SCTE CABLE-TEC EXPO '15
OCTOBER 13-16 NEW ORLEANS

SCTE Society of Cable Telecommunications Engineers

LEVERAGING IP MULTICAST AND ABR TECHNOLOGIES FOR DELIVERY OF LINEAR SERVICES ON CABLE NETWORKS

A Technical Paper Prepared for the Society of Cable Telecommunications Engineers
By

Sangeeta Ramakrishnan
Principal Engineer
Cisco Systems
170 W Tasman Drive, San Jose, CA 95134
rsangeet@cisco.com

John Bevilacqua
Executive Director, Network Architecture
Comcast Cable
4100 E. Dry Creek Road, Centennial, CO 80122
john_bevilacqua@cable.comcast.com

Table of Contents

List of Figures

1. Introduction

Cable operators face a multitude of choices for addressing subscriber demands for more content on more devices in more places. While many cable operators are currently delivering unmanaged IP video services utilizing adaptive bitrate video (ABR) streaming, we will consider the practicality and benefits of using ABR streaming for delivering managed linear TV services.

In recent months, delivery of ABR streams via multicast has gained acceptance as a method to optimize bandwidth requirements over the cable network. In this paper, we will discuss the challenges and benefits of this approach.

Furthermore, we will examine in detail the various choices within such an architecture and the impact of those choices. Finally we will examine the key metrics operators can use to measure the efficacy of their multicast architecture and how to further optimize that efficiency via analytics.

In summary, this paper will assist operators in choosing the most viable architectures for delivering high-quality linear IP video services over their Data Over Cable Service Interface Specifications (DOCSIS®) networks in a bandwidth efficient manner.

2. Background

Over the past several years, video traffic has been growing rapidly and is anticipated to dominate next-generation networks. Globally, consumer Internet video traffic will be 80 percent of all consumer Internet traffic in 2019, up from 64 percent in 2014 [1]. The vast majority of the IP video services deployed over cable operators' networks are generally targeted to on-demand services today. Also in most cases this video is typically consumed on secondary screens such as tablets, smartphones, and laptops.

This video is generally delivered via Hypertext Transfer Protocol (HTTP) adaptive streaming which is also commonly referred to as adaptive bit rate video. In this technology, the video is typically segmented into two (2) second fragments, and each fragment is stored and delivered as a file via HTTP over the Transmission Control Protocol (TCP) as shown in Figure 1. The same video stream is generally encoded at a number of bitrates/resolutions, each of which is called a profile. The Internet Protocol television (IPTV) player such as an Internet Protocol - set-top box (IP-STB), referred to as the player in the rest of this paper, is able to request the appropriate profile based on the device type and bandwidth available to that client. This approach enables the player to adapt to changing network conditions while minimizing video stalls and playback disruptions for the end-user.

