over 100 years. The voice interface of the future may be a video screen, mobile PDA, a utility within a web page, or possibly something that is difficult to imagine today. IP Multimedia Subsystem is a technology that lends itself to integrating Voice to other application types. But is it the best answer for how to do this? This discussion analyzes the known requirements for voice services and how IMS addresses those.

IS IMS THE ANSWER?

If IMS is the answer, what's the question? Its simple and it has nothing to with feature abstraction, common network core, service ubiquity, or any of the flashy promises we have all heard much about. There are many ways to achieve service enrichment goals and just as many advocates and pundits about the right way to do it. So if the question is re-phrased a bit to, "What does IMS do better than any other possible service architecture?", the answer is "Take care of the guy on the other end of the line while all this neat multimedia feature stuff is going on in my network for my subs."

From an architectural perspective, simplification of the call handling must be achieved by minimizing the number of times that call control must be shifted to different applications. In the legacy telephony world the Advanced Intelligent Network service invocation mechanisms that allow applications to manage call state are able to work flawlessly because the rules are very well defined and rigidly inflexible. Unfortunately those same rules limit the communication types supported

to standard voice user scenarios and are not extensible to other media types or session type descriptors such as presence based routing policies and lack effective web integration capability. Voice over IP service invocation and call control based on Session Initiation Protocol lacks the rigor of AIN and consequently shifting call control makes things complex. SIP is a simple and very flexible protocol. It is flexible almost to a fault and the specifications are often interpreted differently by vendors. That's one of the biggest reasons issues still arise in SIP VoIP with features that have worked for decades in the TDM world.

Issue	**Legacy Solution**	**Interim Solution**	**IMS Solution**	**Comment**
Advanced Voice Features	AIN or PRI based call forward	SIP Re-direct and call control handoff	SIP App Server	The IMS SIP Application Server provides features, feature interaction management, and web integration
PSTN Routing and Call integrity	SS7 and TDM Interconnection	SIP CORE using ENUM	CSCF and MGCF	An effective SIP CLASS 4 Tandem easily evolves to MGCF and can provide some CSCF functions. MGCF maintains call state with the PSTN and masks the multimedia functions occurring in the IP domain.
Fixed Mobile Convergence	Dedicated FMC call control agent to move a call from one phone number to another	SIP "Call pull" via re-direct using Application Server	Presence and user preferences tell the network how to connect with the user	In IMS the association of a specific phone number to each voice endpoint isn't required.

Table 1

An ideal scenario for an operator is to map an inbound call to SIP only one time when it has entered their VoIP core and anchor its relationship to the PSTN is managed there. This activity is referred to in IMS terms as Media Gateway Control Function (MGCF).

At Cox we have realized that this approach minimizes the possibility of state mismatches between a Cox customer and the far end PSTN switch that can result in failed or dropped calls. The reality we observe today is that current generation application environments and systems want to handle service requests by taking over control of the call.

A way to visualize this general issue is think about having tried to manually set up a three-way call and then disengage yourself to allow the other two people talk. These often fail because you end up with a different count of call setups versus lines in use. The PSTN switches expect those counts to match so one will initiate a teardown. We have been looking for a way to replicate a trunk release in VoIP the same way a TDM PRI does it for years to no avail because the VoIP protocols are unable how to tell the far end its okay for the counts not to match the way ISDN does this with a channel transfer. This is one of the basic problems of PSTN replication with VoIP.

Cox encountered this issue when executing VoIP interoperability with a Directory Assistance provider at the VoIP level (SIP – SIP) for directory assistance call placement by making it a three way call with a dormant leg since the softswitches involved can't agree how to release it. We pay a direct price there using SIP ports that don't do anything but that's not the real issue. The real problem is this heavy handed approach breaks the general assumptions about call state and subsequent feature invocation becomes clumsy at best, because there is this third call leg involved that started the whole thing which, from a subsequent call treatment point of view, has no business being there. That is the general workaround for releasing a trunk today with VoIP. You don't release.

Simple features become complex because a call leg exists that shouldn't be there.

Fixed Mobile Convergence today is a form of trunk releasing that encounters this same general issue of trunk releasing compromise. This isn't to say that call control based FMC systems don't work, rather, that by not being able to execute a release with a base protocol, it is a heavy handed way to do it that results in making things other than just moving the call from one endpoint to the other incredibly complex. A subsequent consequence of employing a standalone FMC system is that eventually you end up having to host feature applications on this system because it was never designed to support services from another application environment. Providing feature transparency between mobile and fixed endpoints on your existing messaging system systems (or other apps) is where operators will struggle with integration to FMC because up to three call agents(Class 5, Tandem, and FMC)are involved for feature delivery and call control.

Subsequent expansion of the feature capabilities on one or the other platform is now required and portends painful integration until these systems have been augmented past their original design capability. The ultimate consequence is that the operator ends up with another service platform that doesn't perform well but is costly to replace because of all the investment made to integrate to full functionality needed. All this in the name of taking care of the guy on the other end of the line. The issues that IMS technology directly addresses with respect to the PSTN are listed in *Table 1*

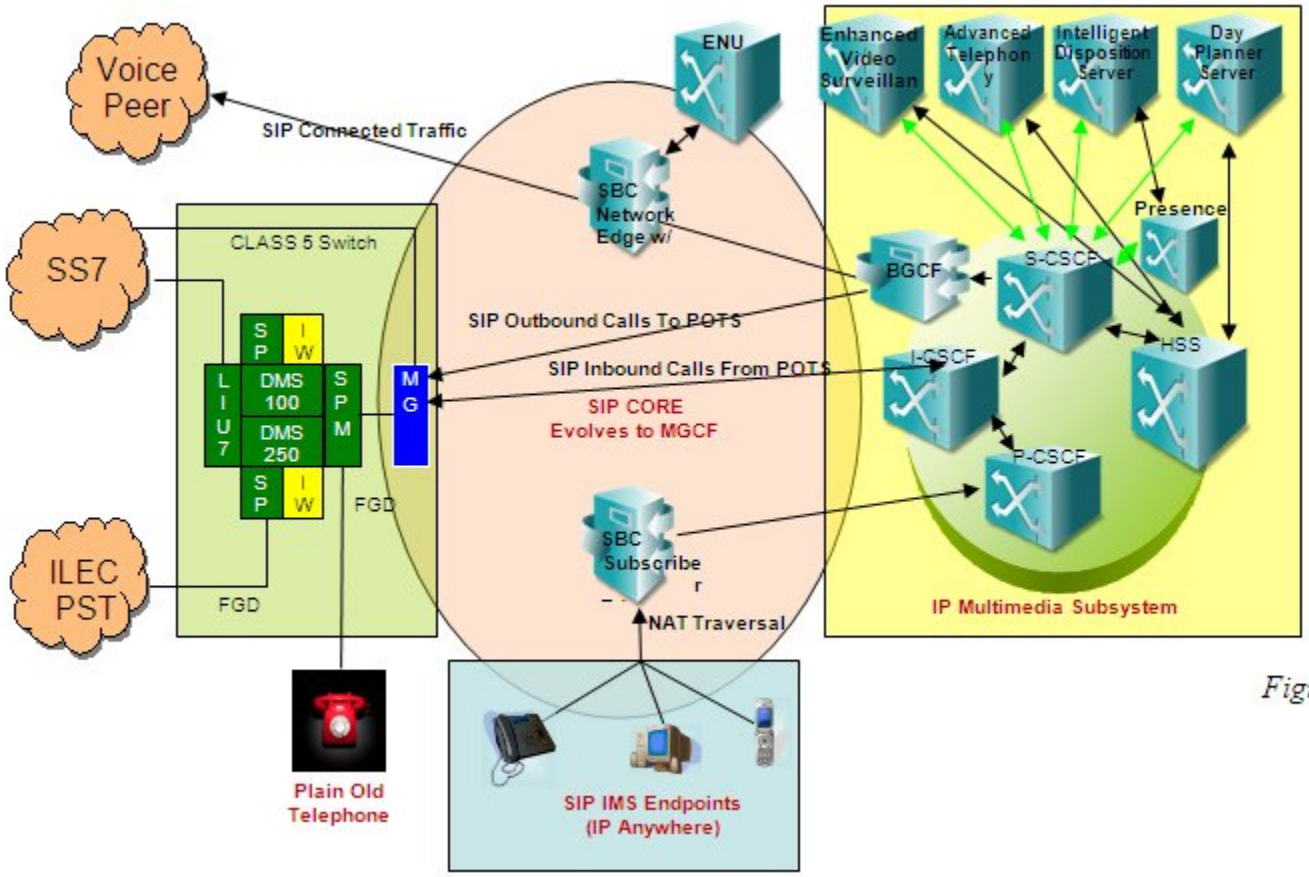


Figure 1

IMS however; addresses this directly by not allowing any one application to dominate call control with respect to the PSTN. It doesn't matter because IMS routes the inbound PSTN call to a subscriber(fully qualified domain name), not an endpoint(phone number).*Fig 1*

The concept of the far end PSTN switch that has to abide by PSTN rules doesn't exist inside of IMS, only at its border(MGCF). A good MGCF looks like just another route for a service to IMS applications. Many of today's available "stovepipe" application systems promising IMS compliance in the future but this will be a real challenge for them because when you closely inspect virtually any current generation VoIP application server you will probably see a device originally designed as a softswitch that tends to behave like one by handling connection state of that far end PSTN switch.

There are exceptions of course, some which rely on AIN and the TDM Class 5 switch it is attached to and some telephony application servers that function well being treated as a giant VoIP PBX until IMS liberates them totally from the PSTN. Good pre-IMS app servers at a minimum mask control of the call by staying in synch with the MGCF.

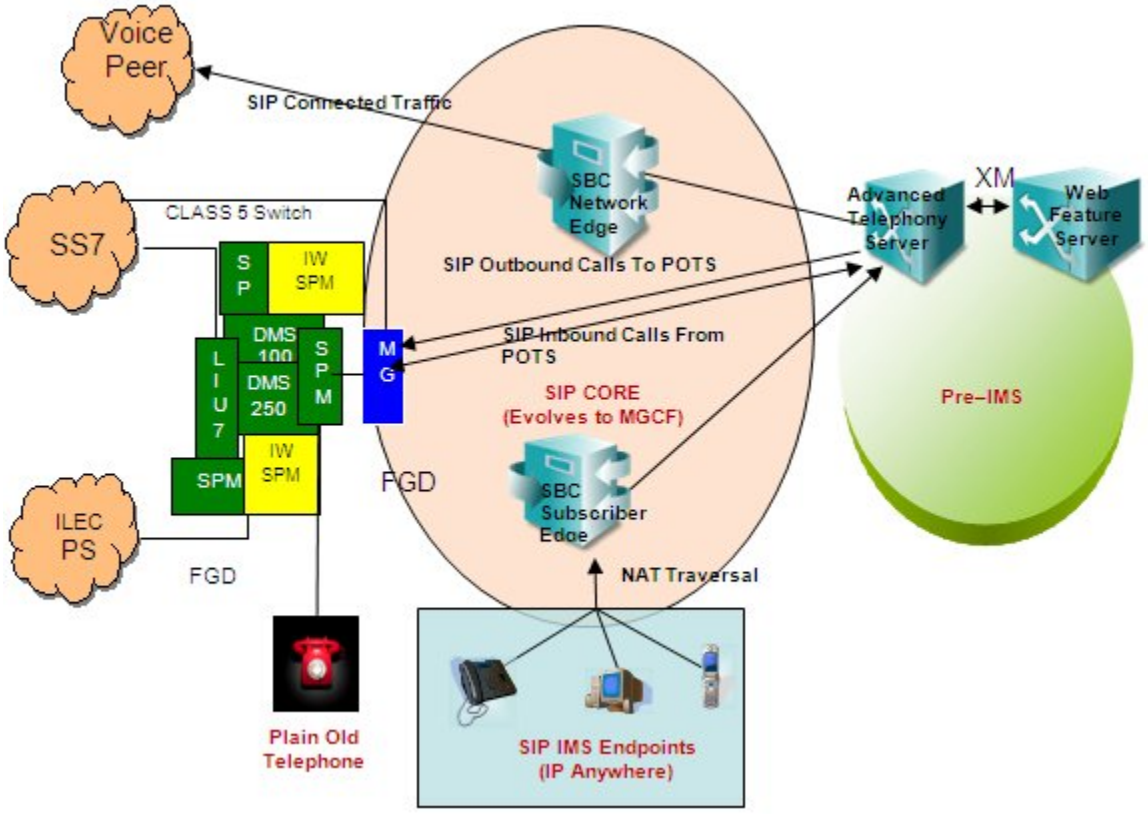


Figure 2

This opens the door to a form of FMC that can be utilized until IMS application integration has reached the level of maturity needed to extend those services to a highly reliable voice offering

This does limit the offering only to endpoints capable of supporting SIP. Inside IMS the distinction of fixed/mobile disappears so FMC based on call control handoff is meaningless. Getting applications to integrate seamlessly now becomes a design quality issue no longer restricted by the limitations imposed at the call routing layer. Feature interaction management is becoming a discipline unto itself that VoIP engineers will be practicing for as long as can be imagined.

The bottom line is that from a technology standpoint any FMC or Unified Messaging design using available application servers that directly interface to the PSTN CLASS 5 switch is a big diversion away from IMS. As such any focus on nailing up a design for FMC and UM under conventional terms becomes a full diversion away from IMS development because the core skills focused on IMS or pre IMS today are the same ones that must be used for interim solutions. Essentially a doubling of effort is required to pursue both paths.

An additional consequence is that to get to IMS, engineering teams will have to expend effort to undo the connected to legacy infrastructure adding yet more development cycles. Our experience at Cox has been that the effort to migrate systems rivals or surpasses replacing them entirely.

If you as an operator ask “Ok, what should I do instead?” the answer today is focus effort on building the MGCF that allows non-possessive application servers to work in your network and start identifying and integrating those application servers. *Fig 2* Does that get an operator to wow factor features as quickly as stovepipe systems directly connected to legacy infrastructure? Probably not but it will get you to where you really want to be faster than taking a big detour and it will allow you to stay there when you get there. Implement a uniform and reliable method of taking care of the guy on the other end of the line and there are only a few steps beyond that needed to realize the service rich environment of IMS.

MIGRATING DIGITAL AD-INSERTION APPLICATIONS FROM MPEG-2 TO AVC (H.264)

Mukta Kar, Ph.D., Cable Television Laboratories, Inc.
Sam Narasimhan, Ph.D., Motorola, Inc.

Abstract

Splicing is the fundamental technique used to insert commercials or short programs in channels, for editing audio/video content in post-production houses, and for channel switching in headends and other broadcast stations. Splicing is currently used in US Cable networks for digital ad-insertion based on MPEG-2 video [3], SCTE and ITU-T standards [4] and there are plans to migrate these applications and develop associated standards based on the emerging H.264/AVC video [1] in the near future.

In these new applications, the splicing equipment (or function) combines two independently encoded AVC streams and is expected to produce a stream for receiving equipment that conforms to both AVC Video [1] and MPEG-2 Systems [2]. To achieve significantly higher compression efficiency than that of MPEG-2 video while providing same or better quality video, AVC compression standard has introduced several new tools, reference picture structures and enhanced MPEG-2 tools all of which can be used adaptively based on the nature of the content. All these make AVC more complex compared to prior compression schemes in addition to being not backward compatible with MPEG-2 video.

Many of the Standards Development Organizations (SDOs) such as DVB/ETSI, SCTE, DVD and ARIB have completed the specifications related to the adoption of AVC in broadcast, VOD and PVR applications. This paper outlines issues related to splicing between two independently coded AVC streams for local Ad-insertion applications and proposes schemes for generating an AVC Video [1] and MPEG-2 Systems [2] conformant output by such splicing equipment so that a seamless or near-seamless splicing can be achieved.

INTRODUCTION

The opportunity for Local commercial insertion has been created to benefit the communities and its businesses on a local, zonal or regional basis since the days of Analog Television. Local commercial / advertisement (Ad) opportunity provides the US broadcast television industry over 35 billion dollars in revenue. The revenue from this opportunity for the cable industry has grown from a few million to a few billion dollars with 5-6 billion in revenue expected in 2008. As it provides a significant cash flow for our MSOs, increasing this revenue further has a prime importance to the industry. In the days of analog television, most local Ad-insertion equipment was proprietary in nature and hence non-interoperable. In moving from the era of analog television to a digital one, the cable industry understood the problems in using proprietary equipment and the advantages in using inter-operable equipment from a multi-vendor market place. To create such a competitive multi-vendor market place, the cable industry took initiative in standardization efforts in both the international (ITU/ISO) and national levels (SCTE) that covered not only Audio-Video but other areas such as cable modem, VOIP, etc. One such application area is the local program/commercial insertion.

Figure 1 displays a block schematic of a typical local ad-insertion system. A timing

signal known as Cue-tone (in analog) or Cue-message [4] (in digital) is embedded with a program and then distributed to the headends or broadcast affiliates via satellite. MSOs receive such a program using an IRD (Integrated Receiver Decoder) at their headend. Then the headend equipment separates Cue-tone/Cue-message from the program. Based on the timing signal in the cue-tone/cue-message, a splicer and Ad-server replaces the national ad with a local ad. This process is known as local ad insertion.

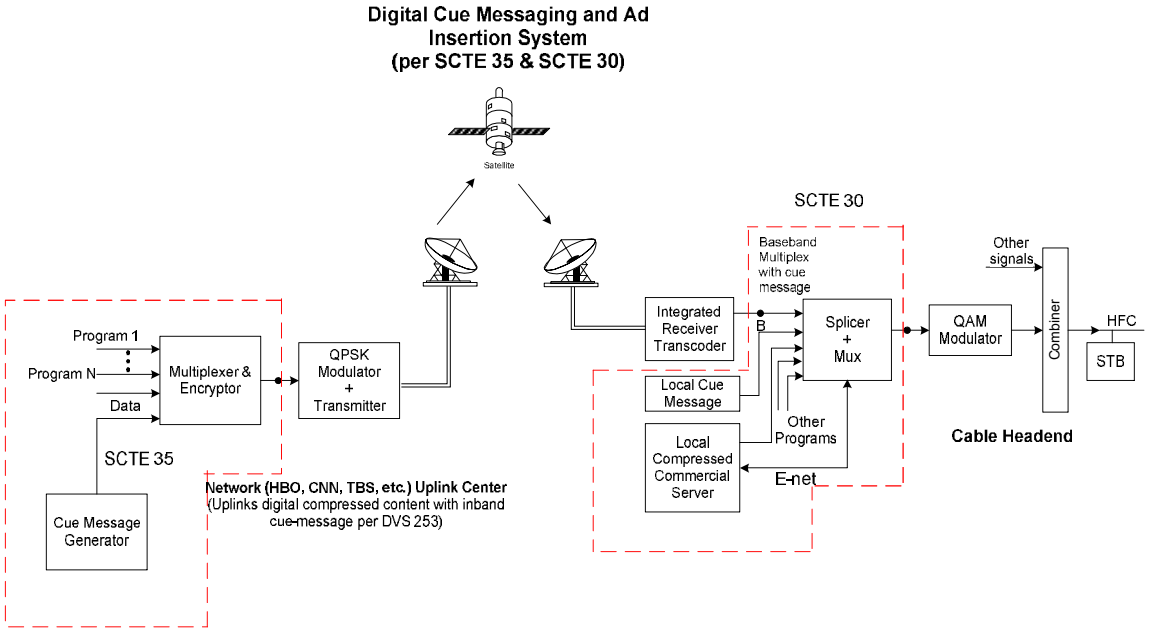


Figure 1. Schematic Block Diagram of Digital Program/Ad Insertion System

As shown in Figure 2, Local Ad Insertion technology in the analog video domain was simple in nature as the transmit order of frames is same as the display order. Splicing the digital uncompressed video is also simple for the same reasons. The process involves frame accurate timing signals indicating the beginning and end of a national advertisement in a program that need to be replaced with a local ad or perhaps with an updated ad. Delivery of analog video or uncompressed video to consumer homes is very inefficient in the usage of bandwidth in addition to many other limitations. Digital video compression technology coupled with digital modulation provides significant efficiency, flexibility and other benefits in delivering digital video and audio to consumer homes. Also digital technology provides superior video quality compared to analog as the former is less prone to noise.

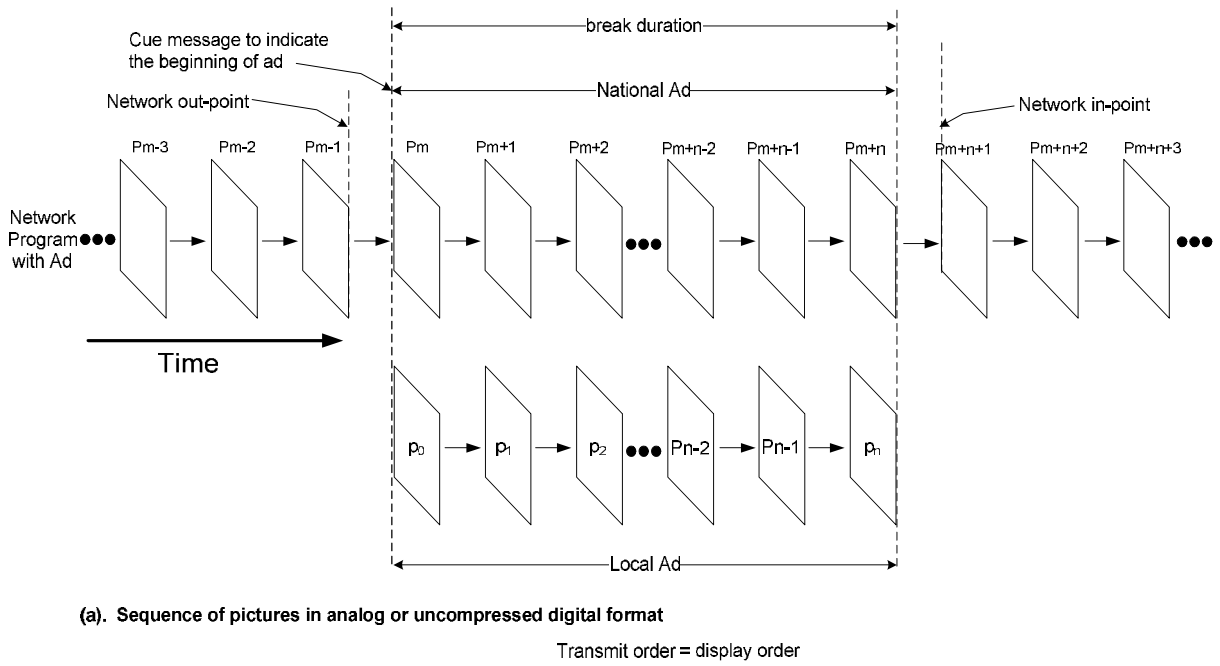


Figure 2. Simplified Diagram of Local Ad Insertion

However, digital video compression technology introduces problems for some of these broadcast applications (e.g. splicing between two MPEG-2 streams or between two AVC streams) primarily due to two reasons – (a) Order of the video frames gets modified in compressed domain (transmission order does not maintain display order) and (b) compressed frames depend on other frame or frames (called as reference frames) for decoding / decompressing them in the decoder.

SPLICING BETWEEN STREAMS WITH MPEG-2 VIDEO

Figure 3(a) illustrates a segment of video where frames are in display order. Figure 3(b) depicts the same segment when compressed in compliance with the MPEG-2 video standard and sent over a transmission channel. One may notice that the transmit/decode order is not same as the display order and hence a splice cannot be done at all picture boundaries. For example if one splices out at any of the B pictures, then this will introduce gaps in the display. For seamless and near seamless splicing between two compressed video streams, MPEG-2 TSTD buffer conformance must also be maintained where decoder buffer (buffer size of 1.8Mbits for MP@ML) must not overflow or underflow as this may result in artifacts. Typical buffer behavior in a MPEG-2 decoder is shown in Figure 4 assuming the decoder receives a constant bitrate channel. MPEG-2 provided tools to achieve seamless or near seamless splicing but it does not tell on how to achieve it. MPEG left it to MPEG-2 product designers for innovation and product differentiation. To achieve seamless or near seamless splicing, some constraints may have to be maintained while creating the streams to be spliced. Such constraints may include GOP structure, an anchor frame at the out-point of the first stream and an I frame with a sequence header and closed GOP at in-point of the second stream.

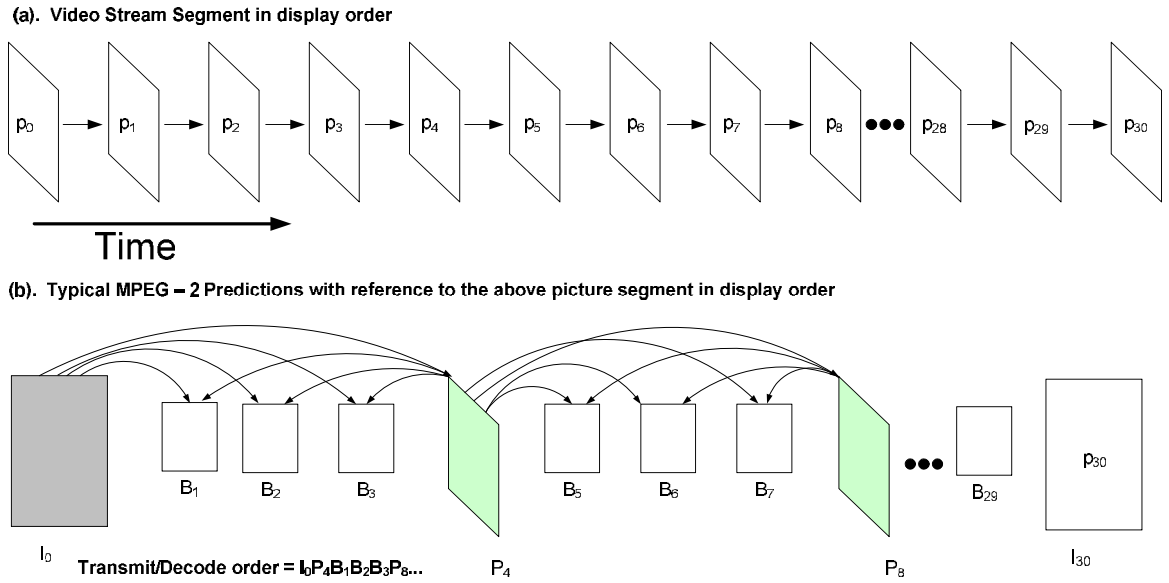


Figure 3. Typical Prediction Methodology used in MPEG-2 Video Compression.

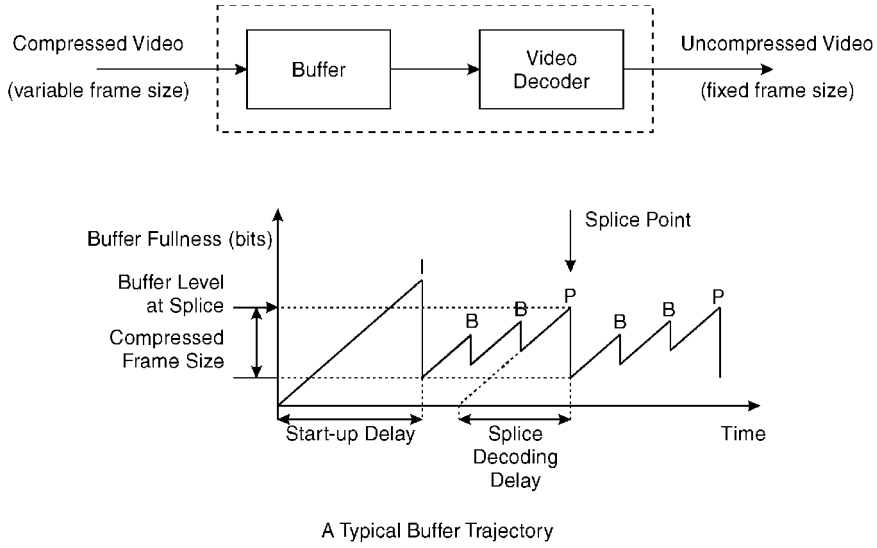


Figure 4. Simplified diagram of an MPEG-2 decoder and level of bits in the decoder buffer.

An additional function that needs to be met by digital ad-insertion systems is the matching of decode delays between the network and splice stream as shown in Figure 4.

Even though it may be easy to splice between two 'well conditioned' MPEG-2 transport streams (standards such as SMPTE 312M specify this stream conditioning), the stream conditioning and matching impose too much of a burden and constraint on the uplink encoders and ad-servers. Hence SCTE developed specifications such as SCTE 35 [4] (digital cue-message standard) to signal the splice opportunities in the compressed video stream and splicers were developed to perform the tasks outlined above so that splicing can occur without imposing too many constraints on

the uplink or ad-servers. Majority of the splicers deployed currently perform the following functions:

- Continuous bitrate transcoding to maintain an average compressed bitrate between the two video streams.
- Time base adjustment to maintain a common PCR, PTS and DTS between the streams without introduction of any discontinuities.
- Matching the vbv-delay between the two streams so that decode and presentation are continuous when the output reaches the settop units at consumer premises.
- Splice between content in film-mode and non-film-mode by maintaining field parity.

In addition, the splicers also maintain conformance to MPEG-2 video and systems standards in their output and make sure that PSI information does not change across the splice so that settop units can present a seamless transition between network program and advertisements.

AVC VIDEO CODING AND HIERARCHICAL GOP STRUCTURES

It has been mentioned earlier that to achieve better compression efficiency than that of MPEG-2 video standard [3], AVC [1] introduced many new tools and enhanced some MPEG-2 tools. One of these tools is in the use of B-pictures as reference and hierarchical use of such B pictures.(MPEG-2 video does not allow B frames to be used as reference). This particular tool introduces an additional complexity for splicing and this will be discussed later. The MPEG/JVT committee also structured AVC in a very flexible way so that it can be implemented in a wide range of applications that includes broadcast, video telephony, video conferencing, and video streaming.

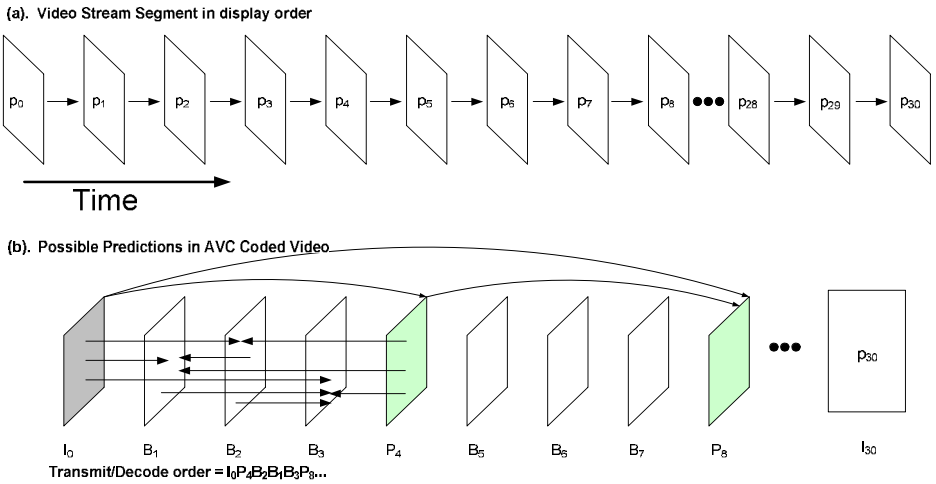


Figure 5. Typical Predication Methodology used in AVC Video Compression.

Use of B frames for reference in AVC and the associated GOP structures that use this are sometimes called 'Hierarchical GOP structures'. Figure 5(b) shows a typical GOP structure in AVC video where I, P and B frames are used. Frames are in display order in Figure 5(a), where as Figure 5(b) shows a typical compression structure of AVC where hierarchical GOP structures are used with B pictures as reference. One may notice that the

difference in transmit/decode order of pictures in Figure 3(b) and Figure 5(b). The use of B pictures as reference and GOP hierarchy makes decoding AVC coded stream more complex compared to MPEG-2 coded video with respect to the management of reference pictures in the decoder memory. AVC also introduced IDR picture and changed traditional definition of I picture which is used in MPEG-2 video. These advanced tools and flexibility of GOP structures make seamless or near-seamless splicing using AVC video [1] for local ad-insertion more challenging compared to MPEG-2 video [3].

ADDITIONAL SPLICING ISSUES WITH AVC VIDEO COMPARED TO MPEG-2

AVC splicers have to implement all the functions that are implemented by MPEG-2 video splicers. These include bitrate transcoding, time base adjustment, CPB delay matching and maintaining field parity between film and non-film modes. In addition, AVC splicers need to manage another function to accommodate the 'Hierarchical GOP' structure variations between the streams being spliced. The following illustrates the issue and proposed solutions.

MPEG-2 provides relatively simple GOP structure involving I, P, and B pictures. Depending on a scene the parameter m (number of B picture between two anchor frames) varies. MPEG-2 allows only two reference pictures at any time which are managed in the decoder memory using FIFO method (also known as bumping process) where arrival of a new reference picture pushes out the older reference picture out of the decoder memory. In MPEG-2 (for non-low-delay mode) the delay between the decode time of first access unit in the sequence and the display time of first access unit in the sequence is always 'one frame period' for both closed GOP (I,P,B,B,P,B..) and open GOP (I,B,B,P,B..) structures. Let us call this as the 'display latency'. This display latency is always 'one frame period' for MPEG-2 sequences that use different values of m and hence MPEG-2 splicer's were able to concatenate any two sequences and still maintain conformance to T-STD, VBV and constant display rate at their output without any difficulty.

In AVC this is not true as the 'display latency' for different video sequences vary based on the hierarchical GOP structures which use B pictures as reference pictures. The display latency can range from 'one picture period' (for non-stored-B structures or MPEG-2 like structures) to several picture periods based on the levels of hierarchy. In addition, unlike MPEG-2 video AVC mandates the maintenance of constant display rate (constant DPB output) when coded video sequences are concatenated. This makes concatenation or splicing two such video sequences with different display latencies difficult as the output cannot conform to AVC specifications (I.E; maintain constant delta in the CPB removal time and DPB output time). If the display latencies do not match, then one will see a missing picture at DPB output time or see 2 access units with the same DPB output time. One solution to this is to only combine sequences that have the same GOP structure and this mandate is not attractive in Cable networks for applications that use splicing. Based on inputs from the US Cable community, AVC has agreed to modify the standard allowing the use of a marker called end_of_stream NAL unit to splice between two AVC coded video sequences or streams with different display latency where the requirement to maintain constant DPB output does not apply. The next section covers proposed solutions for splicing between AVC streams based on this action by AVC.

PROPOSED SOLUTIONS FOR BOTH STREAM CONDITIONING AND SPLICER FUNCTIONS

1. In order to enable seamless splicing between two video sequences at different horizontal

resolutions (at same frame rate and vertical resolution), the application standards such as SCTE 128 [5] are mandating the same number of pictures in the DPB for all the horizontal resolutions (determined by the highest horizontal resolution for the level such as 720 for SD and 1920 for HD). This allows the decoders to keep the same DPB memory management across the resolution change and hence produce seamless output.

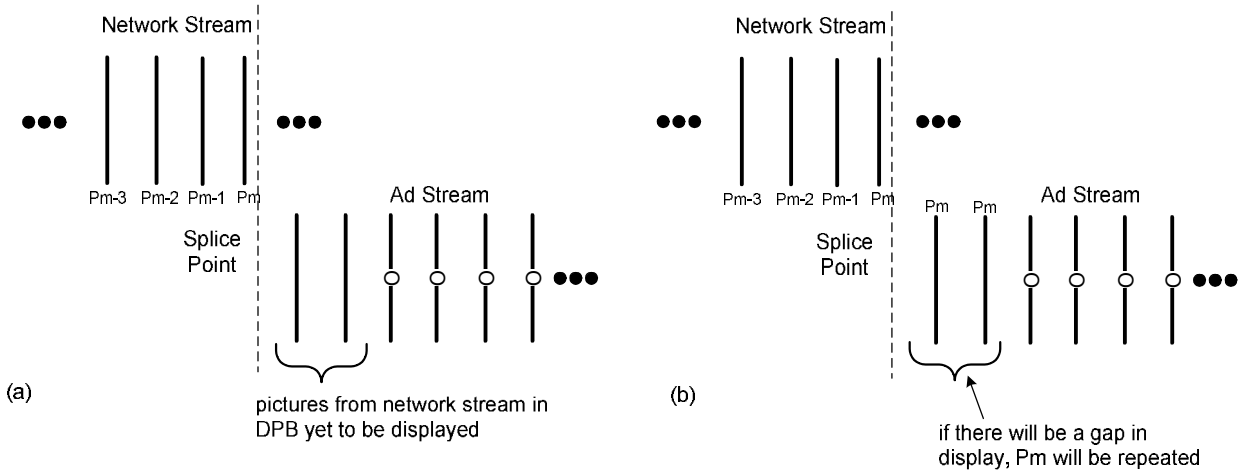


Figure 6. Seamless / near-seamless splicing without the use of end_of_stream NAL unit at the splice point (a) a case where no_output_of_prior_pics_flag=0, (b) a near-seamless case where no_output_of_prior_pics_flag=1.

2. AVC also agreed to loosen the requirement for decoders to infer the DPB management using the no_output_of_prior_pics_flag (I.E., infer this to be '1' and clear the DPB when there is a resolution change). The change allows the application standards to mandate that receivers process this flag correctly so that DPB is managed per the transmission systems intent. Splicers can set this flag correctly at the transition points to achieve 'seamless' splicing as shown in Figure 6(a).

3. The third proposal is the appropriate use of end_of_stream NAL unit at the splice transition points (called Out or In-Point) so that seamless or near-seamless splicing is possible. This is shown in Figure 7. In some combinations, seamless splicing can be achieved without the use of end_of_stream NAL unit. The first example is where the display latencies match between the streams. The second example is where the display latency difference can be adjusted by the use of Picture timing SEI message with an appropriate value for pic_struct. This SEI with the pic_struct value allows repetition of last displayed picture and this can be used to splice a stream with a higher display latency into a stream with a lower display latency. This also requires the first stream to be coded using frame pictures. For all other combinations of streams, the end_of_stream NAL unit should be used with the correct setting of the no_output_of_prior_pics_flag at the transition point to manage the DPB buffer and make sure that two pictures with the same display time are precluded. Seamless splicing with end_of_stream NAL unit can also be achieved by offsetting the decode time of the pictures in Ad Stream appropriately. This mode is not recommended as most receivers expect the decode time to be contiguous between network and ad-streams.

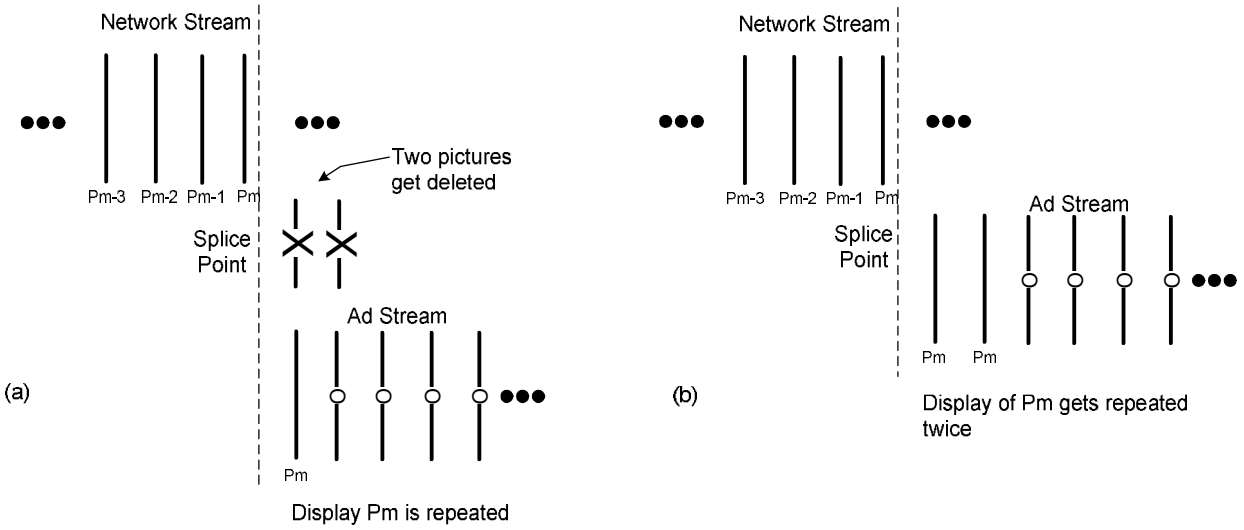


Figure 7. Near-seamless splicing using end_of_stream NAL marker at the splice point (a) a case where no_output_of_prior_pics_flag=1 (b) a case where no_output_of_prior_pics_flag=0.

SUMMARY

In this paper the technology of splicing and its importance to achieve local ad-insertion for the broadcast industry have been discussed. The local ad-insertion application provides significant amount of revenue to the industry, in particular, the cable MSOs. The challenges of splicing two video streams in the compressed domain (e.g. MPEG-2 video) for seamless and near-seamless viewing experience have been discussed. It has been also shown that splicing of two AVC video streams is much more difficult than MPEG-2 video as AVC video coding uses more advanced video tools and complex coding structure to achieve higher compression efficiency. This paper proposes a few methods/solutions to splice AVC streams to achieve seamless or near-seamless ad-insertion as needed in the broadcast industry.

ACKNOWLEDGEMENTS

The authors wish to thank the management of CableLabs and Motorola for their support and encouragement in performing this work.

BIBLIOGRAPHY

1. ITU-T Rec. H.264 | ISO/IEC 14496-10, (2005), "Information Technology – Coding of audio visual objects – Part 10: Advanced Video Coding."

2. ISO/IEC 13818-1, (2007), "Information Technology – Generic coding of moving pictures and associated audio – Part 1: systems."

3. ISO/IEC 13818-1, (2000), "Information Technology – Generic coding of moving pictures and associated audio – Part 2: video."

4. ANSI/SCTE 35-2004 | ITU-T J.181, Digital Program Insertion Cueing Message for Cable.

SCTE 128 (2007), AVC Video Systems and Transport Constraints for Cable Television

MOBILE TV: A TECHNICAL AND ECONOMIC COMPARISON OF BROADCAST, MULTICAST AND UNICAST ALTERNATIVES AND THE IMPLICATIONS FOR CABLE

Michael Eagles, UPC Broadband
Tim Burke, Liberty Global Inc.

Abstract

The growth of mobile user terminals suitable for multi-media consumption, combined with emerging mobile multi-media applications and the increasing capacities of wireless technology, provide a case for understanding facilities-based mobile broadcast, multicast and unicast technologies as a complement to fixed line broadcast video.

In developing a view of mobile TV as a compliment to cable broadcast video; this paper considers the drivers for future facilities-based mobile TV technology, alternative mobile TV distribution platforms, and, compares the economics for the delivery of mobile TV services.

We develop a taxonomy to compare the alternatives, and explore broadcast technologies such as DVB-H, DVH-SH and MediaFLO, multicast technologies such as out-of-band and in-band MBMS, and unicast or streaming platforms.

INTRODUCTION

Cable MSOs operate in an increasingly competitive market with incumbent Telcos and independent wireless operators. Cable's early victories in the voice market led to an aggressive response to offer video products by the Telcos.

The next area for intense Telco competition will likely be mobile television. The addition of television to their mobile voice and data products may be the logical next step ... but it may not be for cable.

We provide a toolkit for the MSO to assess the technical options and the economics of each.

Mobile TV is not a "one-size-fits-all" opportunity; the implications for cable depend on several factors including regional and regulatory variations and the competitive situation.

In this paper, we consider the drivers for mobile TV, compare the mobile TV alternatives and assess the mobile TV business model.

EVALUATING THE DRIVERS FOR MOBILE TV

Technology drivers for adoption of facilities-based mobile TV that will be considered include:

- Innovation in mobile TV user terminals - the feature evolution and growth in mobile TV user terminals, availability of chipsets and handsets, and compression algorithms,
- Availability of spectrum - the state of mobile broadcast standardization, licensing and spectral harmonization,
- Evolution of network technology – the increasing capacity of wireless bandwidth the emerging mobile return path and channel change improvements,
- Usage context and prospects – demographics, viewership, and subscriber willingness to pay.

1. Innovation In Mobile TV User Terminals

As a key driver for mobile TV, advances in user terminals enable new features and usage models that enhance the mobile TV experience. We believe this trend will result in a wide availability of handsets capable of receiving mobile TV over time.

In particular, increasing screen sizes, resolutions, and decreasing power consumption, support longer usage period and usage scenarios and enable a greater number of radio and network alternatives.

We note that mobile TV user terminals supporting mobile TV are predominantly targeting QVGA resolutions today. However, increasing mobile screen resolutions may drive a need for higher bandwidth in the future.

Table 1: User Terminal Resolutions [1]

	Example User Terminal	Width	Height	Total Pixels
VGA	Nokia N800	640	480	307,200
HVGA	Apple iPhone	480	320	153,600
QVGA	Samsung P910	320	240	76,800
QCIF	Motorola V8	176	144	25,344

How much bandwidth is required to support QVGA at 20 frames per second?

We can estimate this by considering QVGA resolution of 76,800 x 24 bit colour x 20 frames per second = 36,864,000 bits per second. Assuming a compression rate of 141[2] provides an approximate video bandwidth of 256 Kbps.

Can today's multi-media user terminals support full frame rate broadcast video?

Over time we believe all mobile TV user terminals will be able to support full frame rate video. Early mobile TV user terminals could not process QVGA resolution at 25 fps or higher[3], typical of most broadcast systems. For example, the Nokia N92 and N77 could not support this frame rate due to processing limitations. This is changing with the new Nokia N96 being capable of up to 30fps at QVGA resolution.

With mobile TV user terminal processing capability improving, it will become an operator decision regarding support for full frame rate video, as the bandwidth required is around 1.5-1.8x higher than today's 256Kbps bit rates, at around 400 Kbps.

Some common data rates for mobile TV are highlighted below. The analysis that follows in this paper will focus on the Class B/Medium data rate.

Table 2: Common Data Rates for Mobile TV [4]

	Data rate (Kbps)	Frames per second (fps)
Class A	128	10 – 12
Class B – Low	256	15 – 20
Class B – Medium	256	30
Class B – High	384	20 – 25
Class C	768	30

How do we know what users consider the minimum acceptable quality when viewing mobile TV user terminals?

Several studies have been conducted into the acceptability of mobile TV content at varying resolutions and varying bit rates.

Figure 1: Mobile TV User Terminal Acceptability of Video [5]

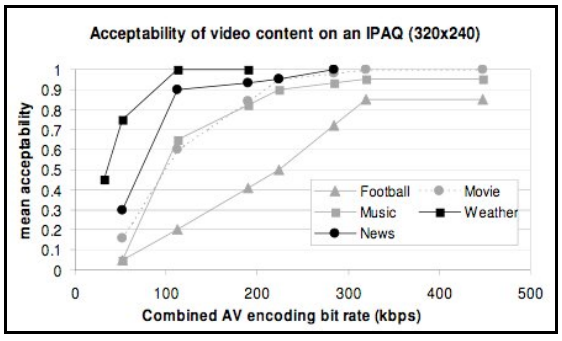


Figure 1, above, indicates that QVGA mobile video acceptability for football reached a plateau for bitrates of 332 Kbps and greater than 84%. In contrast news and weather delivered an

acceptable service to 90% of people at bandwidth of just 112Kbps[6].

What compression improvements are possible with advances in mobile TV user terminals?

As full frame rate handsets become available, requiring higher bandwidth, operators will look to advances in compression technology in order to maximize use of finite mobile TV bandwidth.

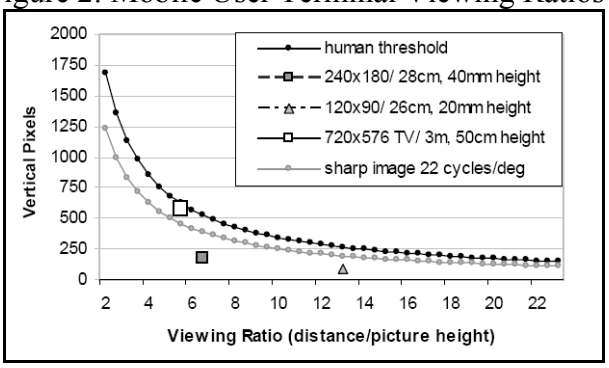
Table 3: Mobile TV Compression Improvements

	2008	2009	2010
Compression Profile	MPEG4 AVC / VC1	MPEG4 AVC / VC1 enhanced	MPEG4 AVC / VC1 improved
Percent Improvement Possible	Today	10 – 15%	10- 15%

Does a smaller screen size make mobile TV user terminals less attractive to the viewer than cable's typical fixed-line TV?

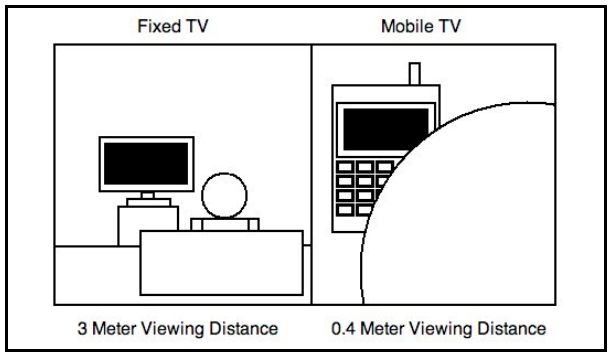
Perhaps counter-intuitively, studies indicate that standard television is much closer to the limits of human perception than mobile TV user terminals.[7]

Figure 2: Mobile User Terminal Viewing Ratios



Interestingly, if we represent this visually in a fixed and mobile scenario, we can see that a typical living-room fixed widescreen TV at 3 meters can be visually similar to a QVGA screen held 40 centimeters from the eyes, as seen in figure 3 below.

Figure 3: Visual Saturation for Fixed and Mobile TV



However, research indicates there are some key perceived issues with mobile TV user terminals for viewing TV, other than 'general detail' and 'image size'. These included 'fatigue' and 'effort', perhaps associated with correct positioning of the user terminal.

Table 4: Mobile TV User Terminal Problems across All Content [8]

Problem	% of General Comments
General Detail	20%
Insufficient Image Size	18%
Fatigue	10%
Effort	8%

What is the availability of mobile TV chipsets and user terminals ?

The availability of user terminals capable of delivering mobile TV is increasing over time. Variations in different regional approaches to standardization and service characteristics have resulted in broader user terminal availability in Asia.

For example, mobile TV user terminals in Japan have reached 20 million units shipped in just under two years since launch, noting that Japan's mobile digital TV service ISDB-T (OneSeg) offers a simulcast of Japanese terrestrial TV

stations at no cost to the end user. The broadcasts are not secured, which facilitates the production of low-cost compatible devices and handset diversity[9].

Other markets appear to be challenged with respect to handset diversity due to the emergence of multiple differing distribution standards (i.e. DVB-H, MediaFLO), the need for diversity and potentially filters (i.e. TDtv), and the need for security support (i.e. OMA Bcast standard) capabilities to protect content.

We believe that as the number of mobile TV distribution standards proliferate, the emergence of handsets with the ability to support multiple technology options will emerge.

2. Availability of Spectrum

Another key driver for mobile TV includes the availability of spectrum, including the state of mobile broadcast standardization, licensing and spectral harmonization.

Considering the standards for mobile TV, we note that today there are five worldwide broadcast TV standards (DVB-H™[10], MediaFLO™[11], ISDB-T, T-DMB, S-DMB) and three more broadcast standards planned (MHP, DVB-SH and CMBB).

When the two most widely known multicast standards (TDtv & MBMS) and the entire category of unicast (in band cellular) are added to the mix it is apparent that the world of mobile TV technology is extremely fragmented[12].

Table 5: Current & Future Standards

	USA	W. Eu	Japan	Global
Current Most Popular Standard	MediaFLO	DVB-H & T-DMB	"one-seg" ISDB-T	T/S-DMB Korea
Options	MPH, DVB-H, MBMS	TDtv, MBMS, DVB-SH	MediaFLO, MBMS, DVB-H	MediaFLO, MBMS, MPH, DVB-H, CMMB (China)
Expected "Winning" Standard	MediaFLO dominates until unicast over 4G	DVB-H will dominate with TDtv, MBMS & DVB-SH emerging.	ISDB-T dominates until unicast over 4G	MediaFLO may emerge in Japan & Hong Kong with large CMBB volumes expected in China

With a view to assessing the availability of spectrum for mobile TV we survey the typical frequency bands available in the summary below and Table 6.

VHF Band: In some European and Asian countries (Korea) narrow slices of the 200 MHz VHF band has become available for terrestrial broadcasters to provide Mobile TV services. T-DMB technology was used in these allocations.

UHF (470 to 870 MHz): In relation to UHF spectrum, we believe that the long wait plus the uncertainty on how much spectrum will be made available for Mobile TV and who will get the spectrum complicates the technology selection for operators[13]. Unfortunately, in many countries this spectrum will not get released until the digital TV transition (Digital Dividend) in the 2012/2013 timeframe.

L-Band (1.452 – 1.492 GHz): Alternatively, the L Band is slated to be made available in some countries (U.K.) in the near future and could offer an alternative for broadcast mobile TV services[18].

UMTS-Bands (1.7 to 2.5 GHz): Because of the broadcast spectrum issue and lack of alternative frequency options it is highly likely that

multicast options like TDtv and MBMS will get deployed in Europe in existing UMTS 3G spectrum bands to begin to relieve unicast capacity problems.

S-Band (2.17 – 2.20 GHz): One interesting alternative is the S-Band satellite spectrum planned for allocation across the entire European continent in 2008. This spectrum will be available earlier and offer a uniform frequency and technology across a large region. The DVB-SH standard is being positioned to serve this frequency range. In the USA, ICO Satellite is looking to promote a similar spectrum and technology allocation[17].

Table 6: Possible Spectrum for Mobile TV

Band	Name	Status
2500 – 2690 MHz	3G Extension Bands	Technology neutral, usable for 3G, DVB-SH, WiMAX, etc.
2170 – 2200MHz	S-band (usable with DVB-SH)	EC decision & Selection process,
1900 – 2170 MHz	UMTS TDD	Usable with MBMS. Possible interference with 3G FDD
1785 – 1805 MHz	UMTS FDD (3G streaming)	Used for mobile TV already today in unicast mode.
1452 – 1492 MHz	L-Band	Possible T-DMB, MediaLFO, DVB-SH.
470 – 860 MHz	UHF (usable DVB-H, others)	Subject to broadcast license laws, used by DTT, analog.

Considering the alternative spectrum options for mobile TV, it is clear that in-band unicast over cellular has a time-to-market advantage.

Other frequency bands are currently either subject to ongoing regulatory approval, competing with alternative technologies or services, or at risk of interference from neighboring services.

The spectrum availability issue may cause technology fragmentation in the near term. Some standards bodies are eager to prevent this outcome by promoting a single specification as the official approved standard for mobile TV. Other regulatory bodies seem to be taking a more technology neutral stance.

What other regulatory factors have an impact on the business case for mobile TV?

Power levels that are permitted in each market have a substantial impact on the number of sites required in a given frequency. For example, MediaFLO in the USA transmit at 50 kW[14], DVB-H in Europe transmit at 5 kW[15]. This can have an impact on the number of sites required and hence the economic viability of a mobile TV network.

For example, in the USA a typical cell radius of the MediaFLO network operating at 50kW transmit power from 150 to 300 meter towers is 19 Km to 27 Km while providing equivalent indoor coverage over similar terrain[16,17].

By comparison, DVB-H technology in Europe has 5 kW power limits imposed due to EMF and interference regulations severely restricts cell site radius. In a Belguim trial an average 3 Km cell radius was typical in suburban locations from 60 meter towers[18].

Additionally, the available heights of transmitter sites will increase or decrease the total broadcast mobile TV site counts and ultimately mobile TV economics[19].

3. Evolution of network technology

Another driver of mobile TV is the evolution of network technology.

Wireless network technologies continue to evolve, with increasing capacity of wireless bandwidth, support by new technologies such as OFDMA modulation and MIMO (Multiple Input Multiple Output) antenna technology, improved compression algorithms, and greater cell densities of mobile operators

Additionally, in-home devices such as femto-cells and Wi-Fi™[20] allow the wireless operator to off-load capacity from its wide area radio network, which will help reduce the need for an overlay network to support mobile TV.

4. Usage Context and Prospects

We consider the usage context for mobile TV; characterizing the demographics and viewership, willingness to subscribe to a pay mobile TV service, and elasticity to the prospects for mobile TV.

What mobile TV viewership and demographics can an operator expect?

A review of literature reveals that perceived mobile TV viewership differs significantly by region. For example, of the markets that have launched mobile TV, France is reported to have the lowest usage with 70 minutes per week, whereas Korea is reported to have the highest consumption with 160 minutes per week, or about 20 minutes per day[21].

Research is showing that consumers tend to watch mobile video in the home more than previously thought … despite the presence of big screen TV's [22]. In addition, content executives have been surprised in the performance of long-form content on mobile devices [23].

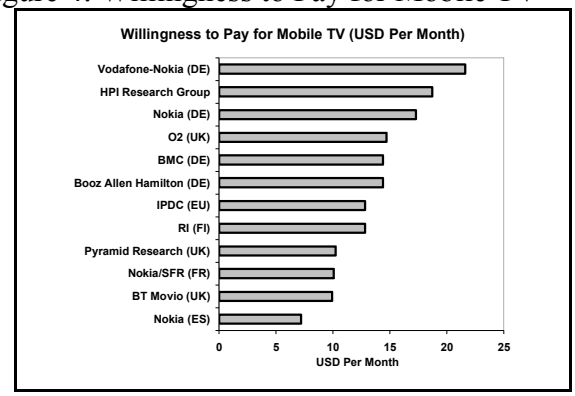
It is a common hypothesis that consumers will use mobile TV to "kill time" leading to the consumption model that mobile TV viewership is a "snacking" phenomena. Data from early research studies[24] indicate otherwise. Namely, that a high proportion of mobile TV viewing (30% to 50% in 3 out of the 4 surveys) is in the home.

Additional research and improved viewership statistics and a better understanding of mobile demographics would assist in refining the technology choices and business model for mobile TV.

What will the mobile TV subscriber be willing to pay for a subscription service?

Recent studies [25] indicate that while a majority of the people were interested in viewing mobile TV 80% of the respondents said they would not pay $15/month for it. The study also concluded that subscribers are more willing to watch mobile TV that is essentially the kind of programming they get on their TV now. Certainly this second conclusion is a positive indication for cable companies regarding the importance of mobile TV to their future business.

Figure 4: Willingness to Pay for Mobile TV [26]



Other studies in European countries assessing the propensity for consumers to pay for mobile TV content indicate a range of US$10 to US$20.

For example, in Italy 3 Italia has 800,000 DVB-H mobile TV subscribers (out of 8 million mobile subs). 3 claims, their mobile TV offering has been instrumental in raising their ARPU 60% over the last year where one-third of the increase has been driven by mobile TV and the remaining two-thirds by voice & data services. 3 Italia may have discovered one willingness to pay pricing model as they offer a popular all inclusive package (voice, data & mobile TV) for US$42 per month [27].

