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

Carol Ansley, Jim Allen, Tom Cloonan
ARRIS

Abstract

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

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

ON VIDEO EVOLUTION

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

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

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

Linear Video Services

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

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

Video On Demand Services

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

IP Video Services

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

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

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

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

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

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

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

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

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

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

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

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

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

UNICAST IP VIDEO DELIVERY

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

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

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

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

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

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

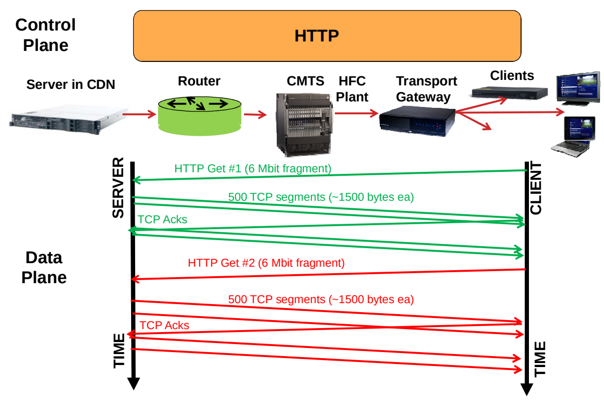

Figure 1 - Unicast over Transport Gateway

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

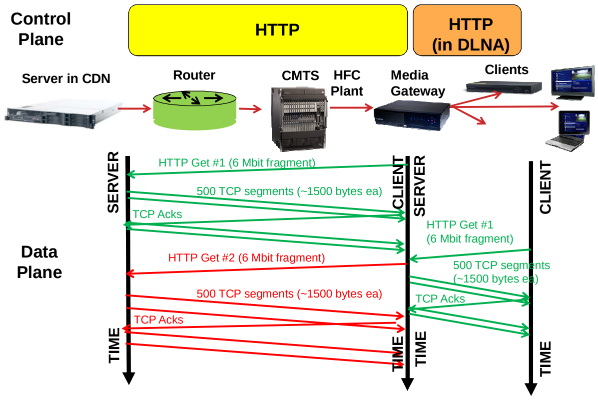

Figure 2 - Unicast over Caching Media Gateway

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

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

MULTICAST IP VIDEO DELIVERY

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

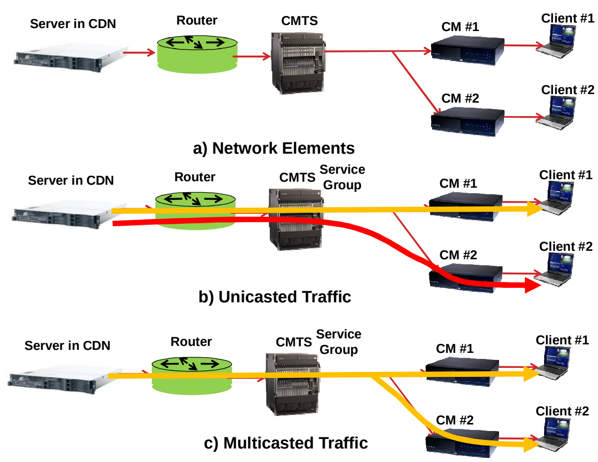

Figure 3 - Unicast vs. Multicast

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

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

Unicast IP Video cannot be utilized for Multicast IP Video.