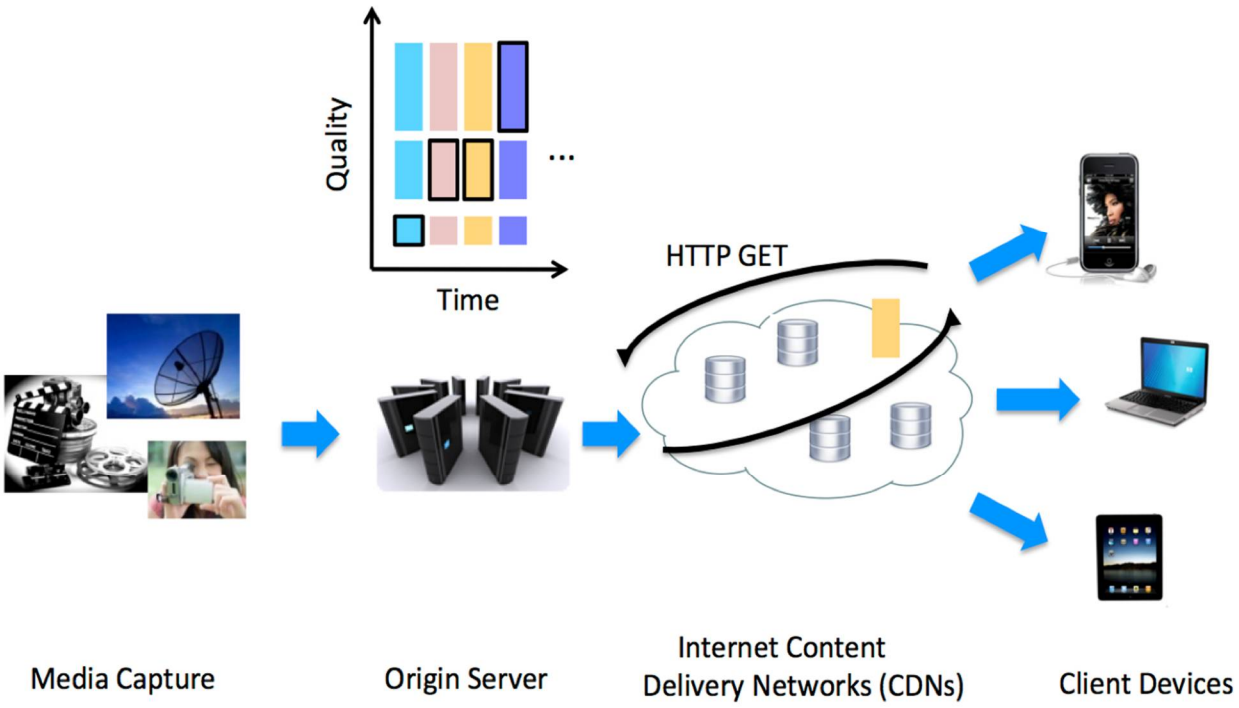

Figure 1 Adaptive Bit Rate Overview

However newer services are starting to be launched where linear services are now being delivered over Internet Protocol (IP) to secondary screen devices. For example, Comcast has just announced a new service called Stream, which enables subscribers to watch live TV on tablets, laptops and phones. Additionally, several operators have plans to deliver linear services to STB/TVs over IP as well. Such moves to deliver linear services over IP to both primary and secondary screens will serve to increase the capacity requirements of the access network. Much of the capacity demand comes from a large number of subscribers viewing the same popular linear content like news, sports, etc. Operators can leverage this concurrency attribute of linear content by delivering these services over multicast to gain significant bandwidth savings. The multicast capabilities in DOCSIS 3.0 specifications [3] have been enhanced significantly over earlier versions of specifications, enabling operators to use IP multicast to deliver linear services.

While one could deliver IP video as Moving Picture Experts Group-transport stream (MPEG-TS) over User Datagram Protocol (UDP) via IP multicast, this paper we focus on the use of ABR with multicast. The main advantage of using ABR for delivery of such video to the home, is that players can receive ABR streams as they normally would, while the gateway in the home converts the multicast streams received over UDP to unicast ABR delivered over TCP/IP to the players.

3. Multicast ABR

The goal with multicast ABR is to deliver live linear IP video fragment files to a small web cache resident in the DOCSIS gateways, with intent to deliver these fragment files to the gateway cache moments before

they are requested by an IPTV player such as an IP-STB. The gateway also implements a transparent web proxy, so it can attempt to fulfill video fragment requests from players within the home from its local cache. An example of such an architecture is described in [4]. In the following sections we describe such an architecture in further detail.

3.1. An Architecture for Switched Multicast Live Linear Delivery

The architecture for delivering live linear multicast IPTV services can have many similarities to the traditional quadrature amplitude modulation (QAM) and MPEG-TS-based switched digital video (SDV) system architecture. For example, in traditional SDV systems, client software in the customer premise equipment (CPE) would communicate with a centralized controller server to signal the desire for the availability of a specific channel. The "switched" concept in both SDV and in the switched multicast live linear IPTV architecture comes about because the operator has more channels in its program library than can be supported by the capacity allocated on the access network. Both systems are designed to make certain that channels, which are actively being watched by customers, are allocated network bandwidth and channels that are not being watched do not have bandwidth allocated to them.