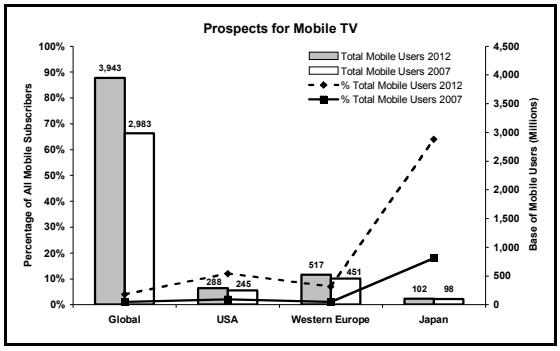
Japan and Korea also offer an interesting benchmark for a mobile TV subscriber's willingness to pay. The ISDB-T (One-Seg) amd T-DMB services, make up a majority of the current mobile TV subscribers worldwide yet are free service offerings, delivering free-to-air content.

We have provided early evidence of operator-based research into end user willingness to pay for premium mobile TV. However, it is not clear at this stage whether a free simulcast model, a pay TV model, or ad supported model will dominate Europe, the United States and Asia. The type of model that emerges could be expected to have a significant impact on mobile TV's prospects.

What does this mean in terms of the prospects for mobile TV?

Based on industry information, the projected take up of mobile TV is estimated at 150 million users or 4% of all mobile users globally by 2012 from a base of almost 4 billion mobile users, with developed economies expected to experience higher penetration rates. For example, by 2012, mobile TV is expected to have 35 million users or 7% mobile user penetration in Western Europe, 35 million users or 12% penetration in the United States and 65 million users or 64% mobile user penetration in Japan by 2012.

Figure 5: Prospects for Mobile TV [28]



We chose a conservative study summarized in Figure 5 to highlight mobile TV's prospects, focused on linear TV programming, rather than all types of video content[29].

Observing the range of industry forecasts, i.e. up to 465 million by 2010 [30] we note that (a) cable operators need to be aware of differing research definitions as to what constitutes mobile TV, (b) methodology of today's forecasts and, (c) as mobile TV is at an early stage of development, variation in forecasts can be expected.

Overall, we note that where a pay TV subscription model is the focus (i.e. USA, Europe), the penetration of mobile TV services is lower than in markets where the service is bundled as a free offering (i.e. Japan). We believe that mobile TV is therefore very price elastic, and significant penetration will most likely come from bundling and cross subsidization with other core mobile or entertainment services.

COMPARING THE MOBILE TV ALTERNATIVES

Which mobile TV technology should a cable MSO consider and what platforms pose the largest threat or present the greatest opportunity?

For the purposes of this paper we are focusing on facilities-based mobile TV technologies and setting to one side alternative non-facilities based alternatives (i.e. in-home radio technologies such as Wi-Fi or storage-based PC to user terminal file transfers).

There are several competing facilities-based platforms for mobile TV. We consider an overall taxonomy based on classifying the technical alternatives into (1) Broadcast, (2) Multicast and (3) Unicast;

Table 7: Mobile TV Delivery Alternatives [31]

	Broadcast	Multicast	Unicast
Network	Broadcast	Cellular	Cellular
Topology	One-many	Mixed	One-one
Return path	No	Yes	Yes
Bandwidth	Dedicated	Mixed	Shared
Throughput	Fixed	Mixed	Variable
Zap speed	1-3 secs	2-5 secs	5 – 8 secs [32]
Technology Example	DVB-H DVB-SH MediaFLO	TDtv MBMS	WiMAX LTE HSPA, HSPA+ 3G (UMTS/ WCDMA)
Advantages	Cost structure, performance	Re-use of existing spectrum	Variety, on-demand
Disadvantages	Variety, additional network	Price, performance	Price, performance

We take a closer look at the technology alternatives to determine what the advantages and disadvantages are, and what this means for the cable MSO.

1. Broadcast

Looking at a typical broadcast architecture for facilities-based mobile TV we note that there are quite a number of similarities to the traditional cable MSO broadcast architecture; including the need for encoders, and electronic program guide, and conditional access systems.

Exploring the broadcast alternatives in more detail we consider MediaFLO, DVB-H, DVB-SH and T/S-DMB.

(a) MediaFLO

The MediaFLO specification was developed by Qualcomm specifically for broadcast mobile TV applications. Consequently, it was optimized for high bandwidth (many simultaneous video channels), high speed mobility, single frequency networks, low power drain CPE devices, large cell radius and fast channel changing capability.[33]

OFDM (Orthogonal Frequency Division Modulation) was chosen as the most effective way to meet these design goals. Fortunately, it was able to leverage other standards such as Wi-Fi™[34], ADSL, DTV, UWB, WiMAX™ [35], LTE and DVB-H that all employ OFDM technology.

Just as TDMA separates communication channels and end user conversations with time division and CDMA segments channels with codes (orthogonal spreading codes), OFDM utilizes frequency. It differs from 1st generation analog cellular frequency division techniques by using very tightly spaced frequencies without overlapping and interfering. It does this by forcing the narrowband FDM carriers (called subchannels or tones) to appear unique or independent from each other. The mathematic concept of orthogonality is the key to maintaining separate communication channels even though the subchannels are very narrow and spaced close to each other.

Qualcomm effectively incorporated time slicing into their specification so that mobile devices used as little power as possible. This technique transmits chunks of data in bursts so that the receiver could be turned on and off during inactive time periods. The result is substantial power savings (90%) over traditional broadcast technologies using fixed high power receivers. Because MediaFLO was designed without the need for compatibility with legacy standards by a company with relevant experience in mobile devices it utilized some very effective techniques. MediaFLO uses a more frequent transmit time interval than DVB-H, which helps in having quicker channel change speeds and improved power saving[36].

The most unique design aspect of the MediaFLO implementation is the ability to have layered modulation. Basically, the data stream bursts are divided into base and enhanced layers. The base layer supports the widest coverage area using

lower quality (15 fps) video that subscribers and receivers in poor signal areas (such as in-building) can decode. The enhancement layer supports high quality 30 fps video and is decoded by the receiver in high SNR (signal to noise) areas. The MediaFLO handset dynamically adapts the video quality based on the signal strength. Consequently, there is a smoother degradation of service as the signal strength varies[37].

An important design tradeoff for a broadcast mobile TV architecture is determining the optimum number of OFDM subcarriers or tones. On the one hand, a large number of subcarriers (8000 in a 5 MHz channel bandwidth) will provide for higher capacity and a larger single frequency network (avoids handover to different frequencies) but will negatively impact high speed mobile performance. Qualcomm and the DVB-H specification both settled on 4,096 subcarriers as the optimum compromise for mobility, capacity and large single frequency networks.

A final unique aspect of the MediaFLO specification is the use of a variable bit rate and statistical multiplexing allocation for the video services. This feature provides a bandwidth efficiency gain of about 30% translating into a higher number of video channels at comparable quality in a channel.

(b) DVB-H

The Digital Video Broadcast standard for handheld devices is based on the existing DVB-T standard for fixed digital TV reception.

Most changes were made to the layer 2 portion of the specification and focused on making improvements so that video transmission would be robust enough for a severe multipath mobile environment and low power mobile devices. Consequently, time slicing and forward error correction elements were added to the specification. Physical layer changes included the use of 4,096 OFDM subcarriers (DVB-T allowed for just 2K or 8K options), better flexibility in using all modulation formats (QPSK, 16 QAM, 64QAM), creating a 5 MHz channel bandwidth and expanded bit interleaving options[38].

For the most part these layer 1 and 2 specification changes put MediaFLO and DVB-H at a similar capability. Overall, MediaFLO has more beneficial performance, coverage and capacity technical characteristics. Conversely, the DVB-H standard is much better positioned as a uniform worldwide standard because of its strong backing in Europe.

(c) DVB-SH

A European wide allocation of satellite spectrum and the vision of a uniform continent wide roaming capability has prompted the creation of the DVB-SH standard. This architecture will provide direct outdoor coverage to handhelds and vehicles (with outdoor antennas) from a satellite. Indoor coverage will require a large number of repeater sites located at existing cellular sites. Utilizing an existing wireless carrier's dense cell site network will be critical for this service to be effective.

DVB-SH provides two key improvements to DVB-H: (1) 3GPP2 Turbo Codes, that improves the quality of reception in tough conditions and (2) Physical layer time interleaving that improves the quality of reception while in motion.

The net result is that, under the same conditions (frequency, channel size, data-rate) signal reception requirements (carrier to noise) are a minimum of 5 to 6 dB lower[39] and up to 6 to 8 dB lower[40] for DVB-SH relative to DVB-H.

Additionally DVB-SH is able to leverage cellular sites to down-convert S-Band Satellite mobile TV content transmissions to maximize coverage and minimize distribution/backhaul costs. All other terrestrial based broadcast technologies require backhaul transport of the mobile TV content to every site. DVB-SH requires a more economical satellite dish and regenerator equipment. On the down side, DVB-SH requires a very small cell radius to get sufficient transmit power at 2.2 GHz to penetrate buildings.

An overall comparison of the three standalone broadcast alternatives indicates performance advantages for DVB-SH.

(d) T/S-DMB

DMB technology was first developed in South Korea and was designed to operate as either a satellite (S-DMB) or terrestrial (T-DMB) mobile TV transmission system.

In some countries, DTT and DAB broadcasters were allowed to utilize narrow bandwidths (1.5 MHz) of their spectrum for mobile TV. To accommodate these opportunities in the VHF spectrum (200 MHz), the T-DMB broadcast mobile TV standard was created by making modifications to the terrestrial broadcasters DAB specification.

Besides the much smaller channel bandwidth T-DMB does not allow for higher modulation formats (16-QAM or 64 QAM) and has less robust coding schemes (lacks either MPE-FEC or turbo coding). Additionally, T-DMB lacks the device power saving advantage of a full time slicing architecture of other broadcast technologies (DVB-H & MediaFLO)[41].

The S-DMB system concept is based on a combination of satellite and terrestrial architecture for the delivery of broadcasting digital multimedia services to mobile end users. Because the satellite coverage provides outdoor only mobile TV service S-DMB is extended indoors with terrestrial repeaters.

Essentially, T-DMB and S-DMB are very similar specifications. The biggest difference is the RF planning and implementation of the network associated with S-DMB as the interference between satellite and terrestrial transmission makes it complicated to design the broadcasting network.

Overall, the success of both T and S DMB technology has been very limited because of capacity constraints (associated with narrow bandwidth allocations), limited CPE and performance/quality issues. The major take-up has occurred where the service offers free to air content to mobile devices (T-DMB).

2. Multicast

Multicast distribution of mobile TV services provides point to multipoint transmission of video and TV media from a single source to a group of users in a specific area. The key distinction between broadcast and multicast is that the end users must have joined the particular multicast group while in broadcast technology all users obtain the content. A classic illustration of multicast is the delivery of radio station content over the internet.

Figure 6: Multicast Network Architecture[42]

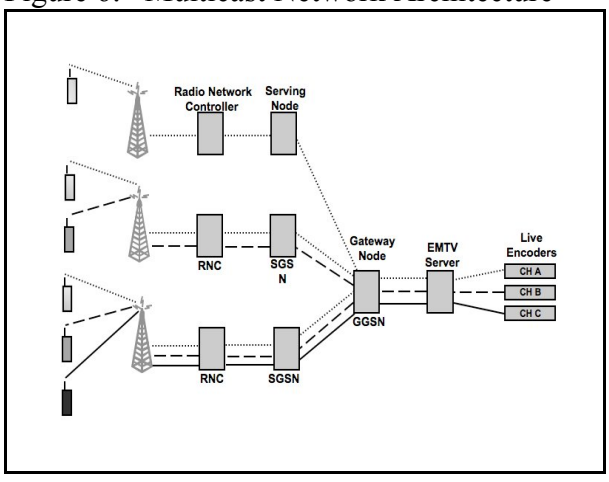


Figure 6 is a visual representation of a typical cellular multicast architecture such as TDtv or MBMS. Only mobiles in a cell site interested in viewing a particular channel (dotted line) join the multicast group in that cell. Other cells obtain additional channels (dashed & solid lines) because of the desire by the mobiles in that cell to view different content at the same time.

A multicast enabled network ensures that content is solely distributed over those links that are serving receivers which belong to the corresponding multicast group. This is a very resource efficient way of delivering services to larger user groups[21].

(a) TDtv

Many mobile operators in Europe were awarded 5 or 10 MHz of TDD unidirectional spectrum (1.9 to 2.0 GHz) as part of the 3G licenses won through auctions and "beauty contests". To date, very few operators have used this spectrum because it is unidirectional and little equipment is available.

TDtv is a multicast technology that uses the existing TDD spectrum in a 3G license. Since there are no TDD capable transmitters on base stations, these have to be added, but they can go on the same towers and use the same power supplies and antenna of the existing base station. Using two antenna in the handset means that only 30% to 50% [43] of base stations require transmitters. Two signals from any base stations in reach can combine through the antenna to give in-building penetration. At present operators have proven 15 channels in 5 MHz of TDD, but claim they can stretch to 28.

Potential interference issues exist as the spectrum sits next to existing 3G spectrum, meaning that not all of the 10 MHz may be available for use and expensive filters may be required in related handsets.

Dynamic channel broadcasting is not envisaged yet but could increase the effectively available number of channels to 90 in the future with the limitation then based on contribution capacity (i.e. E3 at 34,368 Kbps at 75% utilization or 25,776Kbps divided into 256 Kbps of video and 32Kbps of Audio or 286 Kbps).

(b) MBMS

The 3GPP Release 6 specification created the Multimedia Broadcast Multicast Service (MBMS) standard. Only minor changes were made to the existing radio and core network protocols. A new physical bearer channel that carries the media content was created along with logical scheduling and control channels.

The key for MBMS is that it can use all or a portion of an existing 5 MHz HSPA radio channel. TDtv is a multicast configuration that requires dedicated TDD spectrum, cell site equipment and chipset enhancements to the mobiles. MBMS requires none of that additional equipment. But portions of the existing HSPA network must be dedicated to MBMS services. For instance, if 256 Kbps mobile TV channels are planned for then 32 TV channels can be created in a single 5 MHz radio channel. If desired, only a portion of the 5 MHz is allocated to MBMS while the remaining amount is used for voice and data services. Additionally, 128 Kbps or 64 Kbps mobile TV channels can be set.

Overall, MBMS has a capacity advantage over unicast when several subscribers reside in the same sector of a cell and are watching the same mobile TV channel. When there are very few users in a sector then a unicast architecture may make more sense[21].

3. Unicast

Unicast mobile TV technologies stream video to mobile devices over various 3G wireless technologies. Streaming video to handsets in

unicast has some inherent limitations that present challenges relating to performance. `

The most limiting problem for unicast mobile TV has been the overall capacity constraints and end user speeds possible over existing cellular network technologies. Many network upgrades and advancements have been made in recent years that begin to break the capacity limits of unicast mobile TV over 3G cellular technology.

Additionally, channel zap performance can be challenging in relation to unicast mobile TV. In particular, an inherent 15 to 20 seconds delay to move from one specific channel to another because the current session must be closed and a new one must be opened.

Advances have been made in this area, that enable the player to remain "alive" when switching from one channel to another; and keeping the video displayed when switching; and finally, optimization for network conditions, that allows for zap speeds between 3 to 8 seconds[44].

(a) 3G (UMTS/WCDMA, HSPA and HSPA+)

For the purposes of this paper we will focus on the use of mobile TV over the group of mobile standards to come out of the 3GPP (3rd Generation Partnership Project) standards body.

Table 8: 3GPP Specification Releases[45]

Version	Released	Description
Release 99	2000	Original UMTS/WCDMA 3G air interface
Release 4	2001	Added new features including all IP core
Release 5	2002	Added HSDPA (improved Downlink) and IMS
Release 6	2004	Added HSUPA, (improved Uplink) and MBMS … release is called HSPA
Release 7	2007	Added downlink MIMO, improved QOS and VoIP … Release is called HSPA Evolved or HSPA+

The ongoing evolution of the 3G UMTS family of technologies, which builds on the foundations of GSM, are listed in Table 8.

The original promise of UMTS/WCDMA, to provide high speed broadband connectivity, never occurred. Although speeds of 2 Mbps were hoped for, in reality 256 to 384 Kbps were more typical[9]. Recently downlink and uplink software enhancements (HSDPA and HSUPA) have been adopted by operators worldwide that provides much improved performance.

A number of technologies have been deployed to make these improvements including adaptive modulation and coding, fast packet scheduling and Hybrid Automatic Request (HARQ). Adaptive modulation software analyze each end user for signal strength and determines which modulation format (16QAM, QPSK…) and coding scheme will work the best. Fast packet scheduling allows communication between the mobile device and cell site to make the most efficient use of the bandwidth available. In CDMA systems the use of orthogonal CDMA spreading codes and time slots is critical in allowing a device to attain its maximum data rate. In the case of HSPA, devices using 5 codes allows for a maximum theoretical peak speeds of 3.6 Mbps, 10 codes correlates to 7.2 Mbps and 15 codes can theoretically attain 14.1 Mbps.

The combination of HSDPA and HSUPA (called HSPA) is reflected in the 3GPP Release 6 specification. Only 7.2 Mbps capabilities are currently available and most operators claim from 5.0 to 6.0 Mbps speeds are attainable in "real world" operation. More appropriately, the average cell throughput capacity of a sector is the critical design metric as this is the capacity to be shared among all users simultaneously accessing the network. For HSPA the average sector throughput will range from 4 to 6 Mbps.

As is expected, much depends on the cell radius design, indoor or cell edge coverage and signal

strength assumptions. If it is assumed that a large % of users will get great signal strength (for example; 200 meters from site with line of site, outdoor coverage, 16QAM modulation) then the sector capacities will be higher[46].

As an illustration of the variability, a single end user in the sector (no contention with other subscribers) with outdoor coverage could get between 2 and 4 Mbps service depending how far away and if there is line of site to the cell location. A single user indoors will typically get between 800 Kbps and 2 Mbps, again depending on distance from the site, type of building material and how far inside the building the user is located. For this reason most HSPA networks (7.2 Mbps & 10 codes type devices) are designed and offer an average 1 to 3 Mbps product to subscribers. The peak rates advertised are typically marketing buzz as the peak speed is only obtainable if there is a single end user in service on the sector and is operating at the strongest possible signal strength.

Mobile broadband networks with sufficient capacity to provide for some unicast mobile TV applications are becoming more prevalent throughout the world.

For instance, there are 165 HSPA networks in place and a device ecosystem of 465 different devices from 102 suppliers currently available around the world[10]. According to industry research, UMTS/WCDMA/HPSA is the world's most popular 3G cellular technology as it represents more than 200 million customers worldwide. Almost 20 million are already subscribers to high-speed HSPA mobile broadband networks and this number is expected to double by the end of 2008[11]. Nevertheless, penetration of high bandwidth cellular connectivity is still emerging and needs to develop further. For example, only 13% of homes in the U.S. are 3G capable today. [47]

With GSM exceeding 2.6 billion mobile connections worldwide and global subscriptions to all mobile network technologies exceeding 3.3 billion, the 3GPP family of standards (GSM/UMTS/WCDMA/HSPA) represents over 80% of all cellular connections worldwide.

HSPA Evolved or HSPA+ will enhance the downlink and claims to provide a theoretical peak of 42 Mbps by utilizing 64QAM modulation and the uplink to 11.5 Mbps through 16QAM. A further enhancement to help in achieving the increase data rates is the addition of MIMO antennas, usually deployed to enhance the system performance. MIMO increases downlink sector capacity by implementing a technique that transmits multiple desired signals via separate antennas. This has the effect of multiplying the amount of data that is able to be transmitted over a single radio channel.

Other features include reducing latency by keeping the devices in a different state when inactive. Although the lab simulations of MIMO technology are positive, this advancement still needs to be proven out in the challenging RF environment of mobile devices in a live network.

While HSPA+, with its MIMO capability, can provide a capacity improvement over 3G and HSDPA, the upgrade is not cost free to the mobile operator. In order to support MIMO additional capital costs associated with antenna and installation, radio planning; in addition to additional operational expenditure for the site rental associated with the additional antenna. Additionally the upgrade from 3G to higher network bandwidths requires the buildout of larger capacity backhaul.

(b) 4G (LTE and WiMAX)

Long Term Evolution or LTE (3GPP Release 8) and Mobile WiMAX (802.16e) are emerging as 4th generation standards being specified to offer very large average sector throughputs (20 to 40 Mbps) and as such could be impactful to the

application of Unicast Mobile TV applications over broadband cellular technologies. Both technologies utilize OFDMA technology, incorporate MIMO antenna and transmission gains and wide channel bandwidths (10 and 20 MHz) that contribute to large capacity improvements.

Some research forecasts that LTE should be going commercial by 2010 and represent around 24 million subscribers globally by 2012[15].

The research also predicts that HSPA will dominate mobile broadband network deployments by 2012, consistently accounting for about 70% of the total mobile broadband subscriber base. LTE and Mobile WiMAX are expected to achieve only a small proportion of the 1.2 billion total mobile broadband subscriber base[48].

4. Technology comparison across coverage, capacity, and mobility variables

As is typical with wireless technologies, understanding the tradeoffs among coverage, capacity and mobility characteristics is critical when assessing mobile TV technologies. The key to comparing these characteristics is having knowledge of the key assumptions that drive coverage, capacity and mobility performance results.

For instance, the coverage claims of broadcast technologies such as MediaFLO and DVB-H vary widely as a number of assumptions are changed. In our Flanders DVB-H field trial many scenarios were tested. The results of one scenario are illustrated in Table 9 below.

Table 9:DVBH Coverage in Flanders Trial[54]

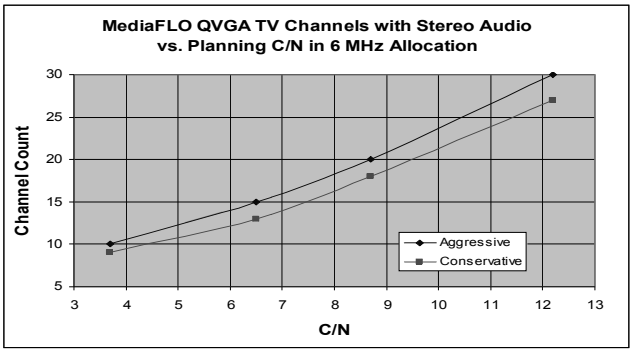
Suburban terrain, 5kW transmit power & 60m antenna heights	Location, CPE Type & Mobility/Portability		
	Outdoor, Handheld, 3Km/hr	Indoor, Handheld, 3 Km/hr	In-car, Handheld, 70 Km/hr
Cell Radius (Km)	9.25	3.35	1.45

The measurement of the cell radius calculations assumed a common modulation scheme being received (16QAM ½) by the device, a fixed maximum quantity of mobile TV channels and an overall probability (90%) of obtaining a minimum signal level. The conclusion from the trial is that coverage can vary widely (from 1.45 Km to 9.25 Km) as different types of mobile devices, their location and speeds are changed. To illustrate further the complexity of the wireless tradeoff's typically encountered we changed the transmit power and antenna height for the Flanders test. Reducing the transmit power to 2 kW and the antenna height to 30 meters for an indoor handheld moving at 3 Km/hr caused a severe reduction in cell radius from 3.35 Km to 1.65 Km[54].

How is the capacity of a mobile TV technology determined?

The ability of the mobile TV device to receive a strong enough signal in the presence of interference is a critical determinant of capacity. Qualcomm has conducted numerous tests illustrating the relationship of sufficient signal strength received by the mobile TV device and overall system capacities for their MediaFLO technology. Figure 7 below shows the relationship between the maximum numbers of 256 Kb/s H.264 mobile TV channels possible for a given received signal at the mobile handheld. The signal strength is represented as a Carrier to Noise ratio (C/N) in decibels.

Figure 7: Capacity vs. Signal Strength[49]



Extensive testing has revealed that a 10 dB C/N ratio is an attainable signal strength under a number of typical mobile TV conditions. As can be seen from the graph a 10 dB C/N results in a capacity of 22 Mobile TV channels. Consequently, the MediaFLO network is designed for a capacity of 22 mobile channels operating at 256 Kb/s per channel. To meet this capacity figure the key design criteria used to determine transmitter locations for the mobile broadcast network will be to meet the 10 dB signal strength level in a majority (90%) of the locations the network serves.

A common question asked of mobile TV services and technologies is what do we really mean by mobility?

Mobile TV terminals are expected to be usable in stationary, pedestrian walking speeds, and high speed in-vehicle applications that encompass both outdoor and indoor environments. Delivering a consistent high speed video signal to a low power, low gain handheld device in motion is a difficult technical challenge[58].

One of the key issues with fast moving mobile devices is the concept of multipath propagation. In effect, the radio signal from the cell site transmitter takes multiple paths to reach the mobile device which results in "echoes" that make it difficult for the mobile to recover the video transmission. These echoes cause the digital video information (or symbols) to "blur" across each other creating severe problems called inter-symbol interference.

Unfortunately, handheld devices in high speed motion amplify the inter-symbol interference reception issue. This effect is called the Doppler shift. Simply put, the faster a mobile is moving the more of a Doppler shift occurs which translates into a greater interference effect on the mobile device receiver. The Doppler shift interference effect is more noticeable at mobile systems operating at higher frequencies (e.g.- > 400 MHz). Mobile TV standards such as MediaFLO and DVB-H attempt to minimize this problem by using various advanced error correction coding, modulation formats and OFDM sub carrier schemes.

A considerable amount of high speed mobility testing has been conducted at 850 MHz for DVB-H mobile TV networks and is illustrated in one design scenario in Figure 8. Figure 8 illustrates the relationship of mobile speeds along the x-axis (represented as Doppler shift frequencies in Hz) and mobile receive signal strength (C/N ratios) along the y-axis for the DVB-H mobile TV specification.

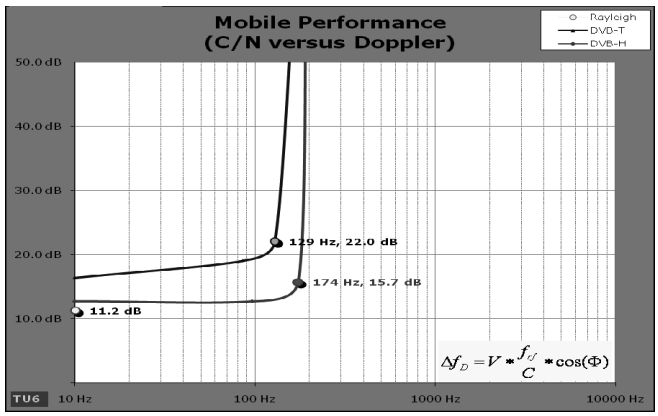


Figure 8: Mobile speeds vs. signal strength[57]

The graph shows that as long as a minimum signal strength can be maintained (11.2 dB) then mobile devices speeds from 3 Km/hr (at 10 Hz on the y-axis) to 126 Km/hr (corresponds to the 100 Hz point on the y-axis) will support mobile TV services. It is interesting to note that at the Doppler frequency of 174 Hz (equivalent to ~200 Km/hr) a sharp increase in C/N signal strength is required. This means that at speeds greater than 200 Km/hr a nearly impossible signal strength would be required to be received by the fast moving mobile.

The major mobile broadcast technologies (DVB-H, MediaFLO and DVB-SH) are designed to operate in a high speed mobile environments (< 200 Km/hr). Likewise, multicast and unicast

technologies, being required to serve traditional mobile voice and data applications, are capable of operating at high speed vehicle and train speeds as well.

THE MOBILE TV BUSINESS MODEL: A COMPLIMENT TO FIXED LINE BROADCAST

What is the cost of extending fixed-line broadcast to mobile broadcast? Are the costs different for a mobile operator? We explore the economics across multiple mobile TV platforms.

Developing the network "Pain Threshold"

Using revenue and operating assumptions from earlier sections of this paper and reference literature we can determine the network "pain threshold" for Mobile TV network economics.

We assume that the average revenue from mobile TV is $20 per month per subscriber[50]. Of this $20 per month we assume that 45%[51], or $9 per month, goes to mobile TV content costs and $8 is required for sales and marketing, billing and G&A per subscriber per month, about half of the amount for mobile data [52]. This leaves $3 per month per subscriber to cover all network related costs.

1. Unicast economics

Initially considering the economics of unicast mobile TV we find that the economics quickly pass the pain threshold of $3! Unicast economics appear suited to low quality, short-clip, long tail content that have low bit-rates and short view times rather than high quality premium content with longer average view times.

Table 10:Unicast Mobile TV Cost/Sub/Month [53]

	HSPA+	HSPA	UMTS/ WCDMA
256 Kbps @ 20m	$7.66	$24.42	$30.21
128 Kbps @ 6m	$1.28	$4.07	$5.04
128 Kbps @ 2m	$0.38	$1.22	$1.55

We assumed quality and viewership parity when comparing unicast mobile TV to multicast or broadcast mobile TV, with all platforms delivering 256Kbps encoded video and 32Kbps encoded audio with average viewing times of 20 minutes to match the viewership studies referenced earlier.

Under differing quality and viewership conditions unicast mobile TV results in varying degrees of network congestion depending on the network evolution technology deployed. We assumed an average sector throughput of 2.5 Mbps for UMTS/WDMA, 6.0 Mbps for HSPA (7.2) and 9.5 Mbps for HSPA+ technologies. Additionally, an urban market density of 1,500 Pop's per Km^2, 0.57 Km cell radius and 3 sector cell sites assumptions were used.

Clearly an early 3G network using UMTS/WCDMA would be unable to support mobile TV services at 256Kbps and 20 minute average view times! Although the evolution of 3G to HSPA, HSPA+ and ultimately LTE reduce the probability of congestion, we need to take into consideration that other services are also operating on the same network. Consumption of 3G and HSPA networks for high speed data is increasing. In the United Kingdom, Vodafone is reporting 50% to 60% of its available 3G capacity is being used for data in dense urban areas[54]. Therefore in this analysis we assume 50% of the sector capacity is allocated to data services.

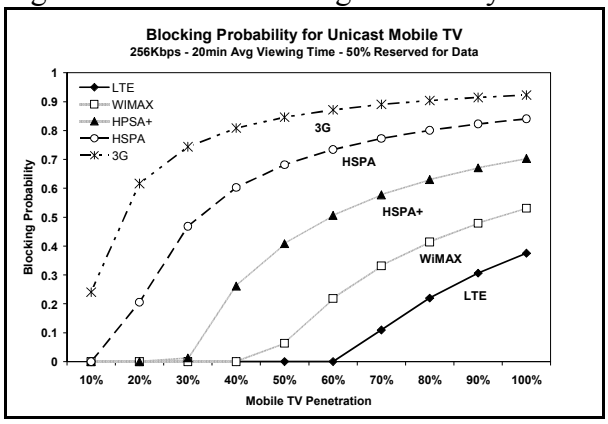
Table 11: Unicast Mobile TV Blocking Probability @ 15% Penetration

	HSPA +	HSPA	UMTS/ WCDMA
384 Kbps @ 20m	0.00	0.26	0.65
256 Kbps @ 20m	0.00	0.00	0.49
128 Kbps @ 6m	0.00	0.00	0.00
128 Kbps @ 2m	0.00	0.00	0.00

We can see from this analysis that low bit-rate, short clips have less impact on unicast mobile network congestion that high bit –rate, long form content, indicating that the mobile network can support short clips such as you-tube like rich-data content.

Because the traffic is unicast as the penetration increases to 30% the blocking probability increases to the point where our 256k, 20 minute scenario has blocking probability of 0.47 for HSPA and 0.74 for 3G … put another way 74% of 3G mobile TV subs could not get the service, in addition to there being no room for growth in data services!

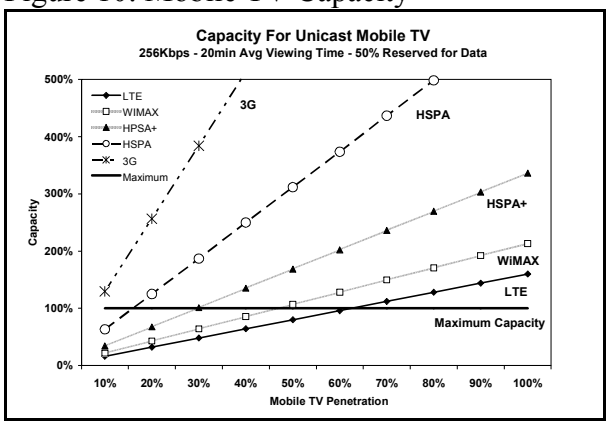
Figure 9: Unicast Blocking Probability



As seen in figure 9, as wireless technologies evolve, reasonable penetration levels of mobile TV could be expected if 15 Mb/s (WiMAX) and 20 Mb/s (LTE) sector capacities are obtainable. Our analysis indicates future WiMAX and LTE technology could provide non-blocking mobile TV service at 40-60% penetration.

Our analysis of blocking probabilities also provide insights in relation to capacity advances for mobile TV. We can see in Figure 10 below that new wireless technologies can enable Telcos and independent wireless providers to deliver mobile TV at greater penetration rates.

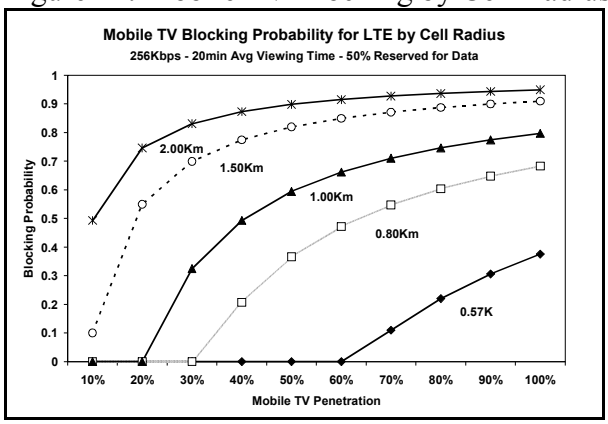
Figure 10: Mobile TV Capacity



It is also important to point out that if a mobile operator has sufficient spectrum to add another 10 MHz channel to their network and reasons that there is sufficient return to allocate it to video services then these capacities can be increased further and therefore provide for improved mobile TV capacities.

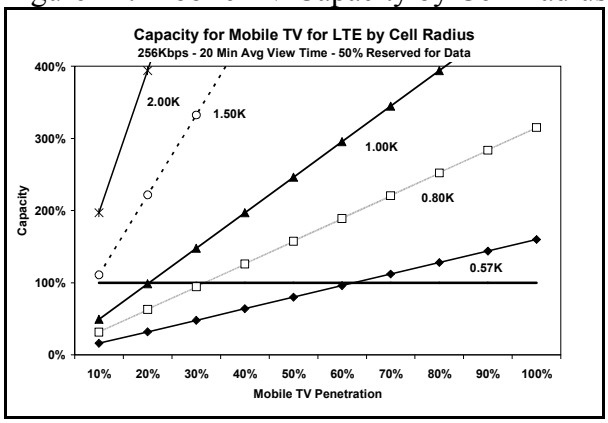
The importance of a dense cellular network for delivering unicast mobile TV is illustrated in figure 11 below. Using LTE as a base we illustrated the impact of larger cell sizes, noting that an increase in cell radius from 0.57 to 0.80 kilometers results in a halving of non-blocking mobile TV penetration potential!

Figure 11: Mobile TV Blocking by Cell Radius



We can also show this in capacity potential, where a cell size of 0.57Km exhausts capacity for mobile TV at 60%, a cell size of 0.80Km exhausts capacity for mobile TV at just 30% penetration (assuming a constant market density).

Figure 12: Mobile TV Capacity by Cell Radius



2. In-band multicast economics

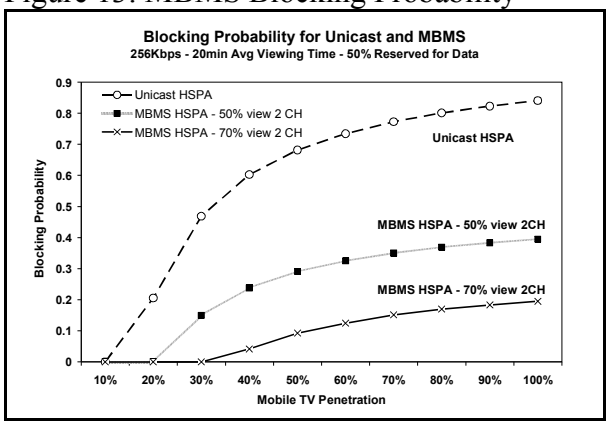
Considering In-band MBMS, or Multicast assumptions to the cellular models without extending available spectrum, we assumed that 50% of the users viewing time was for 2 channels, and accounted for up to 80% of the viewing time for those channels.

We also assume that for in-band multicast over 3G/HSPA/LTE networks that a mobile broadcast bearer requires more radio capacity. We assume that about 13% of the Node-B (Base Station) power is required for Multicast, assuming that soft combining is not enabled (possible reduction to 6%)[55].

Table 12: Multicast Mobile TV Blocking Probability (256Kbps, 20 minutes view time, 30% penetration)

	HSPA+	HSPA	UMTS/ WCDMA
Unicast	0.01	0.47	0.74
50% view 2 Channels	0.0	0.0	0.31
60% view 2 Channels	0.0	0.0	0.21
70% view 2 Channels	0.0	0.0	0.11

Figure 13: MBMS Blocking Probability



As seen in Figure 13 above, unicast HSPA mobile TV begins to suffer from blocking when penetration exceeds 15%, where as adding MBMS can enable support higher penetration rates.

However, because enabling MBMS increases power usage, and the reduction in blocking probability is mitigated by the existence of other services (i.e. mobile data, with its speed and subscriber growth) it appears to be inefficient to deploy in-band with other IP services until LTE becomes available.

Based on our analysis of network congestion, we assume that mobile operators would not consider enabling MBMS as an in-band capability for some time, but would rather focus on deploying MBMS capabilities in separate dedicated unicast spectrum (i.e. TDtv).

3: Broadcast overlay network economics

We used a specific propagation model based on a trial in Gent, Belgium [56] [57] to determine the site radius for typical European DVB-H deployments, and validated this information with other reference deployments in Europe.

Our MediaFLO analysis was based on a reference architecture for Chile. Results were reduced to a cost per kilometer square basis to determine the comparable cost of coverage to Europe.

4. Hybrid network economics

With emerging hybrid mobile TV architectures we used information based on a reference UK model for DVB-SH and assumed a 6dB to 8dB gain over DVB-H providing approximately a factor x2 improvement in coverage area.

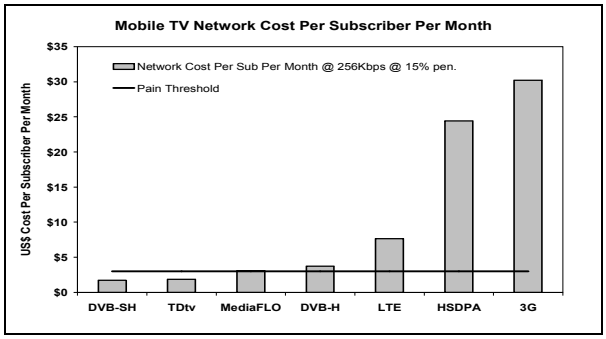
Out TDtv analysis was based on industry information including the assumption that 40% of cell sites require the TDtv transmitter, and used industry literature to determine the capital cost of the network extension.

Comparing the alternatives, what technologies are below out pain threshold of $3 per subscriber per month?

Table 13: Network Cost of Mobile TV

	Network Cost Per Sub Per Month @ 256Kbps @ 15% pen.	Channels Supported @ 256Kbps
MediaFLO	$3.10	22
DVB-H	$3.72	18 [58]
DVB-SH	$1.70	18t & 9ts[59]
TDtv	$1.87	15 – 28, Uni.
HSPA+	$7.66	Unlimited
HSPA	$24.42	Unlimited
3G	$30.21	Unlimited

Figure 14: Mobile TV "Pain Threshold" and Network Costs



CONCLUSION

Based on our analysis we believe that the market for and penetration of mobile TV is relatively elastic to price, with pay markets seeing 10% to 20% penetration and free markets seeing about 60% penetration.

If Mobile Network Operators (MNO) or Multiple Service Operators (MSO) gave away mobile TV as a bundled offering the experience in Japan tells us we would require capacity for 60% penetration and require broadcast overlay solutions such as DVB-H and DVB-SH, or need to wait for high capacity mobile pipes (i.e. LTE) and/or multicast technologies.

On the other hand, cable MSOs or Telcos/MNO considering charging for mobile TV as a pay

offering, could expect penetration between 10% and 20% by 2012. This type of penetration rate is more suited to in-band HSPA, HSPA+, MBMS and ultimately LTE alternatives for mobile TV. Subsequently, today's Telcos with mobile operations, are in a better position to deliver a mobile TV pay service as this alternative does not require a separate overlay network.

We can present the alternatives for the cable MSO as follows:

Table 14: Cable MSOs Mobile TV Choices

	Mobile TV Free Bundle	Mobile TV Pay Service
Penetration	~ 60%	~10 – 20%
Technology Options Eliminated	S/T-DMB **(Limited Capacity)**	UMTS, HSPA **(Limited Capacity)**
Technology Options Today	MediaFLO, DVB-H/SH **(Spectrum)**	HPSA+ TDtv, MBMS **(MNO/Telco)**
Technology Options Emerging	WiMAX, LTE, **(MNO/Telco)**	WiMAX, LTE, **(MNO/Telco)**

In view of Table 14 it is less clear how the many mobile TV distribution choices are suited to a cable operator that does not own suitable spectrum or a wireless network asset, in view of challenging economics and fragmentation of spectrum options.

A summary of the economics of the different facilities-based mobile TV alternatives determined that the economic margin for error is low, flexibility for an sustainable deployment is limited and that the selection of an economically viable distribution technology for the delivery of mobile TV is critical.

Given the inherent low margin aspects of the mobile TV business, the tight linkages to unicast video (and data) consumption over existing cellular data networks and the device centric nature of mobile TV, Mobile Network Operators have a large advantage over MSO's and TV broadcasters.

It is highly unlikely that mobile TV content and viewership will steal subscribers from the in-home TV viewing revenues of cable operators but TV viewing time could be impacted. Any opportunity of competitors to gain a foothold on video and TV viewing time and habits, content aggregation and user interfaces could be deemed a viable mid to long term threat. One area to monitor closely in this area may be the technology advancements such as pico-projectors being built into mobile phones.

Therefore, given the tough barriers to entry in the mobile TV space, combined with advances in mobile technology, it is the opinion of the authors that mobile TV is more of a threat than opportunity for cable operators. In particular, MNOs that serve the low end cable TV base with free to air mobile TV and are able to complement this with premium content delivered using either broadcast or higher capacity in-band wireless could be threatening.

On the other hand, MNO's may need considerable assistance on the technical and economic aspects of content acquisition, management, aggregation, rendering & distribution. An MSO's ability to leverage this strategic advantage into an effective partnership with MNO's is the most viable option for cable operators. This may be particularly attractive to the MNO where the significant cost for mobile TV content can be reduced through partnering with the MSO.

Should MNO partnerships prove challenging, cable operators that strongly desire a mobile TV strategy need to consider the possibility of acquiring or building out a multi-service cellular wireless network to facilitate sustainable mobile TV economics. The build or buy path appears overly aggressive if mobile TV is the only economic driver.

An MSO going it alone in the broadcast area (MediaFLO & DVB-H) appears risky as well. At the end of the day mobile TV using broadcast technology requires devices. Therefore broadcast technologies need a mobile operator's network and mobiles. An MNO operator is a formidable competitor in the mobile TV market, with many distribution options in the toolkit, including established multicast and unicast delivery options that may not be perfect, but allow the MNO to evaluate what content works for mobile TV as the market emerges.

REFERENCES

1. Yoram Soloman, "The Economics of Mobile TV", 2007, p.1
2. Yoram Soloman, "The Economics of Mobile TV", 2007, p.2
3. The "frame rate" of interlaced systems is usually defined as the number of complete frames (pairs of fields) transmitted each second (25 or 30 in most broadcast systems). For example 25p is a video format which runs twenty-five progressive (hence the "P") frames per second. This frame rate is derived from the PAL television standard of 50i (or 25 interlaced frames per second).
4. Yoram Soloman, "The Economics of Mobile TV", 2007, p.2
5. Hendrick Knoche & John D. McCarthy, "Design Requirements for Mobile TV", 2005, p.75
6. Hendrick Knoche & John D. McCarthy, "Design Requirements for Mobile TV", 2005, pp.74-75
7. Hendrick Knocke, John D. McCarthy, M. Angela Sasse, "Can Small Be Beautiful? Assessing Image Resolution Requirements for Mobile TV", MM05, Nov 6-11 2005, p.837
8. Hendrik Knoche, John D. McCarthy & M. Angela Sasse, "Can Small Be Beautiful? Assessing Image Resolution Requirements for Mobile TV", 2005, p.836
9. The Online Reported, 23-29 Feb 2008, Issues 577-I, p.14.
10. DVB-H is a registered trademark of the DVB-H Project
11. MediaFLO USA, and FLO are trademarks of Qualcomm Incorporated
12. Crawford, "Spectrum for Multimedia Services" 2006, page 4
13. Rethink Research, Faultline: European CellCo's act rather than wait for DVB-H spectrum", March 2008
14. Lombardi,Qualcomm, "Presentation to U.S. Subcommittee on Telecom & Internet", May 10, 2007, Pages 6 & 7
15. Joseph, Plets, Martens "DVB-H Broadcast Network Design for indoor reception of DVB-H in Flanders" 2007, page 3
16. Walker, Qualcomm, Presentation to ATSC, "Technical and Business Challenges of Mobile TV" May 17, 2007, Page 3
17. Lombardi,Qualcomm, "Presentation to U.S. Subcommittee on Telecom & Internet", May 10, 2007, Pages 6 & 7
18. Joseph, Plets, Martens "DVB-H Broadcast Network Design for indoor reception of DVB-H in Flanders" 2007, page 3
19. Erkki Aaltonen, Nokia, "DVB-H Radio Network Aspects", July 2005, page 24
20. Wi-Fi® is a registered trademark of the Wi-Fi Alliance
21. Perceived usage rather than measured usage. Ericsson paper, "Changing The Way We Look at Television", p.22
22. Mitch Feinman, Fox Mobile Entertainment, "NATPE 2008 Conference Sessions", "Mobile Content: What's Hot, What's New? What's next?"
23. Salil Delvi, NBC Universal, "NATPE 2008 Conference Sessions", "Mobile Content: What's Hot, What's New? What's next?"

25. Crawford, "Spectrum for Multimedia Services" 2006, page 4
26. Kaufhold , "US Consumers Attitudes About Mobile Communications & Entertainment" , In-Stat, September, 2007
27. All values originally in Euro, converted to USD assuming 1.44 USD/EURO. Urban, "Mobile Television: Is it just Hype or a real Consumer Need" Observatorio Journal,3 (2007) 045-058, Page 53
28. DVB-H Web Site, "3 Italia – Italy Services", May 2007, http://www.dvb-h.org/Services/services-Italy-3Italia.htm. Euro 29 per month at 1.44 EUR to USD
29. David Sidebottom, Understanding & Solutions, Mobile TV Forecast Update based on 3GSM 2008 Review, Feb 22, 2008.
30. Understanding & Solutions define mobile TV in two ways, dedicated mobile TV whereby a dedicated broadcast technology chip is required in the handset to receive programming and includes DVB-H, Mediaflo, ISDB-T and S/T-DMB standards (plus others). The second method is cellular mobile TV delivery, which is linear TV programming delivered direct to the mobile handset via the operators cellular networks. The data and analysis was complied via a mixture of primary and secondary research. Company financials and reports and trade press sources are supplemented with dedicated face to face and telephone research with the key players in these sectors. Total market estimates are derived from a "bottom-up" approach, with individual services totaled to create a total market estimate.
31. A late 2007 ABI research study projects 462 million mobile TV subscribers by 2012. This number assumes a combination of Unicast and Broadcast technologies. Interestingly Asia Pacific has the most growth and represents 56% (260 million) of the 2012 number. The Cable & Satellite Broadcast Association (CSBAA) provides better granularity to these numbers as they track broadcast only mobile TV numbers. Their studies project 76 million broadcast subscribers by 2012. From these two projections one can conclude that a vast majority of Asia Pacific (and probably worldwide) Mobile TV usage will be over unicast technology.
32. Adapted from Yoram Soloman, "The Economics of Mobile TV", 2007, p.10, figure 8.
33. Herbert Mittermayer, "Fast Channel Switching Description", Alcatel-Lucent, p.1
34. Gallouzi "Deal with OFDM, a new old technology" July, 2007, Page 2
35. Wi-Fi is a registered trademark of the Wi-Fi Alliance
36. WiMAX is a trademark of the WiMAX Forum.
37. Gallouzi "Deal with OFDM, a new old technology" July, 2007, Page 2
38. Gallouzi, Brew 2005 Conference,"MediaFLO 101: FLO Technology" June 2005, Pages 23 -25
39. Faria, Henriksson, Stare, Talmola "DVB-H: Digital Broadcast Services to Handheld Devices" 2006, Page 2 & 3
40. Herbert Mittermayer, "Unlimited Mobile TV, DVB-SH A Natural Evolution of DVB-H", Alcatel-Lucent, Jan 2008, p.14
41. Discussion with Juan-Pablo Torres, Alcatel-Lucent, Feb 2008
42. Digital Video Broadcasting Project, "System Comparison of T-DMB vs. DVB-H, 2006, Pages 3-5
43. Bakhuizen, Horn "Mobile broadcast/multicast in mobile networks", 2005
44. Rethink Research "Wireless Watch: In-depth Analysis of WLan, Cellular and Broadband Wireless Markets", Vol. 5, issue 46., Feb 15th 2008, p.19

45. Herbert Mittermayer, “Fast Channel Switching Description”, Alcatel-Lucent, Feb 2008
46. Ericson Technology Paper, “Technical Overview and Performance of HSPA”, June 2007, Page 6
47. Ericsson Presentation, “Mobile Features to increase HSPA performance”, July 2007, Page 1
48. Louis Gump, The Weather Channel Interactive, “NATPE 2008 Conference Sessions”, “Mobile Content: What’s Hot, What’s New? What’s next?”
49. Juniper Research, “Mobile Broadband Markets WIMAX, EV-DO, HSPA & Beyond 2007 -2012” November 28, 2007, Page 7
50. Walker, Qualcomm, Presentation to ATSC.
51. Ericsson paper, “Changing The Way We Look at Television”, p.23
52. Michel Grech, Alcatel-Lucent Mobile TV Business Case (DVB-SH) UK Model” Jan 2008, p.14
53. Qualcomm paper “The Economics of Wireless Data” p.16
54. Economics are based on a snapshot and do not take into account potential declines over time or volume discounts.
55. Iain Morris, “Telecommunication Magazine”, Jan 29, 2008
56. Frank Hartung, Uwe Horn, Jorg Huschke, Markus Kampmann, Thorsten Lomhar, Magnus Lundevall, “Delivery of Broadcast Services in 3G Networks”, IEEE Transactions on Broadcasting, Vol 53, No. 1, Mar 2007, p.194
57. D. Plets, W. Joseph, L. Martens, E. Deventer, and H. Gauderis, “Evaluation and Validation of the Performance of a DVB-H Network”, 2007 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Orlando, Florida, USA, March 28 – 29, 2007.
58. D. Plets, W. Joseph, E. Tanghe, L. Verloock, L. Martens, “Analysis of propagation of actual DVB-H signal in a suburban environment”, IEEE International Symposium on Antennas and Propagation, Honolulu, Hawaii, USA, Paper No. 1386, 10 – 15 June 2007.
59. DVB-H channel based on 16QAM, 18 channels in 8MHz of UHF spectrum.
60. Configuration based on (a) MUX of 5MHz both via satellite and terrestrial, using QPSK = 9channels, nationwide; and either (b) MUX of 5MHz via terrestrial, using QPSK = 18 channels (9channels each MUX), terretrial indoor 3G-like coverage; or (c) MUX of 5MHz via terrestrial, using 16QAM = 36 channels (18channels each MUX), terretrial indoor 3G-like coverage. We assume QPSK in our example to create a comparison to DVH-H. Same number of channels with DVB-SH using half of the repeaters.
61. TeamCast and DiBcom, Mobile Performance Calculator, 2005, http://www.teamcast.com/en/maj-e/c2a2i12445/support/dvb-h-calculator/dvb-h-calculator.htm
62. IBC Show, TeamCast, “DVB-H: Digital TV in the Hands”, Sept. 2005, page 2

MPEG-4 TRANSITION USING SWITCHED DIGITAL VIDEO

John Schlack
Motorola, Inc.

Abstract

During the transition from MPEG-2 to MPEG-4, it will not be possible, due to bandwidth constraints, to broadcast all channels in both MPEG-2 and MPEG-4 format. However, the switched digital video (SDV) system can be used to manage the transition from MPEG-2 to MPEG-4 and to minimize the bandwidth needed by the system until the transition completes.

The switched digital video system delivers streams with an encoding type based on the decoding capabilities of the settop boxes tuned to that channel. The SDV system may need to force tune settop boxes or use transcoding to handle transitions between different encoding types.

TERMINOLOGY

The terms H.264, AVC (advanced video coding), and MPEG-4 Part 10 all refer to the same standard for video compression. These terms may be used interchangeably in this document. Additionally, this document uses the term MPEG-4 to refer to MPEG-4 Part 10. MPEG-4 provides high quality video at substantially lower bit rates than its predecessor, MPEG-2.

BACKGROUND

Cable operators currently deliver a mix of analog video and digital video on the cable plant. The digital video includes a large number of standard definition (SD) channels as well a much smaller number of high definition channels (HD). The operators also provide other services such as high speed data, video on demand (VOD), and voice over IP (VOIP).

The cable industry standardized delivery of digital video content using the MPEG-2 format. This format requires about 3.75 Mbps of bandwidth for SD programs and 15-19 Mbps for HD programs. A 6 MHz QAM channel modulated using QAM256 can carry 10 SD or 2 HD programs. Cable operators may use statistical multiplexing to groom the channels. This allows a 6 MHz QAM channel carry 11-12 SD programs or 2-3 HD programs.

Figure 1 shows an example 860 MHz HFC plant carrying these different services. This example system delivers 50 HD channels. However, many cable plants are 750 MHz or less. These cable systems deliver much less HD programming.

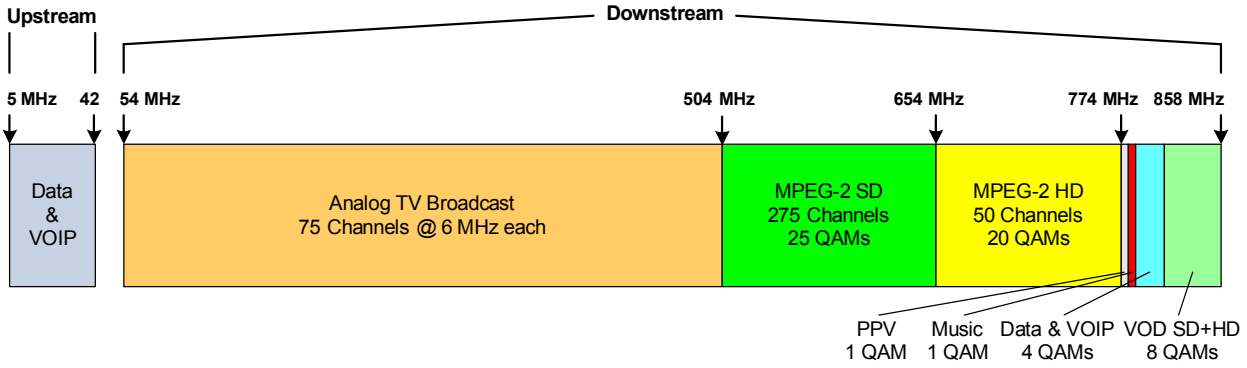


Figure 1: Example 860 MHz HFC Plant

Cable operators are being pressured to deploy additional HD channels. Subscribers are replacing standard definition television with HD sets. This is one source driving the need for additional HD content. Figure 2 shows the current and projected growth of HD sets [1].

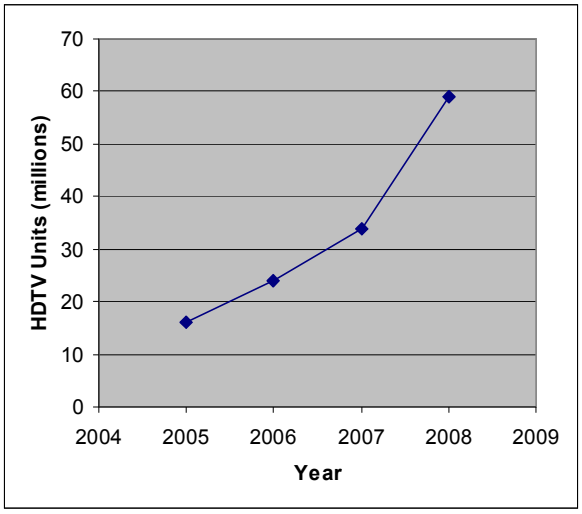


Figure 2: HDTV Growth

Competition is also driving the need to deploy more HD content. Satellite providers currently deliver 70-90 HD channels and plan to increase the number to over 100 channels by the end of 2008. Telco operators plan to deliver a similar number by the end of 2008. Cable operators currently deliver 20-50 HD channels [2]. The cable operators need to deploy more HD channels to keep pace.

The industry has begun a shift towards encoding video using MPEG-4 as a method to reduce bandwidth for delivering HD content via satellite. As an example, a major content provider has announced that it will deliver the HDTV versions of all 26 channels to cable headends using an MPEG-4 encoding [3]. MPEG-4 video requires about 30-50% less bandwidth than comparable MPEG-2 content.

The settop box manufacturers are also beginning to deliver settop boxes that can decode MPEG-4 video. Using MPEG-4 instead of MPEG-2 to encode all SD and HD programs will produce significant bandwidth savings. It will be possible to deliver 100 channels of HD programming and 275 channels of SD programming on an 860 MHz cable plant while still keeping a large analog channel lineup for subscribers without digital settop boxes. The bandwidth calculations are shown in Figure 3.

TRANSITION PERIOD

	Format	Channels	BW
Analog	Analog	75	450
Digital SD	MPEG-2	275	150
Digital SD	AVC	275	78
Digital HD	MPEG-2	100	240
Digital HD	AVC	100	150
VOD SD + HD	MPEG-2		48
VOD SD + HD	AVC		36
Data + VOIP			24
Music	MPEG-2		6
PPC	MPEG-2		6
Total			1188

TODAY – MPEG-2

	Format	Channels	BW
Analog	Analog	75	450
Digital SD	MPEG-2	275	150
Digital HD	MPEG-2	50	120
VOD SD + HD	MPEG-2		48
Data + VOIP			24
Music	MPEG-2		6
PPV	MPEG-2		6
Total			804

FUTURE - AVC

	Format	Channels	BW
Analog	Analog	75	450
Digital SD	AVC	275	78
Digital HD	AVC	100	150
VOD SD + HD	AVC		36
Data + VOIP			24
Music	AVC		6
PPV	AVC		6
Total			750

Figure 3: MPEG-4 Transition without SDV

However, most digital settop boxes deployed today in a cable system can only decode MPEG-2 encoded content. It is not realistic to replace millions of digital settop boxes with MPEG-4 capable settop boxes in a short time frame. A transition period is required where the cable system delivers both MPEG-2 and MPEG-4 content. During the transition, it will not be possible, due to bandwidth constraints, to broadcast all channels in both MPEG-2 and MPEG-4 format. The "Transition Period" table in Figure 3 shows the bandwidth required to dual carry MPEG-2 and MPEG-4 versions of each channel as well as expands the HD channel lineup to 100 channels.