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

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

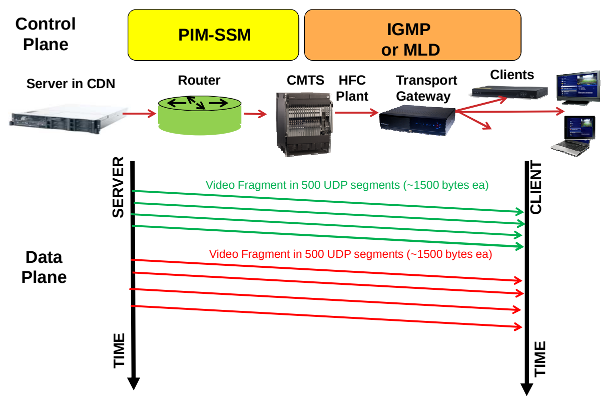

Figure 4 - Multicast with UDP

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

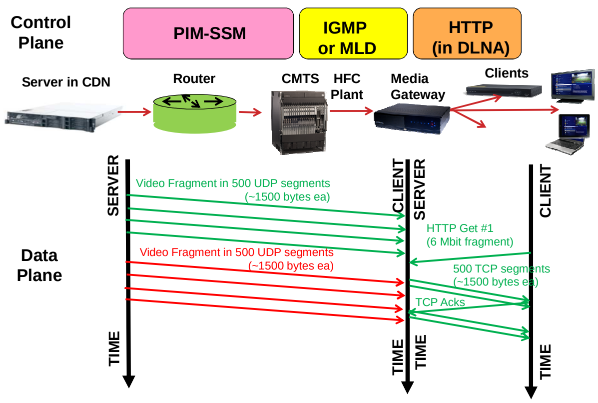

Figure 5 - Multicast with UDP & Conversion to Unicast TCP

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

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

COMPARING UNICAST AND MULTICAST IP VIDEO

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

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

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

Some of the issues to be considered are listed below.

1. Common protocols for multicast IP Video and any optimizations over DOCSIS should be available in an open forum, similar to the TWC ISA or Comcast NGOD SDV specifications

2. Any optimizations for Unicast IP Video that allow robust performance for first screen viewing should be provided in an open forum for maximum benefit

3. Current encoding methods will require multiple choices for unicast and/or multicast stream delivery, new encoding choices, such as SVC, could improve multicast efficiency for multicast and stream management for unicast

4. Reliability concerns in the IP Video packet delivery

5. Distributed Denial Of Service attacks by hackers on head-end equipment

6. Multicast must be tied into a Connection Admission Control algorithm to identify overload conditions when a new Multicast stream cannot be set up