A switched multicast live linear IPTV architecture may include the following logical entities:

- A set of multicast clients
- A set of multicast servers
- A set of multicast controller servers - consisting of:
    - A set of multicast coordinators
    - A set of multicast controller routers (MCRs)

Figure 2 below depicts a high level switched multicast live linear IPTV architecture.

The multicast client can be integrated into a DOCSIS gateway device provided to the customer. Many current DOCSIS gateway devices include an application processor well suited to running the multicast client. The client is responsible for control-plane communication with the MCR and multicast coordinator, Internet Group Management Protocol (IGMP) messaging, reception of the multicast streams, and is responsible for inserting video fragment files delivered via multicast into an onboard HTTP web cache. The client also implements a transparent web proxy, so it can attempt to fulfill video fragment requests from IPTV players within the home from its local cache.

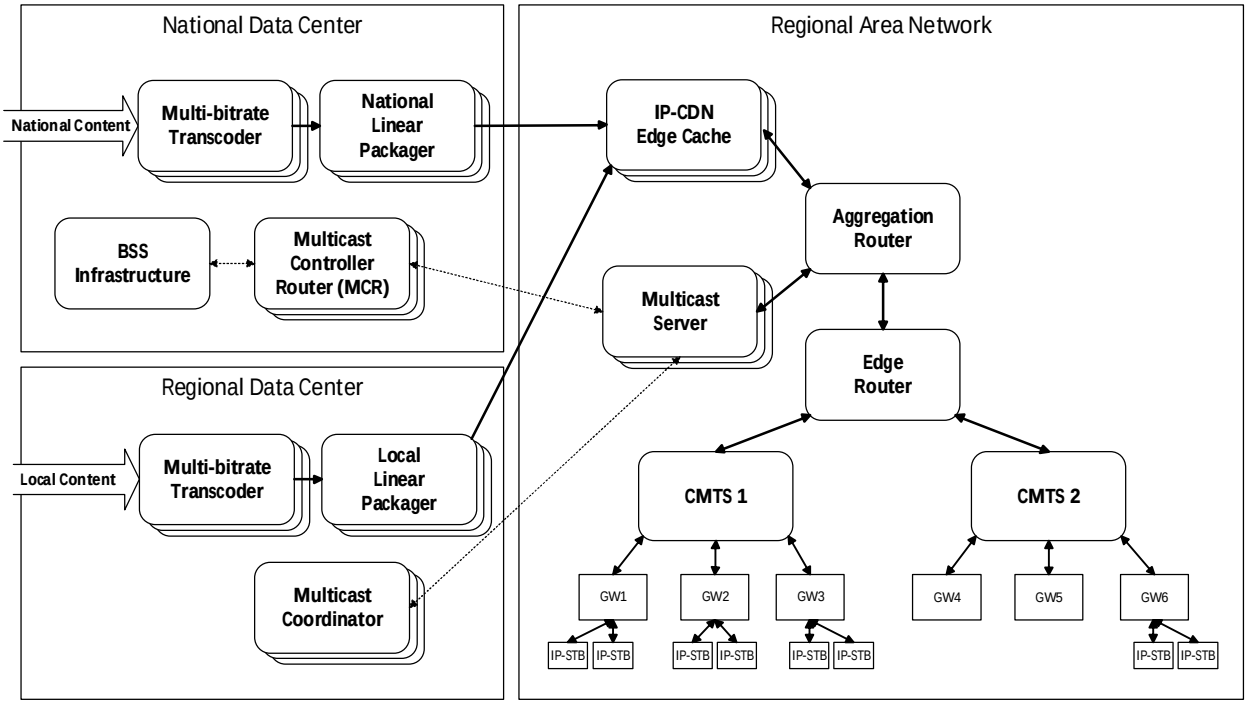

Figure 2 High Level Switched Multicast Live Linear IPTV Architecture

The multicast controller router (MCR) is a multicast controller server component that may be located in a national or regional data center. The MCR's primary task is to be the first point of configuration of a multicast client as it comes online. The MCR can attempt to ascertain the physical location of the client within the operator's network and can hand-off the client to a specific multicast coordinator server that is appropriate for the location of the client.