Many of the cable operators are turning to switched digital video (SDV) as a way to reclaim bandwidth in order to deliver new services, particularly additional high definition content. Switched digital video replaces traditional broadcast programs with a system that only transmits a channel to a service group when requested by a subscriber. The system realizes bandwidth savings since only a subset of the available channels is being watched by subscribers at any given time.

The operator usually places niche or low take rate channels on the SDV tier. When using this "long tail" content, the SDV system can effectively offer at least twice the number of channels than can actually be delivered in a given QAM. The reclaimed QAMs from digital broadcast can be used to offer additional channels or can be assigned to other services.

One can also view MPEG-4 encoded channels as long tail content while the number of MPEG-4 capable settop boxes is less than the number of MPEG-2 capable settop boxes. Thus, moving the MPEG-4 channels onto the switched tier will prove to be a very effective way to introduce MPEG-4 onto the cable plant.

From the example in Figure 1, an operator may choose to move the least watched 100 standard definition (SD) channels and 20 high definition (HD) channels from digital broadcast to the switched digital video tier. These channels originally required 16 QAMs to broadcast. SDV requires approximately 8 QAMs to deliver this content, meaning that 8 QAMs have been reclaimed.

Assume that the 8 reclaimed QAMs will be used for delivering more HD content on the SDV tier. These 8 QAMs can be used to deliver 30-40 additional "long tail" HD channels. Transmitting both MPEG-2 and MPEG-4 format for all channels on the switched tier provides additional bandwidth savings as MPEG-4 capable settop boxes are deployed. Figure 4 shows the example cable plant using SDV to deliver MPEG-2 and MPEG-4 encoded content. The actual mix of SD and HD to deliver on the SDV tier will vary by operator and region based on popularity of the content.

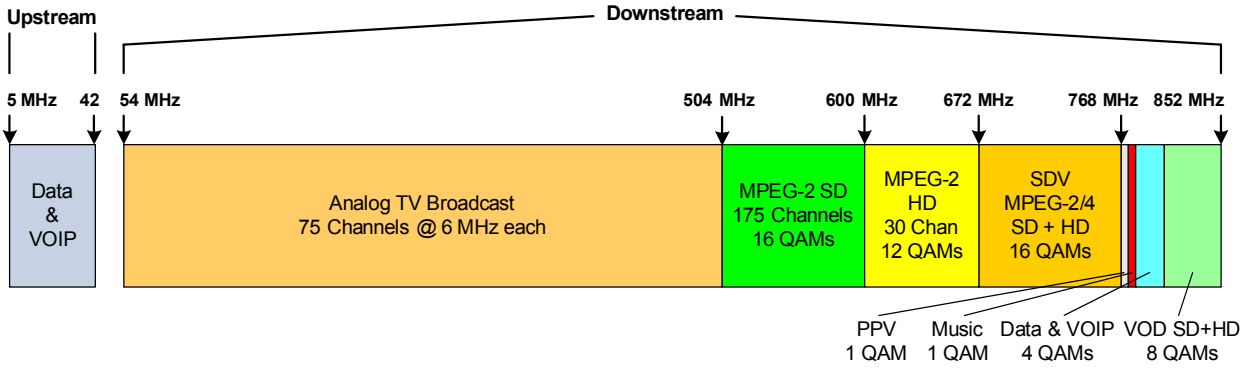


Figure 4: 860 MHz HFC Plant with SDV

The switched digital video system is an ideal platform for launching MPEG-4 channels. However, challenges exist. The SDV system must deliver multiple encoding formats for the same content, ensure that the settop boxes tune to the correct format, and minimize bandwidth usage when there are demands for competing formats of the same content.

MANAGING MULTIPLE ENCODINGS

Architecture

The SDV system has the ability to deliver a stream to the settop box based on the settop box capabilities. For example, if a settop supports decoding only MPEG-2 signals, the SDV system will ensure that the settop is only directed to tune to streams that are encoded using MPEG-2.

The SDV system may be pre-configured with the settop capabilities. For example, the capabilities may be statically tracked by settop model. The operator may track the settop model and settop ID for each fielded settop box. The channel change request from the settop includes the settop ID, which allows the SDV system to discover the settop capabilities and assign the correct stream. Alternatively, when the settop box registers with the SDV system, it may send a message to the SDV system to report its capabilities. The capabilities may include the number of tuners, the video and audio codecs supported, and the communications methods, to name a few parameters.

As the content providers begin delivering MPEG-4 content, the cable system will need to transcode these signals into an MPEG-2 version so that legacy settop boxes will be able to decode the stream. The system may also transcode MPEG-2 content into MPEG-4 content so that additional lower bandwidth content will be available for the MPEG-4 capable settops. This will reduce the overall QAM bandwidth usage on the system if only MPEG-4 capable settops are tuned to a specific channel.

Figure 5 shows an example architecture of a cable plant designed to deliver a stream with multiple encoding types. Transcoders are used to create MPEG-2 and MPEG-4 encodings for the various channels.

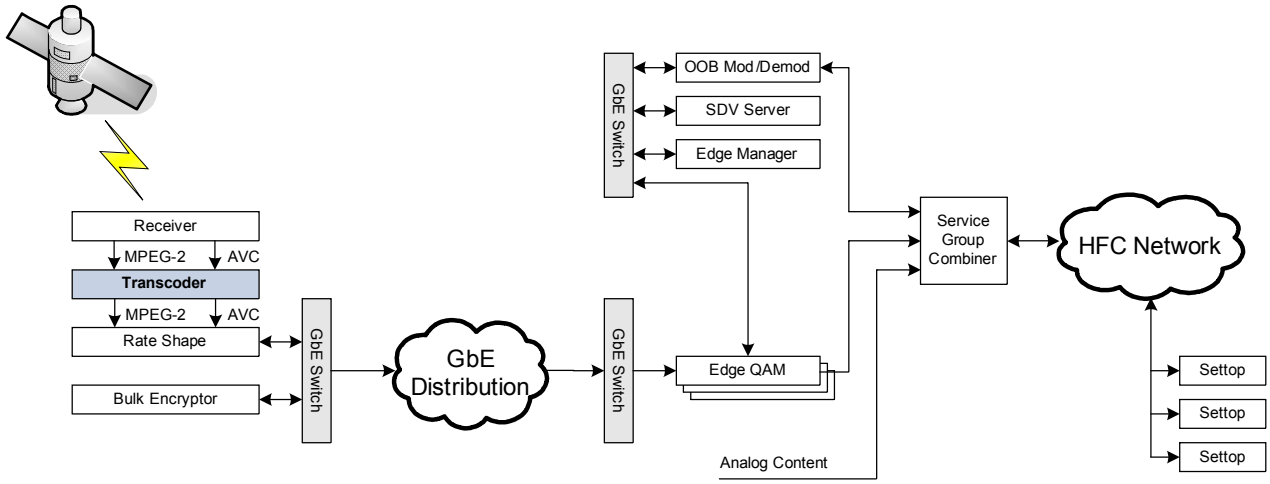


Figure 5: SDV System with MPEG-4 and MPEG-2 Transcoders

In this example, the transcoders are placed in the headend with the other grooming equipment. This increases the load on the distribution network since it will carry multiple copies of some stream to the hubs.

The transcode function is shown performed by a stand alone server. However, it may also be integrated in the edge device, VOD server, Network Encryptor, or any system that ingests and outputs GigE content. The key is that both the MPEG-4 and MPEG-2 versions of the content are available on the plant so that the SDV system can deliver the appropriate version to settop boxes based on their capabilities.

Stream Delivery

The SDV System directs channels to a service group based on requests from the settop box. The SDV system uses the settop capabilities and policy information when determining the encoding format of the stream to deliver to that settop box.

Although it is possible to transcode all content encoded in MPEG-2 into an MPEG-4 encoding to save bandwidth, it is likely that the SDV system will have some content that is only encoded using MPEG-2. In this case, the SDV system always delivers the MPEG-2 version of the stream when requested by an MPEG-2 or MPEG-4 capable settop box. This assumes that the MPEG-4 settop box can decode MPEG-2 encoded content.

The SDV System may have content that is only encoded using MPEG-4. In this case, the SDV system can only deliver such content to MPEG-4 capable settop boxes. MPEG-2 capable settop boxes will be denied service. However, the channel map for the MPEG-2 capable settop box should prevent subscribers from accessing the content if the settop box cannot view it. As an example, the cable operator may have created a special HD tier that is available as a premium package. A subscriber receives an MPEG-4 capable settop box when subscribing to that package so that the subscriber can view that HD content.

For content that is encoded using MPEG-2 and MPEG-4, the SDV system will likely be configured to attempt to deliver only a single encoding of that channel in order to conserve bandwidth.

When a settop box capable of only decoding MPEG-2 content requests an SDV channel that is currently not being delivered to the service group, the SDV system directs the MPEG-2 stream of that SDV channel to that service group and it returns the channel tuning information to the settop box. If an MPEG-4 capable settop box from the same service group subsequently requests the same SDV channel, the SDV system simply directs the settop box to tune the MPEG-2 stream that is already being delivered to the service group. This assumes that the MPEG-4 capable settop can also decode MPEG-2 content.

When a settop box capable of decoding MPEG-4 content requests an SDV channel that is not currently being delivered to the service group, the SDV system directs the MPEG-4 stream of that SDV channel to that service group and it returns the channel tuning information to the settop box. If a settop box capable of only decoding MPEG-2 content that is from the same service group subsequently requests the same SDV channel, the SDV system must direct the MPEG-2 version of the SDV channel to that service group and return the tuning information for the MPEG-2 stream to that settop box. Thus, the channel is being delivered twice to the same service group using different encoding formats.

Carrying both MPEG-2 and MPEG-4 versions of the same SDV channel is inefficient, since it requires at least 50% more bandwidth than carrying the MPEG-2 version of the channel alone.

Managing Multiple Copies

The SDV system can employ several strategies for handling delivery of multiple encodings of the same SDV channel to the same service group. The basic choices are to leave multiple copies of the stream or to force tune all settop boxes onto a single copy of the stream. The chosen strategy depends on the bandwidth

available to the system and the impact of reducing the number of copies of the stream.

The main reason that the SDV system might not always choose to deliver a single encoding of the stream is the force tune operation can be disruptive to the viewing experience. Directing a settop box from one stream to another requires the settop box to tune to a different frequency and/or MPEG program number.

Tuning to a different MPEG program number on the same frequency may take several hundred milliseconds while the tuner waits for the video and audio data to arrive. This can cause jitter or blocking on the display. Tuning to a different frequency requires possibly 1-2 seconds while the tuner waits for information describing the stream contents and then waits for the video and audio data to arrive. This causes either a frozen image or a black screen for the duration of the tune. The tune operation may also impact the DVR causing recordings to fail.

If the SDV system has sufficient unused QAM bandwidth for that service group, the SDV system may simply allow multiple encodings of the channel to remain on the cable plant. The SDV system will still have the ability to service new channel requests, therefore it would be best not to disrupt the viewing experience by merging the streams. The SDV system can defer action until available bandwidth for that service group drops below a "consolidation threshold".

In the case where the SDV system elects to leave both streams on the plant, the SDV system may direct all new tune requests for that channel onto the MPEG-2 version of the stream in hopes of recovering resources when the settop boxes originally tuned to the MPEG-4 version of the channel tune off. This mechanism allows the SDV system to gradually recover resources without force tuning settop boxes to the MPEG-2 version of the channel.

In another case, the SDV System may use recent channel change activity to detect that the MPEG-2 settop is channel surfing. Thus, subsequent tune requests from MPEG-4 capable settops may be placed into the MPEG-4 encoded stream in anticipation of the MPEG-2 settop channeling off shortly. If the MPEG-2 settop stays on the channel or other MPEG-2 settops tune to that channel, the SDV system may revert to directing all new requests to the MPEG-2 channel as discussed above.

If resource availability in a service group is low and both MPEG-4 and MPEG-2 versions of the same channels are being delivered to that service group, the SDV system may choose to force tune viewers to the MPEG-2 stream and recover the resources assigned to the MPEG-4 stream. This force tune operation may be disruptive, but it is the fastest way to reduce resource usage in a service group.

The SDV system executes the force tune operation by sending a message to each settop box currently tuned to the MPEG-4 version of the stream to tune a different frequency and/or program number. If the settop box is still tuned to the stream in question, it will tune to the MPEG-2 version of the stream using a frequency and MPEG program number embedded in the request message. The settop box returns an acknowledgement indicating whether the tune operation was performed. If the settop was already tuned to a different channel, the settop box returns an acknowledgment indicating that the tuner is no longer on the channel in question. The SDV system will retry the force tune operation for settop boxes that fail to acknowledge its request.

The SDV system has the option of either immediately recovering the resources assigned to the MPEG-4 stream or waiting until all settop boxes have been tuned to the MPEG-2 stream. Recovering the resources immediately is even more disruptive than simply the force tune operation since the SDV server directs the edge QAM to stop streaming the MPEG-4 encoded content. This likely occurs before the settop boxes begin their tuning operation to the

MPEG-2 channel. This results in additional time with a frozen picture on the screen or a blank screen.

The more graceful method of handling the resource recovery would be to wait until all active tuners have been directed off the MPEG-4 stream. Although this delays resource recovery by several seconds, it minimizes the impact on the subscribers and provides an opportunity for the settop box to use some advanced features in an attempt to minimize the impact of the force tune.

The SDV system can further attempt to minimize the impact of force tuning by following these guidelines. The SDV system makes every effort to place both the MPEG-4 and MPEG-2 versions of the same channel on the same QAM. This greatly reduces force tune times. The SDV system may schedule force tunes to coincide with transitions between content and advertising. These force tunes would not only require knowledge of the expected time for the advertising, the SDV system would need to receive a trigger from the ad server indicating the exact time of the advertising since the goal would be to place the force tune on the transition into or out of the ad pod. This would minimize the impact on both the subscriber and advertiser since the transition typically includes a black frame or a fade to black sequence.

Recent developments in settop box-based ad splicing are providing the capability that the settop box can perform a seamless splice between programs carried on the same frequency. If the SDV system is able to keep both the MPEG-2 and MPEG-4 stream on the same QAM and it does not recover resources until all settop boxes have acknowledged the force tune operation, then the settop box may be able to jump from the MPEG-4 stream to the MPEG-2 stream with little or no noticeable impact on the subscriber viewing experience.

DYNAMIC TRANSCODING

Overview

The previous section described a static transcoding method in which the headend network carries both MPEG-2 and MPEG-4 copies of the same SDV channel to support both legacy settop boxes and MPEG-4 capable settop boxes.

Another method would be dynamic transcoding. In dynamic transcoding, the SDV System directs a single MPEG stream onto the service group for each requested SDV channel. The MPEG stream is dynamically transitioned between MPEG-2 and MPEG-4 based on the types of settop tuned to that SDV channel.

Dynamic transcoding should be possible to implement. The transport stream structure for MPEG-4 encoded content can be similar to that of MPEG-2. However, the video and audio packets would carry MPEG-4 encoded data instead of MPEG-2 encoded data. Additionally, the stream type identifier will signal that the content carried in the stream is MPEG-4 [4].

MPEG decoders currently exist that can decode both MPEG-2 and MPEG-4 content. These are being deployed into the new MPEG-4 capable settop boxes. However, it is likely that these decoders cannot dynamically transition between the two encoding types. New decoders may need to be deployed.

Decoders capable of handling the dynamic transition between encoding types will be alerted to the transition by a change in the PMT and a change in the stream type identifier. The transition will occur on an I-frame boundary for MPEG-2 and IDR frame for MPEG-4.

Architecture

The transcoding device itself can be a stand alone server, part of an edge QAM, embedded in a video server, or embedded in a network

encrypting device. Note that the transcoding device cannot work on encrypted streams. Therefore, encryption must be applied after the transcode. This can be done by a bulk encryptor in the headend or encryption logic in the edge QAM in the hub.

Figure 6 shows an example system where dynamic transcoding occurs in the edge QAM. The headend and GbE distribution network carry a single copy of each channel. Each channel may be encoded in MPEG-2 or MPEG-4 based on how it was received from the satellite. Then, on a service group by service group basis (i.e. QAM by QAM basis), the edge QAM can be directed to output either the MPEG-2 version of that channel or the MPEG-4 version of that channel.

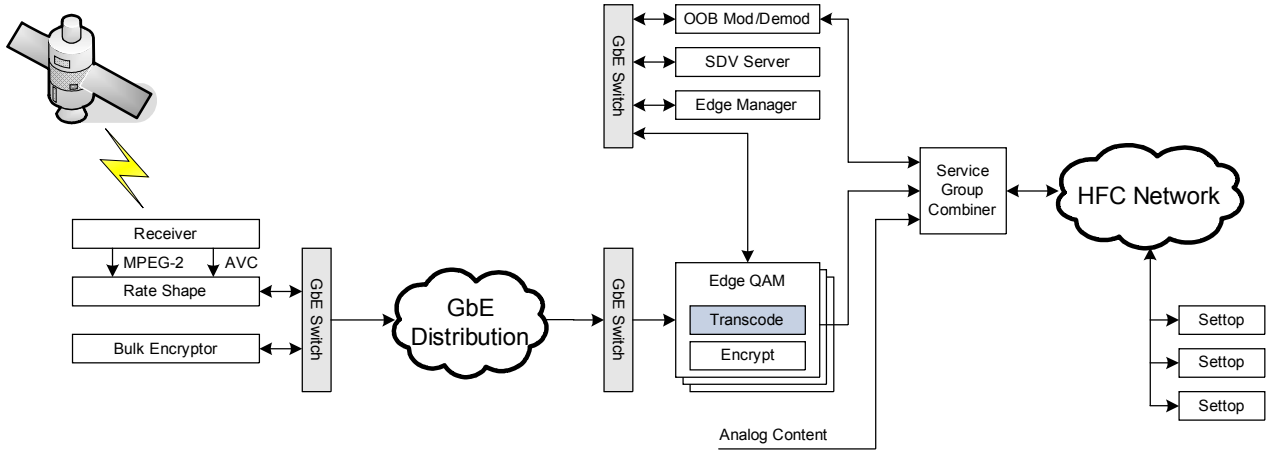


Figure 6: Dynamic Transcoding in the Edge QAM

A drawback of transcoding in the edge QAM is that the stream cannot be encrypted until after the transcode. This means that the stream must be both transcoded and encrypted at the edge. Performing both transcoding and encryption operations at the edge may be costly as it requires dedicated processing power for each service group.

Dynamic transcoding may also occur in a video server. The video server ingests the MPEG-4 or MPEG-2 source content. It can then delivers an MPEG-2 or MPEG-4 version of the stream for each service group with a settop box tuned to that channel. Figure 7 shows dynamic transcoding using a video server.

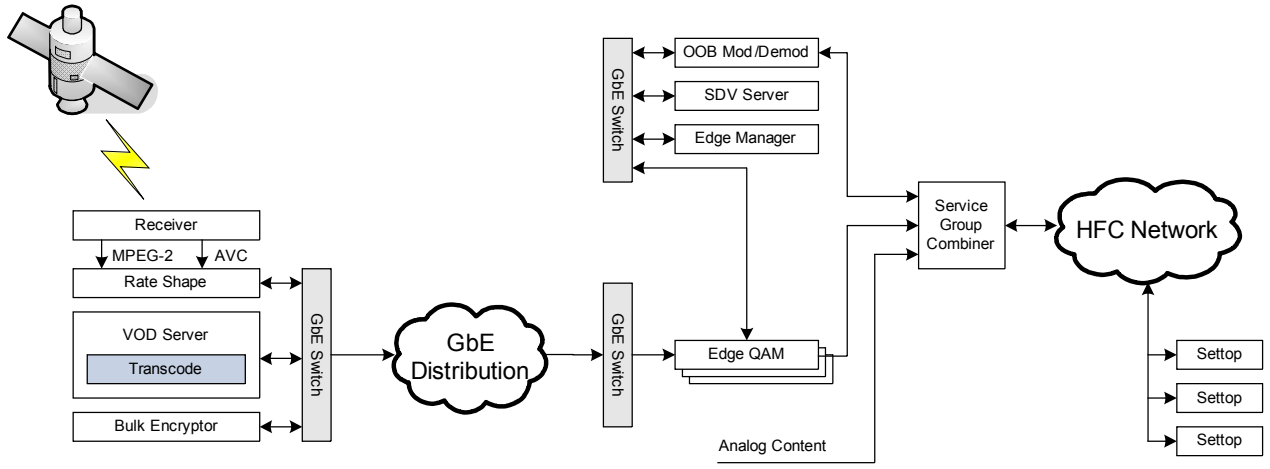


Figure 7: Dynamic Transcoding in the Video Server

By delivering a unicast stream for each service group, the Video Server is free to transition the stream between the MPEG-2 and MPEG-4 versions as requested by the SDV system. Note that this method requires significantly more bandwidth on the GbE distribution network than edge QAM based transcoding.

Stream Management

Stream management when using dynamic transcoding is greatly simplified. The SDV system simply manages the transitions between the MPEG-2 encoding and the MPEG-4 encoding based on the capabilities of the settop boxes that tune to the channel. In most cases, no stream merging concerns exist. That said, there are still issues that need to be handled.

In one case, a particular SDV channel is being delivered in MPEG-2 format to a mix of legacy and MPEG-4 capable settop boxes for a particular service group. If all legacy settop boxes tune to other channels, the SDV Server has the opportunity to transition the stream from an MPEG-2 encoding to an MPEG-4 encoding. The SDV Server sends a message to the transcoding device directing it to transition from the MPEG-2 version of the stream to the MPEG-4 version of the stream for the service group in question.

The SDV System may not immediately transition an MPEG-2 stream to an MPEG-4 stream the instant all legacy settop boxes tune off. The SDV Server may use heuristics to determine the appropriate transition time, since viewer behavior may indicate that some legacy settop boxes are only channel surfing and will return to the original SDV channel shortly. For example, the SDV System may have a database of advertising times for that SDV channel. If legacy boxes tune off during commercial breaks, the SDV System may wait until the commercial break ends before determining whether to transition to the MPEG-4 encoding.

In another case, a particular SDV channel is being delivered in MPEG-4 format to a set of MPEG-4 capable settop boxes. When a legacy settop box requests that SDV channel, the SDV Server must immediately direct the transcoding device to transcode the channel to MPEG-2 format. The transcoding device will begin transcoding at the next I-frame / IDR-frame boundary. The transcoding device then sends a response to the SDV Server when the MPEG-2 transition has completed. The SDV Server returns the tuning information to the legacy settop box.

The SDV Server must wait for confirmation that the transcoding device transitioned to the MPEG-2 version of the stream before the SDV Server returns tuning information to the legacy settop box. This may take several hundred milliseconds or more. By waiting for confirmation, the legacy settop box will be guaranteed it is attempting to tune an MPEG-2 stream and not an MPEG-4 stream.

Note that the SDV Server may deny the request of the legacy settop box to tune to the SDV channel if there is insufficient output bandwidth on that QAM to switch from MPEG-4 to the MPEG-2. Alternatively, the SDV System may need to establish the MPEG-2 version of the stream on a different QAM that has sufficient bandwidth. In this case, one returns to the case where the SDV System delivers multiple streams for the same channel.

COMPLETING THE TRANSITION

Initially with a small number of MPEG-4 capable settop boxes, SDV will be an enabling technology for deploying MPEG-4 into the cable plant, allowing delivery of additional HD content and potential bandwidth savings. As the number of MPEG-4 capable settop boxes grows beyond the deployed legacy settop box count, the cable plant will transition to broadcasting more MPEG-4 content while delivering the MPEG-2 content on the switched tier. This will provide

further bandwidth savings. As the number of legacy settop boxes dwindle due to replacements and upgrades, the cable operator may find it advantageous to replace the remaining legacy settop boxes and remove MPEG-2 content from the cable plant. This would complete the transition to MPEG-4 encoded content.

REFERENCES

[1] Goetzl, D. (2007, November 26) HDTV Set Sales to Soar, *Broadcasting & Cable*; and B&C Staff (2007, December 28) CEA: DTV Penetration Tops 50% Mark in U.S., *Broadcasting & Cable*.

[2] Patterson, T. (2008, February) Fighting for HD bragging rights, *CED Magazine*, 34.

[3] Ellis, E. (2007, August 1) The MPEG Transition, *CEDMagazine.com*.

[4] Robuck, M. (2006, July 1) MPEG-2 Stays Center Stage While MPEG-4 Waits in the Wings, *Cable360.net*.

CONTACT

John Schlack
Distinguished Member of Technical Staff
john.schlack@motorola.com

101 Tournament Drive
Horsham, PA 19044
Phone: (215) 323-1000

OPTICAL SEGMENTATION TECHNOLOGY ALTERNATIVES AND ARCHITECTURES

Phil Miguelez – Director, BAN Advanced Technology
Fred Slowik – Director, ANS Systems Marketing
Motorola Access Networks Solutions - 051808

Abstract

Fiber non-linearity presents serious challenges to fielding multi-wavelength optical systems capable of transporting full broadcast / narrowcast channel loads. Equipment vendors are rising to meet this challenge with new technologies that will allow MSO's to cost effectively segment nodes and harvest fiber for new services or future network segmentation.

In this paper we briefly review the key fiber optic challenges that are driving innovation and examine the different technology choices that are being offered today. We also look at a few implementation models of multi-wavelength on the HFC plant for a variety of different applications.

INTRODUCTION

Market Drivers:

Cable operators are faced with a wide range of opportunities for expansion into adjacent markets such as commercial access and cell tower backhaul. On top of this, competition and customer expectations are creating the need for ever increasing bandwidth capacity in the traditional CATV network. In order to meet the needs of both markets additional fiber or increased capacity of existing fiber is required. In most cases operators prefer to keep business services on a separate fiber network from residential video and data. Cable modems serve small offices well but larger businesses require GigE data rates and dedicated fiber.

Operators are also challenged to minimize CapEx spend and limit system down time. This is especially true in the current unforgiving economic environment. Pulling new fiber is not an option except in green fields and point to point business access situations where the revenue opportunities justify the expense. For all other applications a means to increase capacity using existing fiber is required. Multi-wavelength broadcast + digital transport is an ideal solution to meet this challenge.

Multi-wavelength transport allows node segmentation with minimal touching of the physical plant. More importantly, the increased BW capacity of fibers carrying multiple wavelengths allows surplus fiber to be harvested for other uses. These repurposed fibers can be used for business access or further network segmentation needs.

WDM solutions for digital transport are commonplace. CWDM and DWDM network architectures for baseband digital and QAM data delivery have been in place for 10 years or more. The major barrier for realizing these networks as part of the HFC downstream broadcast system has been the transport of analog video carriers. Early attempts to transport analog broadcast services over a CWDM network yielded poor results. Analog video is extremely susceptible to noise and distortion. Fiber induced distortions add directly to the native distortion of the source laser

transmitter. Additionally, fiber and passive device nonlinearities create a host of potential impairments that must be avoided, minimized, or overcome.

Obstacles to Multi-Wavelength Transport:

When multiple wavelength signals propagate through optical fiber an array of impairments come into play, the most significant of these are Raman crosstalk, four wave mixing, dispersion, and cross phase modulation (XPM). The magnitude of each of these impairments is a function of the laser chirp, the optical launch power, the length of the fiber link, and the dispersion properties of the deployed fiber.

Additional impairments are also possible due to interactions with passive elements in the system such as optical mux and demux filter components.

Detailed descriptions of each of these optical nonlinearities have been presented in numerous articles, technical whitepapers and previous conference presentations on emerging multi-wavelength technology. This paper will provide a brief explanation for each of the critical distortion generators where appropriate to emphasize their impact to analog or digital QAM performance.

Broadcast + Narrowcast Multi-Wavelength Solutions:

Different techniques to mitigate the numerous fiber induced distortions listed above have driven each vendor to create unique, proprietary solutions. In order to take advantage of the wide availability of proven analog capable lasers and keep the complexity low, most vendors have elected to operate in the 1310 nm region. Some vendors have chosen to pursue ITU standard coarse wavelength spaced (CWDM) solutions. Other vendors have promoted dense wavelength spaced (DWDM) solutions. Both approaches permit some of the fiber nonlinearity issues to be minimized while making other optical impairments more difficult to correct.

All of the various solutions have a few common requirements. First among these is the necessity of having identical analog broadcast channel lineups on each wavelength. Analog carriers are the most susceptible to crosstalk distortion. If the signal modulation on each channel is identical, crosstalk susceptibility significantly reduced. Broadcast QAM will also benefit from this same effect. Narrowcast QAM by definition is unique to each wavelength. Narrowcast modulation channels will experience increased noise impairments due to crosstalk but the nature of digital modulation is more robust to these impairments.

Optical power levels for each wavelength should be roughly equal. Mixing high and low power lasers creates the potential for Raman scattering issues.

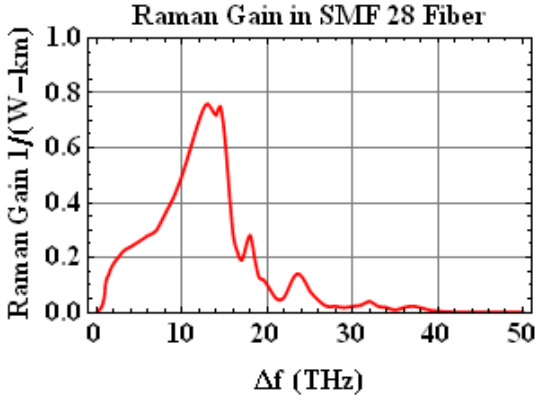
Another requirement in common is that the native laser distortion at each wavelength be as low as possible. Many of the fiber induced distortions will magnify the raw distortion of the laser transmitter. Each distortion parameter in a multi-wavelength system is a composite of the laser distortion plus the distortion generated within the fiber and passive elements as shown in the following example equation to calculate end of line (EOL) CSO performance.

$$\text{EOL (cso)} = 20 \log \left(10^{(\text{Laser CSO}/20)} + 10^{(\text{Fiber CSO}/20)} + 10^{(\text{Mux CSO}/20)} + 10^{(\text{Demux CSO}/20)}\right)$$

The following sections will discuss differences between DWDM and CWDM solutions for multiple wavelength transport of analog broadcast + narrowcast channel loading.

1310 DWDM Solution for Multi-Wavelength Transport:

Stimulated Raman Scattering (SRS) effects have always been the most difficult impairment to conquer. However, Raman gain is predictable based on the optical power, link length, and wavelength spacing. The plot below shows the Raman gain coefficient versus wavelength spacing in THz. Operating with close spaced wavelengths (left side of the plot) is an effective way to minimize SRS.



ITU standards do not exist for DWDM in the 1310 nm O-Band spectrum but translating the 200 GHz or 100 GHz channel spacing commonly used at 1550 nm wavelengths to 1310 nm is easily done.

While DWDM spacing helps to solve Raman crosstalk it enables another impairment, Four Wave Mixing (FWM). This impairment acts in an analogous manner as composite triple beat distortion. With equally spaced wavelengths the interaction of three wavelengths will create a beat that falls on the fourth wavelength. Custom wavelength selection avoiding equally spaced wavelengths is part of the solution to FWM. Distortion from Four Wave Mixing is most pronounced as the wavelengths used are operated near the zero dispersion point (ZDP) of the fiber. Additional crosstalk and CSO with as few as two DWDM wavelengths has been reported when the optical channels were operated in the zero dispersion region.

The selected wavelengths must be located away from the zero dispersion point of the fiber. The ZDP of SMF28 and SMF 28e fiber typically falls near 1310 nm but can vary from fiber lot to fiber lot over a range of +/- 10 nm. The newest version of fiber that Corning plans to introduce this year (SMF28e+) will shift the typical ZDP to 1317 nm. Balancing the choice of wavelength selection to avoid FWM and the ZDP of the deployed fiber is one of the reasons for the different proprietary schemes of the vendors supporting the DWDM approach.

Perhaps the most challenging issue facing DWDM multi-wavelength solutions is related to the optical passives. Mux and demux devices are constructed using thin film optical filters. The broadband response of these filters is usually quite flat but as the filter bandwidth becomes narrow as required for DWDM wavelength spacing the pass band ripple response can increase significantly. This higher ripple creates sloped or tilted regions in the bandpass response which interacts with laser chirp to generate additional CSO beat products. At tilts larger than a few tenths of a dB / nm the CSO generated in the filter will begin to dominate the end of line distortion performance depending on the chirp level of the laser used. To avoid the problem of passband ripple, mux and demux filters must be selected to very tight specifications.

1310 CWDM Solution for Multi-Wavelength Transport:

Maintaining ITU standard CWDM spacing simplifies a number of the challenges that face vendors of DWDM O-Band systems. Four Wave Mixing issues are eliminated since the phasing of the optical wavelengths are de-correlated by fiber dispersion. Optical passives with 20 nm channel spacing provide flat passband response with measured ripple slope of < 0.1 dB / nm. The filter bandwidth is much

greater than the worst case wavelength variation of the cooled laser transmitter, so stability over environmental conditions is generally assured.

The major challenge to CWDM broadcast transport is Stimulated Raman Scattering. CWDM wavelengths are based on 20 nm spacing defined by ITU standards. At this spacing, Raman gain is

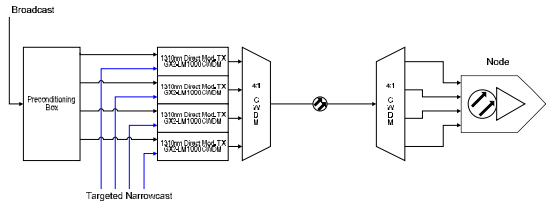
a significant factor and peaks in systems with 3 to 4 sequential channels. CSO distortions generated in the RF and Optical domain by the laser are magnified by Raman gain interactions within the fiber and can dominate the overall system performance. High fiber dispersion such as occurs at 1550 nm would tend to de-correlate the modulated signals (walk off effect) and help reduce the magnitude of Raman crosstalk. Near 1310 nm dispersion is low so walk off is minimal. Optical launch power strongly contributes to the magnitude of the Raman induced CSO distortion. Therefore, limiting laser output levels will minimize the effects of Raman at the expense of link reach.

Adding wavelengths increases the optical power into the fiber and therefore increases the magnitude of Raman crosstalk proportionate to the additive optical level. Without using some external means of correction, multiple CWDM wavelengths with Broadcast + Narrowcast loading muxed onto a single fiber will produce unacceptable CSO distortion in optical links that exceed 12 to 15 Km.

Enhanced Coarse Wavelength Division Multiplexing (E-CWDM)

Enhanced CWDM is a patented technology developed by Motorola to mitigate Raman impairments in multi-wavelength systems. A unique method of conditioning the RF broadcast carriers minimizes Raman distortion along the fiber path. RF conditioning in conjunction with low chirp laser transmitters allows extended link reach of up to 30 km.

E-CWDM Block Diagram



As shown in the block diagram above, the broadcast input channels are conditioned and split to feed the individual CWDM lasers. Narrowcast channels are fed directly to each laser. No custom equipment is required at the node. This solution can be configured with separate forward and return path fibers or combined with a 1310 / 1550 WDM to provide a single fiber solution for upstream and down stream loading.

For short reach applications (<15km) RF conditioning is not a requirement. We have found in these cases that it is possible to reuse currently deployed 1310 transmitters as long as the output power is equal or padded to match the added CWDM lasers.

Multi-wavelength solutions are extremely cost effective compared to the capital expense of pulling new fiber. However, this technology does have limits. Fiber and passive component insertion loss reduces link reach compared to a single transmitter. Since many of the fiber distortions are optical level sensitive, cranking up the power is not effective. Distortion performance is a few dB lower than comparable single transmitter distortion particularly CSO which is the most vulnerable to degradation from Raman and dispersion. Even with these restrictions, multi-wavelength solutions can provide sufficient performance to meet the requirements of typical N+6 cascade architectures.

The next portion of this paper reviews the applications of multi-wavelength technology for Fiber Deep network migration strategies.

Multi-Wavelength Applications

The next wave of network migration for cable operators seems to be focusing upon creating smaller node serving areas in order to provide increased bandwidth capacity to and from fewer numbers of subscribers. Whether accomplished by creating "smaller virtual nodes" via adding physical node segmentation capabilities at existing node locations, or by deploying additional satellite nodes deeper into the network, one fact remains - there may not be sufficient fiber available to support this migration strategy.

Previous sections of this paper address some of the various multi-wavelength technologies that are becoming available to operators to help alleviate fiber constrains. Because node sizes, deployment depth and fiber counts can vary from operator to operator and system to system, this paper refers to node migration in terms of a size reduction factor as opposed to absolute house count per node. In this way, the reader can obtain an appreciation of available multi-wavelength options to meet their end goals. For example, an operator with existing node sizes of 1200 HP might desire a 4X reduction factor whereas existing node sizes of 500 HP may only desire a 2X reduction factor to meet the end goals.

Node Segmentation vs. Fiber Deep

Perhaps we should clarify that node segmentation and fiber deep architectures, both candidates for multi-wavelength solutions, have distinct differences and are not always synonymous.

Early in the evolution of HFC network deployment, cable operators had the choice of designing their coaxial plant emanating from optical nodes in either a balanced or unbalanced fashion. Balanced means that all homes serviced from that node were equally divided among all feeder legs from the node. Due to topology, this balancing often required adding express coaxial cables for segmentation purposes. Although this approach provides a smoother migration path for future node segmentation, cable operators were hesitant to invest in the added material and construction cost to balance their node serving areas. Consequently, many operators chose to opt for the less expensive unbalanced approach where the number of homes passed per feeder leg was random.

Ideally, if the original network design had followed the balanced approach, then virtual node segmentation could occur rather smoothly at existing node locations using segmentation capable nodes. Experience to date seems to indicate that only about 20% of existing nodes are sufficiently balanced to permit this ideal form of segmentation. The remaining 80% of existing nodes may be so unbalanced that some combination of segmentation capable nodes plus the addition of new satellite nodes or adding express coaxial cabling may be required. The latter approach does drive fiber deeper in certain areas,

Tables 1 & 2 illustrate a logical example of node segmentation for both balanced and unbalanced scenarios using a hypothetical 512 home passed node and migrating fiber all the way to the home. Note the unbalanced node creates the need for new fiber deployment during the initial migration process.

Table 1

Motorola HFC Network - Balanced Node Migration Path

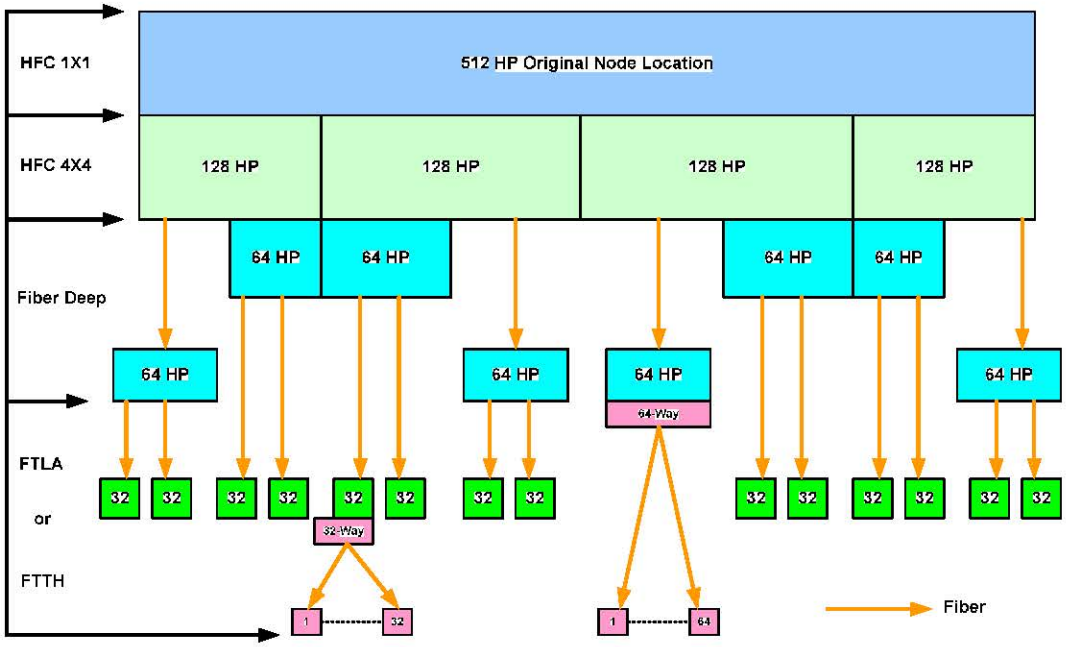
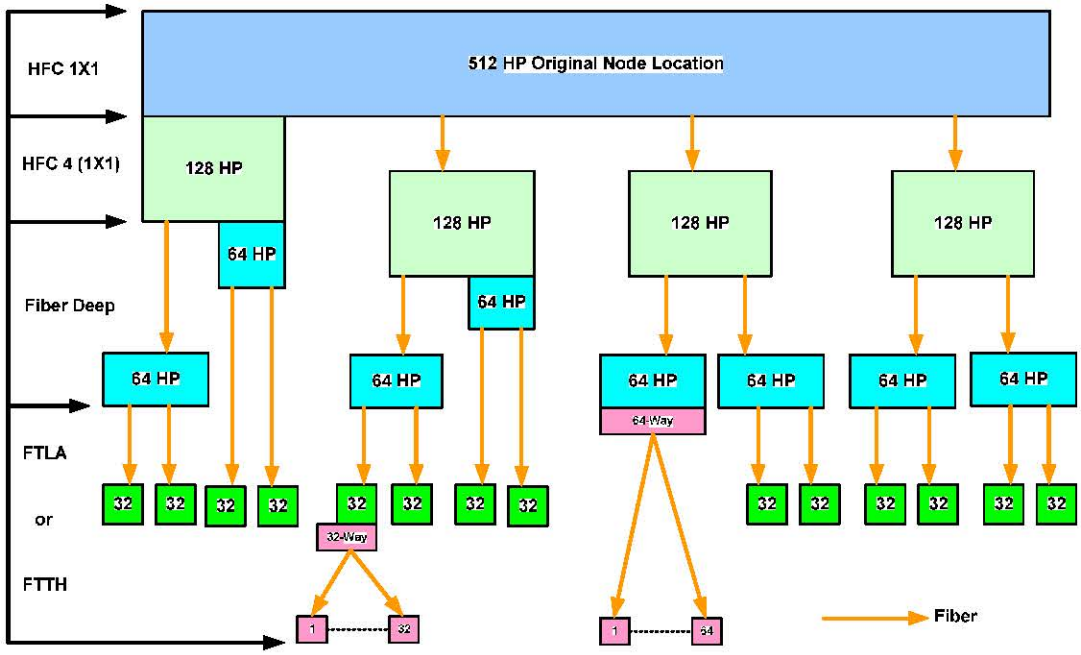


Table 2

Motorola HFC Network – Unbalanced Node Migration Path



Cascade Reduction

Although node segmentation and fiber deep architectures reduce the serving area size with respect to the number of homes/users per virtual node, the amplifier cascade length often remains unchanged. This, due to the fact that certain portions of the segmented node fed from the original node retain their existing footprint while cascade reductions usually take place in those areas where satellite nodes are added.

Figure 1 Existing Node

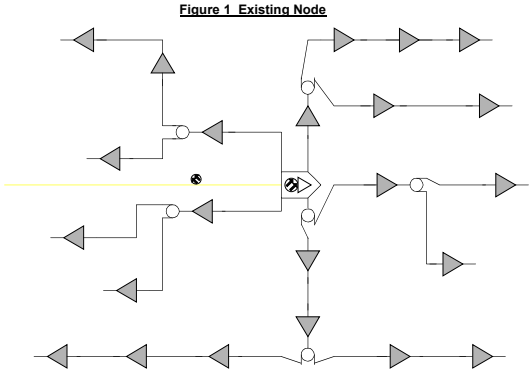
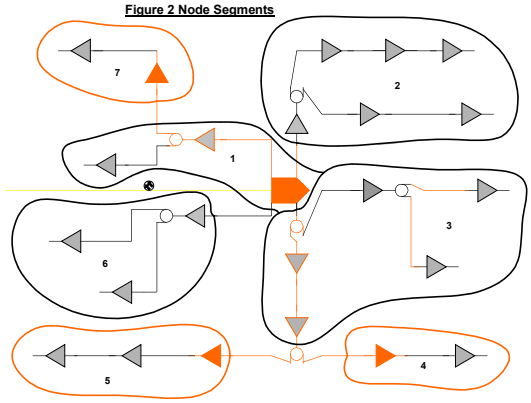


Figure 2 Node Segments



Intentional attempts to drive fiber deeper with the goal of strict cascade limitations often lead to very expensive migration solutions. Maybe a better way of looking at fiber deep would be to size the node to a desirable house count and ignore the cascade length. Pulling fiber to a Node + 0 architecture for example, without a lot of re-plumbing becomes tremendously expensive especially if one merely chooses to drop-in new nodes at all existing amplifier locations.

It is important to understand the cascade impact of fiber migration since different multi-wavelength technologies offer different performance characteristics at the node. Combined optical and RF performance becomes an important consideration in determining which technology will support end-of-line network performance goals. Depending upon RF amplifier cascades, one optical technology might mesh better with reduced amplifier cascades as opposed to another that might be better positioned to support longer amplifier cascades.

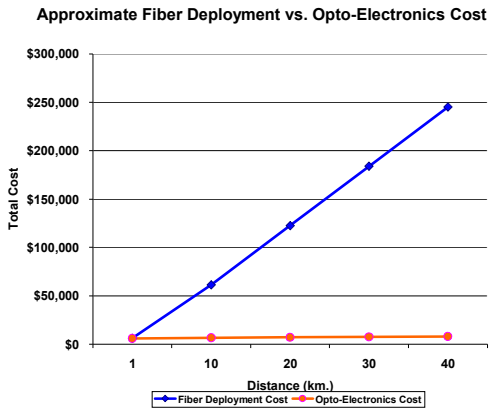
What About Adding Fiber?

If existing fiber counts were unlimited, network migration would be a much simpler task. Unfortunately, fiber counts are constrained in many systems, so operators need to understand new fiber deployment costs versus alternative options such as multi-wavelength technology. A very simple example illustrates.

Fiber Count	Material Per Foot	Aerial Labor per Foot	Aerial Make Ready Per Foot	Total Aerial Per Foot
6	$ 0.27	$ 0.60	$ 1.00	$ 1.87

Using this as an average aerial constructed price per foot, we can easily understand just how expensive installing new fiber can be ($1.87 X 5,280 = $9,873/mile). This cost is far more than the cost of the opto-electronic elements required at the headend / hub, and node location. Multiply this cost by the total distance required to reach an existing fiber starved node location and the cost can become prohibitive. This fiber installation cost does not consider more complex installations such as underground or areas where significant make-ready costs could arise.

The following graph illustrates fiber installation cost on a per km. basis versus the opto-electronic cost per virtual link for various multi-wavelength solutions.



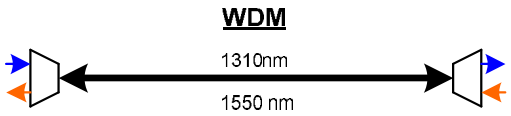
Depending upon the distance and fiber counts to existing nodes, and whether feeder legs are balanced as previously discussed, being able to expand bandwidth capacity via adding wavelengths on existing fiber to existing nodes is advantageous. Less significant are new fiber extensions to satellite nodes that may be required beyond existing node locations. Since these links are usually less than 2 km., the cost becomes much more tolerable, again, depending upon the extent of deployment.

What is the Right Solution?

It now becomes clear that alternative technology is needed to be able to cost effectively drive fiber deeper into the network. Depending upon the particular situation, several multi-wavelength solutions exist or are emerging that may co-exist in the same network. Three basic types of solutions are presented below. All of these options offer significant benefits. These are WDM, E-CWDM, and Broadcast/Narrowcast Overlay. Cost of these bi-directional solutions begin in the $5000/link range (including forward, reverse and nodes, excluding installation and new fiber if needed) and extend upwards based upon specific application needs.

WDM

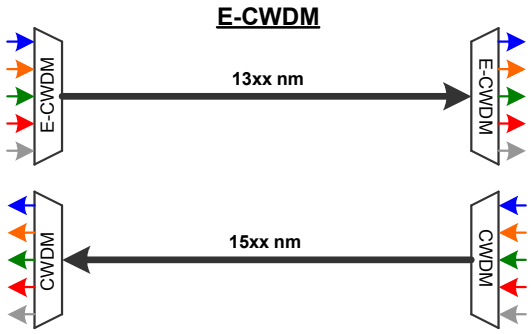
The least complex of the multi-wavelength solutions, this approach simply uses a 1310nm full band (54-1002MHz) downstream wavelength and a 1550nm upstream wavelength placed upon a single fiber. These wavelength directions can be reversed in some applications.



Node segmentation is accomplished by simply lighting up one fiber per wavelength pair. Assuming up to 6X migration is desired and sufficient fibers exist, this method is generally a low cost least complex means to achieve node area segmentation, and can achieve distances greater than 40 km. with excellent performance in the area of 51/-70/-66 dB CCN/CTB/CSO.

E-CWDM

Considered advantageous for fiber constrained applications, this approach, although a bit more complex, enables full band (54-1002MHz) downstream 13xx nm wavelengths upon a single fiber. Depending upon distance requirements, upstream wavelengths may also be deployed upon the same fiber or a second fiber may be required as illustrated in the example below.

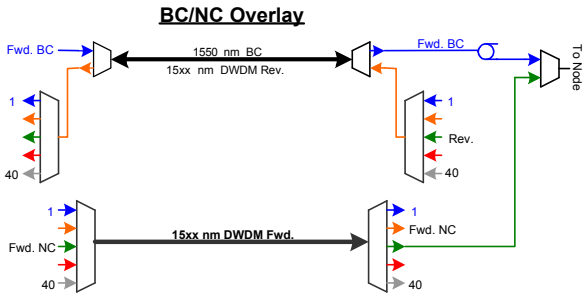


Based upon the number of wavelengths deployed upon a single fiber, this approach can cover distances of up to 30km, providing performance in the area of 50/-68/-60 dB CCN/CTB/CSO. This solution may also permit the ability to re-purpose existing fibers for other business applications.

BC/NC Overlay

A bit more complex than the two previous solutions, this solution offers advantages in networks requiring longer reach as optical amplification is possible. Generally, a two fiber solution, it consists of a single broadcast wavelength typically operating in the 54-550MHz pass band placed on one fiber which also can accommodate upstream
CWDM or DWDM wavelengths.

A second narrowcast fiber is used to transport up to 40 wavelengths typically used for QAM signals in the operating pass band of 550-1002MHz.



Distance, channel loading and performance requirements dictate whether a single or dual downstream optical receiver is required. This solution is also well suited for applications requiring optical path redundancy. Reach of up to 80+ km. are possible producing performance in the area of (49/51)/-66/-66 dB CCN/CTB/CSO.

Which Solution is Best?

Applications vary and so too does the answer to this question. A network analysis is generally required to determine the best fit and in certain instances, more that one solution may be required. Some generic guidelines however, may be helpful in determining where to begin.

A starting point would be to identify existing node sizes in the plant and determine the ultimate node size desired. This is determined by network operator bandwidth requirements based upon service offerings. Once this goal is established, dividing the existing node size by the new desired node size establishes a node reduction factor.

This factor, when considered with the fiber counts to the existing node, the distances required to be covered, and link performance goals, enables a high level selection of which multi-wavelength technologies are most applicable. In some circumstances it may be wiser to just utilize spare fibers if available or convert an existing two fiber solution to a 2X WDM solution (1 DS and 1 US wavelength per fiber).

Ongoing network analysis seems to indicate that the E-CWDM solution will become a dominant short to mid-range tool in the HFC network bandwidth expansion tool kit.

Once this analysis is accomplished, and a few options are selected, it becomes time to put pencil to paper and validate the chosen solution on the network design. At this time, additional decisions may be made to provision for additional future levels of migration should a staged approach over time be desired.

Conclusion

There are many factors to consider when deploying multi-wavelength solutions for node segmentation and fiber deep applications in order to increase network bandwidth.

This paper only presents a high level discussion of some of the technology and options available. Much more detailed analysis of which solution(s) make the most sense for a particular application is required to establish a rational migration strategy.

It is important to note that many operators approach the need to migrate their optical networks as an all or nothing proposition, basing their strategy and CAPEX requirements on an entire network optical migration. In reality, the migration process can and should take place in a phased approach addressing those areas of immediate or impending node congestion and deferring migration of those less endangered nodes to some point in the future if and when needed.

Numerous tools are evolving to expand HFC networks in order to provide increased bandwidth. The tools are growing and are of great interest to the cable industry.

Deployment of these various technologies and architectures can only help in the battle against the competitive forces that threaten the current market.

SCALING MOUNT EVEREST: DELIVERING MULTI-SCREEN VIDEO IN AN 'INFINITE CONTENT' WORLD

John Pickens, Chief Technical Strategist VCNBU
Cisco Systems, Inc.
Sree Kotay, Chief Software Architect
Comcast

Abstract

The consumption paradigm for TV is rapidly changing from pure broadcast to time-shifted unicast. This behavioral model is the driver for the new formula, "Cached Unicast equals Multicast". Supporting this trend is the rapid evolution of the network paradigm from a classic siloed broadcast dominated spectrum to a shared spectrum with converged usage of IP transport for all applications including video. The long range vision is tens of thousands of channels, hundreds of millions of assets, and orders of magnitudes more content producers – all delivered to the device of the consumer's choosing. This paper identifies key characteristics of the next generation solution architecture, such as real time enabled cache distribution hierarchies, in order to deliver an infinite world of content and unlimited scale of subscribers and consumption modalities, while delivering many of the economic benefits of today's architectures.

OVERVIEW

The increasingly rapid user adoption of time shift TV, new HD content (requiring multi-carry) and interactive video services [like video on demand (VOD)], coupled with the exploding popularity of blogging and audio/video podcasting, along with higher delivery data rates (e.g. DOCSIS 3.0) and two-way connectivity, requires a revolution in service delivery for media content. The initiative of Switched Digital Video (SDV) for linear video channel delivery is an early recognition of the emerging paradigm of long tail consumption and niche programming in the core TV market. Time shift TV (even popular linear video becomes unicast), the growing libraries of high quality commercial video (movies, original cable shows, made-for-TV, and straight-to-video) and user generated content is accelerating this paradigm shift, thereby stretching the limits of existing multicast and pitcher/catcher video delivery systems to be competitive.

The formula of "cached unicast = multicast", as embodied by Content Distribution Networks (CDNs) like Akamai, becomes more and more desirable as usage patterns change and different device types proliferate. This proposed shift enables the video delivery system to deliver an extreme scale of available assets, including multiple formats, rates, and resolutions for the same asset, with little economic or operational impact. However, traditional Internet CDNs lack the proper control semantics (e.g. you never need to "rewind" a web page), and scale of solution (latency, throughput and cost scalability).

This paper identifies the next generation technologies and paradigms in real-time media delivery that enable cable operators to migrate to this new world. Points highlighted include a massively scaleable authoritative storage network, transition to more distributed architecture designed for media, dynamic caching in the interior and at the edge of segmented content, and n-screen enabling application paradigm, where resource management and authorization enforcement is built into the next-generation network (NGN) media delivery infrastructure.

VIDEO CONSUMPTION

Consumer video consumption behaviors are undergoing a paradigm change, encapsulated by the concepts of time shifting, place shifting, and device shifting. This phenomenon is portrayed within Figure 1.

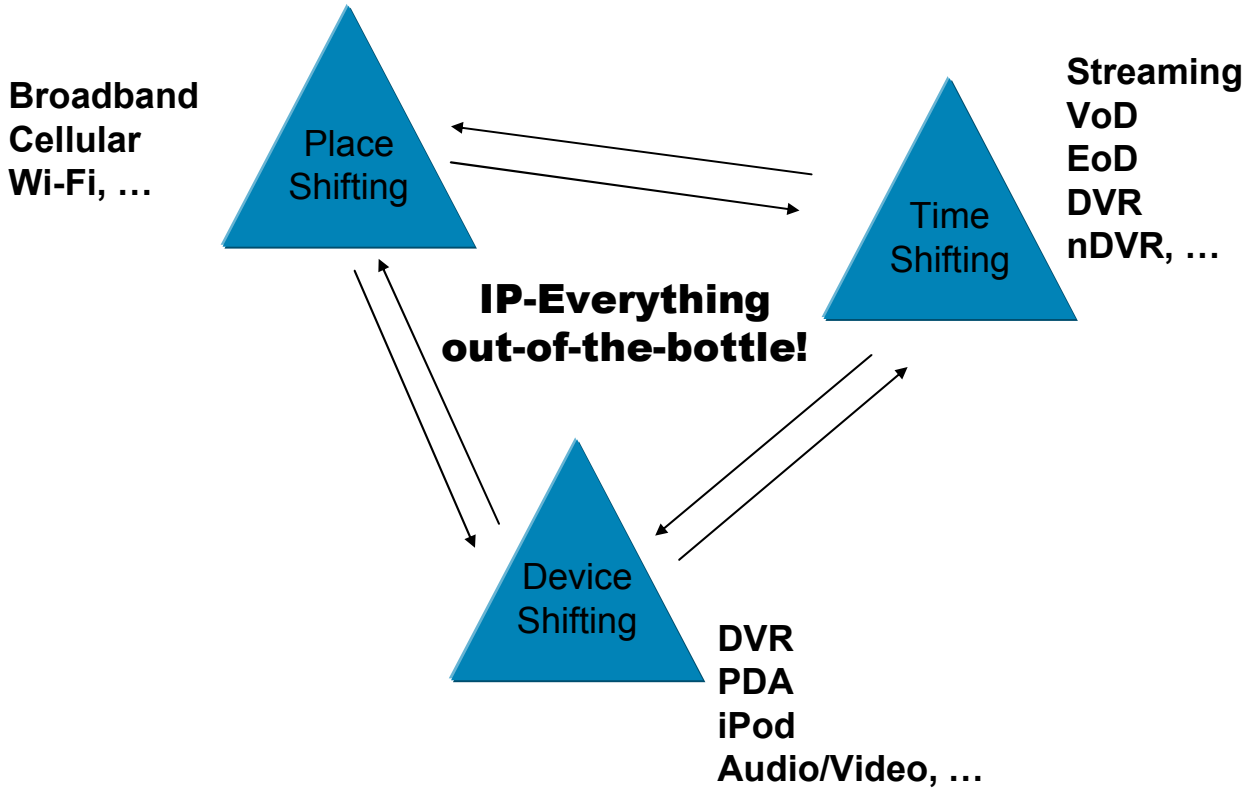


Figure 1 – Consumption Paradigms

For TV, the time shifting phenomenon gained mainstream acceptance with the widespread adoption of digital video recording (DVR). Though a large percentage of homes do not yet have DVR, and those that have a DVR do not have it for all TVs (or a media center), a high percentage of consumers are now very familiar and accustomed to DVR time shift consumption, whether in their own home or in homes of family and friends.