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

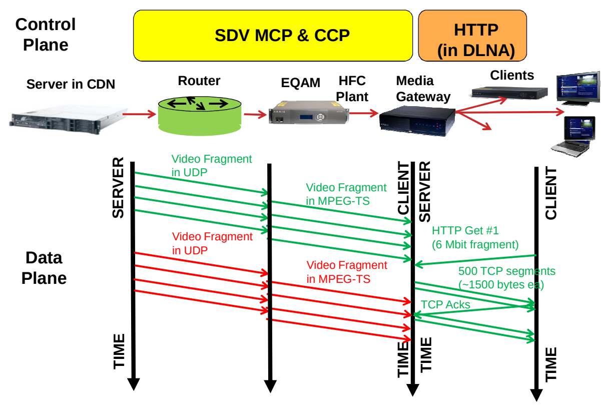

Figure 6 - Multicast with UDP & MPEG-TS & Conversion to Unicast TCP

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

SIMULATION RESULTS

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

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

Viewer Modeling Parameters

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

1. Acceptable Tuning Blockage Probability
2. Program Viewership Popularity

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

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

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

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

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

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

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

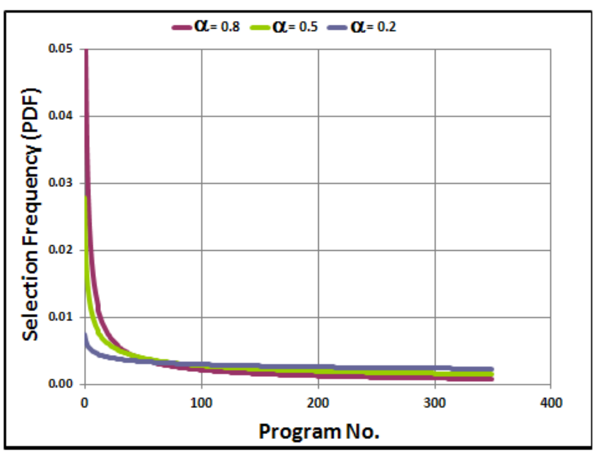

Figure 7 - Simulated Popularity Curves

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

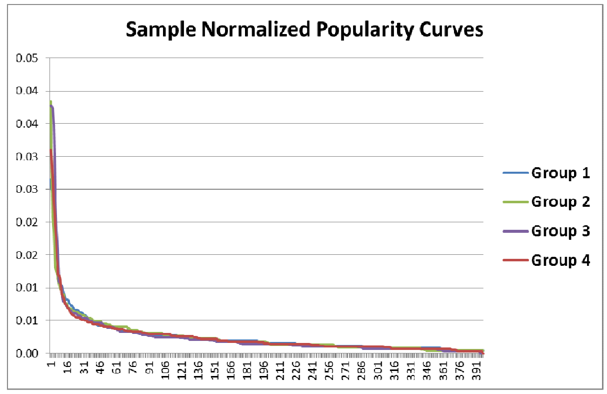

Figure 8 - Sample Popularity Curves

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

<u>Network Modeling Parameters</u>

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

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

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

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

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

Downstream Program Capacity

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

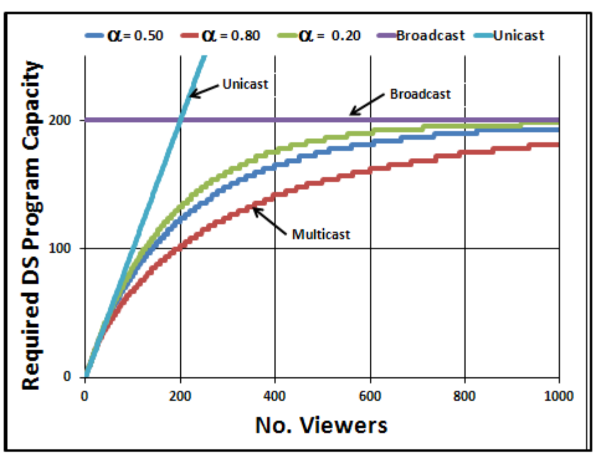

Figure 9 - Required Capacity vs. No. Viewers

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

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

Program Lineup Size

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