The multicast coordinator is a multicast controller server component that may also be located in a national or regional data center. The controller's tasks include providing the specific regionalized configuration of the multicast clients it serves, maintain ongoing control plane communications with the clients to ascertain the popularity of a given linear channel within a specific service area, and providing the clients the mapping of linear stream IDs and bitrates to specific multicast source and group (S,G) IP address and UDP port numbers. The coordinators are also responsible for the ongoing configuration of a set of multicast servers. In particular, when the coordinator determines there is a need to make a specific channel available to the multicast clients in a certain area, the coordinator signals the appropriate multicast server to begin multicasting a specific stream ID and bitrate on a particular S,G and UDP port. Conversely, when the coordinator determines a channel is no longer needed, it signals the multicast server to curtail multicasting that specific stream.

The multicast server component may be located in the headend or hubsite regional area network where it has high capacity and low latency connectivity to the DOCSIS CMTSs. The multicast server operates under the control of the multicast coordinator. When directed to do so by the multicast coordinator, the multicast server begins obtaining video fragment files - via unicast HTTP - from the operator's IP-CDN edge caches and begins inserting these fragments into a specific multicast stream for consumption by the multicast clients. When the multicast coordinator instructs the multicast server to tear down a specific

multicast stream, the server curtails unicast HTTP fragment Gets and curtails transmission of the multicast stream.

The switched multicast live linear IPTV architecture builds on top of an operator's existing ABR unicast video IP-CDN infrastructure. As shown in Figure 3, the unicast and multicast delivery mechanisms co-exist with parallel delivery paths. One advantage of the architecture is that the unicast players' operation remains unchanged even with the introduction of this parallel multicast path. The description of the details of the unicast IP video delivery path from transcoder to fragmenter to packager to HTTP origin server to IP-CDN caches is beyond the scope of this paper.