Place shifting is a new emerging trend increasingly being promoted by service providers, typified by "multi-room DVR" and "whole home VoD", where an individual may choose a content to view on one device, then pause the content (bookmark), and then migrate to another device and resume consumption. TV (HD downstairs, SD upstairs), PC (in office or in hotel), and mobile (while riding the shuttle to the airport) are three well known device types for converged consumption of media.

Device shifting is an increasingly discussed paradigm where, in addition to streaming content from the network, it is possible to download the content to different devices, and view the content on those devices. MP3/MP4 players, mobile phone with storage, and laptop PCs are three examples of such devices.

Use cases demonstrating the combination of all three follows. Within the home the content is consumed on an HDTV, paused, and then

resumed upstairs on an SD TV. Alternatively, the subscriber may travel to a hotel and prefer to view the content on his/her PC. Another example includes a subscriber in a limousine/car who prefers to view it on his/her mobile device. Or the content might be downloaded to his/her laptop and he/she views it on the airplane. In all these cases the content is resumed from wherever it was paused or bookmarked, independent of the type of device or location of consumption.

All three consumption paradigms can already be seen in the internet web browsing model for video consumption. The community of Internet users familiar with these models has grown to a staggering numbers, with over 10 billion unique video views consumed monthly, with YouTube accounting for approximately 30 percent of that number.

Also driving the change in user experience is the explosive growth of HD content and exponential expansion in the number of content producers – a consumption feedback cycle highlighted by the early trends of blogging and podcasting, now extending to video.

All these paradigm shifts require a revolution in mechanisms for enabling service delivery of media content, because traditional multicast pub/sub models lack the cost and operational scalability to compete effectively.

Subscriber Quality of Experience

The TV consumption experience of users served by content within the "broadcast" network is significantly different and higher quality that that of today's users who are served by internet or mobile services.

In the internet model (e.g., YouTube) the consumer has been conditioned to accept lower quality consumption experiences. Experiential examples include long latencies while waiting for the picture to display after the start of streaming, the inability to seamlessly transition into trick mode behavior, forward or rewind, and experiencing random display pauses while repairing under-runs of the elasticity buffer in the PC.

For TV quality consumption, by contrast, the user experience delivered by the network is expected to be extremely high quality. An example is the requirement for low latency (subsecond) delay from stream event to stream action. Examples of stream events are stream start, trick modes (fast forward, rewind), and interactivity (e.g. pause → pause-ad).

Subscriber Infinite Content World

The content universe is growing. Whereas a typical library size for Video on Demand (VoD) used to be a few thousands of hours, it is now targeted to be much larger, on the order of hundreds of thousands of hours and eventually millions of hours. [1]. In early 2007 [2,3] Netflix announced an online library of 70,000 titles. Now the estimated library size is well over 90,000 titles, and a high percentage of newly added titles are HD Blu-ray format reflecting the popularity of high definition programming. Comcast in January 2008 announced plans for Project Infinity to grow the On Demand library to 6000 titles (3000 in HD) in 2009, with that number expected to scale dramatically thereafter [4].

One of the key design differences between today's video delivery systems and those of tomorrow is the split between content discovery (asset metadata, availability, and associated information) and content distribution (the physical movement of the asset from source to consumer.

As the number of available assets grows, it becomes untenable (and undesirable) economically and operationally to scale edge

capacity against the number of assets. Instead, edge capacity must scale against the number of ***unique*** assets consumed. This design criterion demands a separation of data flow from media flow.

An interesting number foreshadowing future content volume growth pertains to the amount of user-generated video content on the Internet. While the quality is not as good as professionally produced content, its growth is explosively accelerating. Based upon unpublished monitoring done by search companies, in early 2007 the estimated number of titles was in the order of 40,000,000. By the beginning of 2008 the number of titles had grown to around 120,000,000.

In addition, operators are beginning to offer managed services that enable users to generate their own content and make it broadly available either downloaded online or as part of user-generated channels.

Increasingly, professional content producers are opening up their content archives to consumers, both directly (called over-the-top) and via managed relationships with service providers (assured quality of experience).

IP NGN VIDEO ARCHITECTURE

Three key initiatives for achieving a video enabled IP NGN architecture are defined. First is a series of infrastructure convergence initiatives required in order to increase the diversity of content delivery and user consumption experiences. Second is a real time enabled caching architecture for content distribution. Third is a transformation to make the content format, place, and device independent in order to deliver n-screen delivery.

Convergence Initiative

In order to deliver the universe of infinite content, and assure the DVR-like experience from within the network, a number of convergence initiatives are underway.

Perhaps most enabling convergence activity is the rapid evolution to a wideband all-IP infrastructure. This transition was foreshadowed in the Video-QAM universe by the migration toward IP enabled QAMs (IP to QAMs, traditional MPEG to home). Switched Digital Video (SDV) [5] for linear video channel delivery was the next step recognizing of the need to rapidly evolve infrastructure in order to free up bandwidth for next generation services. The key insight of SDV, versus traditional broadcast to the home, is that ***it is desirable to scale content against consumption, instead of against the total corpus of availability***. It is now accelerating with the evolution toward DOCSIS 3.0 (wideband all the way to the home) and universal QAMs, which allow service channel sharing across VOD, high speed data, and video services.

As currently portrayed in Figure 2, a DOCSIS enabled wideband infrastructure will enable 6 Gbps aggregate IP enabled spectrum downstream – competitive with other service providers – on a 950 MHz plant.

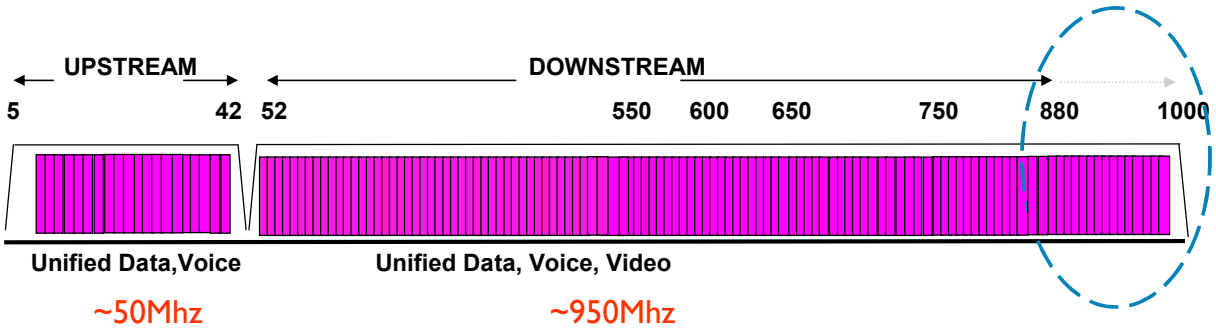


Figure 2 – 6Mbps DOCSIS convergence

A second enabling initiative is the continuing effort to reduce fiber node size. Two driving factors are service enablement and the competitive need to offer higher broadband bitrates.

For service enablement, even at 250 homes passed with 100 percent subscriber penetration, a 6 Gbps infrastructure can serve 750 MPEG-4 HD streams (8Mbps per stream) – 3HD streams per home. The reality is that subscriber penetration is less than 100 percent, and all TVs are not HD, and thus sufficient bandwidth exists even at higher HHP ratios.

For addressing competition a motivation for reduction of fiber node size is the need to further increase peak bandwidth offered per subscriber. For example, optical fiber technologies such as EPON are migrating from today's 2 Gbps:1 Gbps:32 to tomorrow's 10 Gbps:10 Gbps:32 (ratio of down:up:homes). A 6 Gbps DOCSIS® downstream is already higher than the 2 Gbps downstream offered in EPON architectures today and is well in the league of Ethernet Passive Optical Network (EPON) 10 Gbps downstream architectures. The only significant difference between DOCSIS and EPON will be the number of homes sharing the bandwidth, and the amount of spectrum offered for DOCSIS® enabled converged IP delivery.

Other convergence initiatives not addressed in this paper include bandwidth management, metadata, standard advertising interfaces, digital rights management, real time streaming protocol, Digital Living Network Alliance/Universal Plug and Play, conditional access systems, etc. Ultimately these issues need to be addressed as challenges abound. Unlike video services of the past, new services must be delivered to all types of devices in myriad locations -- with high quality and to massive scale. A new architectural approach is needed.

Real Time Caching Initiative

Web caching is a well understood and widely deployed paradigm which features the transient storage of web objects such as HTML documents for subsequent retrieval. Caching enables reduced bandwidth consumption, reduced load on servers with the authoritative storage of content, and reduced interactive latency. Overall it increases the user quality of experience, and reduces network infrastructure cost. [6, 7]

Web caching can be deployed in a variety of modes, from client, to proxy, to arrays of front ending servers. In this paper we focus on caches placed within the network.

Delivery of real time video via caching has similar benefits to delivery of web objects via caching. Consumption characteristics exhibit a Zipf curve phenomenon [8] where more popular content (e.g., a show now playing, though time shifted) is viewed by more people. The first person to consume a video causes it to be downloaded from the authoritative source into cache, and the next person who consumes the video accesses it from cache. No subsequent network transport is consumed upstream of the cache, and the access latency is shorter (by a few hundred milliseconds in worst case). [9]

Given the rapid migration from real time consumption to time shifted consumption, and the existence of the real time caching function, the benefits of caching derive similar benefits to multicast distribution at the edge, with the difference that consumers no longer need to consume content at the same real time timeline. The characteristic that one copy of the content (first user) is distributed across the backbone toward the edge is like multicast. The characteristic that subsequent consumers of the content generate no backbone traffic is also like multicast.

Three significant differences between web caching and real time video caching are identified in this paper. First the bandwidth and size consumed by "objects" is substantially higher. For MPEG-2 HD, the average bandwidth is about 15 Mbps – though it is reduced to approximately 8 Mbps for MPEG-4/AVC. Furthermore, the size of objects (the sum of all object segments) can be in the n*Gigabyte size range. Second the service level expectation of the consumer is higher than it is for web content. The jitter requirement in real time content delivery is much smaller than it is for web services and requires that the consumer experience no visible artifacts or delays. Third, there are multiple correlated object segments being delivered in real time video consumption – the 1x media stream, multiple fast forward renditions of the media stream, and multiple rewind renditions of the media stream.

Therefore additional characteristics are required within the cache delivery infrastructure for video. These are outlined below.

Tiered Hiearchies

A caching hierarchy is defined for real time content distribution. Figure 3 highlights several possible configurations. The number of tiers deployed is arbitrary – it can be minimal depending on the consumption characteristics of devices downstream (e.g. number of subscribers signed up for video service). As consumption demand grows, additional cache storage can be deployed either in parallel or in hierarchies in order to manage the tradeoffs between concurrent usage and latency and resource consumption.

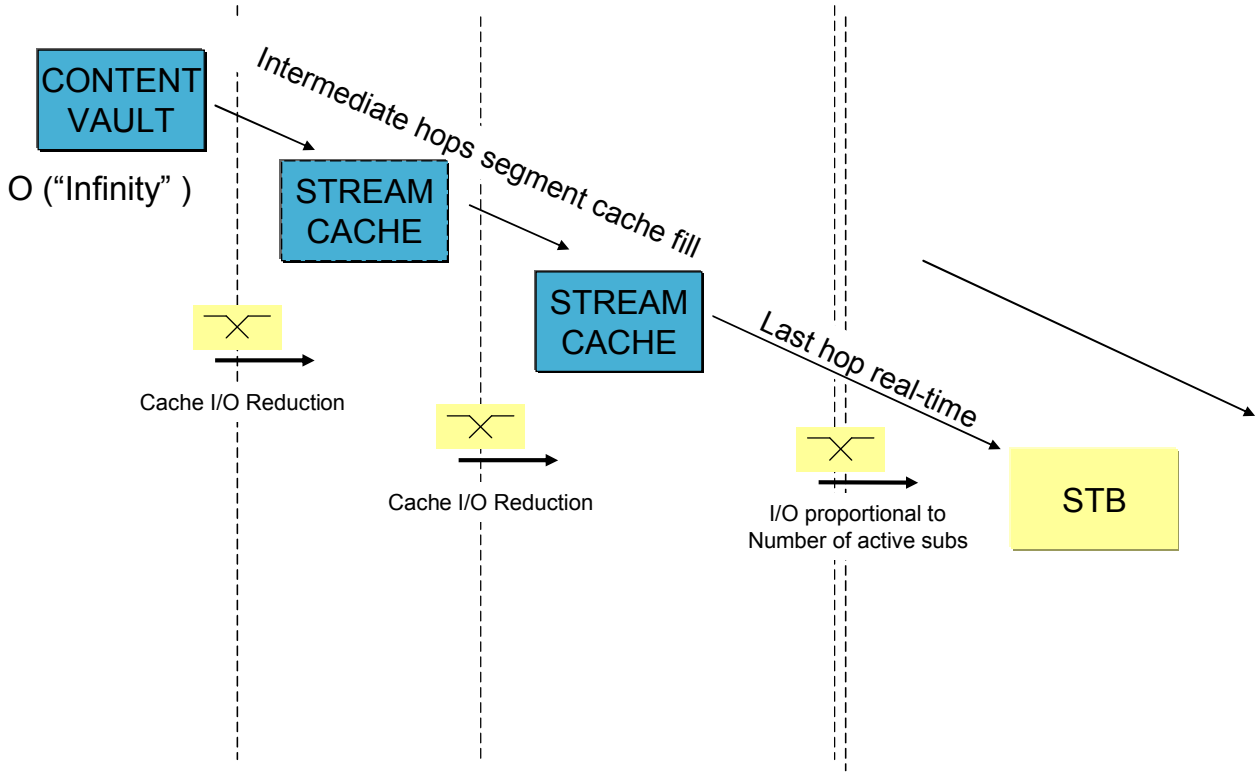


Figure 3 – Real time Caching Hierarchy

At each tier of the hierarchy the latency introduced is minimal – O(n*10ms). Also the ability to transition between consumption modes (1x, n*FF, n*REW, pause) with low latency – O(250ms) is enabled. A key requirement is that, with each of the transition modes, the awareness of frame by frame semantics of the content type is necessary so that the video segments within cache storage can be accurately managed.

Pull Versus Push Distribution

Caching object distribution protocols can be divided into two categories.

The pull category typified by web caching depends on the client that is issuing the requests and intelligently managing the transfer operations. No awareness of server state is communicated to the client. Usually no awareness of the bitrate of the content is communicated either (though it can be communicated by a separate control path).

In the push category, (typified by MPEG transport streams sent over either UDP or RTP transport) the content source is bandwidth aware, and maintains the rate of transfer in order to meet the bandwidth delivery characteristics of the content. The push model avoids significant bidirectional overhead (other than adaptation to network resource constraints) and enables assurance of stream rate for real time content objects.

Either style can be utilized. The pull category requires new mechanisms that assure that the real time caching servers (potentially different servers) being contacted for unpredictable mid-stream distribution operations have ways of learning that the previous operation has been canceled by the downstream cache, and initiate the abort procedure. The push category benefits from this mechanism since communication of client state already exists.

Segmented Object Distribution

Traditional video distribution models exhibit the characteristic that the entire video object must be distributed to the entity that streams the video toward the client. This generates several systemic deficiencies. First is that a significant delay is incurred while awaiting distribution. Second is that the percentage of content consumed is less than 100 percent (especially for long-tail where segments, e.g. famous scenes, are of primary interest). By distributing the entire object the cache is utilized in a non-optimized manner.

Therefore one of the characteristics of TV cache distribution is that object segments are transferred, on demand, if the correlated content segments are not already cached locally. This has the advantage of optimizing bandwidth consumption and cache storage consumption.

Another advantage of segmented object distribution is that it enables new services such as remixing, where arbitrary segments of content can be remixed into a new virtual asset. An example is all the goal shots of the world famous soccer star Pele combined into one segment. This can be achieved without distributing the dozens of entire full-game video objects to real time caching servers. Only the relevant scenes need to be real time cache filled. The object granularity needs to have the ability to identify frame level semantics in all cases of segmented object distribution. Methods for learning and communicating such semantics range from control plane extensions identifying offsets to embedded descriptors highlighted by standards such as TV Anytime [10].

Correlated Object Caching

The functionality of transitioning from 1x content to rewind and fast forward modes of consumption highlights another feature for real time caching that is not present in web caching. This is the capability to transition to whatever object distribution mode is being consumed by the client, assuming that the new object segments from the new mode are not yet cached on the caching entity. It should be noted that this is based on client behaviors driven by operations in the control protocol, e.g., RTSP.

In order to deliver correlated object caching, some structure, such as an indexing database, needs to be conveyed between the authoritative source and downstream caching entities so that all cache entities have accurate awareness of the object content segments contained within the cache storage.

Static-Object Verses Dynamic-Object

Two different types of objects are to be distributed by the caching infrastructure. The first category, here called static-object, pertains to a content item that has been completely ingested into the authoritative source prior to distribution towards the streaming server client. This is typically called VoD, but is not constrained to VoD objects. In this paradigm all ingest and other processing of the content object is completed prior to initiation of cache distribution. In one use case this object is not identified as available to clients until full ingest is complete.

The second category, here called dynamic-object, is a type of object that is dynamically created and ingested by the authoritative source, and is concurrently distributed into the caching infrastructure. In this paradigm the authoritative source is concurrently performing processing on the object (e.g., computing trick files, if required) and making the object available for concurrent distribution toward the destination. One well known use case for the dynamic-object paradigm is time-shift of linear content.

It should be noted that dynamic-object types have an impact on functionality of the correlated object caching indexing database, i.e., dynamic updates concurrent with ingest by the authoritative server.

Source & Sink State Synchronization

The web service caching model exhibits a lack of state synchronization between the client and the server. Each side estimates the projected behavior of the other side. Neither side is aware of any average or instantaneous bottlenecks or constraints of the other side. The real time cache fill protocol should support a method of communicating instantaneous load and state change of both source and client.

One example of state synchronization is the awareness of bandwidth. Each source and sink has a finite aggregate I/O bandwidth limit. Examples of such bandwidth constraints are on-board bus bandwidth, bandwidth to associated storage, and bandwidth between memory and adapters. The real time object caching service should exhibit bilateral awareness of I/O constraints of the source and sink so that unnecessarily high latencies or jitter behaviors are not introduced.

Ingest Overrun Avoidance

Each caching node in the distribution path from the authoritative source to the client has finite bandwidth ingest constraints. Content distribution must not overrun the ingest bandwidth with the aggregate maximum number of active session. This implies the ability to maintain tight tolerances on smooth delivery of the stream, and avoidance of unnecessary bandwidth bursts.

Network Bandwidth Optimization

The path from the content source to downstream caches has finite bandwidth constraints. The caching protocol must be designed so as not to induce either packet loss or excessive buffering jitter in the aggregate number of streams being concurrently delivered.

Opportunistic Resource Utilization

The cache fill protocol should be aware of resources of the source and sink, and also be capable of adapting transfer behavior in response to dynamically changing resource behaviors. If for example, the source, sink, and intervening path are lightly loaded from the perspective of resources, then an optimization is to enable content to be transferred at higher rates opportunistically. Such transfers are not directly correlated to the actual play out state of the content with respect to the subscriber.

Elasticity Assurance

The real time cache fill protocol should also optimize management of cache fill buffer elasticity, while maintaining a short maximum latency for stream event transitions. This requires a distribution mode where content is initially transferred at a rate higher than stream rate, and then, after a short time window, settles down to transfer at stream rate. The transition to higher rates occurs at any point that new content is transferred and short streaming startup latency is required. Examples include session start, splice points (different content objects), interactive transitions to new content, and trick mode transitions (also where different content is transferred). The elasticity buffer accommodates reasonably bounded jitter behavior and retransmission of dropped packets without disrupting streaming behavior to the subscriber.

N-Screen Initiative

A system architecture for enabling N-screen delivery is required. Key enabling characteristics include decoupling of the awareness of delivery infrastructure from the

application layer, and embedding all distribution and resource management into the delivery infrastructure.

Because the model is (a) inherently a "cache-on-demand" model, (b) separates delivery from metadata, and (c) enables "real time ingest" from external storage, sparsely populated media consumption formats (format transcoding) may be generated on-demand, or opportunistically. As with the demands on the central storage systems themselves, this load scales only with the unique assets being consumed, not with the number of streams being watched.

The characteristics of the application layer are expected to be like web services in nature. Figure 4 shows a sample configuration. Subscriber interfaces will be provided for navigation, business logic (purchases, rentals), service configuration, entitlements, etc. Each device type will have control and transport interfaces that are specific to the device, but which are not seen by the application layer.

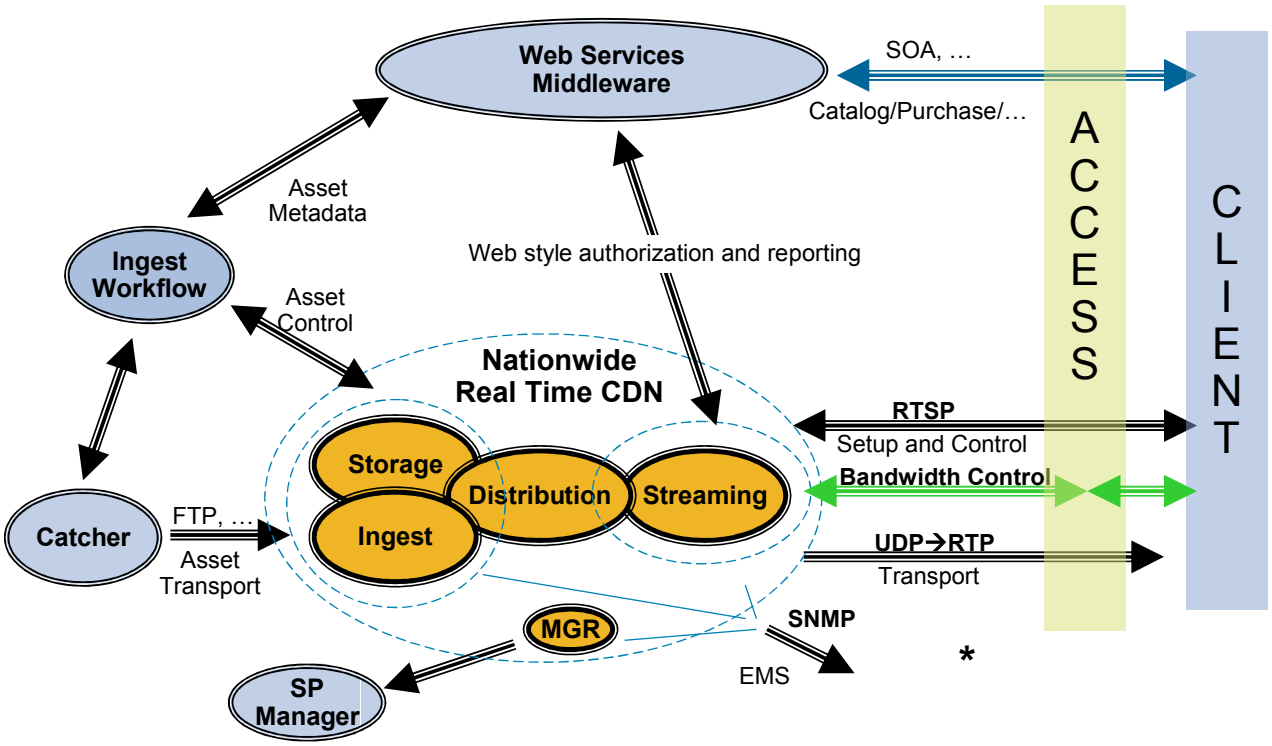


Figure 4 – "n-screen" Convergence

Real Time Streaming Protocol (RTSP) is the dominant control plane signaling for session management. Capabilities discovery and on-path session resource management will be utilized to identify the appropriate version (encoding resolution, bit rate, codec type) of an asset to stream to the device, and is appropriate for more advanced media control. Systems should also expect to provide simplified HTTP semantics (with potentially degraded performance and feature characteristics) to enable the broadest class of device consumption.

The Web services style of interfaces can be defined to allow all streaming infrastructures to consult the authoritative business logic of the application layer. This logic should be authorization oriented, not authentication oriented, as the criteria for playback may be user-centric (commercial entitlements or user

sharing permissions) or publisher-centric (rights management or web availability).

CONCLUSION

This paper identifies the rapid shift in user consumption behavior from traditional consume-on-broadcast-timeline (on TV only) to consume-on-subscriber-timeline with the ability to pause, rewind, and fast forward content (on any device). Infrastructure convergence toward all-IP, wideband edge network transport, and unicast enabled real time cache distribution paradigms are highlighted.

The benefit of having authoritative sources for the content (permanent library storage), as exhibited in the web object distribution model, plus the insertion of real time enabled caching servers in the path between the authoritative source and the destination client, enables the service to be scaled to an unlimited number of consumers, consuming an unlimited library of content (both on-demand and time-shifted live), while preserving the user expectation of DVR-like consumption delivered by the network.

Also highlighted is the decoupling of the application layer from the real time content delivery layer. No specific protocols are detailed in this paper. The primary focus is establishing a framework for scaling to the world of infinite content and infinite number of subscribers.

References

[1] Real Time Video Services & Distributed Architectures: Irreconcilable Differences or a Marriage Made in Heaven, John R. Pickens, SCTE 2006

[2] Netflix 2007 press release http://www.netflix.com/MediaCenter?id=5384

[3] Netflix Current selection - http://www.netflix.com/BrowseSelection

[4] Comcast Project Infinity press release 2008 CES - http://www.comcast.com/ces/infinity_hd.aspx?section=hd

[5] An Open Architecture for Switched Digital Services in HFC Networks, Luis Rovira, Lorenzo Bombelli, SCTE 2006 Conference on Emerging Technologies.

[6] Web Cache Wiki article - http://en.wikipedia.org/wiki/Web_cache

[7] A Survey of Web Caching Schemes for the Internet, Jia Wang, ACM SIGCOMM Computer Communication Review, Volume 29 , Issue 5 (October 1999)

[8] ZIPF Wiki article - http://en.wikipedia.org/wiki/Zipf's_law

[9] *VOD Servers - Equations and Solutions*, Glen Hardin, W. Paul Sherer, NCTA 2005

[10] Metadata - the role of the TV-Anytime specification, Morecraft, C. Storage and Home Networks Seminar, 2004. The IEEE, Volume , Issue , 3 Nov. 2004.

SYSTEM OVERVIEW AND TECHNICAL DATA RESULTS AND ANALYSIS FROM HBO'S FIELD TEST OF DVB-S2 AND MPEG-4 HD DEPLOYMENT

Andrew Levine
Home Box Office, Inc.

Abstract

On June 12, 2007, Home Box Office announced it would make all 26 HBO and Cinemax channels available to HBO distributors in high definition using MPEG-4 compression technology by the end of the second quarter of 2008. The HBO engineering team was tasked with finding a system that would meet HBO's aggressive time frame for deployment. The system would have to be capable of high quality encoding, and robust enough to be able to meet all of HBO's technical requirements. Taking advantage of the newest MPEG-4 (AVC) compression and DVB-S2 satellite modulation, HBO felt it would be able to deploy an efficient and cost effective way to make the 26 HD feeds available.

This paper will describe the technical architecture of the MPEG-4 compression system that HBO has chosen to implement. It will also outline the field test plan that HBO developed for the system, as well as the results of that field testing. The field test results and observations will show that this system is a viable alternative to traditional MPEG-2 compression with QPSK modulation.

INTRODUCTION

In order to meet the aggressive time line for deployment, HBO needed to move rapidly in its selection of a new system. HBO solicited equipment from several vendors for testing at the HBO satellite uplink facility. Once the equipment selection was made, an engineering field test plan was written and implemented so "real world" scenarios could be simulated and observed. The engineering field test provided valuable information and validated the performance of the new system.

SYSTEM SUMMARY

HBO will have three C-band, 36 MHz transponders carrying the 26 HD MPEG-4 channels.

HBO selected Motorola as its MPEG-4 encoding, multiplexer, encryption and satellite modulation vendor. HBO tested several AVC encoders from several vendors, most of which produced good results, but ultimately Motorola was chosen for its ability to provide a complete system including MPEG-4 satellite receivers that would be available for HBO affiliates in early 2008. HBO has a successful history with Motorola as its provider of analog encryption (VideoCipher) and MPEG-2 compression and encryption (DigiCipher II) equipment.

HBO will employ Motorola branded Modulus encoders to perform the MPEG-4 compression.

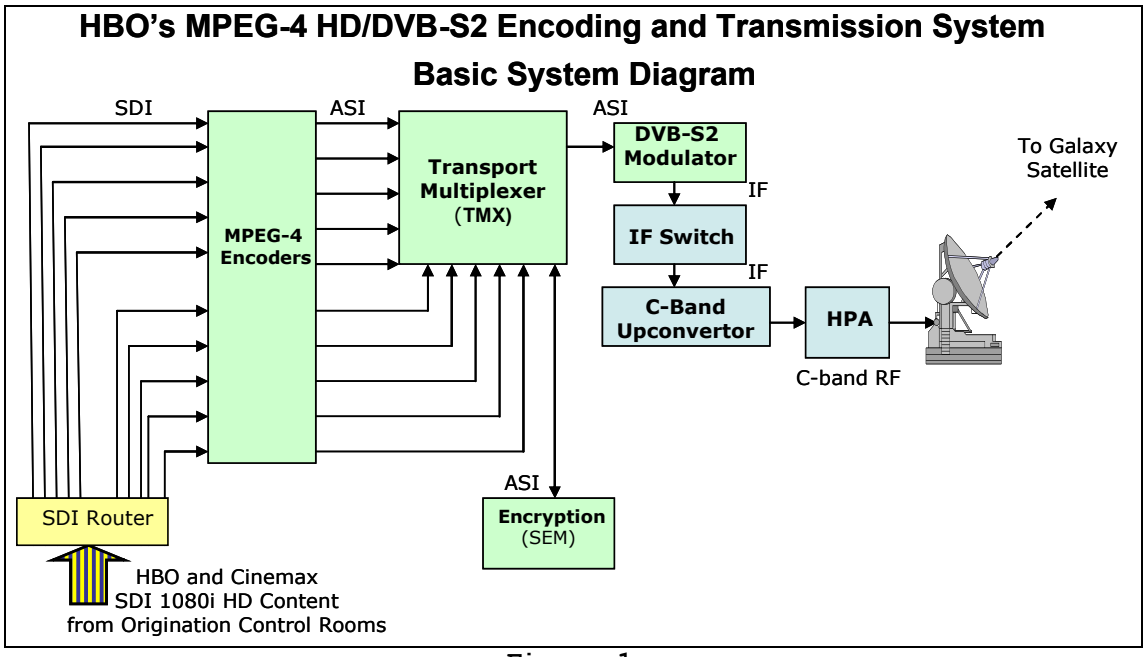


Figure 1

SYSTEM ARCHITECTURE

The MPEG-4 encoding system comprises one rack-unit encoders for each channel of video. The system is very scalable and space efficient. The single channel encoders are easily cascaded and muxed. An identical redundant (back-up) system is configured for each transponder multiplex. HBO has decided to encode the MPEG-4 channels initially at 8 Mbps and determined that this encode rate was the "sweet spot" since quality is very important in addition to optimizing transponder bandwidth as much as possible. It should be noted that 8 Mbps is about half the bit rate of typical broadcast and cable MPEG-2 HD program feeds. With future encoder improvements and MPEG-4 tool enhancements, it may be possible to encode HD lower with no loss in quality. All the MPEG-4 HD services have a single AC-3 English audio program encoded with the video. The audio is 2.0 stereo or 5.1 surround depending on the original source material.

HBO's originated HD programming will feed the MPEG-4 encoders (via SDI) which outputs an ASI stream to feed a Transport Multiplexer (TMX). The TMX muxes the individual ASI streams together, and outputs a single MPEG-4 ASI steam which feeds a SmartStream Encryptor Modulator (SEM). The SEM will encrypt the ASI stream and the TMX will feed ASI to a Newtec DVB-S2 satellite modulator. The last link in the MPEG-4 compression chain is the modulator output to an IF switcher. The IF switcher provides a 70 MHz output which feeds HBO's C-band satellite upconvertors and transmitters. The switcher also provides IF monitoring points prior to satellite uplink.

HBO has chosen to use DVB-S2 satellite modulation. DVB-S2 modulators increase spectral efficiency by using new and advanced high-level coding techniques.

As defined by the DVB organization's DVB-S2 fact sheet, DVB-S2 makes use of the most current modulation and coding techniques to deliver performance that comes close to the Shannon limit, the theoretical maximum

information transfer rate in a channel for a given noise level. DVB-S2 uses a very powerful FEC scheme which is a key factor in allowing the achievement of excellent performance in the presence of high levels of noise and interference. The FEC system is based on the concatenation of BCH (Bose-Chaudhuri-Hocquengham) with LDPC (Low Density Parity Check) inner coding. [1]

The DVB-S2 modulation scheme is not yet widely used in the U.S for satellite distribution feeds. It is, however, quite popular in Europe and Asia and has been proven to be successful domestically with contribution feed users. Extensive testing performed by HBO and Motorola with different FEC (forward error correction) rates led to very positive results. Using a FEC rate of 5/6 yields an available transponder payload of 72 Mbps. Using more aggressive FEC rates vs. required power, the modulation could be traded off to yield more throughput, but 72 Mbps met HBO's payload requirements. 72 Mbps is almost double the transponder payload that most providers currently have with MPEG-2 compression and common QPSK modulation.

HBO utilizes Motorola's Broadcast Network Controller (BNC) software as the interface to control and monitor the modular system, including the TMX and SEM units. The BNC software communicates with HBO's current in-house custom authorization
system, so that no major changes or additional software applications are needed to authorize the new MPEG-4 services. Because no major modifications will be necessary, the HBO authorization hotline will be able to easily authorize services and input pertinent affiliate data just as they currently do for HBO's MPEG-2 linear and HBO On Demand/Cinemax On Demand customers.

One of the most important of HBO's requirements was to have high quality satellite receivers available by early 2008 so that all of HBO's distributors could downlink and process the new HD feeds without having to integrate additional and/or complex gear to their existing headend infrastructure. Working closely with HBO's engineers, Motorola was able to develop two receiver solutions to complete their AVC system.

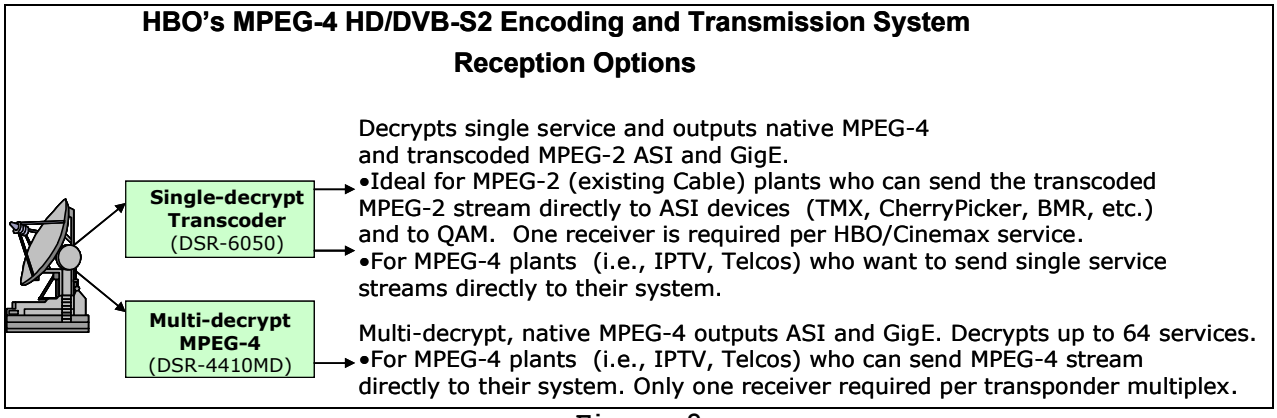


Figure 2

The first is a native MPEG-4 multi-decrypt receiver, the DSR-4410MD. This unit is capable of decrypting up to 64 services using a single one rack-unit box.

The receiver will decrypt all streams on one of HBO's MPEG-4 transponders and output MPEG-4 via ASI and/or Gigabit Ethernet. This will be beneficial to newer MPEG-4 plants who

will be able to send the native MPEG-4 channels through their system without the need for conversions or additional processing equipment. HBO has conducted extensive field testing of this receiver and DVB-S2 modulation with several MSO headends and affiliate labs. The results have been excellent in all parameters, including successful reception links with low power levels in a wide variety of geographical areas and varying weather conditions.

Motorola will also make available an MPEG-4 – to – MPEG-2 single channel transcoding IRD, the DSR-6050, which will output MPEG-4 and/or MPEG-2 via ASI and/or Gigabit Ethernet. This unit will be ideal for conventional cable distributors who have extensive MPEG-2 plants and have numerous field deployed MPEG-2 set tops. This will allow existing HBO/Cinemax affiliates to be able to immediately take advantage of the new HD MPEG-4 channels.

The IRD will decode the HBO MPEG-4 stream, and re-encode it to MPEG-2 all within a single rack-unit device. HBO will set the transcoded MPEG-2 output bit rate to ensure appropriate quality.

MPEG-4 FIELD TEST PLAN

HBO developed a four-phase test plan that encompassed approximately eight weeks. MPEG-4 receivers were sent to 18 test locations consisting of various cable and telco operators and vendor labs. The test locations were chosen based on their ability to process and report on the MPEG-4 signals. It was important to HBO to select a diverse group of test sites (both in location and technical infrastructure), as this would represent several "real world" scenarios which would simulate conditions at HBO's various distributors. HBO specified test site technical requirements that included the following:

1) 3.5m diameter or larger antenna (or equivalent gain and noise temp)
2) C-Band Digital PLL LNB,
Recommended Specifications:
a) Noise Temp 20°K or better
b) Avg gain 60dB
c) LO Stability ±12 kHz or better
d) Extremely low phase noise

HBO provided a testing matrix spreadsheet to all test sites. This matrix document allowed the test sites to record their findings/measurements during the various test phases. Weekly conference calls were held with the test sites, and e-mail alerts were generated to brief test sites on any new developments and/or procedures.

Phase I "Out of Box" Experience and Baseline Measurements. (two week duration)

The first phase of the field test consisted of install, set-up and reporting on the "out of the box" experience. Test sites were instructed to configure the receivers as per the Motorola Operator Guide using satellite transmission and service parameters that were provided in HBO's test schedule. Test sites were instructed to call HBO's authorization hotline (just as they would in a real world scenario) once the receiver was set up and had signal lock on the test satellite coordinates. Test sites were asked to provide anecdotal feedback to HBO in regard to ease of installation and configuration. Once confirmed authorized for service, the test sites were asked to record the following initial baseline measurements:

1) Satellite C+N/No at L-Band (if available)
2) L-Band input signal level at IDR/MRD RF Port
3) Front Panel IRD LED indicators, any alarms present
4) Status Menu/OSD Diagnostics
   a) Signal Strength
   b) "Signal Quality"

c) Eb/No (Energy per bit to noise power spectral density ratio)
d) VCT (Virtual Channel Table number)
e) VCN (Virtual Channel Number)
f) Authorization Status
g) BER (Bit Error Rate)

5) Initial ASI or GbE stream analysis (high level)

Phase II Transmission Performance (one week duration)

The second phase of the field test allowed HBO to vary uplink parameters while the test sites recorded changes to signal lock and Eb/No levels on their receivers. This testing helped HBO to determine the optimal transmission parameters for best real-world link performance with minimal sacrifice of data throughput, while maintaining appropriate downlink margins. HBO used an FEC modulation rate of 5/6.

Specifically, HBO attenuated the uplink transmitter power by -1dB increments every 10 minutes until -6dB was reached. At that point the power was brought back up in -1dB increments every 10 minutes until full power was restored. The total test duration was 80 minutes. The field test sites were asked to record the receiver's Eb/No level at the 10 minute intervals, and also note if the receiver lost signal lock at any time. The identical test was repeated at the same time the next day to ensure all test sites were able to observe the receivers.

Eb/No (or E_b/N_o) is defined as Energy per bit to noise power spectral density ratio, and is considered the measure of signal to noise for a digital satellite/communication or data system. It is normally measured at the input to the satellite receiver and is used as the basic measure of how strong the signal is. It is a normalized signal to noise ration (SNR) measure, also known as the "SNR per bit". It is especially useful when comparing the bit error rate (BER) performance of different digital modulation without taking bandwidth into account. [2]

Phase III Decoder Output Interfaces (one week duration)

In this phase of the testing, field sites were asked to measure, observe and document the various outputs of the decoders in the HBO test schedule. This included taking empirical measurements and recording the responses of terminal equipment (subject to test site capability) including adherence to the following standards as applicable:

- DVB ETSI TR 101-290
- ETSI TR 102-034
- IEEE 488.1-1987
- ITU-T H.264
- ATIS
- SMPTE-292M (HD-SDI Output via external decoder)
- SCTE 40, SCTE 20 and other applicable standards

Phase IV Long Term Stability (four week duration)

HBO transmitted continuous program services to the field test equipment at the test sites during this phase of testing. Test sites were asked to perform qualitative observations of the decoded services at the video and audio outputs of a set top box or other decoding devices (e.g., Sencore/Motorola MRD-3187). Test sites were asked to observe and record (at regular intervals) the picture and aural quality of the signals. Test sites were requested to maintain reception of "existing" MPEG-2 feeds from HBO and compare the feeds "under test" to the existing feeds. No visual differences should be observed in either the decoded MPEG-4 or the transcoded MPEG-2 output. The following items would be monitored:

- Eb/No levels should be steady/consistent. Any variations should be documented.
- Receiver signal status should be "locked." Any variations should be documented.
- Integrity and any irregularities of ASI output stream
- Decoded Picture quality (color, motion, detail, artifacts, disturbances, etc.)
- Decoded Aural quality (levels, dynamics, response, distortion, separation)
- Lip Sync within one frame (30-mSec)
- Closed Captioning (CEA-608 and CEA-708)

Test Completion

At the end of Phase IV, HBO collected and compiled all spreadsheet data as well as all of the field sites' subjective comments and opinions. After several weeks of test data analysis and interpretation, HBO's engineering team made their recommendations to Motorola as to the success of the field-deployed receivers performance.

MPEG-4 FIELD TEST RESULTS

Phase I "Out of Box" Experience and Baseline Measurements

Out of the 18 MPEG-4 (DSR-4410MD) receivers delivered to the test sites, one receiver was DOA (would not power up) and had to be replaced.

Phase I revealed to HBO that there were some "growing pains" to be expected with this new technology. Several sites were not familiar with the MPEG-4 receivers and how to set them up and obtain transponder signal lock. HBO's engineering team and Motorola's IRD product team addressed these issues directly with the test sites. One example of this was whether to set the multi-decrypt MPEG-4 receivers to manual or automatic mode. The difference being, the ACP (Access Control Processor) addresses would be manually assigned by the end user or automatically populated when the anchor (or "in care of") unit address number was authorized. HBO found that it's easier to have the end user set the receiver to automatic mode. This works especially well when the end user is receiving all the channels on the same transponder multiplex.

Another issue (affecting the MPEG-4 receivers) discovered during Phase I testing was a setup step not clearly explained in the operator guide. It seems that after selecting an ACP number, the menu curser must be moved to the corresponding program number and entered by using the up/down arrows for each ACP number that is used. If the program number is not entered, the receiver will not authorize that ACP.

Despite these minor issues, it seemed to be fairly easy for the test sites to set up and have their receivers authorized for the MPEG-4 signals. The following chart includes the receiver's measured signal quality and Eb/No for the 11 field test sites that submitted Phase I data.

DSR-4410 Phase I Results Summary Data

Test Site	Antenna Size	Date	L-Band Sig. Level at RF Input Port	Front Panel LED Signal Lock	Front Panel LED Auth.	Signal Quality	Eb/No
A	4.5m	10/29/07	not available	OK	OK	94	11.6
B	5m	11/1/07	-29.6 dBm	OK	OK	98	14.3
C	7m	10/30/07	-15.6 dBm	OK	OK	99	14.2
D	5m	10/22/07	-47.1 dBm	OK	OK	98	14.8
E	9m	10/25/07	-38.6 dBm	OK	OK	100	10.9
F	4.5m	10/16/07	not available	OK	OK	100	15.0
G	3.8m	10/24/07	-25.3 dBm	OK	OK	92	10.7
H	5m	10/25/07	-22.8 dBm	OK	OK	93	11.2
I	3.8m	10/24/07	-48.2 dBm	OK	OK	100	12.4
J	5m	10/19/07	-47.0 dBm	OK	OK	95	12.4
K	5m	11/5/07	-20.8 dBm	OK	OK	100	14.2
Avg.*	**5.24m**		**-32.8 dBm**			**97.2**	**12.9**

Figure 3

*The calculated averages in figure 3 indicate that with a standard (or common) headend antenna, excellent downlink performance is achieved.

Phase II Transmission Performance

Phase II yielded excellent satellite link results. As the following graphs (figures 4 and 5) will illustrate, 12 test sites were able to participate in the uplink attenuation phase. Considering the different geographical locations and different downlink antenna types and sizes, the average Eb/No at full uplink power was 12.1. When the uplink power was attenuated by -6dB all but three sites still had signal lock with an average Eb/No of 7.35. So, 75% of the downlink sites were able to keep signal lock on the transponder at -6dB power attenuation. This proves the strength of the DVB-S2 modulation. Except in the very rare instance of local weather being unusually extreme and/or the HBO satellite link experiencing a major transmission problem, it would be extremely uncommon for real world conditions to come close to -6dB of uplink attenuation.

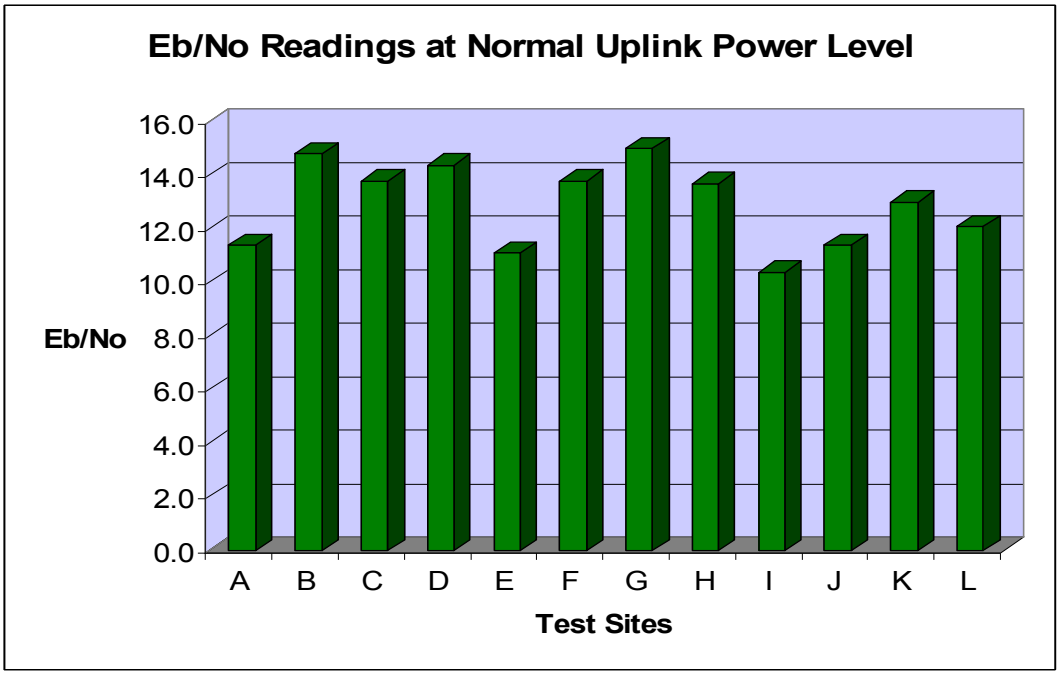


Figure 4

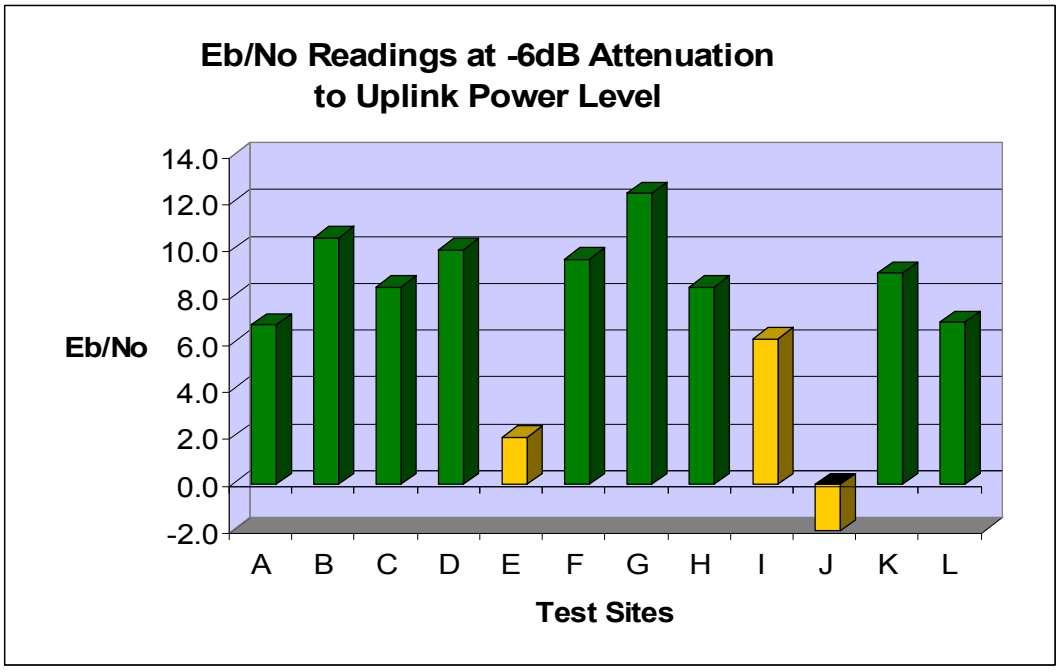


Figure 5

As illustrated in figure 5, only the three test site locations indicated in yellow E, I and J (Eb/No 2.0, 6.2 and -2.0) lost signal lock at-6dB uplink attenuation. The other nine sites were able to maintain transponder signal lock. Site I tested with a small antenna, and this validates that nominal downlink sites can receive DVB-S2 signals with margin.

Phase III Decoder Output Interfaces

In this Phase, the field sites were asked to observe the output quality of the MPEG-4 receivers. Since the test sites all had different

types of measurement equipment (i.e., stream analyzers, etc.), the findings in this phase are somewhat subjective. This actually benefited HBO in determining how the receiver's outputs would perform based on non-consistent readings and stream observations that would vary from site to site. These results again depict “real world" results, as no two headends are exactly the same.

Many of the test sites were not able to look at the MPEG-4 output due to the constraints of having an all MPEG-2 (typical cable plant) infrastructure, but several could and found the outputs to be without any issues or anomalies. Some sites were able to utilize next-generation set top boxes that indicates MPEG-4 technology is rapidly becoming available to cable providers.

The GigE and ASI outputs produced solid MPEG-4 streams. At the time of this testing HBO was broadcasting six HD channels with a total bit rate of about 46 Mbps. One test site fed the ASI output to an active Terayon CherryPicker and logged zero errors. Another site observed legal value video and normal audio as they decoded the ASI stream using a Sencore/Motorola MRD-3187.

The successful Phase III testing indicted the MPEG-4 receiver would be easily deployable in plants that had an MPEG-4 infrastructure, such as IPTV Telco or in plants that would want to employ an external decoding device such as the Sencore/Motorola MRD-3187, or MPEG-4 cable set tops as they become available.

Phase IV Long Term Stability

As with Phase III, many test sites were unable to monitor actual audio and video output of the MPEG-4 receiver, but those that were able to, provided feedback that indicted the receiver was stable and ready for deployment. Some test sites simply observed the Eb/No, signal lock and authorization status during the Phase IV testing.

Eb/No and signal quality readings remained steady as observed for several weeks at different times of day with varying local weather conditions. The observed overall bit rate was consistent with what the HBO uplink was broadcasting for the duration of the testing. All HBO required signal attributes (including closed captions, parental controls/ratings and 5.1 audio) were observed to successfully pass after decoding the MPEG-4 signal. No major MPEG artifacts (including blocking or freezing) were observed during the long term testing. Audio and video quality were said to be very good and there were no instances of receiver failure or need to re-boot or power cycle the units

Transcoding IRD Testing

Directly following the successful testing of the MPEG-4 receiver, HBO (in coordination with Motorola) shipped another receiver to all the field test sites. This unit is a single-decrypt MPEG-4 – to - MPEG-2 transcoding IRD (DSR-6050). This new receiver will enable the end user to provide a transcoded MPEG-2 ASI signal that can be easily put into service in existing MPEG-2 cable plants without the need for any additional equipment. The receiver decodes the MPEG-4 signal and then re-encodes it to MEPG-2. This IRD will output both the native MPEG-4 and the transcoded MPEG-2. The HBO uplink (utilizing Motorola’s BNC software) can control the transcoded output data rate to insure the best possible quality for each HBO/Cinemax service.

The test plan for the transcoding IRD was abbreviated to three phases, with Phase II (uplink attenuation) being eliminated. It was HBO’s feeling that there was no need to duplicate the RF test, since the MPEG-4 receiver performed extremely well. The front end (and RF) portion of the DSR-6050 is identical to the DSR-4410MD, which is based

on Motorola's popular DSR-4400MD (MPEG-2) receiver.

The initial pre-production version of the DSR-6050 had some stability and operational "bugs" which caused the unit to sometimes not enable the MPEG-2 ASI output and also cause the IRD to intermittently "freeze up", requiring a power recycle and/or factory reset. The initial units also did not pass digital closed captions or 5.1 Dolby surround audio. With feedback from HBO and the field test sites, Motorola addressed these problems and issued a firmware upgrade. New DSR-6050s (with the updated firmware code) were sent to the field test sites for evaluation.

Data from the field sites indicates that Motorola successfully corrected all stability and operational issues with the 6050 transcoders. Closed captions and Dolby 5.1 audio pass through the transcoded ASI output without consequence. Many of the test sites reported the overall (video/audio) quality of the transcoded (MPEG-2) output is as good or in some opinions superior to HBO's native MPEG-2 signal. As of this writing, Motorola expects to put the DSR-6050 into full production and make it available to cable distributors by April 2008.

Field Test Conclusions

- The RF transmission link (utilizing DVB-S2 modulation) is extremely stable.
- The Forward Error Correction (FEC) Rate of 5/6 has been chosen and will yield a transponder data payload of 72 Mbps. This FEC rate is sufficient for HBO's MPEG-4 data requirements and is also quite suitable for successful headend reception performance with an average size downlink antenna, and will not cause any downlink reception issues.
- The DSR-4410MD performed well with no major issues and should be recommended to HBO distributors who wish to receive the new HD services with a native MPEG-4 output format.
- Stability and operational issues with the DSR-6050 transcoder have been addressed, and updated (production) firmware proves this receiver should be recommended to HBO cable distributors. This will allow cable MSOs to take advantage of HBO's additional HD channels, without concern about having to alter any existing cable plant infrastructure or incurring the expense of purchasing additional processing equipment.

FINAL THOUGHTS

As HD television is becoming more popular and the related consumer equipment costs are dropping, consumers have a desire for more HD content. HBO has led the charge to increase HD offerings to its distributors and their subscribers. Implementing the Motorola MPEG-4/DVB-S2 solution will address the demand for more HD content by providing a high quality, scalable and cost effective method of providing more HD channels using HBO's existing satellite transponder inventory.

In launching this new system, HBO maintains its presence as a technology leader. HBO was the first programmer to use satellite distribution for its cable affiliates and was the first programmer to use encryption and digital compression. HBO also launched the first satellite-delivered MPEG-2 HD channel and now will be the first programmer to offer all of its channels in HD using advanced DVB-S2 modulation and MPEG-4 compression.

Contact: andy.levine@hbo.com

NOTES

[1] DVB Fact Sheet, August 2007, http://www.dvb.org/technology/fact_sheets/

[2] http://en.wikipedia.org/wiki/EbNo

THE BENEFITS AND CHALLENGES OF DEPLOYING LARGE REGIONAL VOD ASSET LIBRARIES.

Michael W. Pasquinilli, Vice President of Engineering
Sunil Nakrani, Research Scientist
Jaya Devabhaktuni, Systems Architect
Concurrent Computer Corporation

Abstract

Many of the domestic video-on-demand (VOD) systems in service today are being upgraded to ten thousand hours of storage. Much of this storage is for traditional "on-demand" assets. There is also an industry trend towards recording an increasing amount of broadcast content onto the VOD system.

The cost of deploying these very large encode, ingest and storage libraries into each VOD system may prevent the launch of these new services. This paper discusses the economical benefits and technical challenges of introducing regional asset libraries that can support multiple VOD systems. The relationship between network bandwidth and asset caching will also be explored.

OVERVIEW

The current baseline storage level for VOD deployments is ten thousand hours for many of the major domestic MSOs. Ten thousand hours of content, including overhead, equates to approximately 24 terabytes of storage. Some of the more aggressive broadcast models have up to three hundred channels of broadcast programming recorded into the VOD system and retained for up to two weeks. If one hundred percent of this content were retained, the VOD system would require approximately 246 terabytes of storage. Even assuming that the cable operator is able to secure contractual rights for only twenty percent of this content, the storage requirements are still in the 50 terabyte range. (This does not take into consideration high definition (HD) content. These storage values could be three to four times larger for a system with all HD assets.)

Today there is a one-to-one relationship between the VOD system and the digital set-top box control system. The expansion of digital subscriber penetration is causing the digital set-top box systems to fragment into several mirrored digital video systems each with their own dedicated VOD system. This means that a single cable system today with one VOD system and one digital set-top box controller may soon split into three or four mirrored systems as digital subscriber penetration increases. With the current VOD architecture, each of these new digital video systems will require their own VOD library storage.

The economical challenge facing the MSOs of deploying very large VOD asset libraries in each digital set-top box system is further aggravated by the fragmentation of these single digital video systems into multiple mirrored systems. So whether the cable operator has a national footprint of many sites, or is a large single system operator, it is likely that with the current VOD architectures in place today duplication of the VOD asset library will be necessary.

To address this challenge several engineers in the cable industry are working towards developing a shared or "regional" VOD asset library that can serve multiple VOD systems. One of the major differences between the VOD library serving a local VOD system and a regional VOD library is that a local VOD library serves a closed network with dedicated network

resources. In this case the local VOD library may "play" or "stream" the asset directly to the subscriber. However, in a regional VOD library, the network is likely shared with other data and is several Ethernet switches away from the subscriber. Issues such as QoS (quality of service), packet jitter, packet routing and trick mode latency make it less reasonable to expect a remote VOD library to stream across very large distances. In this case the regional VOD library must copy the asset to the local VOD system for play out.