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

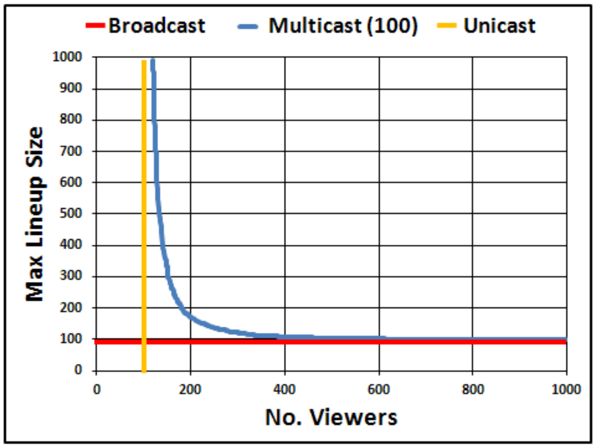

Figure 10 - Maximum Lineup Size

Figure 11 - Lineup Size vs. Downstream Program Capacity

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

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

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

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

Extension with Actual Data

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

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

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

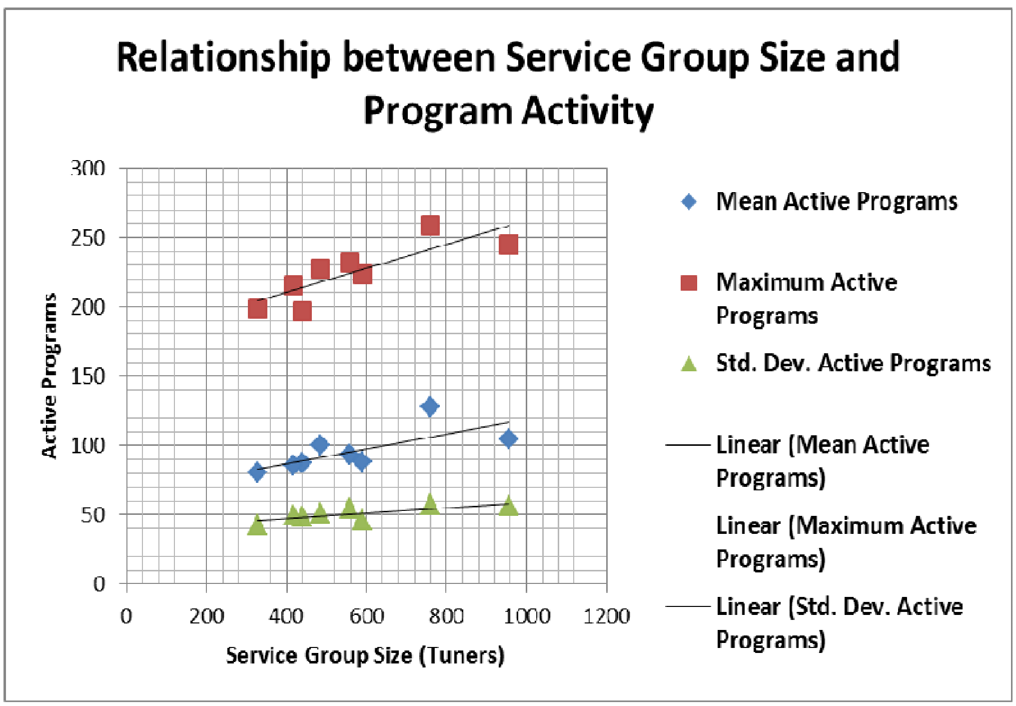

Figure 12 - Comparison of Service Group Size and Viewership

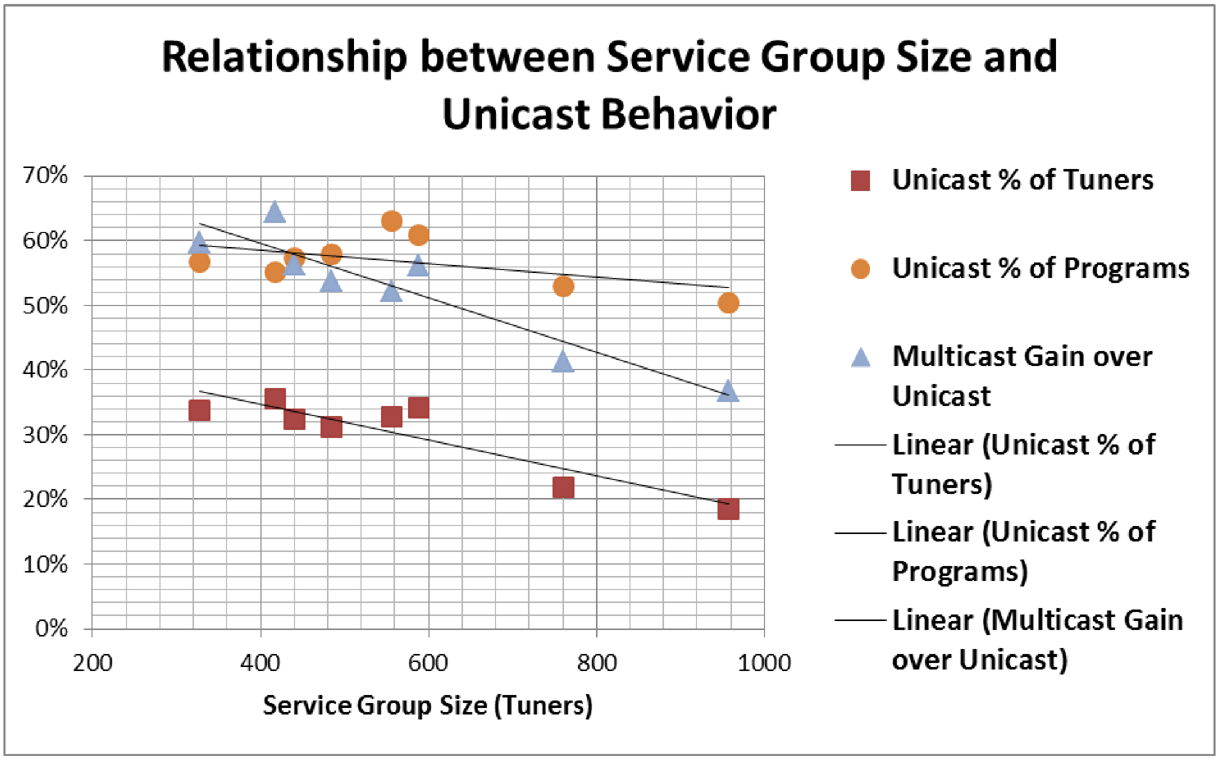

Figure 13 - Relationship of Service Group Size to Unicast Behavior

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

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