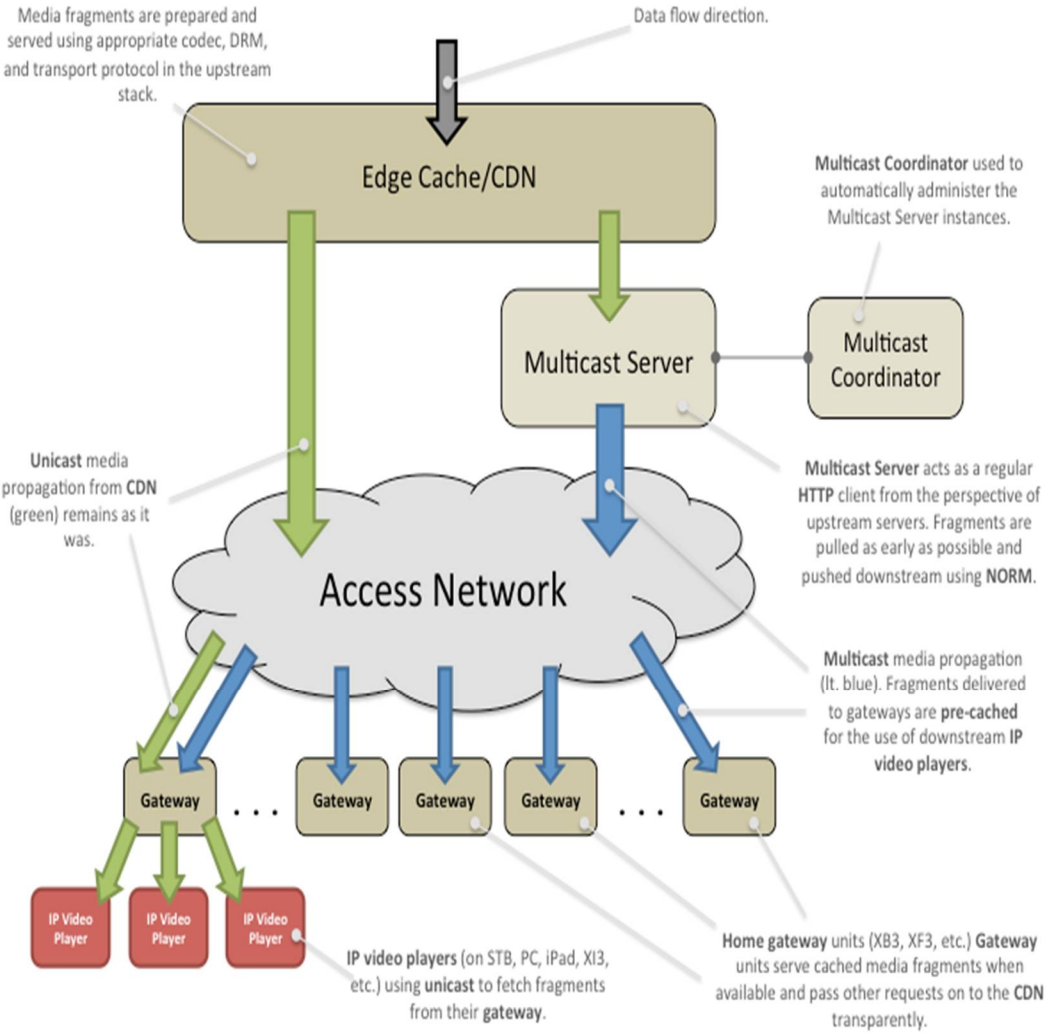

Figure 3 Parallel Unicast and Multicast Delivery Paths

3.2. NACK-Oriented Reliable Multicast (NORM)

NORM is a protocol that can provide end-to-end reliable transport of bulk data objects or streams over generic IP multicast routing and forwarding services. The NORM protocol is specified in [5]. The protocol leverages the use of forward error correction (FEC) object repair. By leveraging NORM for delivery of ABR, we can use the FEC mechanism built into it. With FEC the ability to recover from errors is improved by transmitting additional bits, called FEC bits. NORM allows the number of FEC bits used to be configurable. Operators can therefore choose an appropriate balance between bandwidth overhead and ability to recover errors. The number of FEC bits could also be varied for different streams depending on how critical the stream is, and the quality of the HFC plant.

3.3. What Streams to Multicast?

With the above-described architecture an operator can choose to make their entire linear lineup available via multicast. Such an approach will indeed be bandwidth efficient. However this can come at the cost of multicast server resources. Some operators may choose to only multicast the most popular content. The most popular content may be determined in a static manner and the system simply configured to make those popular streams available via multicast. Alternately the operator can dynamically measure popularity and thereby determine what streams to multicast. Such a dynamic system may have different services delivered via multicast at different times in the day or week or otherwise.

3.4. What ABR Profiles to Multicast?

In a multicast ABR architecture one question facing operators is: what ABR profiles to make available via multicast? Since multicast delivery generally is quite efficient, an operator could choose to make multiple profiles available via multicast. However if a number of profiles have very limited concurrency, it may not be worth the effort to make those profiles available via multicast. In any case, it is recommended that each profile that is multicast use a different source IP address or different multicast group address. This is because IP multicast forwarding on the network only uses the IP source and destination, (often referred to as S,G) address to forward multicast streams. So in order to prevent the forwarding of multiple profiles over the DOCSIS network when only one profile was requested, it is best to keep each profile on different S or G addresses. This ensures that only the requested profile is forwarded, thereby conserving DOCSIS bandwidth.

An operator could choose to multicast the highest profile available for a video stream. In this case, any requests for other profiles are fulfilled via unicast delivery. Such an approach would make sense in cases where an operator means to deliver the highest profile and has other profiles available to address any error conditions. This approach would also be reasonable if the expected concurrency for other profiles is minimal and the multicast server resources available are limited.

At the other extreme, an operator could choose to multicast all possible profiles. Just because all profiles are available via multicast doesn't mean all profiles are forwarded all the time. Only profiles that are requested by multicast clients are forwarded by the CMTS. This ensures that bandwidth on the DOCSIS

network is used efficiently. However this comes at the cost of requiring significant multicast server resources to stream all profiles via multicast.