The design concept discussed in this paper applies the hybrid VOD model in use today between the headend VOD library and hub VOD edge cache devices to a higher level in the VOD architecture. Now instead of a local headend VOD library, the VOD library is a regional library and, instead of hub VOD cache, the entire VOD system served by the regional library equates to the hub cache.

THE BENEFIT

There is clearly an economical benefit in capital savings if a single 100-300 terabyte VOD asset library could serve multiple VOD systems. Just the storage, ingest and streaming costs for such a system would be well over a million dollars. This does not take into account the costs in encoders, MPEG grooming, control systems, powering, cooling and operations.

The primary component that could undermine the economic benefit of a regional VOD library is the network cost to transport the video to each VOD system. This is where the hybrid VOD library architecture model comes into play. The primary benefit of a hybrid VOD library architecture is the savings in headend to hub transport costs. The challenge is to prove that this same savings could be applied to the regional VOD library.

THE COMPUTER MODEL

Since there are no regional VOD libraries in service in today's domestic cable market, it was not possible to gather measured data from an actual regional library. As an alternative, engineers at Concurrent developed a computer model that would simulate the operation of a regional VOD library. Real-world measured VOD asset usage data were used to exercise the model.

The Variables

The model had the following adjustable variables:

System Cache: The amount of memory available at the local VOD system to cache content pulled from the regional VOD library.

Time To Live (TTL): The amount of time that a downloaded library asset resides on the local VOD system cache before deletion.

Cache Management: A Least Recently Used (LRU) methodology was used to manage the local VOD system cache.

Network Bandwidth: This was the bandwidth assigned to each asset being pulled from the regional VOD library. For the sake of simplicity, a value of 3.75Mbps was used as a baseline minimum data rate for each asset pull. (It was assumed that all assets were MPEG 2 standard definition.) Multiples of this data rate were used to simulate a "best effort" data rate model. (It is more likely that a best effort methodology would be applied to this variable in an actual deployment.)

The model assumes that 100% of the VOD assets would initially be delivered to the regional VOD library. It was also assumed that the first copy of an asset delivered to the local VOD system would be cached locally. Assets would

only be purged from the local system if the allocated cache value was exceeded. The TTL of an asset would be directly tied to the active usage of that asset by subscribers. If it was necessary to delete an asset from the local VOD system cache, the least recently used asset would be deleted to make room for the next requested VOD asset. No cached assets would be deleted that were in active use.

It was possible in the model to have denial of service. If the local VOD cache was full of assets being actively played by subscribers, additional requests for regional VOD library plays were denied. Since the objective of the model was to discover the network bandwidth necessary to support the regional VOD library, no restrictions were placed on the network bandwidth between the regional VOD library and each of the local VOD systems.

In several runs of the model the value of the local VOD system cache was changed. We were looking for the amount of local VOD system cache that would allow for no denial of service and have the largest impact on reducing the network bandwidth between the regional VOD library and the local VOD system.

For the purpose of this simulation, actual VOD asset usage data was collected from three large regionally co-located systems that have network connectivity via the MSO's internet backbone. Figure 1 shows a simplistic block diagram of the configuration.

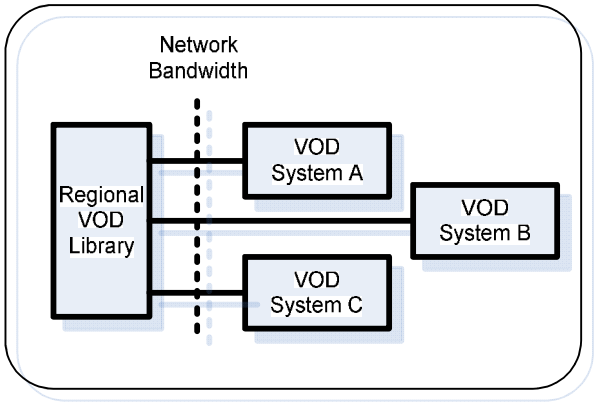


Figure 1. "Three VOD System Model"

The Results

The first run of the model assumed that each individual asset transfer rate from the regional library to the local cache was fixed at 3.75Mbps. No cap was placed on the network bandwidth between the regional library and the local VOD system. The local VOD system cache value started at 250GB and was incremented by 250GB until no Denial of Service (DoS) was encountered due to lack of local VOD system cache storage. Results of the first run are shown in Table 1 below.

System	Cache @ 0% DoS	Peak BW	Avg BW
A	2.25TB	1.67Gbps	0.61Gbps
B	3.0TB	2.47Gbps	0.83Gbps
C	1.0TB	0.90Gbps	0.32Gbps

Table 1. "First Run At 3.75Mbps/Asset"

In the next run all other variables were kept constant, but the asset transfer rate was increased to 15Mbps per asset. Results of the second run are shown in Table 2 below.

System	Cache @ 0% DoS	Peak BW	Avg BW
A	2.25TB	3.7Gbps	0.645Gbps
B	3.0TB	4.4Gbps	0.870Gbps
C	1.0TB	1.4Gbps	0.335Gbps

Table 2. "Second Run At 15Mbps/Asset"

Analysis

Although System C had the highest number of VOD streams during the data collection cycle, and was very near the stream value of System B, it required the least amount of local VOD system cache to store assets pulled from the regional library. This was due to the following: In System B the top 20% of assets accounted for 83% of the views. In System C the top 20% of assets accounted for 88.6% of the views.

This shows that the usage patterns of the subscribers in the system can have a direct and measurable impact on both the local cache and network bandwidth necessary to support a regional VOD library. This is shown to be true even when the two systems have identical content offerings and very similar overall stream usage.

Another important item to note from this data is that a four times increase of the individual asset data transfer rate resulted in a negligible increase in the average data rate and a less than double increase in the peak data rate. This points to the benefits of allowing a best effort transfer of the library assets up to the ingest data rate of the local VOD system. Using the local caching method resulted in a combined peak network bandwidth of only 9.41Gbps. If no local caching were used and instead each asset was streamed directly from the regional library, the combined peak network bandwidth required to support these three VOD systems would be 102Gbps.

TECHNICAL CHALLENGES

The technical challenges facing us today are not in fielding these large VOD asset libraries. We have demonstrated the ability to support large storage systems with high ingest rates and very low encoded content throughput latency. We can see from the data above that both the network bandwidth and cache storage values are manageable. The primary technical challenge facing the deployment of regional VOD libraries is in the fact that the VOD standards and architectures in place today were not designed to accommodate content not entirely under the control of the local VOD back office controller.

The following section will outline some of the major areas that are being addressed in order to support a regional VOD asset library.

Metadata Publishing

The most common way that an asset is received into a VOD system is via a satellite "catcher." These catchers pass over to the VOD system the asset and the metadata file. The VOD back office manages placement of the metadata into the back office database and the transfer of the asset into the VOD server(s). There are back office checks in place today to make sure that the metadata and assets are both accounted for in the VOD system.

With the regional library assets, the ingest point for the asset is at the regional library. Therefore there must be another entity outside of the local back office that is keeping track of the metadata for the assets stored on the regional library. This *other* system must then publish to the local back office the metadata of the assets stored in the regional library. This needs to be done so that the local back office can manage the provisioning, rules and lifecycle of the asset once it is within local control. Most all of the VOD set-top box navigators in use today pull their data from the local back office. This being the case, the local back office must have all the metadata for both local assets and regional library assets within its database.

The problem here is that the local back office does not currently have the means to accept and act on metadata whose associated assets are not locally ingested. (The Comcast NGOD set of specifications begins to address many of these

issues, but as a proprietary specification cannot be discussed in this paper.) A new method needs to be adopted that allows for the regional library asset metadata to be identified as representing remote content and changes must be made to the back office to recognize and manage assets not stored locally.

Trick Modes

It is most often the case that VOD trick mode files and/or indexes are created during asset ingest. This trick mode creation usually happens as the asset file moves from the catcher into the VOD system. Each VOD pump vendor has their own methods of creating trick mode files/indexes.

Since it is envisioned that a regional VOD library would support multiple VOD pump vendors, the trick mode creation would not happen at the ingest point at the regional library, but must instead happen at the time when the asset is delivered from the library to the local VOD system. Depending on how this process is done, it could result in commercially unacceptable latency for the subscriber. (An alternative approach considered is requiring all VOD vendors to adopt a common trick mode standard.)

A second consideration related to trick modes is the effect of the fast forward function. It could be possible that a subscriber tries to *fast forward* to the end of the asset faster than the file transfer speed for that asset. This is why I mentioned earlier in this paper that a best effort methodology of file transfer from the library to the local VOD system would be preferred over the fixed "play rate" of the title. Otherwise the regional library must somehow support the function of trick mode play out especially for the fast forward function.

Latency

An ideal architecture would have the subscriber oblivious to the fact that a VOD title was being served from a regional library versus a local VOD server. But this behavior is not currently guaranteed.

With proper network provisioning and QoS, the path from the regional library to the local VOD system should not contribute to the latency. However, the methods for how the local VOD system ingests and creates trick modes and propagate content will be the most likely cause of latency. Some VOD systems may not be able to play an asset until the entire asset is copied into the VOD server. This restriction may also apply to the creation of trick modes. Depending on the ingestion point into the system and the methods of content propagation, there could be queuing delays before the asset arrives at the VOD server designated for play out.

Asset Lifecycle

Today the lifecycle of a VOD asset is defined by the metadata associated with the asset. The VOD system will retain the asset for as long as is specified by the metadata. Since both the regional library and the local system would have access to the metadata, both systems can continue to use this information. However, for the purpose of managing local storage, a library asset "copy" must be marked eligible for deletion prior to the asset expiration date.

If a local system allocated one thousand hours of storage for caching library content, some content would need to be deleted at times to make room for other requested library content. The regional library may contain many tens of thousands of hours of content and not all content pulled down to the local site could necessarily be retained for the full metadata-specified viewing period. Therefore assets pulled from the regional

library must be "allowed" to be deleted to adequately manage the local cache.

Bookmarks, Active Rentals, Resume Viewing

One of the major challenges of a regional library is supporting the subscriber experience of being able to view an asset multiple times within the rental window (usually 24 hours). To understand the issue consider the following scenario: The subscriber watches all but the last ten minutes of an obscure video that was pulled down from the regional library. The subscriber returns some twenty hours later to resume viewing the remainder of his bookmarked video. Since this was an obscure video it was likely purged from the local cache to make room for more popular or recently requested assets. In this case the subscriber only wants to watch the last twenty minutes of the asset. Does the library download the entire asset again to the local cache or a partial file? Since it may be too difficult to push bookmark information all the way back to the regional library, and it may be too difficult to manage file fragments within the local cache, it is likely that the entire asset must be copied again to the local system.

Ad Zones

The largest regional VOD libraries are likely to be those made up of broadcast content supporting "Look Back" and other network DVR types of services. Ideally the cable operator would like to record just one copy of a regional broadcast channel. Ad zones make this difficult. In our three system examples, each of these systems may have had four or more local advertisement insertion zones. One method of dealing with ad zones is to record one copy of a broadcast for each zone into the VOD system. These broadcast recordings would start to add up as they are multiplied by the number of channels that have local advertisement opportunities, times the number of ad zones, times the number of systems supported by the regional VOD library.

An alternative way of dealing with ad zones is to insert the local advertisement at the point of play out. In our case this would be at the local VOD system's VOD server. Since VOD provides a dedicated session to each subscriber on demand, there is an opportunity to target ads at the subscriber level or to at the very least keep the ad zones intact.

CONCLUSION

Based on the computer model and the measured VOD asset usage data from just three VOD systems it would seem that a LRU managed local VOD system cache of about 1,000 hours per system and total network bandwidth of at least 5Gbps (peak) per system would be a good starting point to support a regional VOD asset library. This does not take into consideration the impact of high definition content on these variables. Also, this specific model does not include "Start Over" type content usage.

It is important to reiterate that a change in subscriber usage patterns can significantly change the resources necessary to support a regional library. This being the case, computer modeling that utilizes actual measured asset usage results will be more important in defining resource requirements than will anecdotal or historical experiences.

There are quite a few technical challenges to be addressed before a regional VOD library is commercially viable and transparent to the subscriber. These challenges represent changes to the VOD back office, VOD system and VOD server. When all of these challenges have been met the VOD system will have evolved into a video delivery platform that will satisfy all of the on-demand and broadcast needs of the MSO at the national, regional and local level.

THE COX NATIONAL BACKBONE: BUILDING A SCALABLE OPTICAL NETWORK FOR FUTURE APPLICATIONS AND NETWORK EVOLUTION

Dan Estes, Cox Communications
Gaylord Hart, Infinera

Abstract

Cox Communications has recently begun building out a national DWDM optical backbone which will run over a combination of owned and leased dark fiber spanning over 12,000 miles. While Cox could have leased transport capacity from a national carrier for this purpose, the build vs. leased capacity analysis showed the higher costs of leased transport would not be economical in the future. In the final analysis, the business case for building and operating this network was based primarily on the rapid growth of Cox's cable modem data services alone. However, this network provides additional incremental economic benefits by allowing cost savings elsewhere in the network (e.g., by building consolidated national headends) and by enabling new revenue generating service opportunities not traditionally addressed by cable operators (e.g., a national footprint for commercial services or cell service backhaul).

Two paramount design considerations for the network were total cost of ownership and service reliability. Other important considerations were network scalability, network flexibility, and the ability to rapidly turn up new bandwidth and services. To meet these requirements, Cox is implementing its national backbone with digital ROADMs (reconfigurable optical add/drop multiplexers).

INTRODUCTION

Cox Communications undertook this project for two main reasons: costs and scalability. Since 2001, with the demise of the @Home consortium, Cox had leased intercity transport from a variety of interexchange carriers. These costs had steadily risen as bandwidth needs increased and as consolidation occurred among the various carriers. The business case for this national backbone network was a classic example of build versus buy with several unknowns thrown in for fun. For example, we had to anticipate the market dynamics of the future lease alternatives while also forecasting the demand for future services that had not been clearly defined.

However, there were some very compelling "knowns." History had shown that our cable modem and business Internet traffic had doubled every twelve to eighteen months with corresponding complexity and cost increases. Speed increases for cable modem and business services were common with a corresponding increase in packets being delivered over the existing leased backbone. Backbone circuits were filling at a rapid pace with long delays in getting new links in service. To compensate for the delay in adding capacity, we were having to order new circuits at about a 65% fill point due to the long lead times that some interexchange carriers had in their order fulfillment processes.

It is interesting to note that while the costs of transporting a megabit for a mile had declined in the period from 2003 to 2007, the bandwidth needs far outpaced this declining cost rate. Figure 1, below, shows the aggregate bandwidth growth of Cox's network for the last 18 months. With demand continuing to increase, the business case was relatively straight forward.

Another driving factor in the decision to build this network was operational simplicity. In the current mode of operating, Cox had to

coordinate multiple entities in the turn-up of a new circuit. First, we would contact the interexchange carrier about three to five months before a cross section was expected to exhaust its existing bandwidth with a circuit order for a new intercity link. This was followed up with equipment orders to up to four vendors for DWDM and SONET last mile connections from the interexchange carrier's POPs to the Cox regional data centers where the circuits terminated. Circuit turn up, interexchange carrier acceptance testing, and end-to-end throughput verification would follow. The entire process would take months from inception to completion. If Cox owned its own national network, especially one in which circuits could be seamlessly engineered from a remote, centralized center, then capacity additions could be enabled quickly and with far less complexity than the leased mode.

A third driving factor in the decision to build a national backbone was one of flexibility. We had experienced unanticipated demand where we could not respond fast enough using leased circuits. One example of this was a market request to connect multiple 10 Gig circuits from a data center in one region to a colo facility in another. Because we had to wait on Type 2 carrier circuits, we missed the opportunity to serve this customer.

One side benefit that has emerged for Cox's national backbone that was not considered in the original business case is rapid disaster recovery. During the California fires in the fall of 2007, several fiber cables were burned, and capacity for Cox's San Diego and Orange County markets was diminished. The question arose as what would have been the impact had this national fiber network been in place. The answer came back that we would have been able to re-home several wavelengths with a few keystrokes that would have bypassed the damaged fiber. The modular design of the DWDM equipment we are deploying in our backbone is such that the minimum bandwidth that is installed in any cross section is 100 Gb/s. In many cross sections, this full capacity is not assigned on day one, making available some

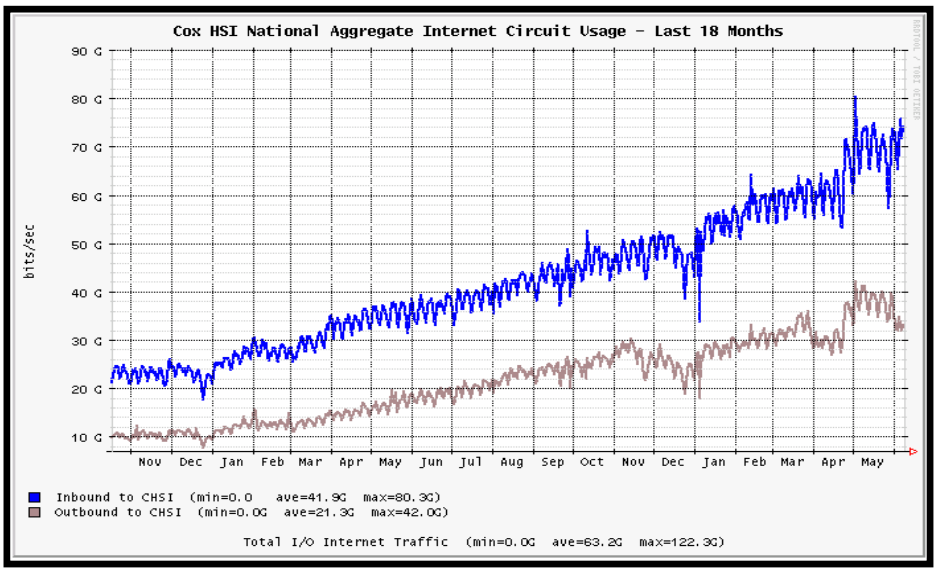


Figure 1 -- Cox Communications Aggregate Internet Bandwidth Growth

Figure 2 -- Map of Cox's Backbone Network

wavelengths that can be used for protected services or that can flexibly be assigned as needed for unusual circumstances. While each potential disaster cannot possibly be anticipated, it is good to note that some outages can be recovered from rapidly with a flexible DWDM network design that can provide end-to-end, any-to-any connectivity between all nodes on the network. The network that emerged from these design considerations is shown in Figure 2, below.

NETWORK APPLICATIONS

Cox's primary use of the national backbone will be for Internet access for our cable modem customers. We have designed our routing tables so that any Internet traffic that can be passed directly at peering centers egresses from our network at these locations. While we still have several locations around the country with Tier 1 Internet transit ports, due to the costs of delivering traffic to these portals, we easily justify the costs of building our backbone into the national and regional peering centers. We have points of presence in Palo Alto, Dallas, Atlanta, Ashburn, New York, Los Angeles, and Chicago.

Cox also uses the national backbone to deliver voice signaling and bearer traffic. We have several soft switches that are used for regional control of the VoIP endpoints. For example, we have a softswitch in Atlanta that controls the VoIP call setup for multiple markets across the country. Signaling traffic from these markets flows over the backbone to Atlanta for hundreds of thousands of calls each day. Network reliability is of the utmost importance since these calls include 911 and other potentially life threatening emergency calls. Long distance traffic also rides the backbone. We have class 4 control points such that any calls that originate in a Cox market and terminate in another Cox market is transported over the backbone's IP infrastructure and are terminated on trunking gateways in the remote endpoint. These trunking gateways not only connect these calls to Cox's local telephony network, they also connect to the local Public Switched Telephony Network. The incremental savings of providing these services over our own facilities amounts to millions of dollars a year.

A very interesting byproduct of having a national backbone is the ability to provide national distribution of high quality video content. One of the needs in a Hybrid Fiber Coax architecture is to maximize the use of the available spectrum on the coax plant. We have begun to digitize channels and distribute as many into a 6 MHz QAM channel as we can at a quality level our customers' expect. Most multisystem operators have settled on statistically multiplexed groups of 12 MPEG2

standard definition channels with a maximum of two MPEG2 high definition channels per QAM. Our marketing groups continue to ask for more channels and higher quality, especially given the marketing hype from the Direct Broadcast Satellite industry about "hundreds" of high definition channels. A technical solution to this competitive threat was needed. While MPEG4 is a possible solution, the millions of MPEG2 capable set top boxes that are in our customers' homes means that solution is still a way off in realistic deployment scenarios. A solution that could be deployed faster was needed. Technology is available today using specialized encoders and closed-loop statistical multiplexers to enhance the capacity of a 6 MHz QAM. The cost of these encoders and multiplexers is prohibitive to deploy in every headend across the country. With available bandwidth on a national backbone, the economics of a couple of centralized digital encoder headends and nationwide IP distribution of multiplexed HD content are very attractive. By centralizing the encoding to only a few locations, Cox will be able to provide hundreds of high definition choices and several hundred standard definition choices to all markets across the country. As well, the higher cost, closed loop encoders would provide a quality that is indistinguishable from the current generation of encoders in use in most headends. The national backbone enables the video feeds to be fixed routed using MPLS so that the path from any one of the national headends arrives at a receive site from two diverse paths at all times.

Cox commercial business services also benefits from this national footprint. Inter-market services such as Ethernet Private Line, High Capacity point-to-point services, and point-to-multipoint WAN services can now be offered to business customers throughout Cox's service area. Cox was already providing some lower capacity services using the leased transport capacity. With a larger capacity network of its own, Cox can offer up to 10 Gigabit per second services on a competitive basis. In partnerships with other carriers and other Local Exchange Carriers, we can serve any location in the United States.

NETWORK DESIGN CONSIDERATIONS

In the selection of the DWDM vendor and the dark fiber provider for this project, there were two overarching design considerations: first, the total cost of ownership of the network, and second, the reliability of the end-to-end services. Each market was to be designed so that no single point of failure exists in the network. This includes dual entrance facilities to each building, bypassing some locations with manhole fiber splices instead of entering the building and patching through a fiber cross connect, redundant power, and back-up power at every location where we deployed electronics. Hut spacing and the quality of the interoffice fiber also made a big difference in the overall design and costs of the network. While closely spaced re-gen sites might seem like they would increase reliability, the trade-off of increased electronics costs did not justify the extra benefits. What emerged was a balanced network that met reliability and throughput needs while maintaining reasonable hut spacing across the country. It is interesting to note the variability of hut spacing as shown in Figure 3, below.

Selection of a dark fiber vendor was pretty limited in that we wanted to use a single provider for as many cross sections as we could. This would build a good relationship with that provider as well as simplify the operational handoffs in fiber restoration and turn-up. Another criterion was the quality of the fiber itself. Significant penalties would accrue if the overall fiber types were such that extra re-gens or extraordinary dispersion compensation would be required. Figure 4, above, shows that the vast majority of fiber for the Cox national

backbone is Corning E-LEAF, the desired fiber from a design standpoint.

There were many available options for optical networking equipment and technologies, but after evaluating these options against Cox's network requirements, Cox decided to build its network with digital ROADMs. Digital ROADMs perform an optical-electrical-optical (OEO) conversion for every DWDM wavelength at every node. While this may seem expensive, modern photonic integrated circuits (PICs) have reduced this cost substantially. For reconfigurability, the OEO conversion allows the use of integrated digital electronic switches in the ROADM instead of all-optical wavelength-only switches.

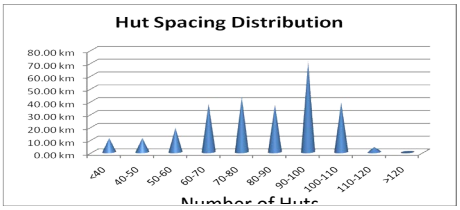


Figure 3 -- Distribution of Hut Spacing Distances

Digital ROADMs typically perform a 3R (regenerate, reshape, retime) operation at every node for every wavelength, and this significantly reduces the optical engineering complexity for digital networks built with these ROADMs. What used to be a large and complex engineering problem for deploying all-optical ROADMs, the necessity of calculating worst-case performance characteristics for a number of optical parameters for every path for every wavelength in the network, is reduced to a simple span engineering solution for digital networks. In a digital network, only individual optical spans between adjacent nodes must be engineered to guarantee any-to-any connectivity of every node in a network. Moreover, the digital ROADM's 3R OEO architecture eliminates the cumulative optical impairment limitations of all-optical ROADMs and permits unlimited node counts and network sizes and flexibly supports multi-degree mesh networks.

Network flexibility was a major engineering criteria for the Cox national backbone. Some DWDM solutions would force a custom design with Dispersion Compensation Modules engineered for the maximum point-to-point span distance anticipated at the time of the initial engineering of the route. While a network could be designed in this manner, it would lock the solution into a specific set of origination and termination points. If there was an unanticipated need, the network would have to be re-designed and possibly reconfigured. This is not a problem with a digital DWDM ROADM that provides 3R regeneration of the optical signal at every re-gen and terminal location. Network re-routing and short-term bandwidth needs can also be accommodated easily with the add/drop nature of the digital network.

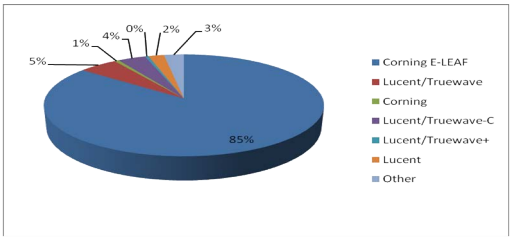


Figure 4 -- Fiber Types in the Cox Backbone

Protection alternatives were hotly debated topics during the initial design phase of the national backbone. Cox made the decision to use layer 3 protocols as the main protection mechanism in the network. While we could build a network that was completely redundant, cost analysis showed that if each wavelength were protected on a 1:1 basis that the equipment costs would increase by 80%. This was primarily due to the fact the protected service would always require an equivalent and dedicated amount of transport capacity for protection as the primary path. One of the

features of a digital ROADM, since it is not based upon transponders for transport, is that it allows you to have one wavelength protect multiple other wavelengths. In general, we have settled on a 9:1 protection scheme such that one wavelength in each cross section is available for maintenance or backup bandwidth. In some cross sections where we have through routes from one market to another that bypass the local add/drop function, we are setting aside additional protection wavelengths, again on a many-to-one protection scheme. While we are incurring some increased costs, the overall protection costs are still much lower than 1:1 protection, and in turn we have increased route survivability and provided flexible bandwidth configuration alternatives that provide operational benefits to the network.

OPERATIONAL CONSIDERATIONS

When you think of the operational implications of a national fiber optic network there are some key concepts that come to mind such as rapid capacity activation, detailed performance monitoring, specific alarm notifications that are pertinent to the problem, the ability to rapidly isolate trouble, and having positive assurance that services are being delivered as expected. Let's delve into these in more depth.

Rapid capacity activation is enabled in this network through a combination of upfront engineering and the use of digital ROADMs. Cox has taken the steps to position the chassis and re-gen sites for one year's forecasted growth. Once the chassis and common equipment such as network management cards and bandwidth multiplexers are installed, adding bandwidth capacity on the digital ROADM is relatively easy and is accomplished in 100 Gb increments by adding line cards. Rarely would any one cross section experience more growth than that in one year.

Client interfaces are only required when services are actually turned up. When additional services are needed, individual client optical modules will be sent to the two endpoints a couple of days ahead of the need. Technicians, either Cox personnel or SmartHands technicians in remote collocation environments, will slot the daughter boards as instructed in a work package. The ROADM will then signal through its data communications channel to the centralized provisioning center of the presence of the new plug-ins. Using point-and-click provisioning, electronic cross connects would then be made to pass the incoming traffic to the appropriate output port, wavelength, and timeslot. End-to-end routing of the new services is accomplished remotely via a GMPLS control plane, and no truck rolls are required to intermediate sites in the path, and no additional optical layer engineering is required. End-to-end connectivity can be established in a matter of days instead of the months it takes in a leased circuit environment.

During a turn-up event, the ability to generate a test signal from the client interface itself would be useful and would enable remote confirmation of circuit continuity even if there was not a technician present with a portable test set. This is very easily possible in a digital world. One of the benefits of a digital ROADM is that performance monitoring is provided for every wavelength at every optical to electrical conversion point. In a digital optical network, this occurs at every node. With a G.709 digital wrapper applied to each service path in the digital ROADM, bit error rate (BER) testing may be carried out prior to service turn-up using an internally generated pseudo-random bit stream or on live services in real time.

Every service interface has a corresponding set of performance statistics derived from the G.709 overhead. Appropriate threshold crossing alerts can then be set to notify the Network Operations Center of any degradation

in services as well as the specific span in the network where this degradation has occurred. From the viewpoint of maintaining a national fiber optic network, this type of visibility is paramount to providing reliable services.

Along with the performance monitoring on a span by span basis, another attribute that is necessary in a national optical network is alarm filtering and suppression. If there is a major outage in the network, there could be hundreds of alarms generated from loss of services in all of the daughter circuits that ride over the system. An intelligent Network Management System is required to filter those alarms, suppressing the non-pertinent ones. Consistent naming conventions and intelligence embedded in the alarm processing system provide information that will assist in pointing the technicians to the correct source of the problem and allow more rapid service restoration.

Finally, a wonderful feature in a digital ROADM is the ability to loop back sub-rate interfaces at any electrical conversion point in the network. This loopback capability can be applied at a re-gen site, add/drop location, or terminal location to quickly isolate intermittent trouble. It will be used in Cox to sectionalize the network between maintenance entities where we have contracted maintenance activity in certain cross sections of the country to an outside party.

NETWORK GROWTH AND EVOLUTION

If optical transport networks simply grew linearly, then network engineering and operations would be a straightforward matter. As we all know, this is rarely the case: services grow faster than expected and not always where expected, new services are introduced, new nodes need to be added, and sometimes whole cable systems are sold to or acquired from other MSOs. All this points to the necessity of engineering a network today for maximum flexibility tomorrow. Key elements of this requirement are scalability, rapid service turn-up and cutover, non-disruptive upgrades and node additions, guaranteed any-to-any connectivity between any two nodes in the network, and an optical layer that requires little or no re-engineering as the network evolves.

Digital ROADMs provide a reasonable solution to all these requirements, and this is possible because a digital ROADM, due to its OEO and digital switching architecture, segregates the optical layer from the service layer. This makes service delivery independent of optical layer engineering and brings unique capabilities to the network. At the optical layer, network design is reduced to simple span-by-span optical engineering, and once the initial span has been engineered, adding additional bandwidth is accomplished by adding line modules with no additional optical layer engineering required.

In the initial network design, one simply provides a pool of available bandwidth at the optical layer. This pool of bandwidth is then used as a resource to be allocated to services as they are turned up. Allocation is implemented at the node through the integral digital switch, and allocation occurs across the network under the control of a GMPLS control plane. This allows services to be routed through the network via any combination of nodes that have available bandwidth along the path. Providing sufficient bandwidth is available at the optical layer, it is possible to connect any two nodes in a digital network with any service.

Treating optical layer bandwidth as an allocatable resource has other benefits, as well, including bandwidth conservation. Because the digital switch in a digital ROADM mediates between the optical layer and the service layer, it can aggregate and groom multiple sub-rate services onto a single optical layer lambda. For example, multiple GigEs can be groomed onto a

single 10G wavelength. And it can do this on a service by service, wavelength by wavelength, and span by span basis, ensuring optimum usage of available bandwidth throughout the network. In a similar fashion, super-lambda services can be created by allocating one high-bandwidth service across multiple optical layer lambdas. For example, a 40G client service can be allocated across four 10G lambdas for transport at the optical layer. This allows 40G services to be transported across a network designed for 10G services, and without any network re-engineering.

Digital ROADMs, because of their inherent ability to switch services over any available bandwidth in the optical network, provide significant network migration capabilities which can be used at any time to reroute traffic through the network for load balancing or bandwidth optimization or to perform maintenance operations such as adding a new node in the network. Digital ROADMs typically provide a bridge-and-roll function that allows a second, parallel path to be created through the network between two end-nodes carrying live traffic. Once this "bridge" is created, traffic is "rolled" onto the second path in under 50 ms, and the first path is then free for re-use or can then have maintenance operations performed on it without disrupting live traffic.

The digital ROADM's ability to reroute services through the network in real time can also be used to provide flexible service protection options that eliminate the need for providing dedicated, redundant optical paths for every service on the network. Using shared protection, sufficient additional optical layer bandwidth is provided for service protection in the initial network design, and this additional capacity can be allocated on the fly to support alternate routing of any failed path. Should a fiber cut or equipment failure occur, the GMPLS control plane recognizes this and immediately and automatically reroutes traffic around the failed path. This provides a cost-effective protection mechanism without the need for dedicating 1:1 protection bandwidth that cannot then otherwise be used.

Finally, digital ROADMs permit multi-degree mesh networks to be efficiently designed and turned up. In this case, the ROADM's integral digital switch is simply used to switch service traffic onto the appropriate optical layer interface for the intended mesh path. No complex optical layer engineering is required because the additional degrees coming off a node are still treated as single spans at the optical layer. This capability allows a national network to be built with integral regional networks and thus allows regional and national traffic to be carried over a common network. The mesh nature of such a network provides more traffic routing and protection options than traditional rings.

SUMMARY AND CONCLUSION

Cox's construction of a national fiber optic based network has positioned Cox to economically provide expanded services for the residential and business communities that it serves. It has also given Cox a competitive advantage in rapidly responding to changes in offers from Incumbent Local Exchange Carriers and Direct Broadcast Satellite companies. The scalability of the Cox national backbone and the digital ROADM used to implement it will also provide Cox a way to deliver business services that have not been offered by an MSO in prior years and is an excellent complement to the Cox local "last-mile" network. Given Cox's operational excellence, this network will provide for outstanding levels of service for higher speed data offerings, expanded video lineups, and reliable voice applications, both wired and wireless, for years to come.

TIMELY AND SECURE:
REAL-TIME PERFORMANCE CHALLENGES OF CONTENT SECURITY

Reza Rassool, Chief Engineer,
Widevine® Technologies Inc.

Abstract

Encryption, authentication, and key distribution are the mainstays of digital rights management (DRM) and conditional access (CA) systems in modern entertainment networks. In the traditional DVB CA security model[1], entitlement control messages (ECM) and entitlement management messages (EMM) are inserted into an encrypted MPEG stream. These messages are received in a timely manner by a subscriber device, to enable it to access the stream data. In more modern delivery networks, watermarking, fingerprinting, and digital copy protection are additional processes that have been inserted into the pipeline to secure the business of on-line entertainment. All these security processes introduce measurable temporal distortions in bandwidth, latency, and jitter to the smooth flowing of entertainment content to subscribers. While basic real-time requirements stem from linear broadcast applications, file-based delivery imposes a new set of constraints that challenge engineers to deliver content in a secure and timely manner. File-based distribution calls for security processing that scales, persists and is faster than real-time. This paper quantifies the potential temporal distortions in the DVB CA security model, detailing the perceptible effects on channel change time, temporal jitter and latency.

INTRODUCTION

Television and movie content is, by its very nature, a temporal experience. Frames and samples are presented to us in rapid succession to give the illusion that we are observing objects in motion and listening to sound. The audio-visual illusion relies on precise timing of the delivery of each sample of content. It is this most basic illusion that must be maintained to ensure that the delivered content is received in the condition intended, and achieves its potential value in maintaining the attention of its audience. Traditional over the air broadcast networks delivered a consistent stream that did not suffer from the temporal distortions of modern packet-based networks.

User Perception of Timing

Psychologists would identify audio video timing as a *hygiene* factor. Herzberg suggests a model for human motivation[2] wherein certain essential factors are considered pre-requisites. Only once these hygiene factors are satisfied can we be motivated by other factors. Herzberg's original work concerned the motivation of employees. Since then, Herzberg's work has been applied to consumer motivation. Consumers are typically motivated to visit restaurants based on the menu rather than the quality of service. A certain level of service is a pre-requisite; in the same way viewers expect to enjoy a movie where each frame and sample is delivered on time.

It turns out that hygiene factors cannot motivate a consumer. But, their deficit can certainly demotivate, in the same way that poor service would negatively affect the enjoyment of the meal - no matter how good the menu. For the audience of this paper it is especially important to understand the hygiene factors of the digital television business. One of these is audio video timing. Surprisingly few studies have explored the area of user perception of temporal distortion in audio and video.

Distortion by Frame rate changes

The illusion of motion can be maintained at quite a low frame rate. It is surprising that utility is found in video conferencing systems operating at less than 10 fps. The content frame rate should not be confused with the display frame rate. Even though the old silent movies had 16 frames of content per second, the projectors would display each frame three times resulting in a display frame rate of 48 fps. The display frame rate is critical and is related to, but not identical to, a physiological concept called the flicker fusion threshold or flicker fusion rate. Light that is pulsating below this rate is perceived by humans as flickering; light that is pulsating above this rate is perceived as being continuous. The exact rate varies depending upon the person, their level of fatigue, the brightness of the light source, and the area of the retina that is excited. Few people perceive flicker above 75 hertz for CRT monitors. A flicker free display is a hygiene factor. The content rate and the display rate must be controlled independently. The display rate must be tightly locked to a steady clock, while the content must be delivered at a rate that was intended to maintain the illusion of motion. It is well known that content played at the wrong content frame rate adversely affects the viewer experience. Even the most dramatic Lillian Gish movies seem comical when played at 24 fps. But at the original 16 fps, the content delivers the intended impact.

Distortion by Audio frequency changes[3]

In many respects, audio timing needs to be more stringent than video timing. In the range 1kHz to 8kHz, the human ear can detect a pitch shift that results from a change of frequency as little as 0.2%. Wow and flutter is particularly audible on music with oboe or piano solo. While wow is perceived clearly as pitch variation, flutter can alter the sound of the music differently, making it sound 'cracked'. There is an interesting reason for this. A 1 kHz tone with a small amount of flutter (around 0.1%) can sound fine in an echo-free environment, but in a reverberant room constant fluctuations will often be clearly heard. [4] These are the result of the current tone 'beating' with its echo. What is heard is quite pronounced amplitude variation, to which the ear is very sensitive.[5]

Distortion by Jitter and Latency

While over-the-air broadcasts deliver isochronous streams in real-time, modern networks burst packets of data that need to be buffered in memory for variable lengths of time and need to be processed to differing extents depending upon the type of data in the packet.

Jitter is caused when a processing element in the pipeline operates in bursts. The result is that the time each data packet spends in the element, from input to output, is not constant. Even though the long term flow rate through the processing element may be constant, the instantaneous rate fluctuates. Jitter causes a problem for subsequent downstream processing elements. Either their buffers overrun due to receiving a burst of several packets, or their buffers run empty due to gaps between the bursts of data packets. Jitter is resolved by larger buffers or by ensuring that each processing element operates in a timely manner. Timestamping each packet on arrival and holding the processed packet, until a fixed period after the timestamp, before outputting it, reduces jitter to the resolution of the timestamp but introduces a fixed latency.

Buffer overrun in an element often results in packet loss, while buffer under-run requires the processing element to deploy an under-run strategy. An MPEG decoder, for instance, would repeat the previous frame at the output if the input buffer under-runs. The viewer sees a freeze frame.

Latency is a fact of life in transmission systems. Latency must be constant so that end-to-end propagation delay can be used to schedule live or real-time events. Provided that all elementary streams undergo the same latency then the content is delivered as intended. Viewers with both cable and satellite systems may notice the different end-to-end delays of each service when they switch from one to another. Buffering, introduced to smooth out jitter, adds to the end-to-end delay.

Distortion by AV Synchronization drift

Reeves and Voelker[6] reported on a Stanford University study. When audio precedes video by 5 video fields, viewers evaluate people on television more negatively (e.g. less interesting, more unpleasant, less influential, more agitated, less successful). This difference is not large, but it is statistically significant. Viewers can accurately distinguish between a television segment that is in perfect synch, and one that is 5 fields out of synch. Viewers cannot accurately tell the same segments are 2.5 fields out of synch but their evaluation of content is negatively affected.

In 2003, an ATSC Implementation Subcommittee (IS)[7] studied the issue of AV synchronization. They said that the overall audio-video synchronization error is the algebraic sum of the individual synchronization errors encountered in the chain. While a given synchronization error may cause either a positive or negative differential shift in audio video timing, the video signal is typically subjected to greater delay than the audio signal, and the tendency is therefore toward video lagging behind audio.

IS finds that under all operational situations, at the inputs to the DTV encoding devices, the sound program should be tightly synchronized to the video program. The sound program should never lead the video program by more than 15 milliseconds, and should never lag the video program by more than 45 milliseconds. In MPEG-2 the end-to-end delay from an encoder's signal input to a decoder's signal output is regarded as constant.

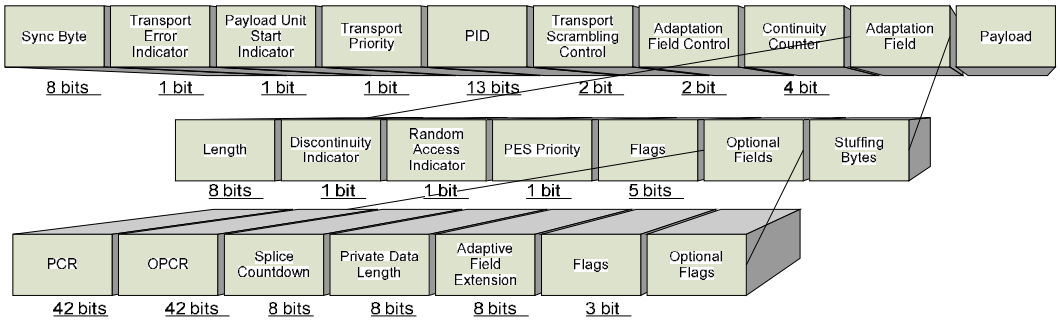


Figure 1 PCR in adaptation field of MTS header

This end-to-end delay is the sum of the delays from encoding, encoder buffering, multiplexing, transmission, de-multiplexing, decoder buffering, decoding, and presentation. Presentation time stamps are required in the MPEG bit stream at intervals not exceeding 700 milliseconds. The MPEG System Target Decoder (STD) model allows a maximum decoder buffer delay of one second. Audio and video presentation units that represent sound and pictures that are to be presented simultaneously may be separated in time within the MPEG transport stream (MTS) by as much as one second. In order to produce synchronized

output, IS finds that the receiver must recover the encoder's System Time Clock (STC) and use the Presentation Time Stamps (PTS) to present the audio-video content to the viewer with a tolerance of +/-15 milliseconds of the time indicated by PTS.

MPEG TIMING MODEL

MPEG supports timing metadata that may be inserted at encoding of the elementary streams and at packetizing of the MTS. These timestamps are read by the decoder to ensure the real-time performance of the stream. An MPEG-2 encoder includes System Time Clock (STC) as a reference time.

The system adds an STC value to the coded AV data as a time stamp for each unit of presented information, and then multiplexes the resultant data. Next, the multiplexing system inserts a reference clock so that the receiver may regenerate the STC on decoding. The receiver places each unit of coded data in a buffer to generate a delay, then decodes and presents the data unit when its time stamp matches the STC. This process corrects the temporal offset between the video and audio streams caused by multiplexing.

Timestamps, tables and their constraints

The MPEG-2 System Standard defines two types of timestamp that are added during encoding: Presentation Time Stamp (PTS), indicating time of presentation, and Decoding Time Stamp (DTS), indicating decoding start time. The multiplexed MPEG transport stream (MTS) includes a Program Clock Reference (PCR), a timestamp marked periodically in the adaptation field of the MTS header.

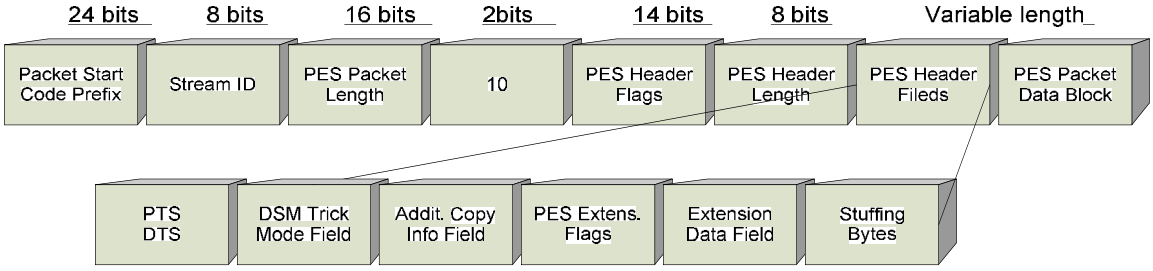


Figure 2 PTS and DTS in PES header

The PCR allows the receiver to regenerate a system time clock to match the timing of the encoding process. The PTS and DTS timestamps are sent in the PES headers. The Program Map Table (PMT) associated PIDs with program(s). The Program Association Table (PAT) associated program number with PMT. The Conditional Access Table (CAT) associated PIDs with private streams.

ATSC and DVB tighten the MPEG-2 constraints on timestamps and tables.

Clock Recovery Schemes[8]

An ideal MPEG decoder would implement a numerically-locked loop (NLL) to regenerate the 27MHz system clock from incoming PCR values. 27MHz was chosen because that is the frequency used to drive the video display electronics. Since the DVB specification requires that PCR values are inserted in the stream, at most, 40ms apart this requires the NLL to operate between 25Hz and the MTS packet frequency, 2500Hz (for a 3.75Mb/s stream).

188 byte packets are received, demuxed, and placed in the buffers according the PIDs. Each frame of MPEG data in the PID buffer contains its own timestamp. Once decoded each frame is stored in the display buffer tagged with its own PTS.

As shown in figure 3, the NLL[9] contains a 27MHz VCXO (voltage controlled crystal oscillator), a variable-frequency oscillator based on a crystal which has a relatively small frequency range. The VCXO drives a forty-eight bit counter. The state of the counter is compared with the contents of the PCR and the difference is used to modify the VCXO frequency. In practice, the transport stream packets will suffer from transmission jitter, and this will create phase noise in the loop. This is removed by the loop filter so that a large number of phase errors are averaged over time before affecting the VCXO. The 48bit counter is divided by 300 to produce a 33bit counter. The 'decode' module retrieves MPEG data from the PID buffer when the 33bit counter matches the DTS of the frame. The 'display' module similarly retrieves a decoded frame from the display buffer when its PTS value matches the 33bit counter.

A heavily damped loop will reject jitter well, but will take a long time to lock. Lock-up time can be reduced when switching to a new program if the counter is jammed with the first PCR value in the new program.

	description	MPEG-2	ATSC	DVB
PTS	90 kHz clock 33bit counter	Interval <0.7s Jitter	Interval <0.7s Jitter <15ms	Interval <0.7s Jitter <15ms
DTS	90 kHz clock 33bit counter	Interval <0.7s	Interval <0.7s Jitter <15ms	Interval <0.7s Jitter <15ms
PCR	27 MHz clock 48bit counter	Interval <0.1s Jitter <4ms	Interval <0.1s Jitter <4ms	Interval <40ms Jitter <0.5ms
PAT	Lists PMT PID	Interval not specified	Interval <0.1s	Interval <0.5s
PMT	Lists prog. PIDs	Interval not specified	Interval <0.4s	Interval <0.5s

Table 1 Timestamps, Tables and their constraints[10]

In legacy receivers, the NLL module could not be implemented in software by the main CPU so it was either implemented in hardware or was radically simplified. Both choices have led to issues in the performance of legacy receivers. One simplification was to replace the NLL with a 48bit counter driven from the video electronics 27MHz clock. This results in a clock that does not dynamically adjust to reproduce the original timing of the encoder.[11]

MPEG TS OVER UDP

A typical IP packet carrying MPEG-2 video-streaming data consists of seven MTS packets, each containing 184 bytes of payload and 4 bytes of header. This results in 1316 bytes, plus the packet overhead – 8 bytes for the UDP header, 20 bytes for the IP header, 14 bytes for the Ethernet header. (Fig.4)

What temporal distortions result from packet loss?

UDP is an unreliable transmission mechanism. Packets can be lost. Loss of IP packets may occur for multiple reasons — bandwidth limitations, network congestion, failed links, and transmission errors. Packet loss usually results in bursty behavior, commonly related to periods of network congestion. Depending on the type of transport protocol

used for the video streaming, a packet loss will have a different impact on the quality of the perceived video. When UDP is used, the lost packets will directly affect the image, as the information cannot be recovered and the image will simply be corrupt or unavailable. When using TCP, a packet loss will generate a retransmission, which can produce a buffer underflow and, consequently, a possible frozen image.

The loss of one UDP packet results in the loss of 7 MTS packets. In a 3.75Mb/s CableLabs stream, one UDP packet represents about 2.8ms of content. Assuming also a 192kb/s audio stream, there will be 18 MTS packets containing video, for each one containing audio. This means that a single UDP packet has a high chance of containing no audio.

The loss of a silent UDP packet results in the video stream jumping forward in time by 2.8ms while the audio stream is undisturbed. Now, a well behaved MPEG-2 decoder should present each video or audio "frame" at the scheduled time – when its PTS/DTS values match the recovered system clock.

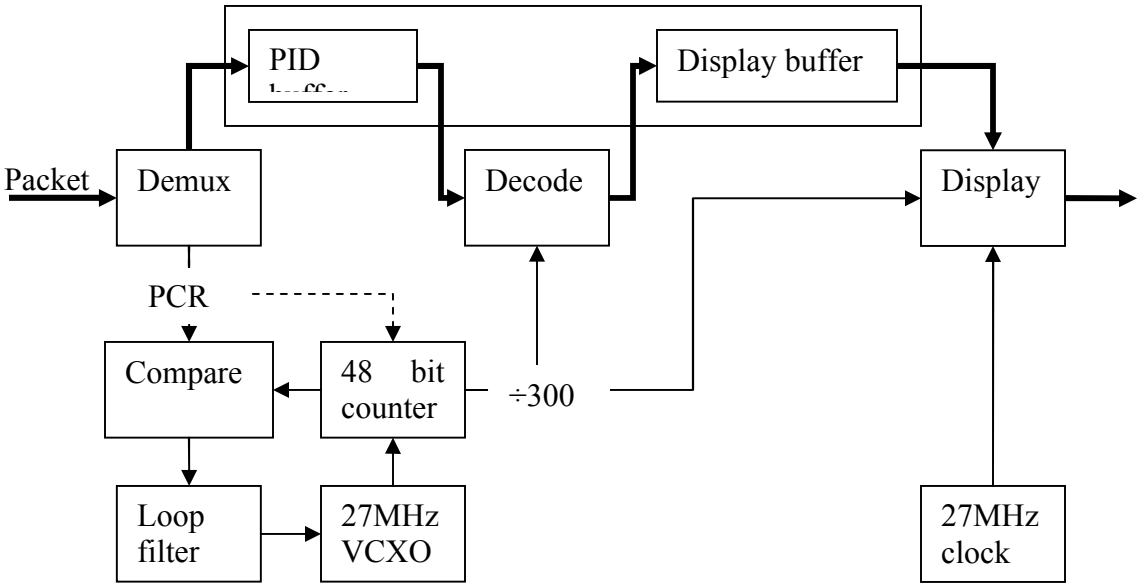


Figure 3 Clock regeneration with NLL

Legacy receivers that omitted the NLL suffer loss of synchronization. The PCR value is looked at only when a program switch occurs and thereafter the system clock runs locked to an internal reference such as the CPU clock or video display clock. This means that packet loss results, inexorably, in loss of AV synchronization. After 15 lost UDP packets, the drift is noticeable. In a network with just 0.01% loss, the sync drift would be noticeable after one hour of continuously viewing the same channel. A simple user controlled remedy is to reset the counter. This is achieved by switching to another channel and then switching back. Try it at home!

Variable MPEG processing delays

The MPEG-2 specification states that video or audio elementary-stream access units that do not contain B pictures are to be transferred immediately from the main buffers to the decoders at the time denoted by its PTS. The STD then decodes and outputs the data in the main buffers when the STC matches the PTS.

However, a video elementary stream that includes B-pictures requires that I and P pictures be decoded before decoding the B-pictures, and it is for this reason that the decoding time and presentation time of I or P pictures differ. In particular, the specification states that I or P picture data are to be transferred immediately from the main buffer to the decoder at the time denoted by DTS. The system decoder then decodes and outputs the I-picture or P-picture in the main buffer when the STC matches the DTS. Thereafter pictures are held in a re-order buffer until its PTS matches the STC.[12]

This means that packet loss can cause drastically different perceived distortions whether the lost packet contains I, B or P data. Higher compression results in greater temporal distortion. A lost UDP packet from a 700kb/s H.264 stream represents 15ms!

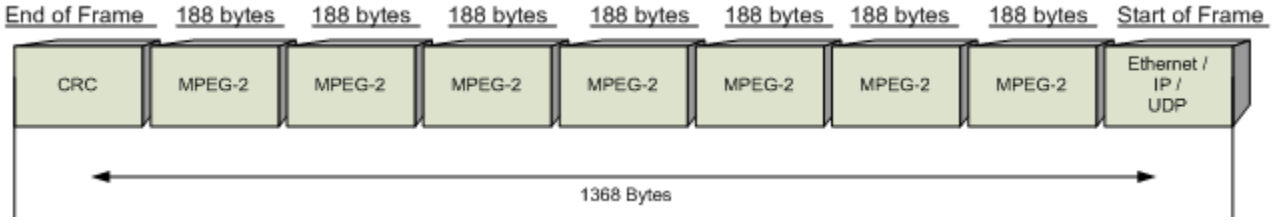


Figure 4 UDP packet contains seven MPEG transport stream packets

DVB-CA SECURITY MODEL

The DVB-CA security model comprises a combination of scrambling and encryption to prevent unauthorized reception. Encryption is the process of protecting the secret keys that are transmitted with a scrambled signal to enable the descrambler to work.

ECM

The scrambler key, called the control word (CW) must, of course, be sent to the receiver in encrypted form within an entitlement control message (ECM). The CW is valid for a particular crypto-period (CP) which is typically 10 seconds long. ECMs must be received and the CW extracted and decrypted in advance of MTS packets in the associated crypto-period. If the ECM is not available for the associated crypto-period in time, then the content cannot be decrypted and the subscriber will suffer service loss.

The ECMs are repeated every 0.1s to ensure that the stream is still decryptable even under severe packet loss. The ECM stream takes up about 1% of the stream bandwidth. The ECMs are transmitted in a separate PID that is multiplexed in with the original stream. The original stream is already time-stamped. The injection of ECMs causes jittering in the PCR values of the original transport stream. An important feature of DVB-CA multiplexer is to perform PCR correction to compensate for this jitter. ECMs of moderns CA systems now carry more that just control words. Watermark metadata, extended copy control information, and other metadata would cause the ECM to grow beyond a single MTS packet.

In the absence of PCR correction the packets could arrive in an untimely manner – outside the 0.5ms jitter spec of the DVB standard.

EMM

The CA subsystem in the receiver will decrypt the control word only when authorized to do so; that authority is sent to the receiver in the form of an entitlement management

message. This layered approach is fundamental to all proprietary CA systems in use today. In a traditional DVB network the EMMs are transmitted in-band, in another PID multiplexed with the content. In IPTV networks the EMMs can be sent out-of-band via a reliable communication channel.

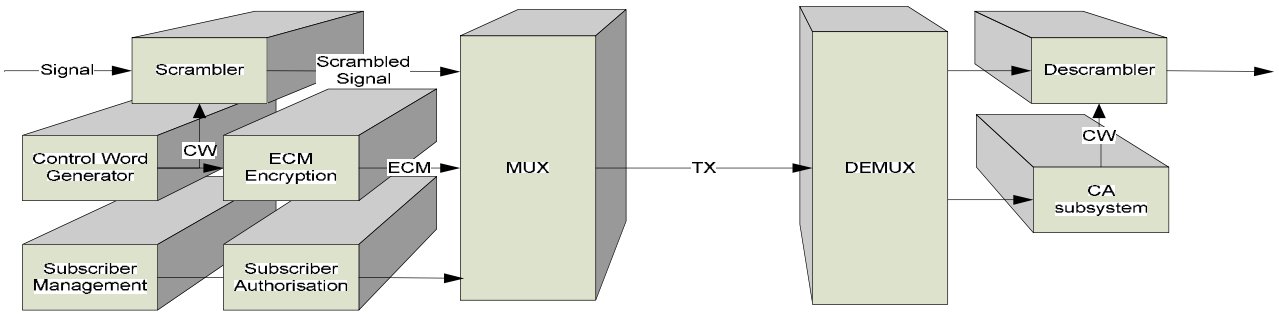


Figure 5 DVB-CA model

SECURITY TIME CHALLENGES

Security affects the timing of streams, by introducing jitter and consuming transmission bandwidth, through the insertion of ECM packets.

Client-side CPU load

The client CPU is burdened with an increasing load to support more sophisticated security systems including processes that insert watermarks into the content, monitor content to generate fingerprints, and monitor the receiving device to ensure that no theft is occurring. This occurs as the client device labors under a six-fold load increase in the transition from standard to high definition. It means that timing is becoming ever more critical in modern client devices.

Timely arrival of keys

In the case of linear content the EMMs are typically sent to the client at the time of subscription to the channel or bouquet of channels. EMMs are revoked and re-issued to rotate entitlement keys -typically on a monthly basis. Unidirectional networks such as satellite need to handle EMMs with special care. The carousel transmission of the EMMs has to strike a balance between reliability and security.

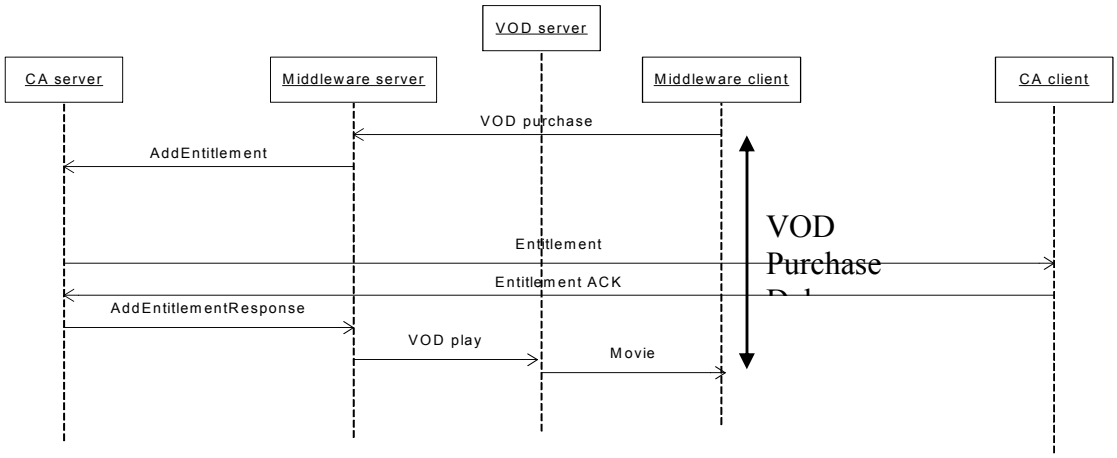
The service operator must ensure that all the subscribers receive the EMMs for the services to which they are entitled, while also being careful not to expose an EMM for too long to hackers.