Comparison of Deployment Scenarios

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

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

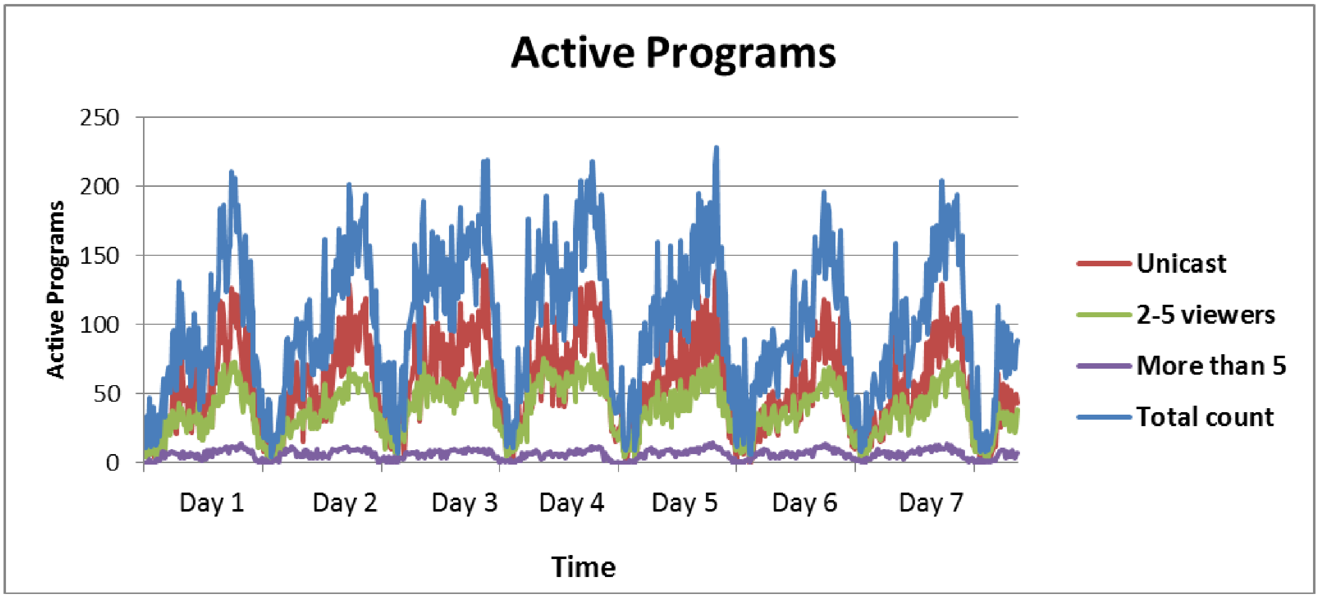

Figure 14 - Classifying Programs by Viewership

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

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

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

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

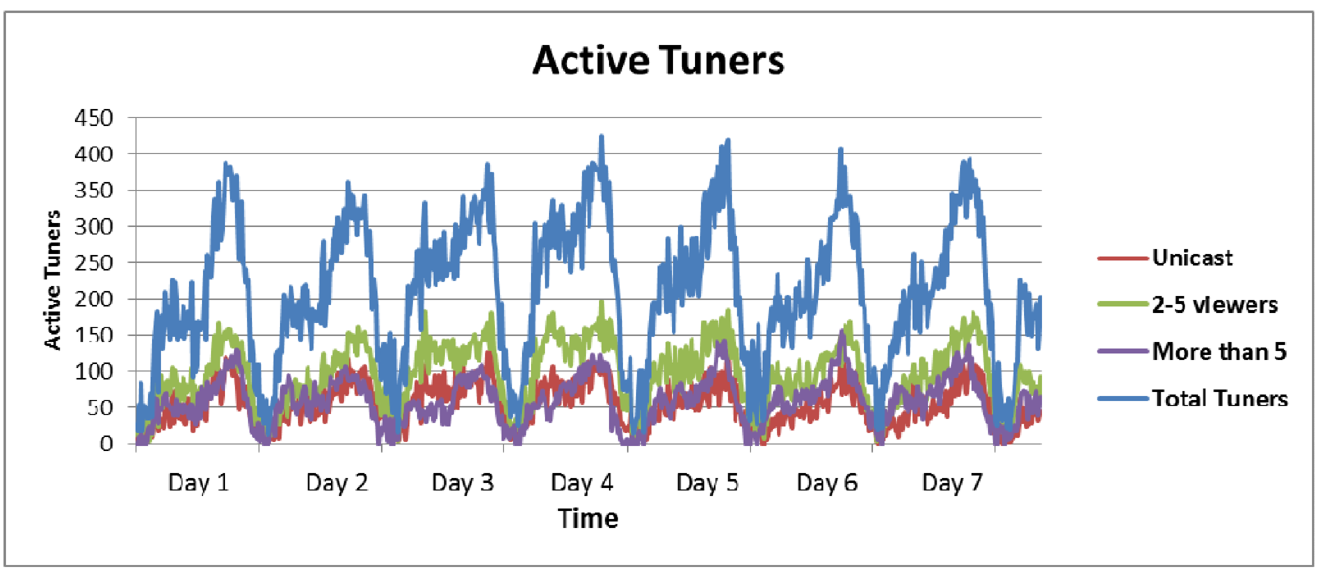

Figure 15 - Distribution of Tuners versus Other Tuners

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

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

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

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

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

In Figure 17, several scenarios are compared