Between the two extremes described above, an operator could also dynamically determine what profiles to multicast or not, based on real-time analytics on what streams and/or profiles are being requested more frequently than others. This is another reason the analytics from the deployment are important and are hence discussed in further detail in section 6. Based on the analytics, the operator could periodically decide what profiles are most popular and statically configure those to be delivered via multicast. Alternately, the determination of what profiles to multicast can be made dynamically by taking both the viewership and concurrency of streams in a single service group, and the popularity of that profile across a set of service groups/CMTSs and the multicast server capacity available to deliver streams via multicast. The rest of the multicast ABR architecture is relatively unaffected by the choices an operator makes regarding which streams and which profiles to multicast.

4. Bandwidth Savings

The proposed approach provides significant bandwidth savings over the DOCSIS network, due to high concurrency typically amongst popular linear content. Due to caching in the home gateway, this can provide bandwidth savings even when users pause and continue watching the video after a brief pause. How much of such “time-shifted” viewing can be served out of the multicast stream is simply a function of the size of the cache in the home gateway and the size of the time shift buffer in the IPTV player device.

Another big advantage of using multicast, is that significant bandwidth savings is also achieved in the regional network. This is because if users in multiple service groups are watching the same video stream, the CMTS has to join the multicast stream only once. One can see from Figure 4 that similarly there are savings even when the same stream is being watched on multiple CMTSs, since the aggregation router needs to receive only one copy of the multicast stream and can then replicate it across the multiple CMTSs. Overall the savings in the regional network is significant.