Blocking EMMs revocation, an obvious hack to thwart key rotation, would allow a subscriber to access a service long after the subscription has expired.

In bidirectional IPTV networks, with reliable TCP/IP communication, the EMMs can be issued on a just-in-time schedule. EMM acknowledgements can also become part of the business logic of the service.

In the case of an impulse VOD purchase, the EMM cannot be pre-staged on the subscriber's set top box. To reduce 'VOD Purchase Delay,' the latency between the client's purchase request and the start of the movie, the timing of the security communications needs attention.

Figure 6 VOD sequence diagram showing unmitigated VOD Purchase Delay



Ordinarily the delay would be several seconds. One method to mitigate the delay is to leave a leader of the movie in the clear. The leader duration is just longer than the maximum VOD purchase delay. An even more secure solution is to encrypt the leader with a key that is only issued to those clients that have subscribed to the VOD service. This approach means that the Middleware server can issue the 'VOD play' command to the VOD server as soon as it receives the VOD purchase message. The subsequent CA communication will then occur in parallel with the playing of the leader. The EMM for the movie will arrive at the STB in time to decrypt the remaining duration of the movie.

Timed entitlement

Normally EMMs are issued and revoked by the CA server. In this case the time is maintained by the server. In the advent of secure processors, secure memory, and secure clocks, the CA client can be safely implemented to operate with a higher degree of autonomy. In these more secure devices the client can receive EMMs with richer rights expressions. A simple timed entitlement includes the start and end time. The client will only use the entitlement after the start time and will purge it after the end time. This would allow a customer to download a movie onto a mobile device and watch it on a plane or boat, disconnected from the CA server.

Secure time

Manipulation of the client clock is a well-known hack to retain expired service. Secure clocks are tamper proof and are protected against unauthorized changes. The clock can be set through a secure protocol each time the client connects with the CA server. Time should be maintained independently of local time-zone, using UTC/GMT, to avoid errors caused by the CA client and CA server operating in different time zones. The CA server should itself obtain time from a trusted NTP source.

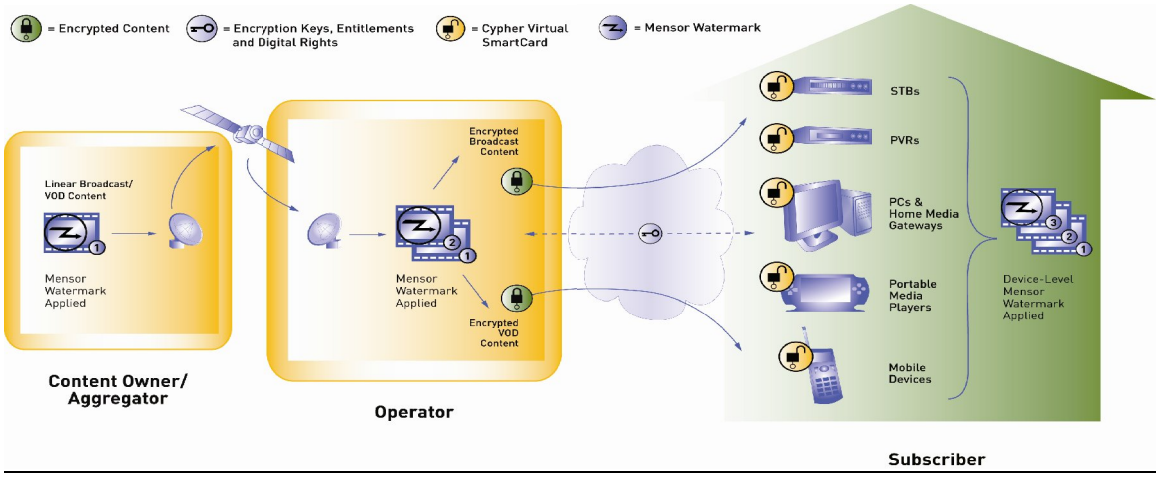


Figure 7 Hybrid CDN

File-based transmission

Modern content delivery networks (CDN) transmit linear and file-based content. These hybrid networks aggregate content and distribute files to remote service operators that each serve separate populations of subscribers. In a network, as in figure 7, the security processing has the traditional real-time requirements at the point of displaying the content on the client device. However the rest of the network treats the content as files. Files are transmitted from the aggregator to the operators as fast as the satellite transponder allows. In this environment files may be secured by different conditional access systems in different legs of the pipeline. Each file may need to be encrypted at the aggregator, decrypted at the operator and then re-encrypted with the operators CA of choice. Then the file is served to the subscriber where it is decrypted, decoded, and displayed in real-time. At the operator, however, bulk crypto processing should happen as fast as possible as files are pitched from the aggregator at 20 times faster than real-time. As described earlier in the paper, legacy CA systems that have implemented their stream parsing and crypto-processing in hardware have built their systems around the real-time clocking requirements. These real-time scramblers and descramblers cannot be easily retooled for the task of bulk encryption or decryption required to secure a file-based CDN.

TVN Entertainment[13], a VOD service operator, delivers 5000 hours of file-based content per month to affiliate operators around the country.

This network is both encrypted and watermarked. In 2007, benchmarking tests TVN Entertainment showed that an off-the shelf single rack unit server, running Widevine Cypher® DRM software, can encrypt or decrypt one gigabyte of MPEG file in 1.5 minutes while a traditional real-time DVB scrambler takes 35 minutes.

CONCLUSION

At IBC07, SMPTE and EBU[14] jointly declared:

The current methods of timing and synchronization for television, audio and other moving picture signals rely on standards that have been in place for more than 30 years. While these standards have proven to be robust solutions that have served the industry well, they are predicated on technologies that are becoming increasingly inappropriate for the

digital age with, for example, networked content sharing or higher frame rate HDTV image formats; they now impose unacceptable limitations for the future.

The time constraints on MPEG streams are onerous. The addition of security processing provides an extra timing challenge for control logic of servers and clients. Legacy CA systems suffer from a lack of flexibility in dealing with the timing of a faster than real-time CDN. The oversimplification of the timing logic or its implementation in hardware precludes traditional CA crypto-processors from reaching the performance of modern software bulk encryptor / decryptors. The evolving security landscape is challenged with emerging requirements for watermarking, fingerprinting and copy protection. A software solution is best placed to rise to these challenges - both timely and securely.

References

[1] ETSI (1997), TS 101 197-1 Digital Video Broadcasting (DVB) DVB SimulCrypt Part 1: Head-end architecture and synchronization.

[2] Frederick Herzberg, 'The Motivation to Work' (1959), Work and the Nature of Man (1966), The Managerial Choice (1982); and Herzberg on Motivation (1983).

[3] E. Alexandra Athos et al (2007), Dichotomy and perceptual distortions in absolute pitch ability, PNAS, September 11, 2007, vol. 104, no. 37,14799

[4] Audition, by Pierre Buser and Michel Imbert, English translation by R. H. Kay, MIT Press, Cambridge MA, 1992

[5] CD Audio Demonstrations, by A. J. M. Houtsma, T. D. Rossing, W. M. Wagenaars, Philips 1126-061.

[6] Reeves and Voelker (1993), Effects of Audio-Video Asynchrony on Viewer's Memory, Evaluation of Content and Detection Ability, Stanford University.

[7] ATSC (2003), ATSC Implementation Subcommittee Finding: Relative Timing of Sound and Vision for Broadcast Operations, Doc. IS-191, 26 June 2003.

[8] Tryfonas and Varma, Timestamping Schemes for MPEG-2 Systems Layer and Their Effect on Receiver Clock Recovery, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 1, NO. 3, SEPTEMBER 1999, Page 251

[9] Watkinson (2001), The MPEG Handbook, Page 333, Focal Press

[10] Isnardi (1999), MPEG-2 Systems, Sarnoff Corporation, August 25, 1999

[11] SS. Bindra (2006), Studio Systems, July – August 2006

[12] Yoshimura (2002), Technologies and Services on Digital Broadcasting (5), Broadcast Technology no.11, Summer 2002

[13] Dom Stasi (2007), Broadband Business, CED Magazine, Sept 2007

[14] EBU, SMPTE announce joint task force on time, synchronization, Broadcast Engineering, Sep 15, 2007

TRANSCODING AND STATISTICAL MULTIPLEXING OF MPEG4 (H.264) BROADCAST VIDEO

John Hartung, Ph.D.
EGT

Santhana Krishnamachari, Ph.D.
EGT

Abstract

The bandwidth demand of HD content is driving the use of more efficient video compression such as MPEG4 (H.264) encoding for satellite distribution, and statistical multiplexing for MSO access networks. Transcoding and statistical multiplexing are usually implemented independently; however, in this paper we show that this is not the best approach. We show that integrating the transcoding and statistical multiplexing operations will result in improved video quality, reduced operational complexity and lower cost. The paper is organized into four sections: Introduction, Transcoding, Statistical Multiplexing, and Conclusion.

INTRODUCTION

MSOs are planning to increase the number of HD programs they offer from around 25 today to more than 100 over the next couple of years. This increase is placing a tremendous strain on the available access bandwidth in the MSO HFC networks. Three approaches are being taken to solve this bandwidth problem: analog channels are being converted to more efficient digital transmission, switched digital broadcast is being deployed, and HD content is being statistically multiplexed so that 3 or more HD channels are carried in a QAM. The statistical multiplexing approach has the advantage of not requiring the additional network infrastructure and software needed to support switched broadcast or of turning off existing analog services. In addition, the cost of statistical multiplexing can be shared across multiple service groups and nodes locally, regionally, or nationally.

The new HD channels will be received in various encoding formats and bit rates from satellite distribution and terrestrial broadcasters. MPEG2 format is typically received at rates around 15 Mbps or higher, and satellite distributors are beginning to use MPEG4 encoding at 8 Mbps for new programming. Although MPEG4 capable set top boxes are beginning to be deployed, the large numbers of legacy MPEG2 set top boxes require the conversion of all content into MPEG2 format. This paper describes various approaches for transcoding and statistical multiplexing with quantitative comparisons. A novel approach that combines transcoding with statistical multiplexing is shown to have the best compression efficiency and quality.

TRANSCODING

Overview

In general, transcoding from MPEG4 to MPEG2 requires a full decode and re-encode because many of the tools available in the MPEG4 standard are incompatible with MPEG2. Examples of these tools include advanced prediction algorithms such as the use of multiple reference frames and intraframe prediction, and filtering in the prediction loop to reduce blocking artifacts. In some specific instances the MPEG2 parameters can be determined or estimated from the MPEG4 parameters leading to higher quality and lower complexity.

Independent Decode-Encode

One approach to transcoding from MPEG4 to MPEG2 is to fully decode the MPEG4 frames and then re-encode with an MPEG2 encoder. This can be implemented with an entirely separate decoder and encoder; however, this approach does not produce the highest possible MPEG2 encoding quality and is computationally expensive. One reason that quality is compromised is frame coding types are not preserved, and therefore high quality reference frames, such as I and P frames, are not re-encoded with the same types in MPEG2. This results in lower quality reference frames and propagation of coding distortion when they are used for prediction of P and B frames. Separation of decode and encode functions also prevents the original MPEG4 encoding parameters from being re-used as initial MPEG2 encoding parameter estimates to reduce complexity. Reuse of these parameters is especially useful in motion estimation where initial estimates can be used to reduce the search complexity by limiting the search range.

Integrated Transcoding

An alternative approach is to decode the MPEG4 input, and at the same time pass the MPEG4 encoding parameters to the MPEG2 encoding stage. This is shown in Figure 1. In addition to preserving frame types and reducing encoding complexity, the MPEG4 parameters are also used by the First Pass

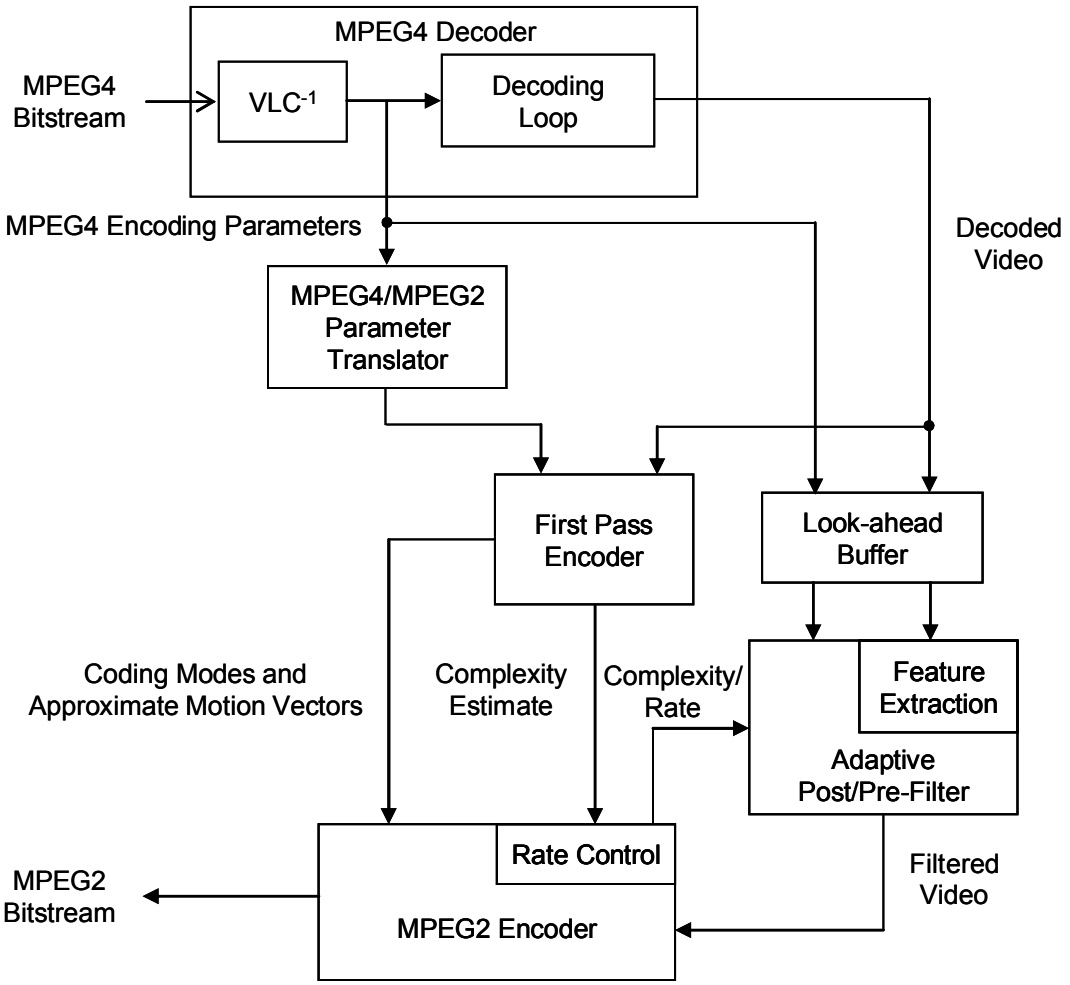


Figure 1

Encoder and Adaptive Post/Pre-Filter. The First Pass Encoder determines the MPEG2 encoding modes and approximate prediction residuals from both the decoded video and MPEG4 encoding parameters, where possible. A relative complexity is determined for each frame within a group of pictures (GOP), and this in turn is used by the second pass MPEG2 encoder Rate Control to determine an optimal target encoding rate for each frame within the Look-Ahead Buffer, thereby achieving the best overall quality.

Integration of the decoding and encoding functions also enables advanced Adaptive Post/Pre-Filtering of the decoded video. This filtering serves two purposes: removal of encoding artifacts from the decoded MPEG4 bit stream, and filtering to reduce MPEG2 encoding artifacts. For both types of filtering feature extraction is used to identify areas having characteristics that mask distortion due to the response of the human visual system (HVS). For example, distortion in textured areas is difficult to perceive so those areas can be highly filtered to reduce the required number of coding bits, while areas with edges need to be preserved in order to retain image details. The original MPEG4 encoding parameters are used to adaptively remove encoding artifacts by estimating the encoding distortion from prediction parameters and quantization step sizes. The MPEG2 Encoder allocates a coding rate to each frame based on its' complexity and the bits available for all frames within the GOP. This ratio of complexity /rate indicates the amount of pre-filtering needed to minimize artifacts in the MPEG2 encoded frame. Optimal filtering and reduced complexity result from the availability of both MPEG4 and MPEG2 encoding parameters along with feature extraction of the decoded video.

Comparison

Figure 2 shows a plot comparing the quality of Independent Decode-Encode with Integrated Transcoding for 1920 x 1080i HD video. The original MPEG4 video is encoded at 10 Mbps. Peak signal to noise ratio (PSNR) is used as an objective measure of the difference between the original and encoded video. A higher PSNR represents better

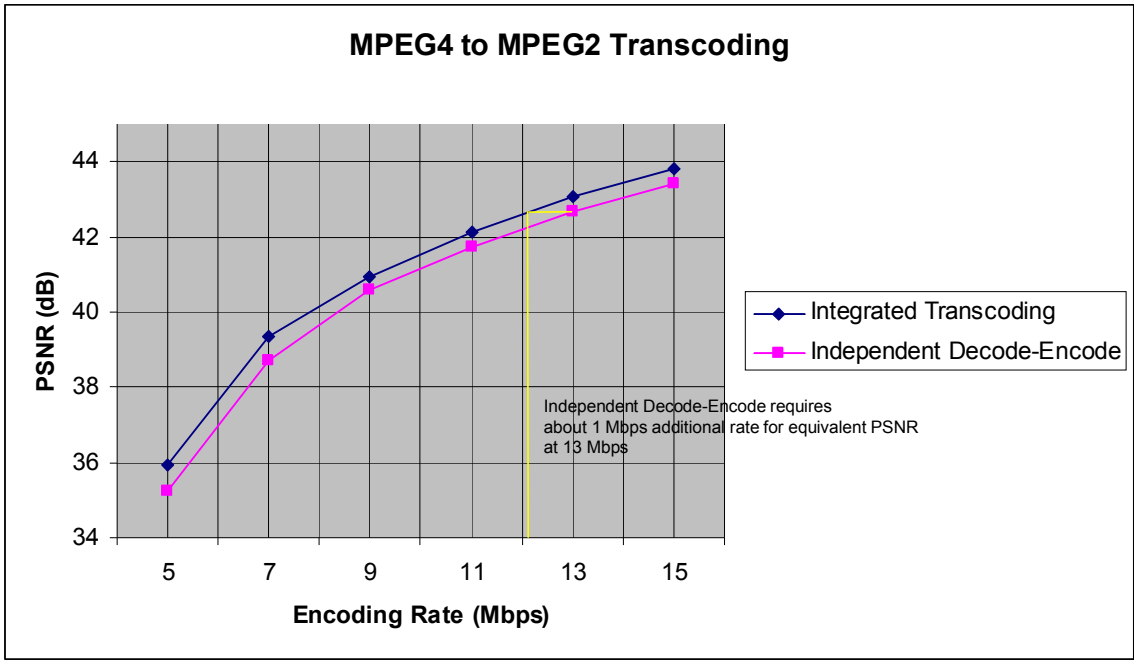


Figure 2

quality with about a .5 dB change resulting in a perceived quality difference. It can be seen from the plot that Independent Decode-Encode achieves, on average, about .5 dB lower PSNR than Integrated Transcoding. This translates into about a 1 Mbps higher rate to achieve equivalent performance. The next section also shows that the average PSNR is not the whole story when it comes to statistical multiplexing. Integrated Transcoding also results in lower frame to frame PSNR variance and therefore a more uniform and lower rate to achieve a constant quality.

STATISTICAL MULTIPLEXING

Overview

HD channels are delivered to a head end at a constant bit rate using either MPEG4 or MPEG2 encoding. The bit rate is chosen to produce good quality for the most difficult sequences, even though a lower rate would be sufficient most of the time. For MPEG2 HD content this rate is 15 Mbps, or higher, allowing only two channels to be transmitted within a 6 MHz QAM channel. Statistical multiplexing increases the number of programs that can be carried by re-encoding each input at a lower rate that varies as a function of the channel's complexity. The individual rates are controlled in order to maintain the original video quality at an aggregate rate that allows additional channels to be carried within a QAM. For a 38.8 Mbps QAM channel this corresponds to an average encoding rate of below 13 Mbps for three or more HD channels.

The challenge to achieving multiplexing gain is to combine channels such that their instantaneous encoding rate remains close to their average rate. For SD this requirement is met because of the large number (>12) of channels transmitted in a QAM. However, with only three HD channels transmitted in a QAM, the channel characteristics must also be considered. One approach is to combine two low complexity channels, such as progressive movie content, with one high action channel, such as sports. A second factor that limits the number of channels that can be multiplexed is the efficiency of transcoding and rate shaping. These translate directly into the rate required to encode individual channels at high quality.

Two methods have been used to transcode and

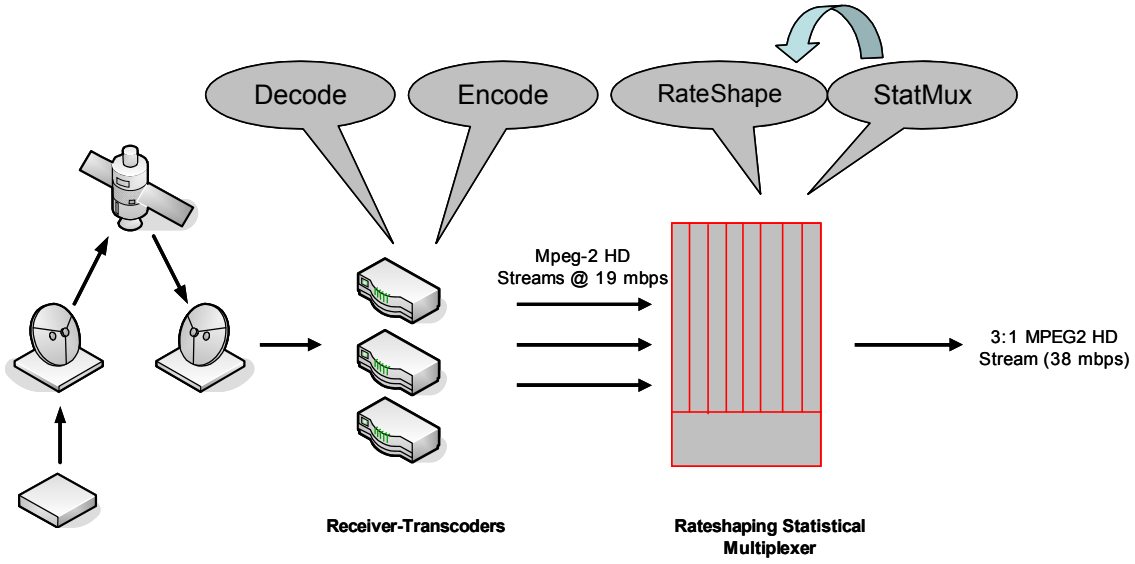


Figure 3

rate shape content for statistical multiplexing as described below. These are transcoding from MPEG4 to MPEG2 followed by rate shaping, and MPEG4 decoding and MPEG2 re-encoding with a closed loop statistical multiplexer. A third approach, MPEG4 to MPEG2 transcoding integrated with closed loop statistical multiplexing, is shown to produce the best quality.

Transcoding and Rate Shaping

In this architecture the MPEG4 input is first transcoded to MPEG2 in the Receiver-Transcoders. For HD MPEG4 delivered at 8 Mbps this first stage of transcoding produces an MPEG2 output of about 15 Mbps. The MPEG2 programs are then statistically multiplexed in a second stage of rate shaping to form an MPTS meeting the QAM rate as shown in Figure 3. The second stage is usually implemented using a rate shaper that modifies the original MPEG2 encoding parameters without performing a full decode and re-encode. This approach runs into problems when the rate reduction for any channel exceeds about 15%; significant video quality reduction occurs under these conditions. A rate reduction of significantly greater than 15% is fairly common, and occurs whenever two of the channels need a bandwidth above 13 Mbps to achieve adequate quality. If we consider the case where two channels require 14 Mbps, then the third channel must be reduced by greater than 33% (15 mbps to 10 mbps) to meet the total rate of 38.8 Mbps. The performance of this approach is fundamentally limited by the fact that is uses two stages of MPEG processing, transcoding followed by rate shaping. In the comparison section we show that the performance falls well below the two other approaches.

Decoding and Closed Loop Encoding

A second approach begins by decoding the input MPEG4 bitstreams using Receiver-Decoders as shown in Figure 4. This output is then re-encoded using MPEG2 encoders within a closed loop statistical multiplexer. A single stage of re-encoding, decode followed by encode, introduces less distortion than the previous method, however separation of the decoder and encoder prevents reuse of the original MPEG4 encoding parameters. This in turn leads to a lower PSNR for the target rates determined by the statistical multiplexer. The Comparison section also shows that both this approach and the previous one, introduce greater variance in the frame to frame PSNR. This shows up as artifacts in the statistically multiplexed video that degrade the video more than would be reflected in the average PSNR comparisons.

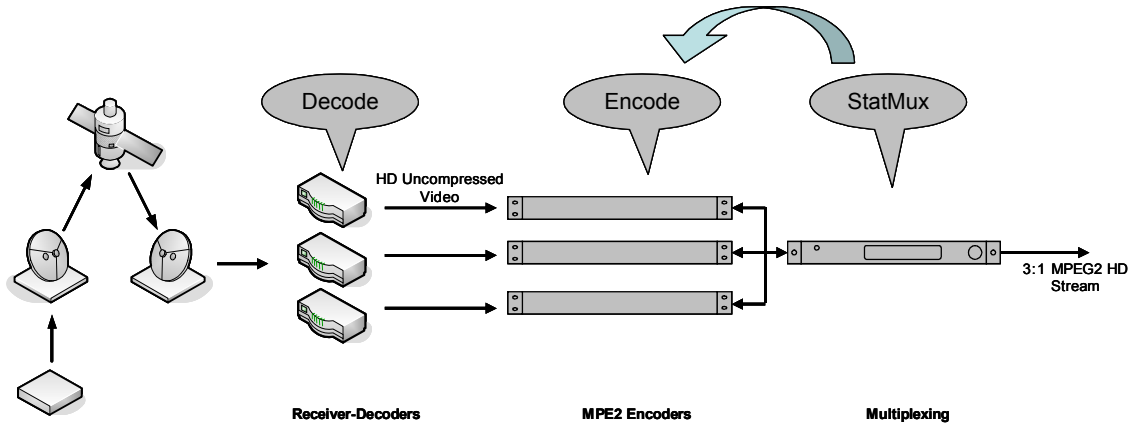


Figure 4

Closed Loop Transcoding

The third approach converts the MPEG4 output from a receiver directly to MPEG2 using an Integrated Transcoder, as shown in Figure 5. This approach achieves the best performance by transcoding directly to the statistical multiplexing rate in a single stage of MPEG encoding. The rate feedback also enables the integrated transcoder to adapt the post/pre filters for the target rate, rather than an intermediate rate. The perceptual quality is also improved by the lower encoding rate variance achieved in this implementation as shown in the next section.

one channel peaks. Lower target rates can be chosen for the easier channels, allowing a higher rate to be allocated for the complex channel.

Figure 7 shows the MPEG2 output frame PSNRs for 1920 x 1080i video transcoded using the three approaches. The original MPEG4 video was encoded at 10 Mbps and the output is at 13 Mbps. The important consideration here is the variance of the PSNR for each frame. A lower PSNR indicates that the frame is more complex and would need to be coded at a higher bit rate to achieve equal quality. Rate shapers having a

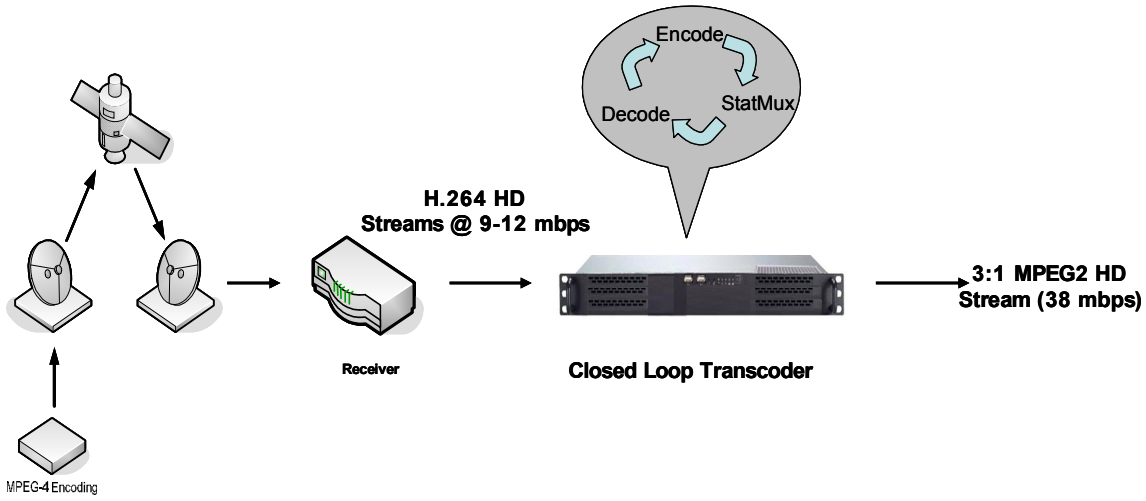


Figure 5

Comparisons

Figures 6 and 7 show the single channel PSNR performance for the three statistical multiplexing approaches described above. The results are for 1920 x 1080i HD video originally encoded using MPEG4 at 10 Mbps. Integrated Transcoding achieves a 2.75 Mbps advantage over Transcoding and Rate Shaping at rates around 13 Mbps as shown in Figure 6. It achieves a 1 Mbps advantage over Decoding and Closed Loop Encoding as shown in a previous section. These gains produce higher overall quality, but are particularly important when the complexity of high variance produce frequent artifacts because there is a higher probability that individual target rates exceed the aggregate available rate. The plot shows that the Closed Loop Transcoder achieves the lowest variance, followed by the Decoder with Closed Loop Encoding and Transcoding and Rate Shaping implementations.

CONCLUSION

Integrating transcoding and statistical multiplexing produces several benefits over competing approaches, the most important being optimum compression efficiency and

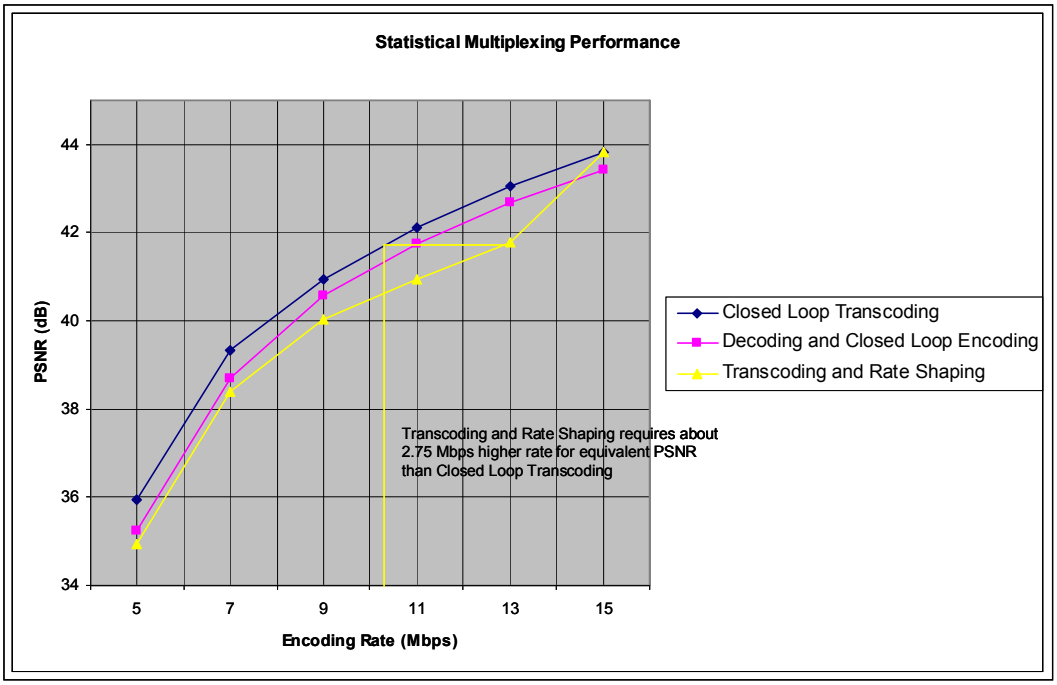


Figure 6

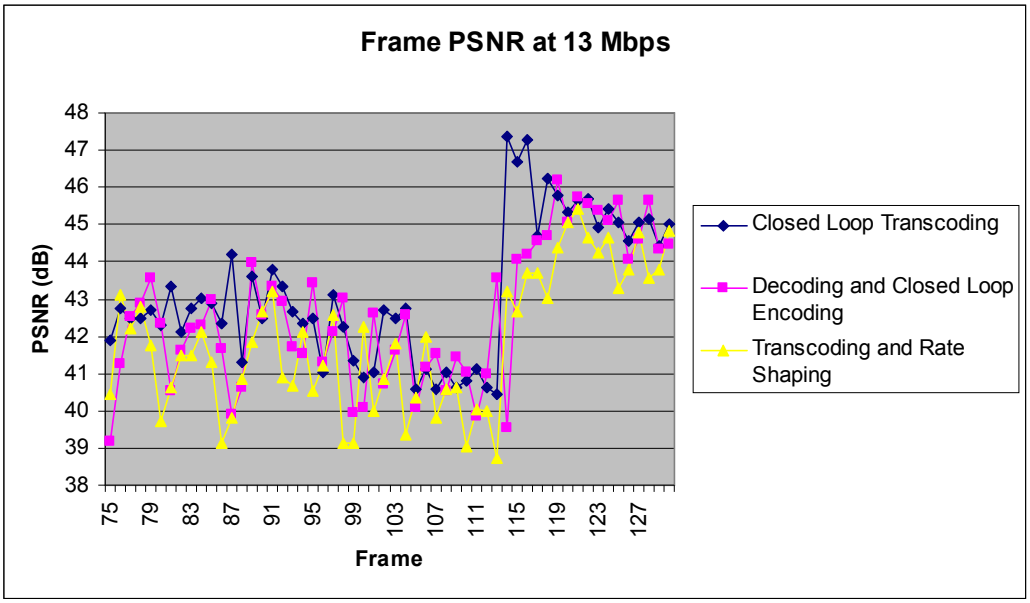


Figure 7

video quality. This architecture achieves these benefits through the reuse of the original encoding parameters, both for conversion between input and output formats, and for encoding at the statistical multiplexing rate. Integrating the two functions also enables a single stage of encoding thereby avoiding the distortion due to two generations of processing. Although this paper has focused on transcoding from MPEG4 to MPEG2 standards, similar gains are achieved when constant bit rate MPEG2 content is statistically multiplexed into an MPEG2 output.

UNICAST VIDEO WITHOUT BREAKING THE BANK: ECONOMICS, STRATEGIES, AND ARCHITECTURE

S.V. Vasudevan, Xiaomei Liu, and Robert Kidd
Cisco

Abstract

Driven by competition and consumer demands, linear video delivery is following a trajectory from broadcast to multicast and ultimately to unicast. Traditionally, video delivery has been broadcast only. Today, cable operators are deploying switched digital video (SDV), which uses multicast technology to improve the bandwidth efficiency of HFC networks. The next logical progression to unicast delivery is on the horizon and is positioned to become tomorrow's video delivery mechanism.

Unicast delivery of linear content is an incremental extension of the multicasting approach used in SDV implementations. The incremental investment in bandwidth resources to support unicast delivery can be offset by the contribution of preferentially valued advertising opportunities, reduced subscriber churn, and the ability to attract new subscribers through differentiated service offerings.

This paper analyzes the unicast value proposition, including cost, revenue potential and return on investment. SDV field trial viewership statistics will be reviewed, and used to shed light on the cost sensitivities related to channel popularity and HD penetration. Best case and worst case scenarios for HFC bandwidth consumption will be explored and analyzed, along with the cost structures associated with each of them. Cost mitigation and revenue improvement strategies will be explored, demonstrating how cable operators can optimally combine unicast and multicast approaches in order to maximize overall return on investment.

Based on the results of this analysis, a switched architecture will be presented for cable operators to smoothly migrate their networks to support unicast delivery mechanisms for linear video services. The proposed architecture accomplishes the strategies for cost-effective unicast delivery and supports:

- *A flexible combination of multicast and unicast delivery mechanisms*
- *Traditional ad insertion based on geographic ad zones or a new generation of targeted ad insertion based on demographic profiles*
- *Fast channel change and personalization of unicast content*

INTRODUCTION

Switched Digital Video (SDV) is now a mainstream technology that is delivering on its promise to dramatically improve upon the bandwidth efficiency of the traditional linear broadcast model. Aggressive SDV oversubscription ratios (the ratio between the number of SDV programs offered and the number of stream resources provisioned) have been observed and will continue to increase as more niche and HD content is added to cable MSO service tiers. Switching technology has proven to be a powerful addition the cable operator's bandwidth management capability.

Yet there remains unlocked potential within the SDV infrastructure. Current generation systems support open standards, allowing the insertion of new technologies and applications. The session and resource managers (SRMs) that

respond to subscriber channel-change requests provide a level of intelligence and network awareness that was previously unavailable. SDV systems maintain a real-time accounting of programs being viewed as well as the number of subscribers viewing those programs. This awareness of program usage comes concomitant with knowledge of bandwidth allocation. This knowledge is powerful, for even with existing switched multicast, there remains significant room for improvement in bandwidth utilization. Furthermore, *Switched Unicast*, an advanced form of SDV, is drawing increasing interest. This emerging SDV architecture offers exciting opportunities to introduce new revenue-generating services, but it also has the potential to overwhelm available access bandwidth. Only with the knowledge of user demand provided by advanced SRMs is this latest architectural challenge tractable.

It is first helpful to clarify the requisite terminology. *Switched multicast* refers to a video delivery architecture where an MPEG program, typically in the form of a single program transport stream (SPTS), is IP-encapsulated and transported on a distribution network via IP multicast. A system session and resource manager (SRM), acting upon channel change requests from subscribers, may then instruct an IP-attached edge QAM to join the multicast. The edge QAM rebuilds a multi-program transport stream (MPTS), containing content requested by multiple viewers, and modulates the content onto the HFC network. The key differentiator of this approach is that multiple set top boxes within a service group can share a stream that is active within that service group; if N viewers within a service group watch the program *MTV*, only one instance of *MTV* is switched into that service group. Thus, the content is delivered over the IP network using IP multicast and over the HFC plant with a stream sharing, RF multicast mechanism resembling that on the IP network.

Switched unicast refers to a delivery mechanism in which, regardless of the IP transport and routing mechanisms deployed, the stream on the HFC side of the plant is destined for a single tuner within a single set top; i.e. set tops within a service group no longer share MPEG streams; if N viewers within a service group request *MTV*, N instances of *MTV* are switched into that service group. Typically the transport mechanisms on the IP network will also be via IP unicast, but hybrid solutions can be envisioned in which, for example, ad servers at the edge join multicast IP content, insert a targeted ad, and then send the new stream via IP unicast to the edge QAM.

The drivers for deploying switched unicast are compelling. Switched unicast offers the opportunity to personalize video. Since each tuner now receives an individual video stream, media processing techniques can be used to modify the stream to suit the preferences of an individual subscriber. These modifications may include graphics overlays on the screen that are tailored to the subscriber's preferences. For example, a ticker that displays preferred stock quotes, sports scores, localized weather information, etc.

But perhaps the most compelling driver for switched unicast is in the ability to personalize advertising. The North American cable industry has a long history in spot advertising as a revenue source, and many systems perform a limited level of localization by dividing a cable system into zones and offering spot insertion on a zone by zone basis. However, the ability to transcend beyond zones and offer advertising on a *personalized* basis offers MSOs the opportunity to charge a premium for these slots – indeed the success of Google in its ability to personalize advertising in the online space has made all participants in the advertising delivery chain sit up and take notice. Furthermore, with personalized advertising, ads need not be restricted to 30-second spots – they can additionally take the form of graphics logo

overlays, with additional possibilities enabled from the incremental ability to launch interactive ads.

This raises an interesting question which is the subject of this paper – switched unicast has the ability to increase MSO revenue through the incremental benefits of personalized advertising, yet to reap this benefit requires an investment in delivery infrastructure to enable these personalized capabilities. Is the investment worth it?

This turns out to be a challenging question to answer, as the answer depends on a number of variables, including subscriber viewing patterns, service group sizing, equipment costs, and the anticipated premium to be expected from placing a personalized ad as opposed to a zoned ad. To better understand this topic, we propose to study the subject from 3 perspectives: analyzing data from a real SDV deployment to understand multicast and unicast resource sensitivity based on program popularity, service group size, and other factors; developing a financial model of switched unicast, which allows ROI to be analyzed based on a number of factors; and exploring alternate models of SDV that enable unicasting on an opportunistic basis.

SUBSCRIBER VIEWING PATTERNS AND SWITCHED UNICAST

It is difficult to make purely analytical predictions regarding SDV efficiency, since efficiencies ultimately depend on subscriber viewing patterns, which in turn is driven by the behavior of human beings. However, by analyzing the pattern of channel change messages from SDV, it is not difficult to infer what the expected system behavior would be if the system were unicast instead of multicast.

To better understand this relationship, viewership information was extracted and analyzed as part of a SDV trial with a major North American MSO. By aggregating and post-processing channel change information from server log files, it is analytically straightforward to determine not only the relative difference in resource requirements between multicast and unicast, but it is also possible to analytically modify the "virtual" size of the service group to better appreciate the sensitivity between service group size and resource utilization for both the multicast and unicast scenarios.

The viewership study was conducted over several weeks and included channel-change data for 247 broadcast video services comprising 228 standard definition programs and 19 high definition programs delivered to 680 tuners in a service group. The study included four major steps: (1) computation of the viewership long-tail; (2) segmentation of the long-tail into popularity quintiles; (3) segmentation of the settops into "virtual service groups", and (4) computation of unicast (total) and multicast (unique) streams required to deliver cumulative quintiles of programming to a range of service group sizes.

The concept of a "virtual service group" warrants some discussion. One important result desired from the analysis was the expected variation in video stream resource requirements with service group size. In most production systems, a target service group size is established and the inside and outside plants designed to that target size. If an analysis is performed only with the existing service group structure, bandwidth requirements can be predicted for only a very narrow range of service group sizes. Therefore, instead of using the existing system's service group structure, the nodes were regrouped into sets of virtual service groups containing tuner counts that span the range of interest.

Figure 1 provides viewership results for one week of data and illustrates the classic long-tail viewership phenomenon. The graph is generated by summing the number of seconds viewed for each broadcast video program and then ranking the channels in order of decreasing total viewership [1]. Once the programs are ranked, the viewership curve is segmented into five sections, each of which includes 20% of the offered programs. These quintiles are used to evaluate stream requirements as a function of program popularity.

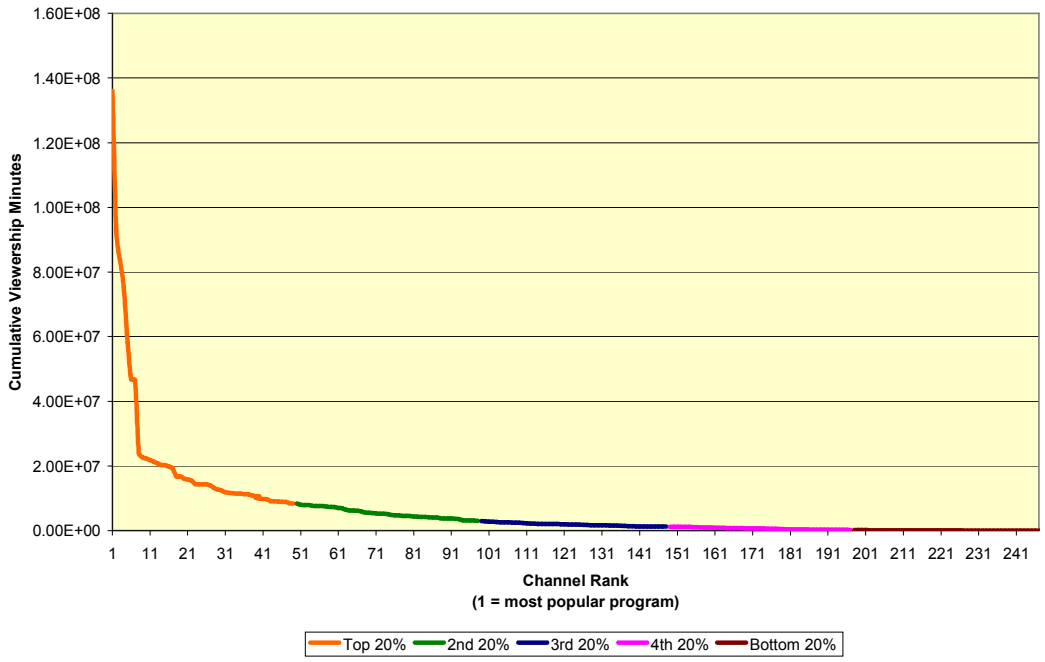


Figure 1. Program Viewership Ranking

Figure 2 illustrates switched multicast stream usage and displays the number of peak unique simultaneous streams required as a function of service group size and popularity of content. The horizontal axis displays the number of tuners per service group, and the vertical axis displays the number of peak unique simultaneous streams. Five curves are included on the plot, each of which illustrates peak stream requirements for a particular grouping of content: the bottom curve illustrates the peak unique streams required for the least viewed 20% of content; the next-to-bottom curve illustrates peak unique streams required for the least viewed 40% of content; and so on up to the top curve that illustrates the peak unique streams required for the entire broadcast video lineup.

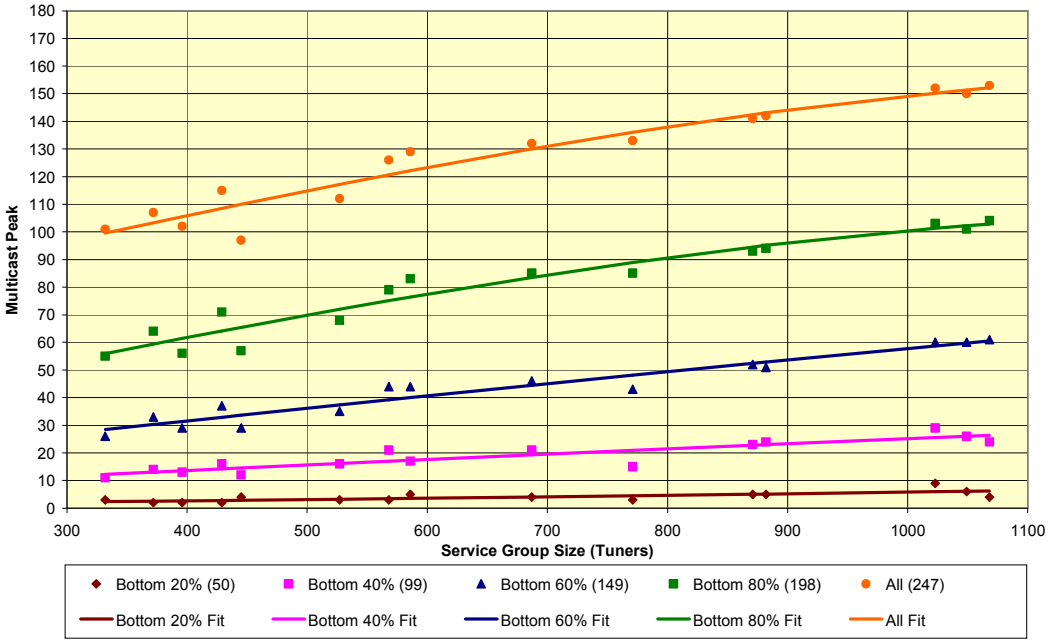


Figure 2. Switched Multicast Peak Stream Usage

Each curve is generated from a regression analysis of the raw data points, also included on the plots, that result from processing each virtual service group for a specific grouping of content. Two-sided 95% prediction intervals (not shown in the figure) were also computed. Future individual peak stream count results are expected to fall below the upper bound of this prediction interval 97.5% of the time.

Figure 3 is the unicast equivalent of Figure 2. In this case, the vertical axis displays the number of peak total simultaneous streams as opposed to the number of peak unique simulcast streams; otherwise, the data analysis is equivalent to that described above for the multicast case.

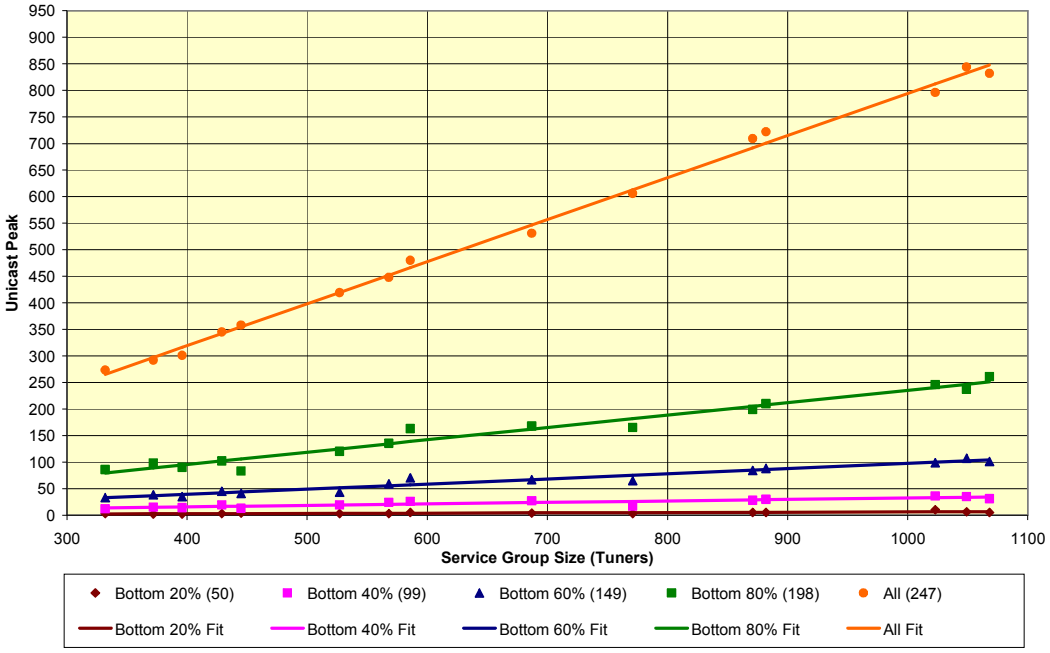


Figure 3. Switched Unicast Peak Stream Usage

Figure 4 summarizes the stream count dynamics between unicast, multicast, and basic broadcast for the entire 247-program lineup.

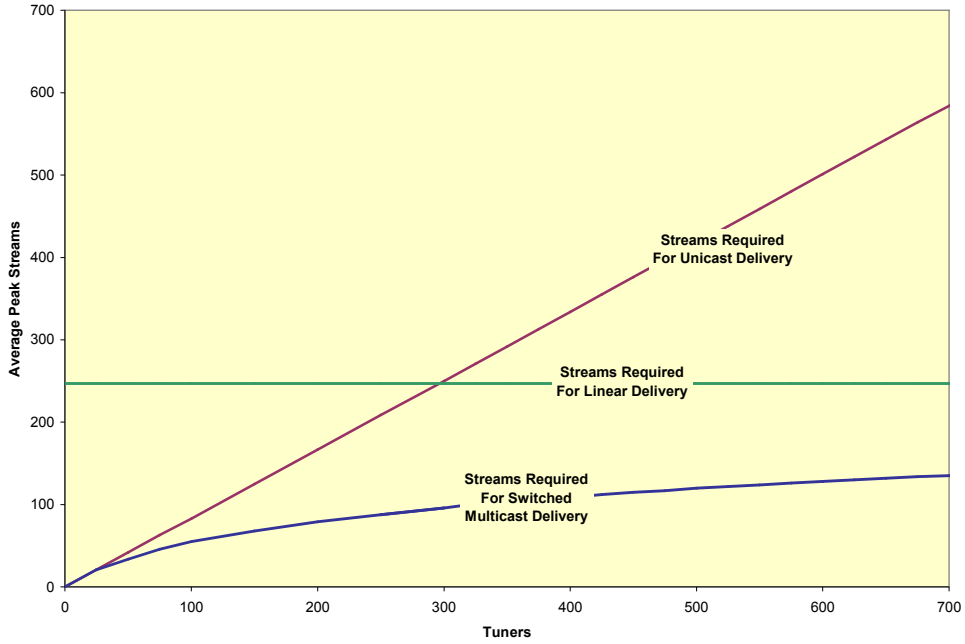


Figure 4. Stream Requirement Overview

As expected, the different delivery mechanisms require dramatically different HFC bandwidth allocations. The number of streams required to deliver the lineup via traditional linear broadcast is of course constant and therefore independent of service group size. The number of streams required to deliver the same lineup via switched multicast is significantly less since settops may share streams within an HFC service group. The number of streams required to deliver the lineup using switched unicast is linearly proportional to service group size. This latter viewership curve is similar to that for VOD except that the slope of the unicast demand curve for broadcast content is much steeper than that experienced with VOD.

For delivery of the most popular content to larger service groups, unicast requires significantly more bandwidth than not only multicast SDV but also simple linear broadcast. For example, in order to offer the entire broadcast video lineup to today's typical 500 tuner service group, the unicast model of Figure 4 requires approximately 417 peak streams, the multicast model requires approximately 120 peak streams, and the broadcast model requires 247 streams. Making the simplifying assumption of a 50/50 SD/HD split for the unicast streams, approximately 104 QAM carriers would be required to carry the unicast content.

Clearly the capacity to unicast the entire lineup is not available on a typical service group in today's hybrid digital/analog HFC systems; however, there are a number of revenue-enhancing opportunities that can be supported today by the surgical insertion of unicast technologies. These surgical deployments can be more fully developed as the industry continues the inexorable push toward smaller service groups and increased digital delivery. Service group sizes are trending towards a future where 250 tuner service groups will be the norm, and only 209 streams will be required to unicast the entire the 247 channel broadcast lineup, fewer than the number required for simple linear broadcast.

One key caveat should be raised at this point regarding unicast: viewership statistics are nonstationary from a statistical standpoint, that is, they change with time, and this fact has significant practical implications. As an example, consider the case of a weather or news channel, a service with average viewership sufficiently far down the long-tail to support its inclusion into a switched tier. As time goes by, a hurricane or other major news event will inevitably emerge, and the popularity of this previously moderately-viewed programming skyrockets. If the channel is offered on a multicast tier, viewers that flock to the channel share a single stream; however, if the channel is offered on a unicast tier, viewers receive their own stream, and the required edge bandwidth mushrooms beyond that which may have been predicted based upon prior viewership studies. A unicast tier is much more sensitive to the choice of selected content than a multicast tier and is therefore less stable from a bandwidth planning perspective. In order to mitigate this risk, a key potential feature of a unicast system would be the ability of the system to automatically promote and demote between the unicast and multicast tiers.

Finally, in the above discussion stream counts are used as a proxy for bandwidth requirements and the two are often used interchangeably. However, if different streams have different bandwidths, aggregate stream counts and their associated total bandwidths may not be directly proportional. For example, if viewership of HD services is consistently higher than that for SD services, the most viewed 20% of content may require substantially more bandwidth than the stream estimates alone would predict. In the system under consideration, 19 of the 247 broadcast video programs offered are HD. The percentages of these HD services in the 20%, 40%, 60%, 80%, and 100% quintiles of content

were 0%, 21%, 42%, 11%, and 26% respectively. Thus the spread of HD content is slightly skewed towards the most viewed groupings; however, the effect of this skew on bandwidth requirements is muted given the relatively small number of HD programs offered at the time of the study.

SWITCHED UNICAST ARCHITECTURE AND ROI MODELING

It was previously mentioned that perhaps the incremental revenue generated from switched unicast could fund the investment in the necessary delivery infrastructure. In reality, this is a complicated problem with numerous factors contributing to the analysis. Before going into details of ROI analysis, we will start with some basic assumptions of a switched unicast architecture.

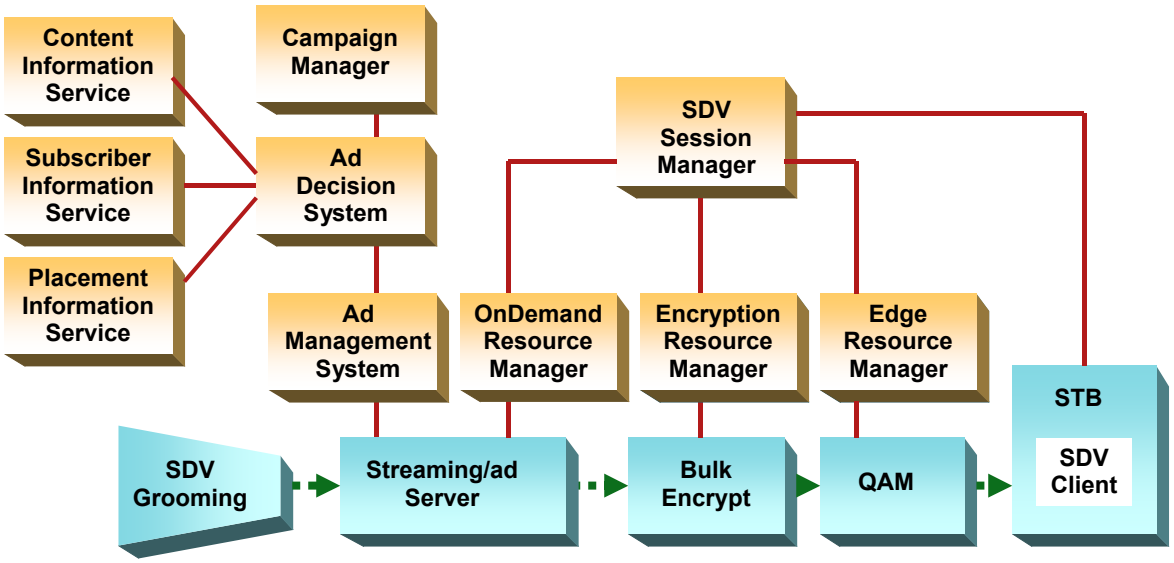


Figure 5. Switched Unicast Architecture

A switched unicast system architecture builds upon the existing, widely deployed switched multicast architecture and represents an evolutionary path. Existing components may be retained and augmented, minimizing the incremental investment. In the example switched unicast architecture shown in Figure 5, switched video content flows from a SDV groomer (performing VBR-CBR rate clamping) to STBs passing through streaming/ad servers, bulk encryptors and QAMs. The key difference from traditional switched digital video is the introduction of a streaming/ad server in the data path. The streaming/ad server is defined as a component that constantly ingests live linear content and streams it out as requested. The streaming/ad server also detects ad placement opportunities and splices ads. In the control plane, the SDV session manager manages SDV channel changes and orchestrates SDV resource allocations through various resource managers. The targeted advertisement control plane components are shown and are defined by the SCTE-130 specification. An ad decision system makes an intelligent decision on which ads to insert for an ad placement opportunity based on information from the functional components labeled content information service, subscriber

information service and placement opportunity information service. The streaming/ad server detects placement opportunities, notifies the ad decision system about the placement opportunities via the ad management system and inserts ads based on decisions made by the ad decision system.