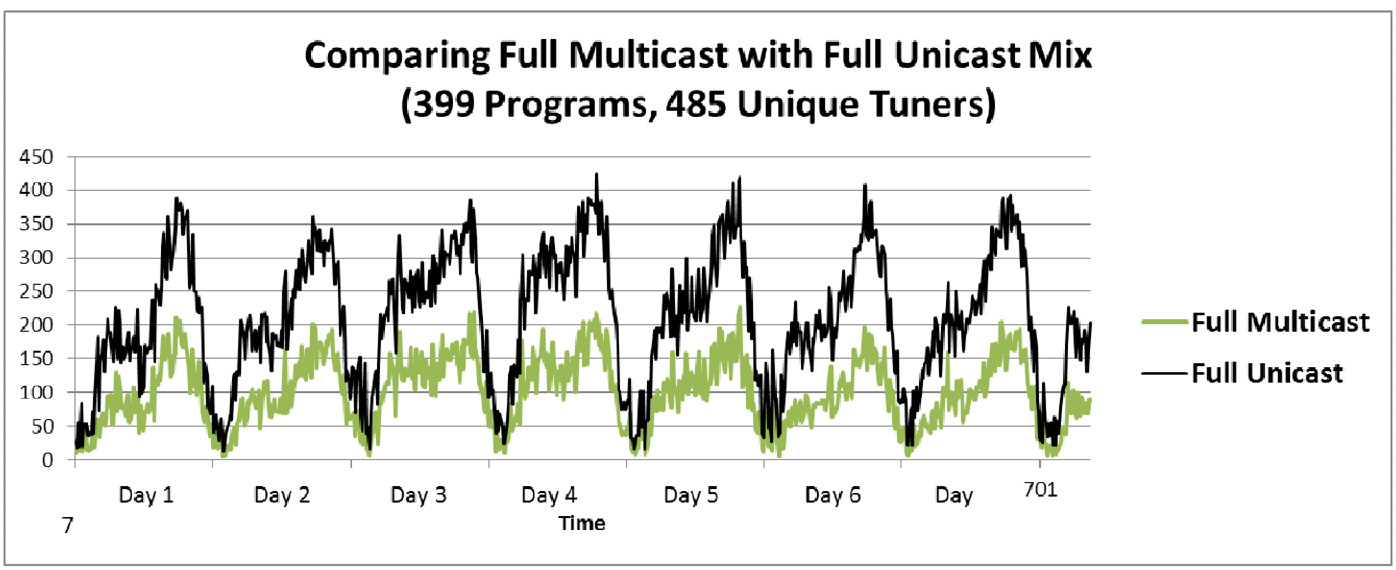

Figure 16 - Comparison of Multicast and Unicast

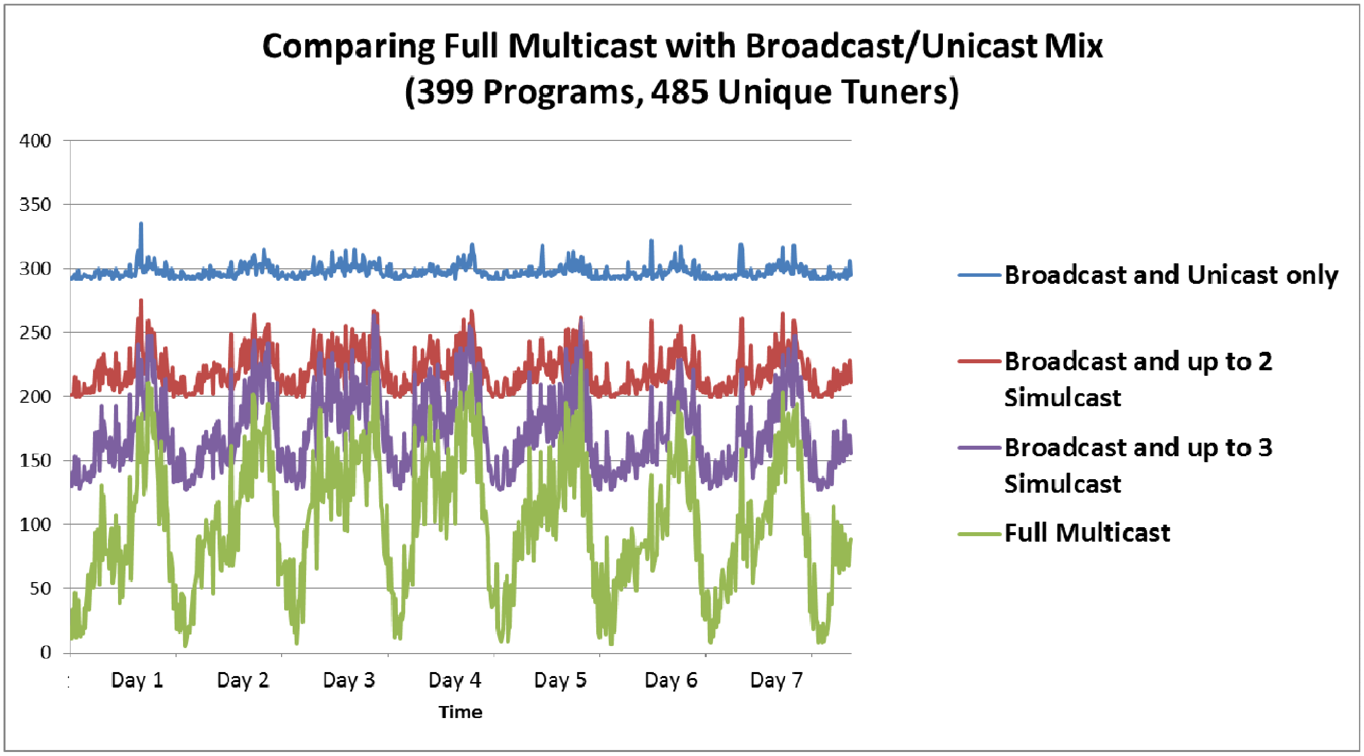

Figure 17 - Comparison of Multicast with Hybrid Broadcast/Unicast Model

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

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

<u>Other Considerations</u>

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

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

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

The Path Forward

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

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

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

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

In Summary

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

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

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