An analytical model representing switched unicast sizing and pricing was developed to evaluate optimal dimensioning parameters for future deployments, and quantitatively examine targeted advertising development opportunities that could provide an attractive return for the infrastructure investment.

The analytical model for switched unicast considered the following parameters on the "expense side", shown in Table 1. The graphs that appear later are computed from the default values indicated in the tables.

Cost Modeling Parameters	Default Value
Service group size	1000 HHP
Subscriber penetration	60%
Digital penetration	60%
Tuners per household	1.8
Total channel offered	249
HD channel percentage	20%
Channel popularity	Extracted from trial log data
SD channel bandwidth	3.75 Mbps
HD channel bandwidth	15 Mbps
Spectrum available for SDV	30 RF channels
QAM channel bandwidth	38.8 Mbps
AVC STB penetration	20%
Node split cost	$15,000
Transport cost	$9/Mbps
QAM cost	$13/Mbps
Streaming server cost	$15/Mbps

Table 1. Cost Modeling Parameters and Default Values

The following parameters were considered on the "revenue side", shown in Table 2.

Revenue Modeling Parameters	Default Value
Ad revenue per subscriber	$300
MSO per sub ad revenue	$60
Possible ad revenue share with programmers and networks	30%
Targeted ad percentage	30%
CPM improvement of targeted ads	2
Targeted advertising operating margin	80%

Table 2. Revenue Modeling Parameters and Default Values

The strategy for the analysis was as follows: using the number of channels, the extracted program popularity, and the bandwidth of the programming, the total aggregate bandwidth of the expected unicast streams was calculated. Based on this information, the optimal service group size was calculated. If this optimal value was less than the existing service group size, the model would factor in the price of node splits to compute the infrastructure costs to achieve the proper service group size. Once this value was known, industry-current figures for QAM, streaming server, and transport costs (normalized to a $/Mbps factor) were used to calculate expected investment costs.

To migrate to switched unicast, operators incur both a data plane cost and control plane investment cost. The data plane costs largely consist of capital equipment (QAMs, streaming servers, etc.) The control plane costs largely consist of software enhancements to existing SDV server platforms to enable unicast signaling. Taking advantage of the fact that control plane components such as the SDV session managers and resource managers are already essential ingredients of switched multicast video services, the additional cost for providing switched unicast control plane infrastructure is not significant. For our analysis,

the cost of SCTE-130 targeted advertisement control plane components was factored in with an advertising operating margin of 80%. Major data plane expenditures came from spectrum, QAMs, streaming servers and other transport costs.

Comprehensive analysis has been done evaluating the cost and benefits of various technologies that can squeeze more bandwidth out of the cable plant [2]. Among 1 GHz node upgrades, all digital conversions, node splits, advanced video coding, 1024QAM etc, node splits stand out as the attractive approach with an appealing cost to benefit ratio. In our cost analysis, we assume the plant starts with a 1000 HHP service group size, which is the equivalent of 500 tuners per service group with 60% subscriber penetration and 60% digital penetration. Assuming a maximum 30 RF channels are available for digital linear video service in a service group, node splitting to smaller service groups will be used until there is enough spectrum for providing the switched unicast service. The estimated cost of node splitting is $15,000 per split.

Although advanced video coding saves 50% bandwidth when compared with MPEG-2 video, it is not currently widely used in the linear video service because of the broadcast nature of the service today. A lowest common denominator is chosen in terms of STB capabilities. In other words, as long as legacy MPEG-2-only STBs are in the field, the video must be offered in the MPEG-2 format. This situation is changed with the introduction of switched unicast. With switched unicast, AVC coding can be used when delivering unicast video to newer AVC-capable STBs, while the MPEG-2-version of the program can be used when delivering to legacy MPEG-2-only STBs. In the cost analysis, we assumed 20% of STBs as AVC capable.

The result of per subscriber cost analysis is shown in Figure 6.

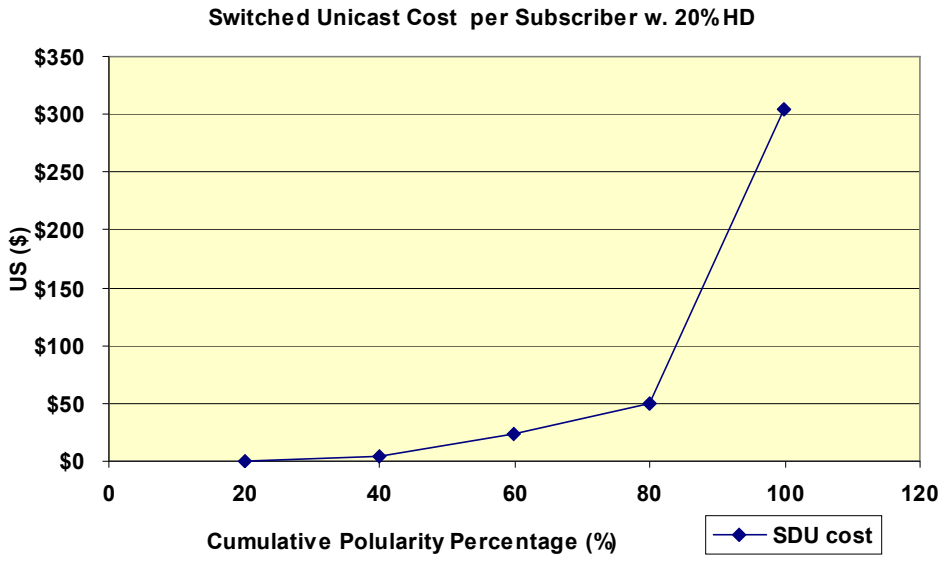


Figure 6. Switched Unicast costs per popularity quintile

As one would expect, switched unicast per subscriber costs increase as more popular content is offered as unicast. The biggest jump comes when the top 20% most popular channels are offered as unicast.

Targeted advertising is often touted as the next revenue growth engine for cable MSOs. Switched unicast is an enabling platform for targeted advertising. In 2006, worldwide cable TV advertisement spending totaled $24 billion, with nearly $5 billion contributing to MSO revenues. With only a 20% share of the total spend, targeted advertising could be an effective vehicle to improve this number by providing a higher-value product. [3]

When estimating the ad revenue potential of switched unicast, we apply the advertisement CPM (cost per thousand impressions) improvement of 2x for targeted advertising. Currently, in a typical cable network [4], per subscriber advertising revenue is $60. Since the MSO's share is just 20% of per subscriber ad revenue, the total per subscriber ad revenue is calculated as $300. The ad revenue modeling assumes that 30% of placement opportunities in switched unicast are targeted and assumes that MSOs have a 30% placement opportunity split with broadcasters and cable programmers for non-local ads. Then the potential MSO per subscriber ad revenue increase can be derived. However, the uneven distribution of unicast viewers with regard to the channel popularity complicates the calculation. The additional ad revenue of a node is calculated next as the potential per subscriber ad revenue times the unicast viewers in the node. Lastly, this ad revenue of the node is averaged to compute the switched unicast ad revenue per subscriber. The projected ad revenue is shown in Figure 7.

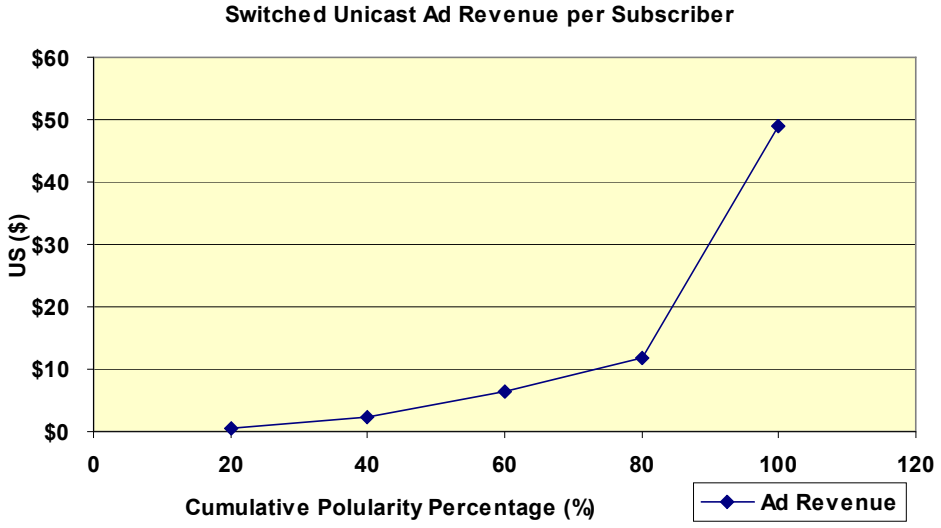


Figure 7. Switched Unicast potential revenue per popularity quintile

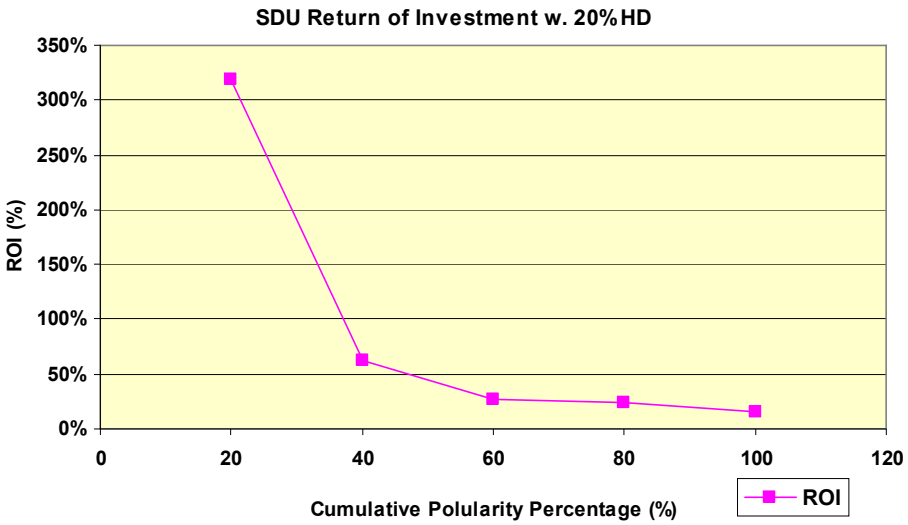


Figure 8. Switched Unicast ROI per popularity quintile

The return on investment results are illustrated in Figure 8. The diagram reveals that the ROI decreases as more popular channels are offered as unicast. In spite of the decrease, even when all channels are offered as switched unicast assuming 20% HD channels, with the submitted parameters the ROI can be demonstrated to be as good as 16%.

One interesting fact to notice is the sensitivity to HD channel percentage in the channel lineup. Figure 9 clearly demonstrates that as the percentage of HD channels in the offering increases, the cost for switched unicast increases dramatically. In fact, if more than 30% of the 247 channels are HD channels, bandwidth and spectrum requirement will push the service group size to below 125 tuners if the majority of the STBs are legacy MPEG-2 only STBs. This can be intuitively understood by the fact that an HD program can consume 4 times the bandwidth of an SD program, but may not necessarily command a 4x premium for spot ad insertion.

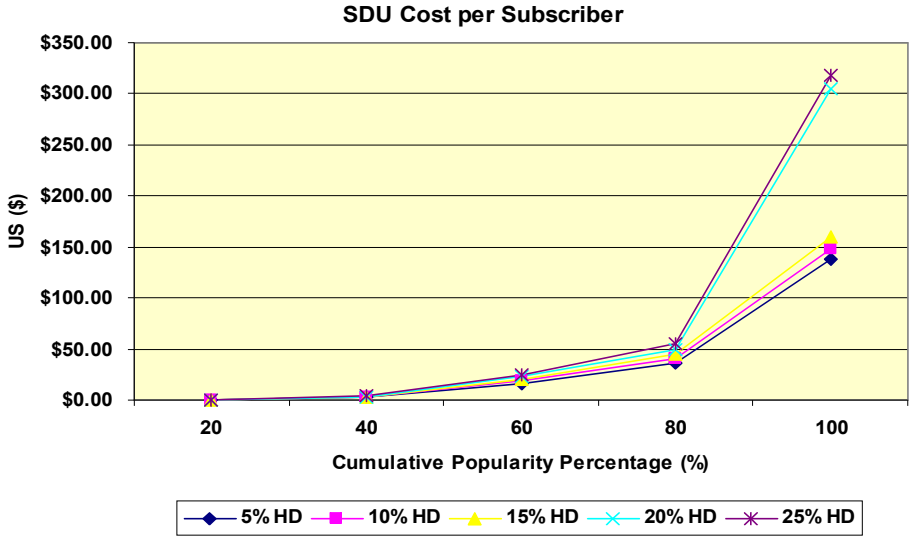


Figure 9. Switched Unicast costs with varying percentage of HD programming

What is the best strategy for MSOs?

The previous analysis provides insight that supports a recommended switched unicast strategy. First, although there is little doubt that the ad revenue potential of switched unicast is well worth the investment, switched unicast can be offered gradually with the least popular channels offered in unicast first. This approach is even more appealing if we consider that the least popular channels also have more local ad inventory accessible to MSOs. In a switched video architecture, a switched digital video session manager could implement a policy control to prioritize the less popular channels in unicast first.

Second, the number of HD channels offered in the switched unicast must be carefully evaluated to maximize the ROI. On one hand, it is tempting to offer more HD channels in switched unicast. On the other hand, HD channels consume much more bandwidth than SD channels. For MPEG-2 video, the HD bandwidth is roughly four times the SD bandwidth. Unfortunately, in today's advertisement arrangement, there is no revenue premium to insert ads in HD channels as opposed to SD channels. The proliferation of AVC STBs does alleviate the problem by reducing the HD bandwidth need by half. Additionally, MSOs may find creative ways of getting more revenue from HD ad insertion.

Third, even though the spectrum capacity can be increased by node splitting to get an HFC plant unicast ready, a better way might be to reclaim analog channels first to free up some spectrum. In the cost analysis, we assume that there are 30 RF channels available for linear digital video services since most of the RF spectrum is used by the analog tier. With analog reclamation, more spectrum will be available for linear digital video services and fewer node splits would be needed.

HYBRID UNICAST-MULTICAST DELIVERY SYSTEMS FOR MAXIMUM BANDWIDTH OPTIMIZATION

It has been observed that an appropriate premium incentive for a targeted advertisement service can justify the infrastructure investment. What guidance does such a statement provide to an MSO? Should an operator wait until a certain CPM threshold is crossed before deploying a targeted advertising solution?

It is possible to consider another variation of switched unicast. This variation can be easily understood by examining Figure 10.

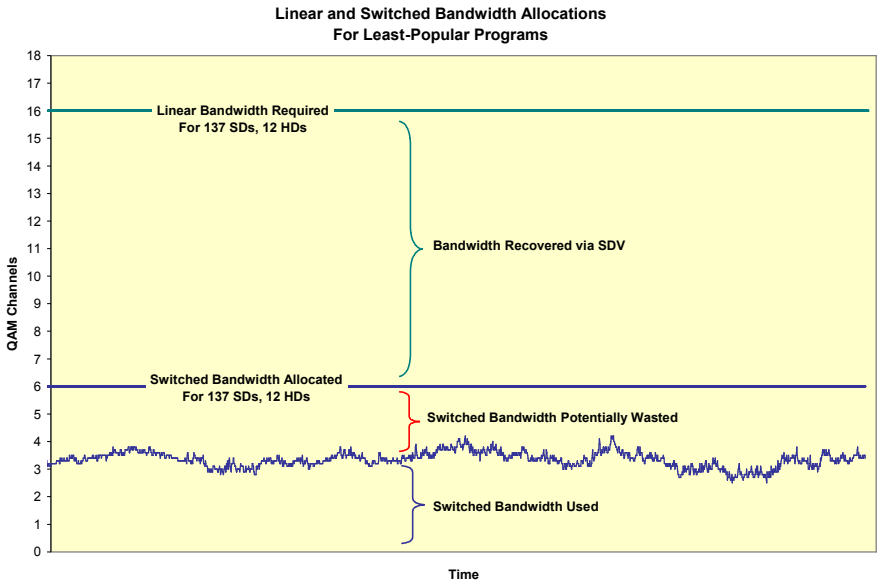


Figure 10. Switched Multicast Bandwidth Utilization

Indeed, Figure 10 shows that SDV saves bandwidth over its linear video equivalent, but what of the spectrum capacity that is allocated but not used? This represents a lost opportunity for better bandwidth utilization. To be more demonstrative, one could argue that this too is *wasted* bandwidth. The counter-perspective is that the unused bandwidth provides a "buffer zone" to protect against possible resource overflows. This is true, but the design of SDV systems demands that enough QAM resource be provisioned to handle the peak SDV consumption periods (which is typically during the evening/prime time hours). During non-peak hours, the bandwidth utilization can be quite low.

A hybrid multicast-unicast switched digital delivery system can provide the optimum blend of content personalization and bandwidth utilization in every situation. In such a system, a SDV server monitors resource utilization, and uses this indication to determine whether to respond to a channel request with tuning parameters for a unicast or multicast stream. When bandwidth utilization is low, for example in non-peak hours, the system can respond to channel requests by creating a personalized unicast channel. If bandwidth resource utilization reaches a (configurable) threshold, the SDV server can respond to subsequent channel requests with tuning information for a non-personalized multicast or "shared" stream. Since the multicast stream is shared by all of the subsequent users, this provides an effective "safety valve" to cap the stream usage when necessary, while offering maximum opportunities for personalization. As the subscriber population churns through channel changes, the aggregate number of active unicast channel streams will reduce by attrition, and the system can vary the unicast/multicast stream mix through a natural feedback process to manage the bandwidth utilization and the personalization opportunities to an optimal level.

What would a personalized vs. non-personalized channel look like? An example (one possible embodiment) is shown in Figure 11 [5].

Figure 11. Screenshot of linear channel vs. personalized channel

Multicast/unicast stream mix is an example of policy-driven resource management. In this example, allocation of edge resources is driven by a policy that seeks to optimize the insertion of unicast streams while observing rules on maximum bandwidth limits. But the rules do not need to be this simple. Future evolutions of this stream selection function could use more deeply sophisticated decision-making algorithms, weighing factors such as subscriber profile, program type, advertising value, time of day, and other factors to hyperoptimize the allocation of shared resources.

CONCLUSION

Switched Unicast is an extension of Switched Digital Video that enables content personalization and targeted ad insertion. While the main goal of most SDV deployments is centered around cost-effective programming expansion, switched unicast offers a revenue generation opportunity through targeted advertising and interactive services. Analysis of actual subscriber channel-change log data can provide valuable insight to the viewing patterns that might be expected in a switched unicast environment. Hybrid multicast/unicast implementations offer an opportunity to incrementally explore the value proposition of interactive programming by enabling fractional levels of personalization with a modest incremental investment to current SDV systems.

REFERENCES

[1] “A Comparison of Edge Bandwidth Requirements For Unicast versus Multicast Delivery of Switched Digital Video Services”, Robert Kidd and Michael Shannon.

[2] ”An Evaluation of Alternative Technologies for Increasing Network Information Capacity, Ron Shani and David Large”, NCTA 2005

[3] ”Irreconcilable Differences or a Match Made in Heaven? The Future of Advanced TV Advertising”, Ben Hollin and John Morrow, SCTE ET 2008

[4] Comcast 2007 Financial Results

[5] Screenshot courtesy ICTV and Turner. Use of this screenshot does not necessarily imply endorsement by ICTV or Turner of Cisco and/or the concepts presented in this paper.

VARIABLE BIT RATE VIDEO SERVICES IN DOCSIS 3.0 NETWORKS

Xiaomei Liu, Cisco Systems
Alon Bernstein, Cisco Systems

Abstract

DOCSIS 3.0 and Modular CMTS promise to provide ten times the bandwidth at one tenth of the cost, compared to existing CMTS technology. With these forward-looking trends, it is becoming increasingly viable to consider a channel-bonded DOCSIS network as a fully-converged network to transport video, voice and data. A bonded, converged and asynchronous data pipe, married with variable bit rate (VBR) video coding, can deliver the full potential of IPTV over cable.

This paper examines the technical and economic implications of VBR over DOCSIS. It proposes an IP-level VBR network statmux to deliver VBR video over channel-bonded DOCSIS and quantifies the efficiency of the network statmux with the results of lab tests. It provides insights into various architecture issues related to VBR delivery. Finally, it explores mechanisms to improve robustness and enhance the subscriber viewing experience.

INTRODUCTION

Cable networks are experiencing an explosion in demand for increased bandwidth. A significant amount of bandwidth pressure comes from High Definition Television (HDTV) service expansion, which MSOs have used as a strategic move to compete with satellite and telco video service providers. Today, the 100+ HD channel service is on the horizon as more HD content is offered. Meanwhile, content personalization and targeted advertisement are gradually tranforming the video delivery vehicle from broadcast to unicast. Yet over-the-top video services and user-generated video content simultaneously drive bandwidth demand with millions of video assets streamed or downloaded to PCs.

Cable operators have many tools to address the overwhelming bandwidth crunch problem. Some of these approaches in the MSO toolkit include: analog channel reclamation, switched digital video, node splitting, plant upgrades to 1GHz, and MPEG-4 part 10 video coding. Although cable operators can drill for additional bandwidth in HFC networks with major capital expenditure, there is work that can be done to eliminate any bandwidth inefficiency in HFC networks first.

Starting with video sources, it is common knowledge that variable bit rate (VBR) encoding of video is significantly more efficient than constant bit rate (CBR) encoding. In MPEG video encoding, while the CBR video keeps the bitrate constant, the VBR video attempts to keep the video quality constant. The nature of MPEG video encoding allows encoders to use fewer bits for simple scenes and more bits for complicated and motion rich scenes. With comparable video quality, VBR can yield 40 percent or more bandwidth savings over CBR [1]. As a result of its coding efficiency, VBR video is widely used in DVD and in broadcast video applications such as digital satellite and cable.

The introduction of broadcast-oriented MPEG statmuxes paved the way for delivery of VBR streams in broadcast video. Since most transmission channels have fixed bandwidth, MPEG statmuxes combine a number of VBR streams into a single aggregated constant bitrate channel. The statistical distribution of bitrate peaks and valleys allows the combined streams to use less bandwidth than what is needed if each VBR stream is sent individually. At any

given point in time, if the bandwidth of a VBR bundle exceeds the capacity of an MPEG transmission channel, the MPEG statmux applies requantization at the MPEG level to reduce the instantaneous bitrate of video streams to fit the transmission pipe. This action does come at the expense of a non-zero impact to video quality.

Ironically, in advanced video services such as Switched Digital Video (SDV) and Video on Demand (VOD) where the last mile bandwidth efficiency is needed the most, CBR instead of VBR video is deployed universally today. This is because SDV and VOD present a challenging case for traditional broadcast oriented MPEG statmuxes:

1) MPEG statmuxes are computationally intensive and costly. MPEG statmuxes achieve rate reduction by transrating selected MPEG frames. Transrating is an expensive operation as macroblocks in pictures are re-quantized and re-encoded at the MPEG level. Although the cost of MPEG statmuxes is not a concern in a broadcast network as the per stream statmux cost is shared among all the subscribers in the network, quite the contrary is true in the SDV and VOD world. In an increasingly unicast-based video delivery network, the per stream statmux cost now becomes a per-subscriber cost, which makes current MPEG statmuxes economically impractical to deploy at the network edge.

It is already a complicated operation to statistically multiplex MPEG2 encoded VBR streams; it is an even more daunting task to perform statmux on MPEG4 encoded VBR streams because of the incrementally intensive video computations involved.

2) MPEG statmuxes apply extensive stream analysis in order to mitigate the video quality degradation caused by the transrating operation. The stream analysis as well as the transcoding operation introduces delays, typically on the order of 1 second. This long latency is more noticeable and undesirable for on-demand and interactive video services.

3) It is also no surprise that traditional MPEG statmuxes have difficulties dealing with encrypted content considering that the rate reduction techniques involved need to analyze and re-encode the stream content. For example, pre-encrypted VOD content makes the elementary MPEG stream inaccessible for transrating.

The business and technical issues pointed out above have forced cable operators to give up VBR efficiencies and opt instead for CBR video delivery in switched and on demand video services.

These assumptions change as video over DOCSIS becomes a reality. Not only does wideband DOCSIS provide an IP transport to MPEG video, it also brings along a promising new way of VBR statistical multiplexing.

DOCSIS 3.0 AND NETWORK STATMUXES

DOCSIS 3.0 is perhaps the most anticipated technology of the year in cable industry. DOCSIS 3.0 takes the DOCSIS beyond just an IP transport for data and voice services. IP video over DOCSIS is rapidly gaining traction with DOCSIS 3.0. In fact, some MSOs already have started market trials and deployments of IPTV over DOCSIS are in the planning stages.

Partly because of DOCSIS 3.0 channel bonding and DOCSIS 3.0 enhanced multicast, IP video over DOCSIS becomes a feasible technical possibility. Channel bonding, the most important feature of DOCSIS 3.0, makes the channel bandwidth a magnitude higher than before by allowing CMTSs to bond multiple downstream and/or multiple upstream RF carriers in order to deliver higher bandwidth to

the home. Today, eight channel bonded cable modems are already available in the market, enabling downstream bandwidth rates of around 300Mbps. This increased bandwidth capacity is essential to bandwidth-hungry applications such as standard definition (SD) and high definition (HD) video.

DOCSIS 3.0 enhanced multicast adds source-specific multicast (SSM) and Internet group management protocol (IGMPv3) support. In addition, multicast sessions can also be managed with quality of service (QoS) guarantees. Switched multicast in an IP/DOCSIS transport increases bandwidth utilization efficiency in the same way as switched digital video does in an MPEG transport.

Equally important in terms of their potential technological impacts on the industry are modular CMTS (M-CMTS) and universal quadrature amplitude modulation devices **(**QAMs). The separation of the DOCSIS media access control (MAC) and physical layer protocol (PHY) allows independent scaling of upstream and downstream bandwidth. Economics of scale will drive down the costs of universal QAMs, lower the overall solution cost of M-CMTS and make DOCSIS economically viable for IP video delivery.

IP Transport vs. MPEG Transport

IP transport distinguishes itself from MPEG transport in a number of ways. IP transport is asynchronous packet-oriented transport. IP networks also introduce jitter. MPEG transport, on the other hand, is synchronous transport. When MPEG transport streams are delivered over IP networks, receivers must remove network jitter in order to recover the original video source clock. IP set-top boxes (STBs) normally have dejittering buffers that can tolerate 100ms of network jitter.

IP/DOCSIS 3.0 transport supports wideband transmission with bandwidth upper capacity limited only by cable modem technologies. Downstream bandwidth speeds of 300Mbps are enabled by today's eight-channel cable modems. It is only a matter of time before much higher bandwidths are available as Moore's Law keeps bringing down the cable modem cost. MPEG transport, on the other hand, which does not support channel bonding, has a bandwidth limitation of a single QAM channel. In North America, the bandwidth of a single QAM channel is capped at 38.8Mbps with QAM256 modulation. The Law of Large Numbers implies that statistical multiplexing efficiency improves as the number of VBR streams in the transmission channel increases. The proliferation of HDTV in households and the increasing number of HD streams in the network make wideband transport much more attractive for the purpose of statistical multiplexing. While the narrowband MPEG transport struggles to provide efficient statistical multiplexing of HD streams without compromising the video quality, the bonded DOCSIS 3.0 transport can easily support statistical multiplexing of HD streams with good statistical multiplexing gains and video quality.

IP networks also have built-in quality of service (QoS) capabilities. Cable modem termination systems (CMTSs) implement advanced DOCSIS QoS features such as admission and policy control, priority queuing, traffic policing, traffic shaping etc. These IP network features are readily applicable for VBR video delivery.

IP networks also are converged networks. They allow data, voice and video to be simultaneously delivered through the network. Converged networks provide great bandwidth savings just by enabling all

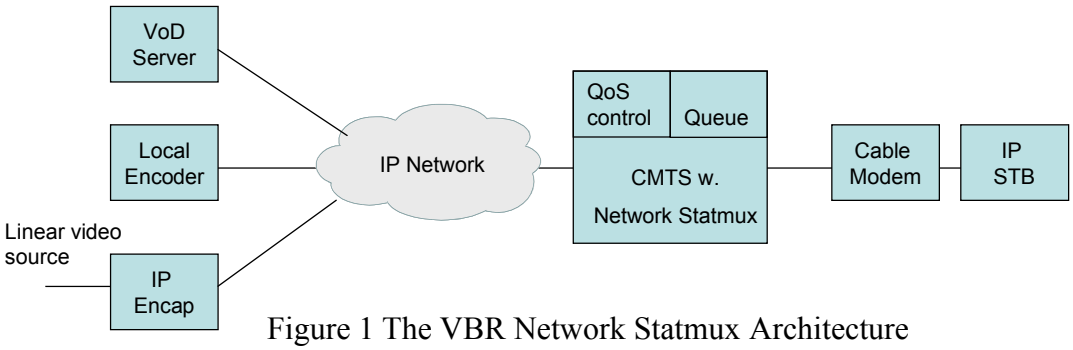


Figure 1 The VBR Network Statmux Architecture

services to share a single bandwidth resource pool. If the slightly different peak hours of these different services are also considered, the bandwidth savings are even greater. Field data indicates that bandwidth savings of 20 to 30 percent are achievable with a truly converged network. Even better is the fact that these savings are realized on top of the VBR over CBR bandwidth savings. In a converged IP network, if at a certain instant there is leftover bandwidth after all the VBR video traffic, then lower priority data traffic can consume the unused bandwidth. In other words, not a single bit of bandwidth is wasted! MPEG transport networks, however, are special-purpose networks used for video delivery exclusively. In an MPEG transport network, either due to MPEG virtual buffer constraints or low instantaneous video bitrate, MPEG statmuxes must insert NULL packets to fill the MPEG transmission channel. The NULL packet filled bandwidth is simply wasted.

VBR Network Statmuxes

The characteristics of IP transport make it a perfect match for VBR statistical multiplexing. The essence of an IP-level VBR network statmux, or simply network statmux, is to avoid transrating at the MPEG-level in an attempt to solve the bandwidth oversubscription problem. Instead, queuing and buffering are used at the network edge. Figure 1 depicts a high level system diagram of network statmuxes.

In this architecture, all IP video sources transmit VBR streams. VOD servers store and stream VOD content in single program transport stream (SPTS) VBR format directly. The coding efficiency improvement of VBR over CBR bodes well for VOD as it brings 40 percent or more storage and streaming capacity savings to VOD servers. Local encoders encode real time video and send out SPTSs with desired VBR mean rates and peak rates. Linear video sources from satellite are converted from multi-program transport stream (MPTS) to SPTS and IP encapsulation is added at the same time.

The IP video streams then travel from the video sources through the network and arrive at the cable network edges where the last mile is the DOCSIS path. The CMTS, be it modular or integrated CMTS, is the starting point of the DOCSIS path. Powered with QoS control and advanced queuing features, the CMTS is an ideal candidate to implement VBR network statmuxing. At the CMTS, if there is no congestion, i.e. when the combined VBR stream bandwidth is less than the bandwidth limitation of the IP/DOCSIS pipe, all video streams pass through packet buffers inside the CMTS with minimum delays. When congestion occurs, the bursty VBR streams will be queued up at the

CMTS temporarily. In the extreme case that the CMTS buffer is full, packets will be dropped by the CMTS, though the probability of such drop to occur is controlled by an admission control function that limits the number of video flows admitted to a single pipe. The CMTS queuing will be further discussed in more details.

The IP video streams eventually terminate at IP STBs. The IP STB can either be a standalone IP STB behind a cable modem, or a hybrid STB with an embedded cable modem. As a result of CMTS queuing, additional network jitter is introduced to video streams. This jitter is absorbed by the IP STB as it dejitters and buffers packets before video is sent to video decoders. In order for VBR network statmuxing to work, there must be a limitation of the jitter introduced by the CMTS queuing so that the end-to-end network jitter is tolerable to the IP STB. In today's well managed service provider network, end-to-end network jitter is a magnitude lower than the dejittering capability of IP STBs. This leaves a big jitter budget for the CMTS to implement VBR statmuxing. For instance, the CMTS introduced jitter can be limited to 60ms. The jitter limitation is enforced on the CMTS by restricting the CMTS queue buffer size. The buffer size is chosen so that packets will stay in the buffer for less than the jitter limit time. The CMTS queue can be drained at the maximum bandwidth of the bonded DOCSIS channel.

The VBR video traffic should be marked with higher priority than data traffic either by differential services control point (DSCP) marking, or video flows should be explicitly identified through flow specs. In a converged IP/DOCSIS pipe, any underutilized bandwidth resources after the VBR video traffic can be consumed by lower priority data traffic.

Besides the number of streams participating in statistical multiplexing, another crucial factor that affects the statistical multiplexing efficiency is the VBR peak (bandwidth) to mean (bandwidth) ratio. The higher the peak to mean ratio, the less efficient the statistical multiplexing is. Early research [2] has shown that the MPEG-2 encoded broadcast quality VBR video has a typical peak to mean ratio from 1.3 to 2.4. IP network statmux design assumes video sources have peak to mean ratios within the limit of 2.4. This peak to mean ratio should be enforced at the video encoder for best video quality. Although service providers can use rate clampers along the video transmission path, the method of rate clamping at the network is inferior to the encoder peak rate enforcement solution as a consequence of additional video processing cost and degraded video quality.

The buffering and queuing scheme brings significant advantages to IP network statmuxes over traditional MPEG statmuxes. First, network statmuxes preserve original video quality keeping the video unchanged at MPEG level. Video is produced by video sources and consumed at STB receivers. No network components within the transmission path re-encode the video content between video sources and STBs. For the same reason, pre-encrypted content can now be easily multiplexed by network statmuxes. Second, network statmuxes introduce delays bounded by the network jitter limit, which make them ideal for low latency video services such as VOD and interactive TV. Lastly, by avoiding the expensive transrating operation and by leveraging the built-in QoS capabilities of IP networks, network statmuxing turns out to be a cost effective approach to VBR video delivery.

As a summary, Table 1 below highlights the key differences between network statmuxes and traditional MPEG statmuxes.

	Network Statmux	Traditional MPEG Statmux
Statmux Efficiency	More efficient w. wideband DOCSIS	Limited by the narrowband MPEG QAM
Bandwidth Overflow	Buffer and delay	Transrate
Video Quality	As good as original stream	Quality degradation from transrating
Latency	Less than 100ms, e.g. 60ms	0.5-1 second
Pre-encryption	Transparent	Have difficulty
Cost	Buffering and QoS already built into IP transport No deep packet processing	Additional system components Deep packet MPEG level transrating
Bandwidth Utilization	100% with converged network	Null packet filling, suboptimal

Table 1. Key Differences - Network and Traditional MPEG Statmuxes

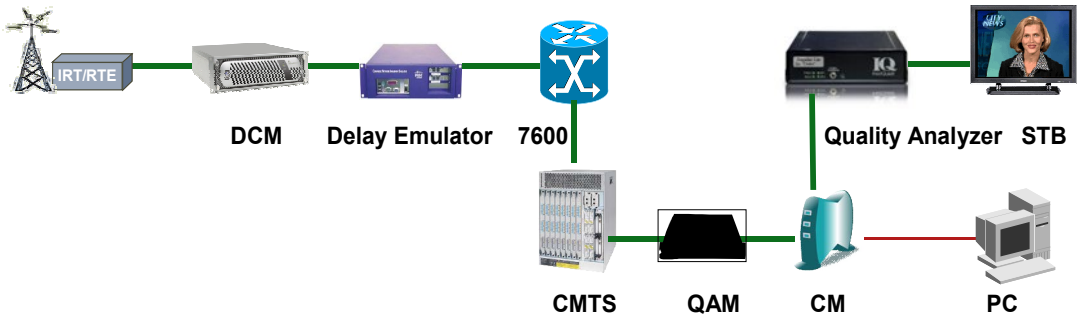


Figure 2 VBR over DOCSIS System Test Diagram

THE EFFICIENCY OF NETWORK STATMUXES

Exactly how efficient are network statmuxes? Perhaps nothing answers the question better than lab results from a proof of concept project. The basic design is to deliver VBR video streams into a controlled IP/DOCSIS channel with embedded QoS features. Streams are added to the channel one by one until the video quality is affected by packet drops. Based on the maximum number of VBR streams supported and the transmission channel bandwidth capacity, VBR statmux efficiency and bandwidth utilization improvement over CBR are calculated. To further study the effect of QoS control on VBR statmux efficiency, network buffers with different sizes are used to smooth out the VBR traffic.

Figure 2 presents the system diagram of VBR over DOCSIS testing. In this experiment, video sources are obtained from satellite feeds. Video streams are converted from MPTS to VBR SPTS and IP encapsulation is added by a Digital Content Manager (DCM). The VBR video streams then pass through a network delay emulator before they reach the DOCSIS CMTS. In the DOCSIS path, a pre-DOCSIS 3.0 modular CMTS based wideband solution is used. In this solution, the CMTS, the EQAM and the cable modem together form the DOCSIS last mile. VBR video streams are terminated at the IP STB.

A video quality analyzer is added to the path to monitor video impairments. A PC is used as a receiver of best effort data traffic.

Video Source

Live SD-only satellite feeds are used as the VBR sources. In the video industry, SD CBR MPEG2 video streams are encoded at a nominal rate of 3.75Mbps. The video quality associated with 3.75 Mbps CBR coding is well accepted as the standard for broadcast video. Assuming 40 percent VBR coding efficiency improvement over CBR, the VBR stream with equivalent video quality should have an average bitrate around 2.25Mbps. Two HITS satellite feeds selected for this experiment have just the right video characteristics. HITS1 and HITS9 each comes as an MPTS bundle from the satellite at 27Mbps. Each bundle has 12 video programs. The average bitrate is 2.25Mbps and the peak to mean ratio is 2 to 2.4.

Since these VBR streams originate from MPTS bundles, they are not good sources due to the correlated bitrate peaks and valleys. To create independent VBR streams with uncorrelated bitrate peaks and valleys, the network emulator is used to insert different delays to individual streams. For example, the first stream of the MPTS bundle is delayed 300ms, the second stream is delayed 600ms, the third stream is delayed 900ms and so on until the last stream is reached. The emulated delays combined with the live feeds generate the desired VBR sources.

Video Quality Measure

The IP video delivery quality can be measured with both packet drops and network jitter. As discussed earlier in the paper, the buffer sizes in the CMTS are chosen so that the maximum jitter introduced is bounded. The CMTS-introduced jitter will be removed by IP STBs without affecting video quality. Unfortunately, IP packet drops will cause considerable video quality impairments. To ensure that video quality degradation due to IP transport network is negligible from a subscribers point of view, most carriers allow the transport network to introduce at most one visible degradation in video quality every two hours. This criteria translates to 1E-6 maximum packet drop probability. Packet drops are detected from QoS counters either in the CMTS or in the video quality analyzer. In the test, VBR streams are inserted to the DOCSIS path one by one while packet drops are monitored. A VBR stream can only be added if this addition will not cause any packet drops for a two hour period. Besides, the probability of packet drop is measured over long term tests and is close to 5E-7, which is better than the well accepted 1E-6 criteria.

DOCSIS 3.0 Channel and QoS Options

Different DOCSIS 3.0 bonding group sizes are used to investigate the statmuxing efficiency with regard to channel bandwidth. To make it simple, the bandwidth increment is 38.8Mbps – the bandwidth equivalent of a single QAM.

The CMTS wideband default class queue is assigned to the VBR streams. Since all VBR video streams are delivered through multicast in the test, the classifier at the CMTS classifies all multicast video traffic to the default class queue. The default class queue size is adjusted to reflect the maximum jitter introduced by the CMTS. Best-effort IP data traffic is mixed with video traffic to drive the IP bandwidth utilization to 100%. While the default class is used to transmit video traffic, the best effort class queue is used to transmit unicast traffic to PC.

Results and Discussions

To facilitate discussions, two quantitative measures are defined. VBR statmux efficiency is defined as

Efficiency (%) = (#VBR streams * VBR average rate) / channel capacity

where *#VBR streams* is defined as the maximum number of VBR streams that can fit in a transmission channel without causing packet drops for a two hour window. The theoretical limitation of VBR statmux efficiency is 100% if one considers CBR as a special case of VBR with a peak-to-mean ratio of one.

The bandwidth utilization improvement is also derived by comparing VBR statmuxing with a CBR based solution. The bandwidth utilization improvement is defined as

Improvement (%) = (#VBR streams - #CBR streams) / #CBR streams

where the *#CBR streams* is defined as the maximum number of CBR streams with equivalent video quality that can fit in the same channel.

The experiment results are displayed in Figure 3 and Figure 4. Figure 3 highlights the VBR network statmux efficiency vs. DOCSIS channel bandwidth. Figure 4 shows the VBR network statmux bandwidth utilization improvement over CBR vs. DOCSIS channel bandwidth.

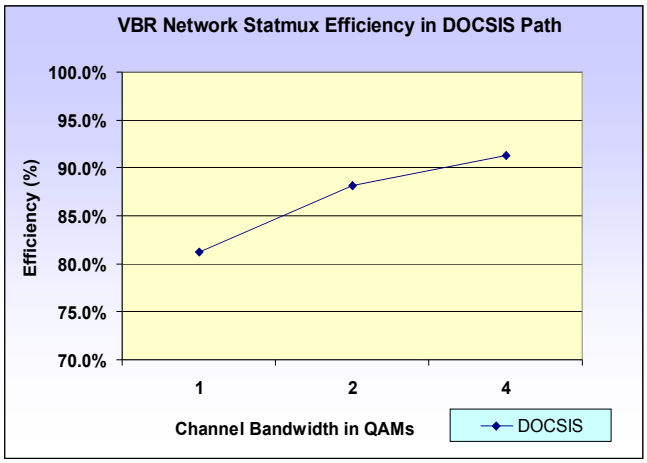


Figure 3. VBR Network Statmux Efficiency in DOCSIS

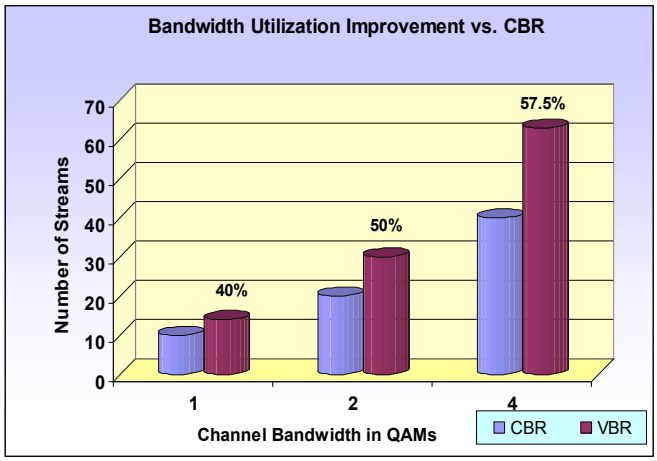


Figure 4. VBR Network Statmux Bandwidth Utilization Improvement

Several conclusions related to network statmux efficiency can be readily drawn from these results. First, the network statmux efficiency improves when the transmission channel bandwidth increases. With 60ms QoS buffers, the statmux efficiency is 81.2 percent for a DOCSIS channel of 38.8Mbps and the efficiency rapidly reaches 91.3 percent when the DOCSIS channel bandwidth is 232.8Mbps with four bonded QAM channels. Quite contrary to our original assumptions, VBR network statmuxes are efficient for SD-only content even when used with unbonded DOCSIS channels.

Next, network statmuxes dramatically improve the bandwidth utilization over the CBR solution. While the unbonded DOCSIS channel delivers 40 percent more streams if VBR and network statmuxing are utilized, the four channel bonded DOCSIS can boast a 57.5 percent enhancement. This improvement is superior to what traditional MPEG statmuxes can achieve in the MPEG transport path. Because network statmuxes only require queuing and buffering instead of the heavy MPEG level processing required by traditional MPEG statmuxes, the same or more bandwidth savings are achieved with only a fraction of the cost.

In addition, as the QoS buffer size increases, the network statmux efficiency further improves. This aspect of the testing was implemented on a simulated Gigabit Ethernet (GE) path with QoS buffer control. To demonstrate the QoS buffer effect, the buffer size is represented in terms of the maximum jitter introduced by the buffer. In Figure 5, when the bandwidth is two QAM equivalent, the statmux efficiency is 75.4 percent with a 4ms buffer and the efficiency is 86.9 percent with a 60ms buffer. However, as the channel bandwidth increases, the buffer size introduced improvement is reduced. For instance, if the channel bandwidth is 38.8 Mbps and the buffer size is increased from 4ms to 60ms, the statmux efficiency improves 11.5 percent. In contrast, when the channel bandwidth is 232.8Mbps, increasing the buffer size from 4ms to 60ms only yields about 4 percent improvement in statmux efficiency.

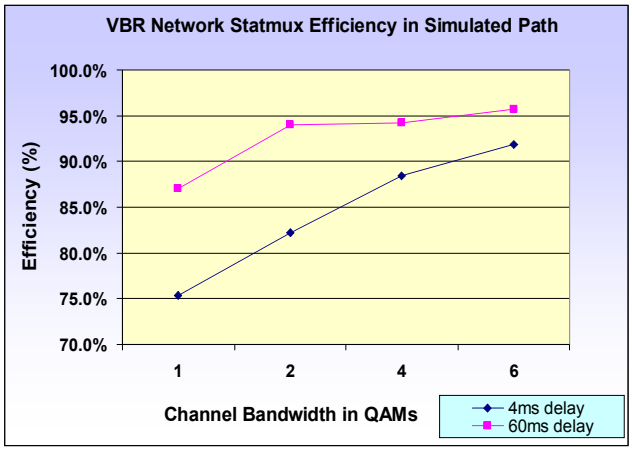


Figure 5. The Buffer Size Effect

This is not unexpected since the statmux efficiency is already very high when the channel bandwidth reaches a high threshold, thus, room for additional improvement is limited. Note that the statmux efficiency of an IP/DOCSIS path is slightly worse than that of the simulated IP/GE path as a result of additional DOCSIS overhead at layer one and layer two.

Finally, bandwidth utilization should not be confused with VBR statistical multiplexing efficiency. Bandwidth utilization of a transmission channel can reach 100 percent as long as the lower priority traffic can be mixed with video data and no single bit of bandwidth is wasted. There is no doubt that a converged IP pipe holds the promise to fully utilize the transmission channel. To prove the point, the PC behind the cable modem pulls big files from the CMTS through the bonded DOCSIS channel when the VBR video bandwidth utilization is as high as 90 percent. No packet drops are detected during the process due to the QoS features of the CMTS and the priority treatment of VBR video traffic.

RELIABLE VIDEO DELIVERY

Video streams are particularly sensitive to packet drops. Because of the high level of compression used in video delivery, even a single packet drop could result in significant video artifacts. There are three main sources of packet drops in IP networks:

- Because the core of the network is usually rich with bandwidth, packet drops at the core are usually not related to congestion. Instead, load balancing actions, route changes and/or temporary equipment failures could cause packet drops.

- The edge is relatively bandwidth-poor. It is the pipe to the subscriber which is at risk of being congested and as a result, it is where packets are most likely to be dropped.

- Media errors: though technically packets might be dropped on the Ethernet part of the network, most drops occur on the HFC network itself due to RF issues.

To minimize and possibly eliminate the video artifacts caused by packet drops, we propose a

three-tier approach which addresses all major sources of packet drops:

- Admission control: the role of admission control is to make sure that the network, or in the context of this paper, the CMTS specifically, can deliver content reliably.

- Scheduling: while admission control makes sure that we deliver content reliability, it is the scheduler that does the actual work of delivering the content in real time in a reliable way.

- Error repair: Error repair was designed to help in cases where packets are dropped because of media errors and/or network flaps, however, it could be used to help in cases where both admission control and scheduling could still not guarantee packet delivery.

ADMISSION CONTROL AND SCHEDULING

With CBR services, admission control is trivial. If a CBR flow requires X mbps, and the bandwidth of the channel is 10X Mbps, then 10 flows can be admitted. The CMTS would track the number of flows that the cable segment has to carry, and reject any request to activate an 11th flow.

When it comes to VBR the picture is more complex. As explained previously in this paper, VBRs can be oversubscribed because it is "reasonable" to assume that not all flows send at peak traffic rate all at the same time. But what does "reasonable" mean? There is a probability that enough traffic peaks occur at the same time and in such an event the channel will not have enough capacity to carry all the traffic. In such an event the CMTS scheduling queuing and admission control disciplines will help to minimize (or eliminate) packet drops in the event of traffic congestion.

To illustrate how a scheduler works, we can start with a simple example: assume a CMTS with a backhaul link of 1Gbps, and two video flows, each one with a traffic peak of 6mbps. Furthermore, we can assume that admission control limits the CMTS to accept only two flows for this case (though more could possibly have been admitted). Both flows drain onto a single output that is capable of 10mbps. We assume that the targets of these flows are two cable modems as depicted in Figure 6.

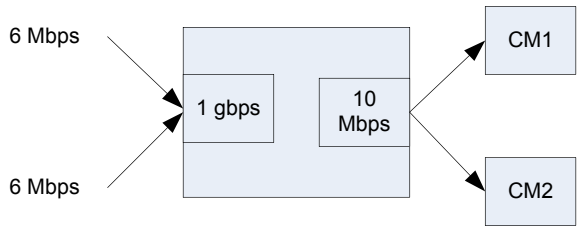


Figure 6. Queuing example

Since the video flows are 6Mbps each there is no congestion risk on the backhaul. However, when they get to the cable interface the worst case aggregate rate they can reach is 12Mbps while the cable interface in this example can support only 10Mbps. The tools the CMTS can use are queuing and scheduling:

- queuing will buffer up the packets in a "queue" until the 10 Mbps channel is available again to send them.

- scheduling will decide which queue to service and in what order

The CMTS can use the DOCSIS tools to define the queuing/scheduling structure needed to deliver these flows reliably:

- A classifier that will uniquely identify the video flow. For example, the combination of a destination IP address of the client device, and a destination user datagram protocol (UDP) port are a good way to identify a

packet stream that belongs to a single video flow. This approach can also be used for multicast flows. The detailed discussion of how this classifier is created is outside the scope of this paper.

- A service level agreement that defines how to queue the flow. For example, in our case it's a flow that has a 6 Mbps peak rate.

The CMTS manages queuing by dedicating a queue to each one of the video flows and by controlling queue scheduling.

Naturally all the queuing/scheduling can do is to mitigate the cases where the aggregate traffic bursts are above 10mbps. If the bursts are too long then packets will experience an unacceptable delay in the CMTS queues (and eventually will get dropped as the CMTS queues have limited size). By putting a limit on the number of flow admitted by admission control, the MSO can control the tradeoff of how many flows can be admitted to a channel vs. what the packet drop probability would be.

An additional tool that can reduce the risk of packet drops (at the expense of having less video flows committed) is the use of "guaranteed minimum rate". This parameter in DOCSIS defines the rate that a scheduler MUST deliver even in a case of congestion. In a way, one can view CBR as a case where the the peak rate equals the committed rate. Based on this, the smaller the difference between the committed rate and the peak rate, the smaller the risk of packet drops.

Another tool in the DOCSIS toolkit is "priority". This parameter is critical in an environment where we have mixed video with other services such as data. The priority parameter, along with the "guaranteed minimum rate" parameter, gives an assurance that even if data services in a given channel are congested to the point where packets are dropped, there is still enough bandwidth dedicated to high priority video flows.

However, even with proper admission control and scheduling, packet drops could still occur. Facilitated with extremely low packet drop rate, error retransmission and IP-based packet-level forward error correction (FEC) are promising cost-effective solutions for the packet drop problem. Both technologies are well understood and have been applied to IP video applications to protect video streams from common impairments of IP networks, such as packet loss.

MPEG over UDP/IP is widely used to transmit real time video traffic through IP networks. With UDP transmission, packet drops are not reliably detected due to the lack of a sequence number at the UDP layer and due to the limited capabilities of the MPEG transport stream level continuity counter. By adding a real time protocol RTP above UDP, packet drops are easily identified through a 16-bit sequence number at the RTP layer. Error retransmission and FEC leverage the RTP encapsulation of the video stream and repair dropped packets at the network edge. An example architecture for error retransmission and FEC and is shown in Figure 7.

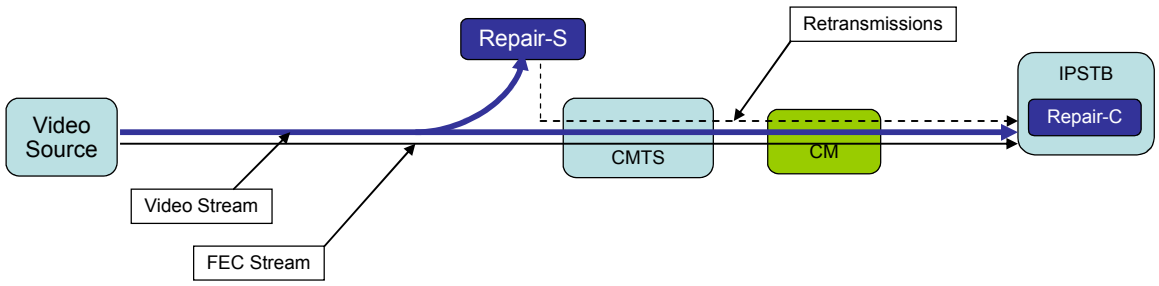


Figure 7. Error Repair Architecture

In the above error repair architecture (Figure 7), an error repair client is located at an IP STB. When FEC is applied, the video source sends out an FEC stream along with the video stream. In this scheme, periodically selected media packets are used to generate FEC packets. The error repair client is responsible for detecting packet loss and recovering the lost packets utilizing the additional FEC stream. When error retransmission is utilized, a repair server at the network edge caches video content. The error repair client utilizes standard based RTP/RTCP toolkit defined by IETF to request retransmission of lost packets. The same toolkit can also be used to accelerate channel changes in IPTV.

CONCLUSION

The ever increasing demand for bandwidth requires efficient HFC bandwidth utilization. DOCSIS 3.0 and IP video are shifting the VBR video delivery to a new paradigm. IP level VBR network statmuxes overcome the shortcomings of traditional MPEG statmuxes and provide the least expensive, low latency and best video quality approach to reap the benefits of VBR video.

The proof of concept work introduced in this paper not only proves the feasibility of this VBR network statmuxing in today's DOCSIS 3.0 networks, but also demonstrates the tremendous value and potentials of wideband DOCSIS in video delivery. DOCSIS CMTSs, with their built-in advanced QoS capabilities play an important role in VBR network statmuxing.

The trend in tomorrow's video delivery is more HD content and more advance coded video content. VBR network statmuxes respond to this trend by leveraging the channel bonding capability of DOCSIS 3.0 and generate unprecedented multiplexing efficiency as the wide channel is promising to get wider. By avoiding deep packet processing, VBR network statmuxes scale easily to future video coding technologies.

REFERENCES

1. Si Jun Huang, "Principle, Applications of Variable Bit Rate Coding for Digital Video Broadcasting, w. Statistical Multiplexing Extension", NAB 1999

2. Daniel P. Heyman and T.V. Lakshman, "Source Models for VBR Broadcast Video Traffic", IEEE Transactions on Networking, Feb. 1996

VIDEO LAYER QUALITY OF SERVICE: UNPRECEDENTED CONTROL AND THE BEST VIDEO QUALITY AT ANY GIVEN BIT RATE

Ron Gutman,
Marc Tayer

Imagine Communications, Inc.

Abstract

Spurred by DirecTV's 2007 declaration that it will be the world's first television service provider to reach 100 HD channels, cable operators are moving rapidly to create additional bandwidth not only to carry dozens more linear HD channels, but also to provide hundreds, and eventually thousands, of HD-VOD titles. The video quality bar is simultaneously rising due to the mass consumer adoption of large HDTV displays and the growing popularity of Blu-ray.

This paper discusses the fundamental and elusive paradox of how to cost-effectively increase bandwidth efficiency without sacrificing video quality. Leveraging a concept from IP networking, Video Layer Quality of Service (Video Layer QoS) involves creating minimum and maximum video quality values at the service level, while adding the technical dimension of true Variable Bit Rate (VBR) constant quality video coding.

Similarly, Video Layer QoS allows the optimization of bandwidth efficiency while guaranteeing the quality of service in a sustainable manner throughout the various switching, multiplexing and splicing stages of video communications networks. The paper also discusses the human visual perceptual system as well as related video processing and delivery aspects for a variety of digital video services.

INTRODUCTION

The North American market for video subscribers is becoming increasingly competitive and fragmented, with cable, DBS, telco and Internet service providers all jockeying to gain a bigger piece of a growing pie. After a long gestation period, the HDTV market is finally hitting its stride. The most successful service providers will offer libraries of virtually unlimited content delivered conveniently and with the highest possible video quality.

An important emerging element of this infrastructure is Video Layer QoS, defined as the establishment of video quality levels at the content origination or delivery site, combined with the process of sustaining these levels all the way to the consumer viewing environment in the most bandwidth efficient manner possible.

A Video Layer QoS solution provides:

1. Excellent MPEG-2 video quality at 3:1 HD and 15:1 SD (per 256 QAM channel) on an end-to-end basis, from content origination all the way to the set-top box.
2. Consistency (equalization) of a service provider's video quality across the Digital Broadcast, SDV, VOD and Network PVR (e.g., Start Over) categories.

3. The ability to assign different quality levels to different classes of assets (e.g., HD-VOD PPV), or even individual assets (e.g., the Super Bowl), at the discretion and under the control of the content provider or operator.
4. Sustenance of the pre-calibrated video quality levels in a cost effective, non-disruptive and backward compatible manner throughout the various multiplexing stages, including local and addressable ad insertion.

THE HUMAN VISUAL SYSTEM AND PERCEIVED VIDEO QUALITY

A logical place to begin a discussion of video quality is the area of human visual quality perception. The visual and perceptual system can not merely be construed in the context of resolution, frame rate and bit rate since these factors alone do not explain the phenomenon in which two streams with equivalent parameter settings can appear very differently to the human eye. The two streams may look quite similar most of the time, but the majority of subjects in a typical focus group will still select one sequence over the other.

When standing close to two identical screens positioned side-by-side, a trained set of eyes can begin observing the traditional compression artifacts. To mention a few notorious examples, many of us have observed blockiness at facial edges, in sky-dominated backgrounds, and during scene changes; random noise on football field grass or basketball courts; lack of detail, softness or the absence of a "pop" effect in a complex or colorful image; or tiling around logos or scrolling text areas. In many cases, the discerning viewer may need to wait for a period of high activity in the video stream in order to see artifacts, such as rapid motion, panning, scene changes, fades or flashes. This instability and unpredictability of quality over time can be quite annoying, and is highly correlated to consumer complaints regarding delivered video quality.

The following images show the same picture compressed with three different methodologies. In the first image, a "Compression by Quality" technique is used, in which an objective video quality measurement system is involved to minimize compression artifacts relative to the source. In the second image, a typical MPEG-2 encoder is used. In the third image, pre-processing and high frequency pixel filtering techniques are used in addition to the traditional compression methods.

The first image appears noticeably sharper and cleaner than the other two, showing neither the blocking artifacts of the second image nor the blurriness or loss in detail of the third image. All other things being equal, such an improvement in perceived quality can be made possible if the video quality has been exhaustively analyzed and measured as part of the video processing and multiplexing solution.

Figure 1 – Compression by Quality

Figure 2 – Typical Compression

Figure 3 – Compression by Pre-Processing and High Frequency Pixel Filtering

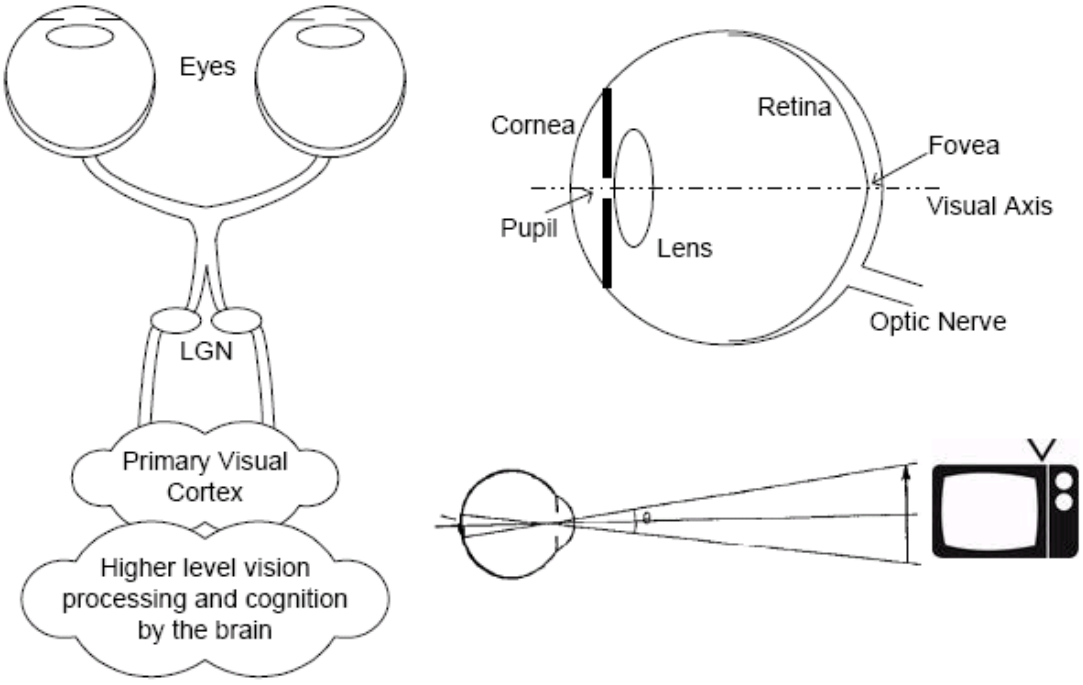


Figure 4 – Human Visual System (HVS)

In this manner, it is finally possible to guarantee the Video Layer QoS at any second, at any frame and even at any pixel. But before jumping more deeply into video quality measurement, we must first briefly discuss the anatomical and psycho-visual features of the human visual system (HVS).

Visual stimuli, in the form of light and images coming from a TV screen, are focused by the optical components of the eye, including the cornea, pupil, lens and eye fluids. These stimuli are then translated into electrical signals by the neurons and photoreceptor cells in the retina, before being received and organized by the lateral geniculate nucleus (LGN) in the brain. The LGN then transmits these signals to the primary visual cortex which tunes and processes them into spatial and temporal frequencies, orientations and motion. Higher levels of visual processing, cognition and memory associations subconsciously analyze these complex information streams while the viewer relaxes comfortably on his or her couch watching television.

Each layer of visual processing or compression removes unnecessary levels of informational redundancy, forming the video signal into data that are essential for human interpretation, i.e., entertainment or viewing satisfaction.

The more redundant information that can be removed during the video compression process, the fewer bits are required to be delivered over a communications network or stored on a storage medium. The ubiquitous audiovisual coding standards of MPEG-2, and increasingly MPEG-4 AVC (H.264), are designed specifically for this purpose. However, with TV screens becoming increasingly larger, compression-related

Human Visual System (HVS) Feature	**Compression & Quality Measurement Guidance**
Eye optics modeled by a low-pass point spread function (PSF)	Caveat: making tradeoffs against picture sharpness is a risky area
Non-uniform retinal sampling	"Compress the perimeter more" trick has some utility but is not working as well as in previous video quality tests
Luminance masking	Potentially ripe area; extreme darks and lights can be compressed more
Spatial frequency, temporal frequency and contrast sensitivity functions	Another ripe area, but needs adaptation to specific MPEG-2 and MPEG-4 AVC compression impairments
Masking and facilitation	Some image components do a good job of masking the visibility of others. Very difficult area to model and compute.
Neural pooling (cognition layer)	Perceptible distortion is more annoying in some areas of the scene (human faces, text, sea or sky background) than in others.

Table 1 – Compression Tricks Relative to the Human Visual System (HVS)

artifacts which may previously have been relegated to the category of acceptable or imperceptible marginal noise are now distinctly observable and in many cases even annoying to ordinary consumers.

Table 1 shows some known characteristics of the human visual system, and then comments on their potential effectiveness with respect to video quality measurement and image compression.

VIDEO PROCESSING AND MULTIPLEXING BY QUALITY

In this section we describe a method that allows "closing the loop" with respect to objective video quality measurement systems, significantly increasing the signal quality (and bit rate efficiency), and providing Video Layer QoS through the re-processing, re-multiplexing, VBR to CBR conversion (or vice versa) and splicing stages.

Step 1: Select or devise a video quality measurement technique

There are several subjective video quality testing methods that are accepted by industry professionals, such as the Double Stimulus Impairment Scale, the Double Stimulus Continuous Quality Scale, and other methods described in ITU-R BT.500-11. In contrast, objective video quality measurement methods, by attempting to correlate as closely as possible to subjective test results, are very elusive by definition. For this reason, through the history of digital video, subjective video methods have been heavily relied upon, with objective video quality method serving more as a sanity

Objective Video Quality Measurement Method	Computational Complexity	Correlation with Subjective Test Results
PSNR Peak-Signal-to-Noise-Ratio Most common. Based on mean squared error (MSE).	Simple	Poor , even imperceptible pixel errors contribute negatively to the measured result
MPQM Moving Pictures Quality Metric	Complex	Mixed. Certain parameters are incorporated.
VQM Video Quality Metric ANSI T1.801.03-2003	Very Complex	Good, measures perceptible impairments such as blurring, jerkiness and distortion.
SSIM Structural Similarity Index [3]	Complex	Fair, uses a structural distortion measure instead of error.
ICE-Q™ Interchangeable Compressed Elements-Quality	Very Complex	Excellent. Accounts for numerous visual impairments; designed and optimized specifically for MPEG-2 and MPEG-4 AVC (H.264)

Table 2 – Objective Video Quality Measurement Systems

check or rationale for adding certain compression tools to a standard. In other words, subjective video quality testing has been the litmus test up until now.

All objective video quality measurement methods use some form of HVS modeling. The more successful methods are backed by correlation with subjective video quality test results and have endured long periods of tuning. They are also very computationally intensive, in effect representing a form of artificial intelligence. Table 2 shows a summary of the known objective video quality measurement methods and their correlation

to subjective tests results based on personal experience as well as available information:

Note that all of these objective video quality measurement methods involve the comparison of two signals. For example, they may involve an uncompressed source vs. a compressed/decompressed signal, or a satellite-received compressed signal vs. a re-encoded signal. This remark becomes more relevant and important in subsequent stages of a signal path, in which the stream is re-multiplexed, potentially multiple times, before arriving at the consumer's set-top box.

Step 2: Video Processing or Encoding

For this step, a video processing device is required capable of "closing the loop" with the selected objective video quality measurement method. It requires processing of every frame and every macroblock of every frame, as part of the selection of a constant video quality requirement, level or "grade." Once the decisions are made, by iteratively comparing the re-processed options to the source, using the objective video quality measurement system as the arbiter, the resultant reconstructed signal is essentially guaranteed to be a constant quality signal at levels or grades which are known in advance. This signal is Variable Bit Rate (VBR) by definition since the activity and complexity vary over time. High complexity scenes will automatically be processed at higher bit rates than low complexity scenes, with both types of scenes being coded at the same measured quality level, hence the notion constant quality.

In great contrast to today's encoding or rate-shaping methods, the video processor is configurable to a pre-calibrated quality level rather than a maximum, minimum, or average bit rate. A recommended method to guide this process is to use a mathematical scale, such as 1 to 100, rather than more crude or subjective groupings such as "good," "bad," or "average."

Step 3: Calibration

During this stage, all of the available signals or video assets need to be processed (i.e., intelligently compressed using an effective objective video quality measurement system), using the video processor from Step 2, employed at various selected quality grades. The system should be calibrated in such a way that the service provider is reasonably comfortable with the constant quality experience at any grade. If this is not the case, then the previous steps should be repeated. One can define a minimum of two quality grades in a similar fashion to the QoS utilized in IP networks as follows:

1. QG_a – target average video quality grade, for example "96"
2. QG_b – Guaranteed or minimum allowed video quality grade, for example "90"

In some deployments, QG_a can be defined as "just noticeable difference" (JND), which means

even expert viewers (i.e., "golden eyes"), cannot see substantial differences from the source. Then, QG_b can be defined as the quality level or grade at which, the vast majority of the time, ordinary viewers can't discern differences from the source.

A good practice for delivering the signals, including packing density, suggests a target of no more than 1% of the time the video stream will contain QG_b.

Step 4: Statistics

Process all of the target channels or video assets at QG_a and QG_b and gather statistics for at least 24 hours, or preferably for one week. Measure the respective bit rates per second and create two vectors, one for each quality grade.

$B_a(t)$ – bit rate measured per second at QG_a
$B_b(t)$ – bit rate measured per second QG_b

Per channel, calculate your global (time tested) average bit rate at QG_a and your global maximum bit rate at QG_b.

BA_a = Average ($B_a(t)$)
BM_b = Maximum ($B_b(t)$)

BQ = Maximum (BA_a, BM_b) – defined as the channel effective bit rate for lineup allocation, utilized statically for digital broadcast and dynamically for VOD, SDV and Internet video.

It is also possible to correlate the bit rate statistics to time of day or type of program. Interestingly enough, the quality requirements during prime time are generally higher than average. In other words, the average bit rate of QG_a and the maximum bit rate of QG_b are higher in prime time; therefore, BQ should be calculated during this time window.

Step 5: Lineup

Determining the digital service combination per multiplex contains a goal of providing QG_a quality on average and never less than QG_b. The following equations can help optimize the multiplex lineup using the bit rate measurement statistics first. For example in 3:1 HD within a 256 QAM channel at 38.8Mbps:

$$\sum_{c=1}^{3} (BQ)_c <= 38.8\text{Mbps}$$

In order to guarantee the quality it is possible to simulate the statmux by repeating this calculation for every second in the database

$$\sum_{c=1,t}^{3} (B_a(t), B_b(t))_c <= 38.8\text{Mbps}$$

Select $B_b(t)$ only when needed and by measuring $B_b(t)$ usage at less than 1%.

Because of the natural statistical behavior of constant quality signals, it is advisable to have the largest number of signals per mux as possible.

Step 6: Statistical Multiplexing

Using the lineup as defined in Step 5, it is now time to actively statistical multiplex the streams. The encoders should be able to encode at multiple quality grades in real time and the statmux should choose the highest quality grade possible under the maximum channel bit rate constraint. The grades are expected to extend to the entire range between QG_b ("90") or even lower, through QG_a ("96"), and up to "100." The proportion of null packets should be very close to 0% at any grade under "100."

The statmux device should report the eventual quality grades utilized in the stream. Some of the channels may change their content type over time. HD channels currently using

upconverted SD content will use an increasing proportion of native HD content over time. Certain movie channels may rarely show concerts or sports events that are more difficult to compress, while other channels may alternate between movies, sports and concerts. It is important to monitor the average and instantaneous video quality grade for every mux, including the percentage of time the system is running at a grade under QG_a (expected to be less than 1%) and the percentage of time the system is running at a grade under QG_b (expected to be less than 0.1%).

A service provider may also choose to completely skip Steps 4 and 5 and base the lineup selections entirely on quality statistics rather than on the bit rate statistics. In this case, the process involves adding or subtracting one SD channel at a time to output muxes that are over or under the video quality requirement, respectively.

Although the statmux uses the entire mux bit rate to provide the highest quality, it is possible to assess the effective available bit rate according to the desired calibrated thresholds. If a certain mux consistently has average grades above the QG_a, there may be some available bandwidth for other services. The available bit rate can be computed by monitoring BA_a and BM_b in real-time even when the statmux is selecting other quality grades.

Guaranteed available mux bit rate = 38.8Mbps - $\sum_{c=1}^{3}(BA_a, BM_b)_c$

Average available bit rate (opportunistic data) = 38.8 Mbps – Average ($\sum_{c=1,t}^{3}(B_a(t), B_b(t))_c$)

The statmux device can also calculate in advance what it would take to convert any of the streams to a CBR. In this case, the minimum CBR rate would be BM_b under the CBR buffer model calculation, but when converting the signal into CBR the percentage of time at which it is running under quality grade QG_a might be significantly higher than 1%. It is possible to iteratively and heuristically determine the optimal CBR rate for QG_a and QG_b. Note that this bit rate is significantly higher than BQ, which is the effective bit rate in VBR. Since the CBR rates are generally expected to be 3.75Mbps for SD and 15Mbps HD, it is possible to calculate, in advance, the average quality and percentage of time at which streams are running at quality grades under QG_a and QG_b.

THE BOTTOM LINE RESULT: VIDEO LAYER QoS

Video Layer QoS provides an unprecedented level of control for a system operator or content provider, all the way from content origination to the set-top box. Assuming the IP and MPEG-2 transport layers are intact, this capability opens up new possibilities for ensuring video quality, not available with previous digital or analog delivery solutions. Technically, Video Layer QoS means maintaining the pre-determined quality requirements (QG_a and $QG_{b)}$ through the communications delivery network, including sustainability through the various re-multiplexing, splicing, encryption, edge statistical multiplexing, and VBR to CBR conversion for services such as Start Over and SDV.

In order to take advantage of this capability, the statmux device from Step 6 needs to convey the following information per service:

1. QG(t) – instantaneous quality per frame
2. QG_a – target average quality grade, for example “96”
3. QG_b – Minimum allowed quality grade, for example “90”
4. QCBR – target CBR rate in a multi-rate CBR switched environment, the rate that

will support QG_a on average and QG_b no more than 1% of the time.

5. BQ – channel effective bit rate for VBR lineup allocation in real time

The importance of sending QG(t) indications is crucial for maintaining ultimate video quality. As noted above, objective video quality measurement techniques compare two signals and it will be impossible to compare the target to the original stream at a receive site at the terminal of the network. Given the instantaneous quality per frame QG(t), it becomes possible to keep the quality within the target range, where it requires re-multiplexing, by repeating Steps 4-6.

AD INSERTION

There are two main approaches for ensuring video quality of advertisements during ad insertion. The first approach involves pursuing the highest quality possible for the ad, even at the expense of the underlying digital services not containing ads at the same time. In this approach, during the splicing period (the ad avail), the other streams are constrained to being multiplexed at QG_a and not higher.

The second approach involves equalizing the ad quality to the underlying stream quality to the extent possible. In this case, the ad is multiplexed at QG_a and not higher, or at the eventual average quality grade of the primary stream. In any case, the ads should be processed and stored on the ad server at the maximum possible bit rate and quality level, providing downstream flexibility. A third approach is to provide the advertiser with QG_a and QG_b on a per asset or group of assets basis.

CBR FOR VOD AND SDV

Applying Video Layer QoS to the conversion of VBR signals to CBR (for SDV and nPVR applications) is relatively straightforward, with the quality levels being calculated in advance at the content origination site as discussed in Step 6.

In some cases, due to content complexity and also the inherent nature of CBR, the selected CBR rates may need to be higher than the standard SD and HD rates of 3.75 Mbps and 15Mbps, respectively. With respect to VOD, since VOD assets are originally encoded in CBR, it is possible to insert the QG(t) information into the stored stream for downstream edge statistical multiplexing.

EDGE STATMUX

A state-of-the-art edge statistical multiplexer can increase, by up to 50%, the number of streams per QAM channel without quality degradation for VOD and SDV applications.

In order to simultaneously maintain the Video Layer QoS and optimize the bandwidth efficiency, it is important to also involve the Edge or Session Resource Managers (ERM/SRM). A brute force method involves simply allocating 15 SD "blocks" of 3.75 Mbps each per QAM channel (or 3 HD "blocks" of 15 Mbps each), i.e., tricking the system into thinking each QAM channel has available up to 56.25 Mbps.

A more intelligent design can allocate the service bandwidth according to each service's effective bit rate (BQ) as suggested in Step 4, and then load balancing the quality across the switched QAM channels, thereby guaranteeing Video Layer QoS. This method ensures the best quality at any given bit rate for edge and switched applications including VOD, SDV, nPVR, Switched Unicast and addressable ad insertion. The effective video quality will be significantly higher than today's capped quality at 3.75 Mbps. SDTV CBR and overall network

efficiency will be 50% better allowing 15 SD VBR streams per edge QAM channel.

INTERNET VIDEO

Using this approach in conjunction with the standard IP QoS mechanism ensures constant video quality and Quality of Experience (QoE) for video services over the Internet and to mobile device. The IP QoS guaranteed bit rate should be set to BQ and the maximum required bit rate for the service should be set to QCBR. In some preliminary assessment, it is shown that this approach not only provides the best video quality at any bit rate, but it also consumes 25% less bandwidth and storage.

CONCLUSION

The cable industry is in the midst of a dramatic transformation toward an increasingly competitive and complex environment. Multiple categories of digital television services will co-exist on a unified platform, including digital broadcast, VOD, SDV, nPVR, and Internet video, each of which will encompass standard definition and high definition signals.

This evolving comprehensive suite of services and architectures must be presented in a transparent and convenient manner to consumers, who now have multiple choices for their service provider. In this new environment, a key consideration and a competitive differentiator is the ability to provide true Video Layer QoS, combining control and optimal video quality across all categories with the utmost in bandwidth efficiency.

References:

1. *Objective Video Quality Assessment, by Zhou Wang, Hamid R. Sheikh and Alan C. Bovik , Department of Electrical and Computer Engineering , The University of Texas at Austin*
2. *Survey of Objective Video Quality Measurements, by Yubing Wang , EMC Corporation Hopkinton, MA 01748, USA*
3. *Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004*

NCTA, The Cable Show
May 18-20, 2008
New Orleans

Ron Gutman, CTO and Co-Founder
ron@imaginecommunications.com

Marc Tayer, SVP Marketing &
Business Development
marc@imaginecommunications.com

Imagine Communications, Inc.
2053 San Elijo Ave
Cardiff-by-the-Sea, CA 92007
(760) 230-0110

WHAT TECHNOLOGY WILL WIN IN THE BATTLE TO DELIVER BROADBAND VIDEO TO CUSTOMER DEVICES?

By Dave Lively, Cisco Systems
Marty Roberts, thePlatform

Abstract

Today, with consumers increasing their consumption of broadband video and with cable operators and programmers continuing their entry into the online video space, the need to understand content delivery options is paramount. Cable operators already have the network capacity for delivering the content. The question is one of where to store the content and stream it from. The first issue is whether to build an infrastructure using generic web streaming and download servers, or to build a content delivery network (CDN) to handle the job. Cable programmers often have relationships with commercial CDNs but they may not be efficiently leveraging their internal digital storage and streaming servers.

Peer-to-peer (P2P) also presents another option. Cable operators can build their own application that leverages P2P protocols. P2P eliminates the cost of storage and Gigabit Ethernet ports required when building a CDN by pushing that cost to the individual users (the service provider is essentially co-opting their users' PCs for the storage and streaming). But, this method incurs additional costs for more upstream bandwidth, and potentially dealing with network congestion. Plus, what's the incentive for users to "donate" a portion of their bandwidth and computing and storage resources on their PC? Hybrid models also exist, allowing operators to potentially leverage the best aspects of all technologies.

A media management and publishing system can give the cable operator or programmer more control over their delivery options. Traffic can by dynamically directed to files on different CDNs without consumers experiencing any quality impacts. Policies may be applied to media to automate the management and storage of old or unpopular media files. As decisions to switch to a new content delivery option arise, a media management solution can ensure the transition is easy for production staff and seamless for viewers.

This paper will look at the impacts on the network for both downloading content and streaming content, as well as using CDN technology versus P2P technology to actually deliver the content (whether it's being streamed "live," or downloaded for future viewing). Media management systems may be applied to provide additional control over delivery policies. Virtualization of content, storage, and applications can also be leveraged by cable operators and programmers for delivery of content and even web-based applications in the future.

INTRODUCTION

More and more, consumers are looking to the Internet to get their content. And while user-generated / contributed content sites like YouTube.com continue to dominate the online video market, the Internet is rapidly becoming a viable means to distribute premium studio content as well.

Top U.S. Online Video Properties by Videos Viewed Jan. 2008[i] Total U.S. – Home/Work/University Locations		
Property	**Videos (000)**	**Share (%) of Videos**
Total Internet	*9,814,010*	*100.0%*
Google Sites	3,363,335	34.3%
Fox Interactive Media	584,132	6.0%
Yahoo! Sites	315,001	3.2%
Microsoft Sites	199,288	2.0%
Viacom Digital	197,737	2.0%
AOL LLC	118,033	1.2%
Disney Online	95,041	1.0%
Time Warner - Excl. AOL	85,467	0.9%
ESPN	81,402	0.8%
ABC.COM	49,017	0.5%

Figure 1: Top U.S. Online Video Properties by Videos Viewed Jan. 2008

Top U.S. Online Video Properties by Unique Viewers Jan 2008[ii] Total U.S. – Home/Work/University Locations		
Property	**Unique Viewers (000)**	**Average Minutes per Viewer**
Total Internet	*139,521*	*206.3*
Google Sites	80,056	109.9
Fox Interactive Media	53,913	11.7
Yahoo! Sites	36,362	18.0
AOL LLC	21,859	7.4
Viacom Digital	21,690	33.0
Microsoft Sites	20,842	30.0
Time Warner - Excl. AOL	13,914	18.2
Disney Online	13,005	8.9
ESPN	8,798	15.9
Apple Inc.	8,743	21.2

Figure 2: Top U.S. Online Video Properties by Unique Viewers Jan. 2008

Major studios and content owners are looking to the web as a new outlet and method to monetize their content. With more consumers turning to the Internet as a source for video, the increasing load will force both the content owners and service providers to examine new technologies to handle the distribution effectively while maintaining a high quality and reliable consumer experience.

Today, "Internet Video" is largely a computer-only phenomenon. But we are rapidly approaching a time when consumers will have a choice on how the will receive their "television and movie" content, from whom, and on what device. Digital Media Adapters (DMAs) and Digital Media Servers (DMSs) are available from major consumer electronics vendors today,

with new models that drive consumer trends being announced every month. When consumers have a choice of getting premium content through their cable operators, as well as a variety of traditional and online competitors, the providers with the most accommodating overall solution will come out on top. Given their incumbent position with both television and broadband Internet service, cable operators are in a prime position to be that provider.

CONSUMER NEEDS

Hundreds of television channels, thousands of websites, millions of videos, all a few clicks away. This is an era where there are very few barriers to making content available to consumers, leading to an unprecedented amount of entertainment options from a rapidly growing number of sources. This has great promise, because of the high likelihood that something very tailored to every consumer's taste is available out there somewhere.

This simple consumer proposition places a heavy burden on content providers. It requires a balance between providing a breadth of content choices and enabling easy discovery of that content by the consumer. Breadth is increasingly measured in the hundreds of thousands or even millions of titles. Discovery of those titles has to be intuitive and fast. And the personal connection between a person and "their shows" requires that the service is always available.

At the same time, the growing popularity of digital video recorders (DVRs) and portable media players (such as the iPod) is causing consumers to demand a much more personalized and portable video experience. They want to view content highly relevant to them, on their terms. They want it on whatever device they happen to be using at that particular moment, and they want the content to be available all the time.

Engagement

According to a 2007 viewer study, nearly 60 percent of adult consumers surveyed stated that they watch online video[iii]. More businesses are recognizing that broadband video can help them target audiences, generate real revenues, and gain creative control of the user experience. Broadband video supports brands in a way that breathes vitality into and can extend the life of a media or service provider business. It adds flavor, perspective, and additional information to existing pages, increasing audience engagement. Figure 1 shows audience engagement for a top destination site on the web

.

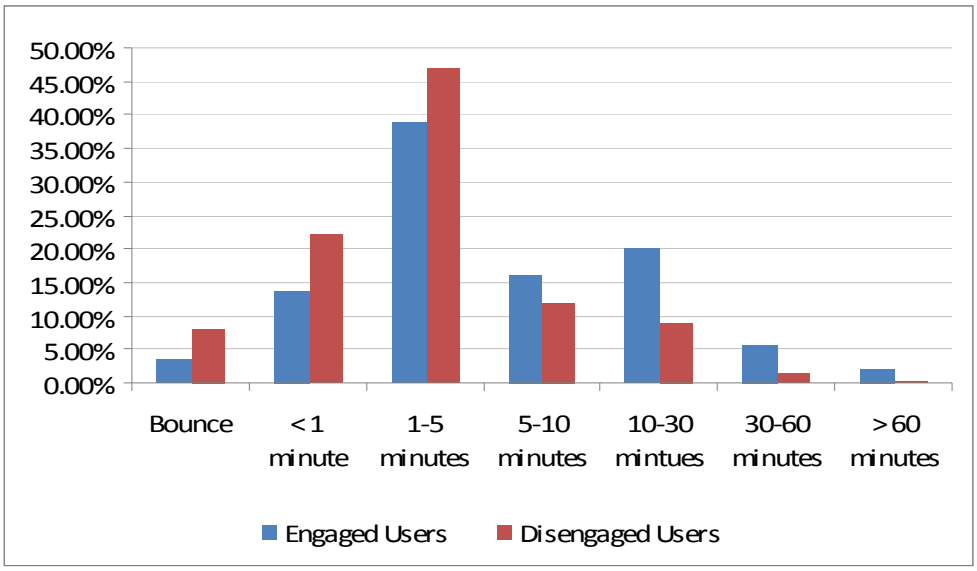


Figure 3: Sample Engagement Data for an Online Video Site

Relevancy

Programming has a valuable place in the entertainment universe. On TV, a network orders its shows to naturally lead the viewer from one to the next. This provides convenience and a sense of flow that can make the experience more enjoyable with less work. The challenge in front of us is how to apply the notion of programming to a much broader set of content while presenting a very personalized and relevant experience. Clearly, the traditional model of people making editorial decisions that result in a broadcast schedule begins to break down as the model moves increasingly towards a 1:1 engagement with consumers and includes a much larger body of content. Technology will have to play a bigger part in determining what is presented to a consumer if the promise of personalized programming is to be realized. The payoff for figuring this out is a better consumer experience, driven by the presentation of more relevant content.

Layering better-informed programming with technology designed to improve and measure the effectiveness of the content guide, the results are very smart, proactive recommendations that give consumers what they want - the convenience of low-touch programming with the benefit of highly personalized, relevant entertainment. This approach allows everyone to win--consumers get a better experience and content owners increase viewership, and ultimately, service providers can improve the experience and value of their service. But the consumer experience doesn't stop with content sourced from the network. Increasingly, consumers are contributing content as well. It might be highly personalized content meant for family and friends such as video from a family event, or it could be quickly captured video destined for a larger audience. Consumers want the user-generated content experience to be seamlessly integrated with the television and movie programming experience. This increasingly complex mash-up of personal content, friend and family content, professional programming from multiple content owners, and new independent content optimized for web distribution places even more demands on the cable operators and content owners looking to distribute content.

CONTENT OWNER NEEDS

At the same time that consumer requirements are increasing, content owners are asking for more. Media companies have an unprecedented ability to reach their audience through multiple outlets. This includes cable, Video on Demand, Digital Video Recording, DVD, the Internet, and mobile devices. Aggregators that have the ability to distribute content across platforms while maintaining a consistent, high quality experience will have an advantage.

Branding and Cross-Platform Promotion

In a world that isn't necessarily anchored by a channel on the TV, providing branding opportunities around the content becomes critical. Associating the show with the provider allows content owners to leverage their brand investments and connection to an audience. As soon as a content owner feels comfortable with their brand attribution in one medium, cross-platform promotion becomes a requirement. Leading a consumer from promotional clips on their mobile phone to the full show on the web or a video on demand (VOD) system and finally to a linear channel to find other, similar shows will become the norm.

Content owners' final requirement is to decrease their distribution costs. When each incremental audience member costs the content provider more, aggregators that can reduce this expense become very attractive partners.

Distribution Costs

Service and content providers concerned over costs of distribution have recourse to new management capabilities that can significantly lower the expense of doing business with content delivery network (CDN) suppliers.

Traditionally content providers have relied on one or even two CDN suppliers under multi-year contracts, which limits their ability to take advantage of recent changes in the CDN market. But, with over a dozen global CDN suppliers now vying for business in the U.S. alone, there's no longer any reason for content providers to tolerate onerous contractual terms with built-in cost escalators and other unnecessary cost burdens. Fortunately, the costs of implementing dynamic control over CDN services are miniscule compared to the potential savings. By turning to highly automated operations support tools and services, content distributors of all types and sizes can achieve cost breaks across the CDN domain, avoiding over-priced services and remaining flexible enough to take advantage of advances in CDN technology wherever they occur.

Unnecessary Cost Drivers

CDN providers typically set terms that require customers to pay minimum monthly fees for storage and distribution, covering all traffic volume up to a certain limit. After exceeding that limit, the charges are then based on incremental storage and distribution volume, which can add up quickly. There are ways to shift that traffic to cheaper distribution channels, and there are typically no contractual barriers to doing so. What can be termed "success-based" cost escalators also come into play with delivery of advertising with content over CDNs. As users access more ad content in the service stream, content providers have to pay more to CDNs, sometimes upwards of $7 per thousand views, which translates to a large share of the $14-$25[iv] cost per thousand (CPM) rate content suppliers typically charge advertisers. Some media companies are reducing CDN costs by leveraging management tools that dynamically execute their CDN requirements across multiple providers. This is often done by coordinating use of CDN options with in-house storage capabilities.

Cost-Cutting Strategies

There are many ways media management infrastructure can be used to help bring down CDN costs. Some approaches coordinate external CDN resources with sizeable in-house infrastructures. Other approaches rely more heavily on external support. In all cases, the content provider must be in a position to manage CDN fulfillment in a streamlined, dashboard-based environment with mechanisms in place to seamlessly switch content flows from one CDN to another.

Companies utilizing their own digital media storage systems need to actively manage files by pulling them from the CDN when those files are not in use. This is especially relevant for those companies that have large content libraries or already have invested in storage devices.

Some firms may also want to direct consumer video requests in non-peak hours to in-house distribution servers, only using CDNs for peak traffic. Policies that govern how traffic is allocated between internal resources and outside CDNs can be very simple, for example setting a weighted control on the flow of user requests for content where 60 percent may be directed to the internal servers and 40 percent to CDNs. In some cases, the content provider might build a system to monitor flows and manage traffic so that, if traffic hits a certain threshold, it is switched over to outside CDNs.

Achieving such capabilities requires highly sophisticated management software and infrastructure controls capable of switching traffic across multiple CDNs without disruptions to end users. Whether or not a company has internal storage and delivery resources, it will want to use CDN switching resources to direct traffic to multiple CDNs to ensure contractual caps are never surpassed.

This allows content providers to avoid paying inflated rates for files already duplicated and stored in the network that are on the downside of their usage curves. By moving such files out of the high-cost CDN storage environment, the content provider pays storage rates that square with current usage trends.

The need to ensure that quality parameters on each content stream meet end user requirements also has an impact on the CDN selection process. As content providers enable flows that support full screen viewing, the file storage and distribution volumes escalate. The content provider must be responsive to situations where high bandwidth requests are pushing traffic volume over a particular CDN cap faster than expected.

Along with sophisticated operations functionalities, successful management of CDN services requires a savvy approach to contracting services. While long-term contracts might appear to offer aggressive rates, CDN providers know that by locking customers in they will make out very well as traffic and storage demand exceeds caps. Even if a content provider has recourse to capabilities discussed above that can use multi-CDN access to avoid over-cap costs, it's prudent to use those management capabilities to facilitate working in an environment where one-year contracts are available.

Short-term contracts will allow a broadband video business to capitalize on lower bandwidth costs and the latest delivery technologies. In fact, broadband video providers with low volumes can gain leverage in their negotiations by looking at very short-term contracts of just a few months duration. The company can then bargain for reduced fees in exchange for agreeing to a longer-term contract.

The key for content providers is to configure their systems to optimize traffic distribution,

taking into account such factors as the extent to which in-house resources can be leveraged and how the provider can work within existing CDN contract terms to improve efficiency.

Implementing architecture that supports ever-changing dynamics of the business, including expansions of fields covered by the system, extensions of metadata categories and application rules, types of security to be applied to various content categories and ongoing variations in end user pricing and access rules is also a critical component.

Selecting Content Libraries

In addition to making premium content available online, television networks and cable programmers have mined their stockpiles of content, knowing that supporting video and audio clips make newer offerings more interesting to their audience. Very often businesses have a lot of content in their archives that could pull in a very large online viewership, either by itself, or when used as supporting material for related content.

There are several major content sources from which companies can harvest media that paves the way for broadband video business success. The best way to maximize content is to get it into a digital library and add value with commerce and advertising solutions that use IP based-communication and web-based presentation.

Multiple Outlets/Distribution

As big television networks and broadcasters move their content online, one big challenge that presents itself is that unless you have incredible brand awareness, exclusively focusing on building your own destination site is only one small part of building a successful online strategy.

Going halfway is not a winning proposition. Content providers have to commit to getting the content to all of the online destinations (i.e. syndication) that make sense: that's distribution. It will entail headaches. It will involve dealing with multiple video formats, different policies, and different advertising models. To get content out there and to monetize it is a lot of work.

This strategy involves thinking approaching online video holistically. ABC, CBS, NBC, Fox have all built their own Websites to host their television content, but they've also invested in distribution.

Previously, broadcasters have seen the Internet as supplemental to the TV business; it's something they've used as a promotional tool. Online, broadcasters can go out and find an audience rather than rely on an audience to find them. Beyond promotional deals, it's that long tail that allows broadcasters to connect with an audience.

Here's an example: If a consumer watches *Heroes* for the first time on TV and really wants to catch up on what she missed, she can go online and watch back episodes. In essence that library of long-tail content becomes a destination. This is a nice driver for the market: Broadcasters can find out where there is demand and do some real-time determinations of what kind of content is popular.

Broadcasters, cable channels and operators need to create a destination to catch all of these viewers. But they also have to be willing to follow the consumer. If a consumer can't get what they want through one outlet, then they will go elsewhere because it doesn't end with having just one destination to catch these kinds of viewers. Brands have to reach different kinds of people across the Internet, with different preferences on where they want to go to get their content and how they want to consume it.

If a media company cuts the right distribution deals, its shows can land in the places where consumers are landing natively. They can reach customers in a way that's not quite the same as in the operator environment where walled gardens are the norm. This, in fact, flies in the face of a walled garden.

As more network TV content moves online, there will be increased pressure to open up the walled garden a little bit and allow content to flow to more than one place. Ultimately, that helps create the seamless experience the consumer is after.

Success will mean interoperability across platforms and distribution of content far and wide. Standards are now emerging that will ease interpretability between different platforms and enable seamless and even greater multi-platform delivery of content.

CONTENT DELIVERY METHODS

There are two primary methods for getting content to consumers: the consumer downloads the content for either immediate or later viewing, or streams it for viewing "live". Streaming is generally defined as content being delivered to the subscriber "just in time" for viewing, typically without the ability for the user to "record" or keep a copy of the content. Downloading is generally defined as transferring the content to the subscriber's device and then viewed locally. However, the boundaries between these two methods are blurring as new technologies come on to the market. Protocols typically used for downloading can be used to simulate a streaming experience, and vice versa.

Streaming

Today's protocols for streaming video depend on a fairly reliable transport network. The video is sent at a constant rate, and any variance in delay (jitter) in the stream is compensated for by buffering at the client. This results in delayed startup of streams while buffering, as well as delays when switching from one stream to another (which requires buffering of the new stream).

Streaming via traditional IP streaming protocols presents a problem for video delivery across the "generic" uncontrolled bandwidth of the Internet, as it is subject to congestion and choke points across the network. The solution for this is to stream the video from a source as close to the subscriber as possible. By eliminating as many potential points of congestion from the network as possible the video can often be streamed at a higher rate with better quality for the subscriber. A second improvement for delivering video via streaming protocols is to apply Quality of Service (QoS) to the stream, giving the video packets higher priority in the event of network congestion. However, this is typically only viable on a single, controlled network, not across the Internet in general. In fact, this is typically the way that cable operators transport both broadcast and on demand video across their regional networks.

Streaming protocols typically limit the bandwidth to the actual stream rate. On one hand, this can prevent the client from taking advantage of "pre-buffering" the video when extra bandwidth is available. For example, by enabling fast start capabilities native to some streaming protocols. On the other hand, streaming protocols are very efficient in that they only send packets if the client is viewing them. For example, if a user starts watching a 10-minute video but decides to stop one minute

in, only the first minute of the video is transferred.

Downloading

Downloading content gives the user more control over when and where they want to view the video, as once it's downloaded to the client device, no network connectivity is needed to view the video. This allows for a very high quality video experience regardless of available network bandwidth or connectivity, assuming the subscriber has downloaded the video in advance. Digital Rights Management (DRM) can be used to limit transfer of the content, the number of times the content can be viewed, the viewing window, etc. Downloading video via broadband is most often done using the HTTP protocol, but can be done via proprietary protocols with dedicated clients as well. Dedicated video clients also provide more capabilities for the subscriber by managing where videos are stored on the device, preventing screen captures of the content, providing a common navigation engine, or allowing users to set up subscriptions to download multiple videos.

Progressive download is the ability to start viewing a downloaded video as soon as there is enough video "in the buffer" to continually play the asset given the transfer speeds that are seen during the initial buffering. Progressively downloaded content allows the subscriber to start viewing the video as soon as possible while simultaneously saving the video for later viewing. Within the progressive download model, downloading can simulate a streaming experience, but using downloading protocols instead of streaming protocols.

Whole asset downloads, and proprietary methods of downloading small "chunks" of the video are common methods utilized today. These methods provide the benefit of using download-type protocol architectures while utilizing burst transfers of the video at faster than stream rate for quicker delivery. Some methods also provide the efficiency of streaming protocols by only transferring just the video that is being watched in addition to a small buffer.

Peer to Peer (P2P)

P2P protocols can be used to enable download type services or streaming type services. Where both traditional downloading and streaming protocols work get their content in a fairly linear fashion (start at the beginning, keep going until the end) from a single source, P2P clients can get content simultaneously from a large number of sources, and do so in a non-linear fashion (getting the last part of the file first for example). However, while P2P clients can source the data from multiple locations, it still must all traverse the same broadband access link. Thus, P2P will have the same broadband download bandwidth issues as traditional streaming and downloading protocols will. P2P protocols have routing metrics to determine the nearest, highest bandwidth, most reliable sources to source content from, increasing the subscriber's utilization of the access network. However, because there are not commonly accepted standards for P2P networking protocols, they are not integrated into popular browsers, and typically require custom clients.

P2P protocols have a great advantage for content owners in that they don't require any streaming or downloading infrastructure. Content owners need only to seed the content once into the P2P network. As more clients download that content from the initial peers, they in turn make that content available to other peers. Thus, the content owner's individual subscribers are using their own storage, streaming resources (their PC) and broadband network connection bandwidth to distribute the content for them, all at no cost to the content owner. Since P2P protocols work best when

there are a large number of peers in the network with the content a client is asking to download, it tends to work best for the most popular content, as that is likely to reside on the greatest number of peers.

P2P clients have downsides as well. For long tail or niche content, the number of peers with the content might be very low, resulting in slow download performance and a poor experience for the recipient. Since most P2P networks rely on individual users who may or may not be online at any given time, niche content may not even be available to requestors who wish to view it. Most P2P networks are also not managed by any central source or entity, and thus may end up being less reliable overall as they rely on the individual users to each do their part.

P2P relies on individual users to source the content over their asymmetric broadband connections. While download speeds can be very fast, the upload bandwidth is typically a much lower speed, and more subject to congestion. A P2P user can easily consume a large percentage of both the upstream and downstream available bandwidth of a given network segment, depending on the popularity of content they host, and the volume of content they seek to obtain. The effects of popular P2P applications that are used to distribute copyrighted content is the best example recognized by operators today.

The use of P2P protocols isn't limited just to clients trying to use the Internet to distribute content. Service providers could leverage P2P for distribution of content within their network, as well as potentially leveraging P2P protocols to "peer" with other service providers for content, in a similar way as they peer for Internet packet transport today. Instead of publishing their content to 3rd party CDN providers (described in the next section), or to client-based P2P networks, content owners could publish their content to a select number of service providers they peer with at a content level, who would in turn distribute that content as needed either to other service providers for delivery to their customers, or direct to consumers who are using service providers that don't have content peering relationships.

Content Delivery Networks

To help scale the delivery of content across the Internet, CDN providers have built infrastructures that help virtualize that content across the network. Conceptually, CDNs work by ingesting the content from a source (such as a content owner's website) into a network of intelligent caches distributed throughout the network. As subscribers request to view content, a copy of that content is stored in a cache closest to that subscriber. When the second subscriber requests to view that same piece of content, the request is redirected to the local cache, with no need to go back to the original source. As local caches "fill up" with content requested from subscribers, the least popular content is purged to make room for more popular content. CDN caches can be placed in a tiered hierarchy, allowing for content population to match the interests of its local subscribers. This demand-based method keeps the most popular content closest to the users requesting it, resulting in the best performance for those subscribers. Niche content, however, is always available from the source. CDN caches simultaneously serve the content to the requestor while the content is populating the CDN cache's local storage repository.

But as described in the previous section, the more popular the content is, the more it is downloaded and the more the content owner is charged for distribution. This is in stark contrast to P2P, which performs best (and least expensively) for the most popular content. However, also in contrast to P2P, CDNs are

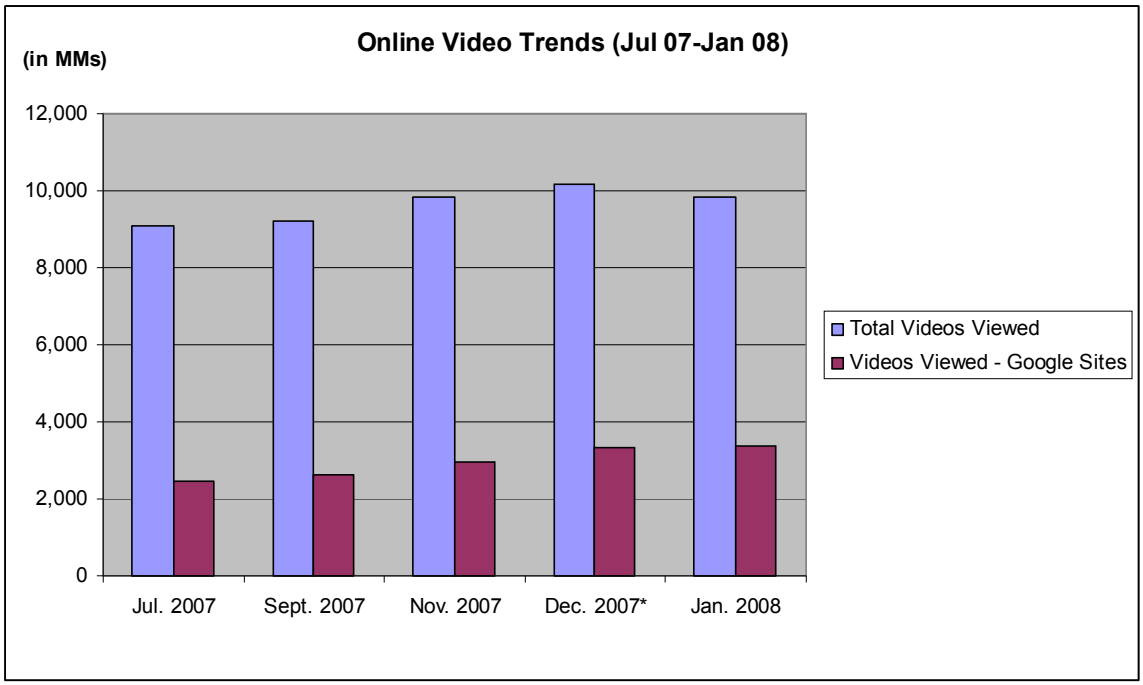
managed networks and can provide service level guarantees to content owners, and deliver all content regardless of how popular it is, from within the operator's own network. The most popular content is more widely distributed throughout the network, and thus typically resides close to the subscriber, minimizing the number of "hops" or links through the network the content must traverse to reach the subscriber. The closer to the subscriber, the better - proximity minimizes the potential for congestion and enhances the overall customer experience.

Hybrid Delivery Options

Some companies are starting to look at hybrid P2P and CDN delivery options. A client would first look to see if the content is available from peers, and if not, start sourcing that content from a CDN provider. Thus, the most popular and expensive content traditionally provided by a CDN would be delivered via the "free" P2P network instead.

Role of the Cable Operator in Online Video

Today, when it comes to content destined for the television, the cable operator acts as the primary, and often only, aggregator and distributor of content from multiple content owners to the operator's subscribers. Consumers navigate linear broadcast content via a program guide, with a menu-based navigation portal for accessing popular movies, TV shows, and niche content on demand. But when it comes to broadband video, many cable operators don't have a presence at all. Today's business models are still developing around what and how consumers will pay for online video. But as the comScore numbers in figure 1 show, consumers aren't waiting for their cable company to figure out. Cable operators have the opportunity to play the same aggregation and distribution role for online content as they do for traditional television content, and in some cases, controlled delivery of the content may almost pay for itself.



*Single heaviest month for online video consumption since comScore initiated its tracking service

Figure 4: Online Video Trending July 2007-Jan 2008

Impact of Over the Top Video Delivery on Cable Operators

Subscribers watching online video "over the top (OTT)," or using the cable operators bandwidth for video not distributed by the cable operator, is effecting the operator both technically and financially. The most obvious is the substantial traffic growth
Cable operators are already transporting all of the online video that their subscribers are watching – they're just not getting any revenue from it aside from the revenue they are receiving for providing broadband Internet connectivity. For popular content, the cable operator is probably transporting the same exact content countless times across their backbones. In addition to the hit that cable operator backbones are taking from a bandwidth perspective, operators are also facing a service substitution challenge. This is happening today in the voice market with OTT providers such as Vonage and Skype. Once premium video is available online, and viewable on the television set via retail DMAs, subscribers can also start shifting their content spend from the cable operator to other providers. A subscriber shift is starting to happen today through such devices and services as Apple TV and iTunes, Vudu, and Microsoft Xbox LIVE Marketplace. And as services continue to shift, so too will the advertising revenue associated with those services.

Cable Operators as Online Content Distribution Partners

By becoming an active participant in the distribution of online video, cable operators can accomplish multiple goals while providing more advantages to both the content owner and consumer at lower cost than traditional CDN providers. Cable operators already own their regional network and broadband infrastructure, and have fundamentally lower cost structures for building a CDN than providers which need to lease those facilities. In addition, this infrastructure is already being used for the transport of other services that are largely funding its build-out; such as traditional cable TV, voice services, business services, and even OTT online video itself.

By leveraging CDN technology within an existing infrastructure, cable operators can cache popular content at the edge and eliminate duplication of bandwidth across the network to help alleviate some of the costs already incurred. Because the operator controls the infrastructure, cable operators can also leverage QoS capabilities to prioritize video traffic in the event of network congestion, providing a better experience for consumers. With fewer potential congestion points and the ability to prioritize video streams, cable operators can delivery higher quality, higher bit-rate video to the consumer, further enhancing the experience vs. OTT delivery, which contributes to continued subscriber loyalty and brand awareness.

Cross-Platform Service Capability

Beyond the advantages of more efficient delivery of content, cable operators can add significant value in the services they can bundle together and offer consumers. As cable networks transition to all-digital with set top boxes, digital video recorders have become commonplace in nearly every cable household. A common platform establishes a foundation for cable operators to provide the same content to multiple screens for the consumer. By partnering with the content owners, cable operators can provide customers who are subscribers to premium tiers access to that same content online. Consumers could access that content through a portal that both maintains branding for the content owner and gives the consumer a single destination for all things video-related.

As cable operators begin to deliver content across platforms with a single infrastructure, they can also start working with advertisers on cross-platform advertising capabilities. Advertising campaigns can span from television content to long-form online advertising. Targeted banner and bumper ads can accompany online video, and different ads can be shown each time the consumer watches the content. By integrating online video with communication services such as email and VoIP, cable operators can enhance the video experience by becoming involved in the subscriber's social experience.

Capitalizing on the Online Portal

Opportunities for additional advertising or click-through revenue rise once mainstream, premium content is presented along with niche, long tail content and user-generated content. The goal for cable operators is to become both the distribution network and the jump off point for all things video and content related. By integrating all of a consumer's content needs, across multiple content owners, premium and free content, and personal content the cable operator can become the primary video experience provider for their consumers. Beyond a single portal for video content, the cable operator has the opportunity to integrate management of all video, content, and communication services for the subscriber into a single experience, allowing the subscriber to access any service regardless of the device they're using.

NEED FOR MEDIA MANAGEMENT AND PUBLISHING

Once a cable operator/content provider has their distribution strategy worked out, the next step is figuring out the media management. Everyone in the value chain benefits from a system that can provide the control and automation necessary as content is provided across platforms. Figure 5 provides an overall sense of the data flow involved in publishing online video.

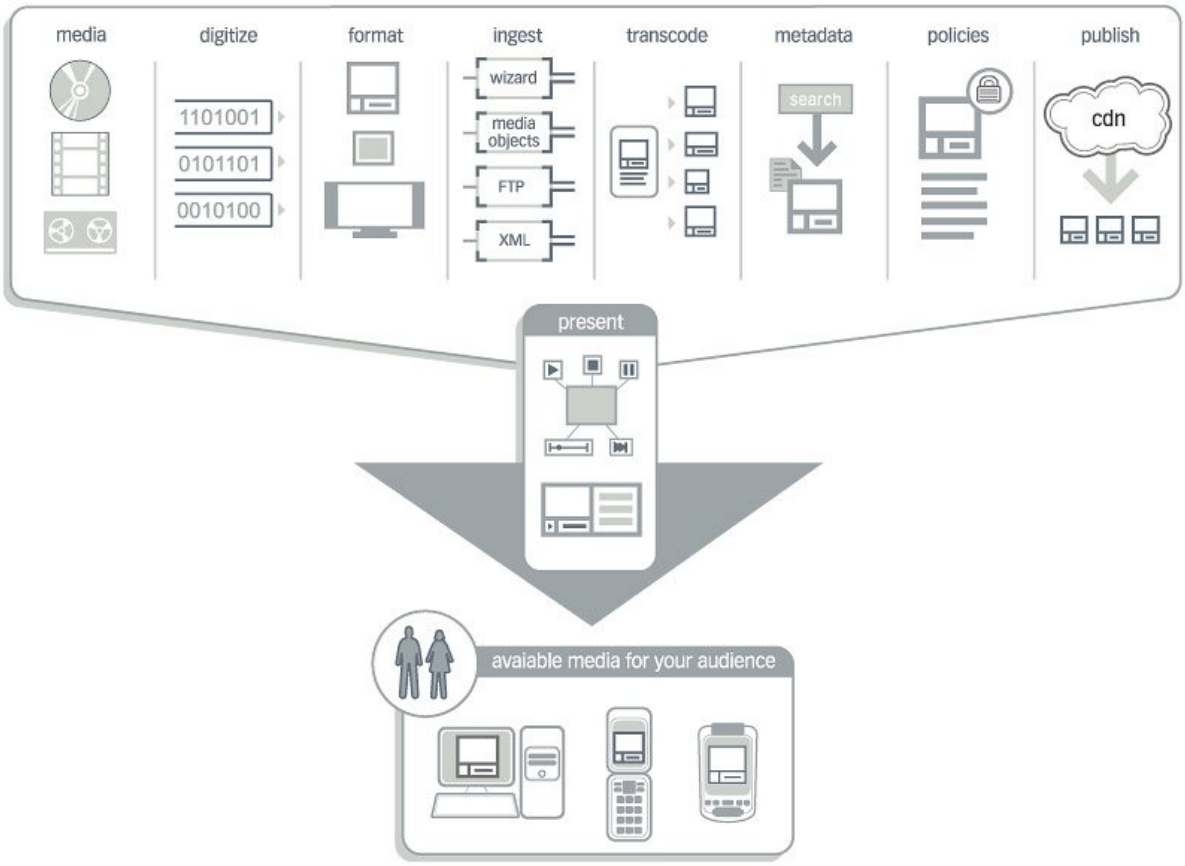


Figure 5: Online Video Workflow

Applying Business Policies/Control

Control leads to the application of business policies for each piece of content. A media management system must be able to reflect the carriage deal including geographic restrictions, air dates, end-user restrictions, digital rights management, pricing models and advertising policies. These set parameters that control when content is available, who can access it, when it expires, and what delivery methods are allowed. As these policies are applied to a video, they must be enforced in every medium including VOD, the web and mobile.

Common types of broadband video controls include:

- **Content restrictions.** Scope content usage to reflect business requirements. Examples include restricting media so it can be accessed only on certain dates or in certain geographies.
- **End-user restrictions. C**ontrol user access via an integration with existing "single sign-on" authentication application. These controls enable the system to perform all the response tasks (prompting the user for ID and password if necessary), and either generating a license directly, or using a management system to pass the license information and grant access.
- **Pricing policies.** Help monetize video content by supporting pricing schemes such as free trial periods, pay-per-view, and pay-per-download.

- **Advertising policies.** Ensure a tight integration with advertising campaign management systems to target and track ads.

Automation

The only solution to complex content distribution strategies is automation. A media management system should gracefully meet various metadata schema and video format requirements. The heavy-lifting tasks of transcoding, file transfers and encryption are best accomplished programmatically, eliminating the need for personnel to manually start each chore.

Dynamic Entitlement

The final mission of a media management system is to enable the monetization of content. Entitlements should be dynamic, allowing operators to combine consumer segments with pricing models and content restrictions. Advertising needs to merge the emotional connection of the television with the targeting of the web. An open approach to trying new advertising campaign management systems and relentless testing will result in consumers receiving more relevant ads that are actually appreciated for their educational value.

Choosing a Content Delivery Network

If content owners have a large library of video they may choose to select a CDN to host their files once they are published. The selection process isn't one that can be addressed briefly. There are a lot of factors to consider in determining the best value: quality of service, the number of sites the content is uploaded to, and what reports and alerts are offered.

Applying policies that govern how traffic is allocated between internal resources and outside CDNs, or directing traffic to multiple CDNs to ensure that contractual caps are never surpassed are a couple ways to reduce CND costs. Another approach is to move older files out of high-cost CDN storage to an in-house storage as the audience moves on to newer content.

Determining Formats

The next step after content ingest is formatting content. The formats needed depend on where the media is being sent, as different media have disparate technical requirements. This often means creating more compressed versions of a video for viewing within a browser or for faster download for viewers without broadband connections. Some examples include media companies that want viewers to be able to watch video as it downloads or content providers who wish to support both Windows and Mac players.

In addition to selecting the appropriate formats for a content or service provider's own site(s), there are also additional formats required for syndicating media. For example, when syndicating to a mobile carrier, files must be provides in formats that work on their devices. Considerations in selecting format can be impacted by the target audience, the video/audio quality required, file security and the content being posting to live or on-demand media.

Format	Description	Compatible platforms
3GPP, 3G2	Specifications for creating, delivering, and playing back media over high-speed broadband mobile networks to multimedia-enabled cell phones. Intended for mobile, but playable on desk/laptops.	• Mac • Windows • mobile
Flash Video (.flv)	Flash video is compact and supports both progressive and streaming downloads.	• Mac • Windows • mobile
MPEG-4	Used for streaming media video on the web, CDs, and broadcast television.	• Mac • Windows • mobile
Windows Media Video (.wmv)	Used for several proprietary codecs developed by Microsoft.	• Non-Windows computers require that an extra (free) component be installed. Some newer-format Windows Media video cannot play on non-Windows computers. • Supports mobile, depending on the codec used.

Figure 6: Common Broadband Video Formats

Most videos files have at least two types of file formats, the container and the codec. The video file container holds data like audio and video, which have been compressed using codecs. Codecs compress files so they take up less storage space on a computer and can more quickly be transmitted across the Internet. Figure 6 lists some common container formats for broadband video. Note that there may be exceptions— some formats will play on certain platforms only if a special plug-in is installed, or some videos will play on a device only if the correct codecs are in the container file.

SUMMARY

More, more, more: More content choices, more devices to view that content on, and more ways to find the content. Consumers only have so much time during the day to view video content, and want that time to be spent viewing relevant content, not figuring out what to watch, where to find it, and how to get it on the device they want. Content owners want more ways to monetize their content. They want to sell it direct to the consumer when they can, but also leverage the most popular distribution outlets to make sure their content is easily accessible to consumers.

This means streaming the content live to consumers, allowing them to download it for either immediate or later viewing, and giving them the flexibility to move that content around and share the experience with friends and family. Cable operators are one of the primary distribution outlets for television content today, but certainly not the only one. The same television show is available initially via broadcast, and then online via the programmer's website, online through retailers such as iTunes and Amazon.com, on demand through cable operator's VoD platform, and ultimately on DVD and syndication as well.

The Internet gives both content owners and consumers more avenues for receiving and viewing that content. Cable operators have the opportunity to continue to play their current TV distribution role for online video as well. Multiple partnership and engagement models exist, ranging from acting as a wholesale CDN provider up to and including being the aggregated retail storefront for content. And by integrating the online video experience with the current bundle of digital video, voice, and broadband data cable operators can further enhance their value proposition for consumers.

But managing all of this content from so many different sources to so many different destinations requires a sophisticated system for media management to help automate and scale the process. Additionally, it provides mechanism for extending content and the consumer experience to broadband with the flexibility to assign specific policies and business rules applicable to operators and programmers, and most importantly, accomplishes this without relinquishing control over specific media objects. As the market continues to evolve and grow, the management of online media will continue to grow in complexity with sheer volume and types of content available. As new formats emerge and the delivery mechanism, type of content and business rules change, media management systems need to serve as an extension of the programmer/operator product teams and dependably deliver video that meets the expectations of viewers accustomed to high quality video over the TV.

REFERENCES

[i] comScore Video Metrix, Press Release, March 2008

[ii] comScore Video Metrix, Press Release, March 2008

[iii] Online Video: A New Local Advertising Paradigm, The Kelsey Group, Inc., 2007

[iv] Based on today's average CDN industry pricing models, internal company resources

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