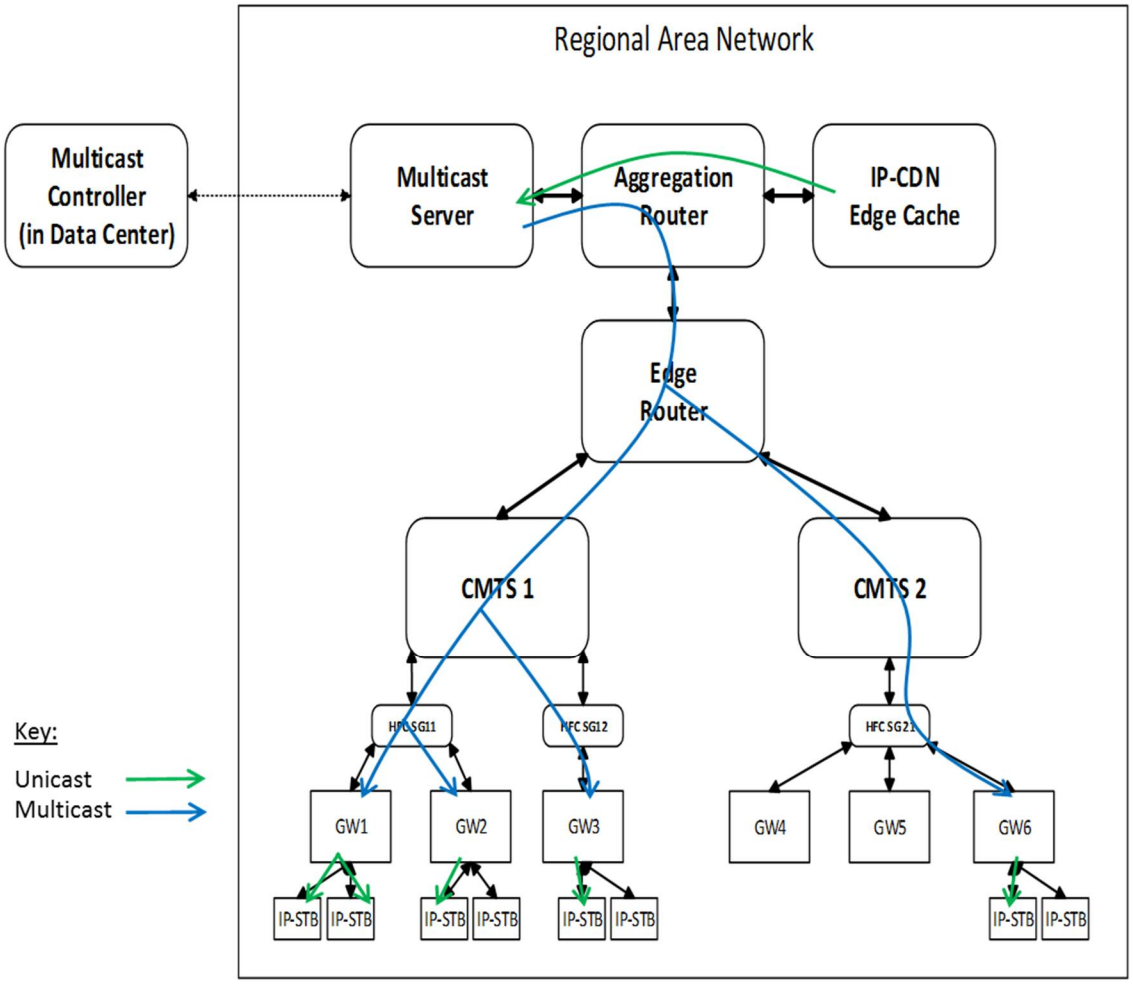

Figure 4 Bandwidth Savings in a Regional Network

The actual amount of bandwidth required to deliver the linear service via ABR multicast is out of scope of this paper. However there are previously published papers that the reader can refer to, to get an understanding of the bandwidth required for such a service when using multicast delivery. Reference [6] presents data gathered from an IPTV deployment that uses multicast delivery. Although that paper primarily focuses on linear video delivered via MPEG-TS over UDP via multicast, the concurrency and bandwidth savings observations should equally apply to ABR delivered over multicast. Reference [7] provides an in-depth estimate of capacity required for delivery of IP video services including both linear and on-demand. Much of its modeling is based on data gathered from switched digital video (SDV) deployments, and since multicast ABR is very similar to SDV in subscriber behavior the bandwidth estimates and bandwidth savings outlined in [7] should apply very well to the multicast ABR use case.

5. Multicast ABR System Considerations

In order to ensure that the above described multicast ABR architecture functions as expected and delivers on the promise of bandwidth savings, a number of system characteristics have to be taken into consideration. In this section we discuss those aspects and highlight recommended configurations to maximize the performance of the system overall.

5.1. Congestion Handling

Before we examine how congestion is handled in the case of multicast ABR, let us first take a look at how congestion handling occurs with standard ABR delivered via unicast. Standard ABR is delivered via TCP, so it leverages TCP/IP's congestion avoidance mechanism, to detect and adapt to congestion. Moreover TCP has an in-built re-transmit mechanism that ensures that packets are delivered reliably to the client. Finally ABR itself has a rate adaptation algorithm that enables it to adjust the rate profile selected to network bandwidth available to it.

Multicast ABR is not able to easily leverage any of the above-described mechanisms. Hence congestion avoidance and congestion handling must be addressed carefully for multicast ABR delivery.

It is important to deliver these multicast streams in a reliable manner because any delay or dropped packets now impact multiple subscribers instead of a single subscriber, as in the case of unicast. Additionally, cache misses that occur due to delayed/lost packets would result in clients' requests being fulfilled via unicast. Since a large number of clients might be viewing the same video, this may result in a significant increase in bandwidth demand due to delivery of the same video fragment multiple times – one copy per subscriber.

Irrespective of all the proper configuration of the various components, there is always a possibility of packet drops and uncorrectable errors. In such cases we recommend the multicast client recover the lost packet via unicast request. One method would be to request the entire ABR fragment in which one or more packets were lost. However that can be quite wasteful in terms of bandwidth, since a fragment may be several Megabytes in size while only 1-2 packets worth a few thousand bytes might be missing. So we recommend that for optimal performance, clients use HTTP Byte Range Get Requests to get only the missing packets.

5.2. Multicast Server

Unlike typical real-time streaming applications, these multicast streams are really small files streamed out of a server. Hence the streams may be a lot more bursty than typical real-time streaming video. In fact when delivered as unicast, this burstiness is easily accommodated by TCP's congestion window adapting to lost/delayed ACKs from the client, and the re-transmission capabilities of TCP. However when these ABR fragments are delivered as UDP multicast, due to the presence of multiple receivers there isn't an easy way for the multicast client to signal the server to slow down. In fact, different multicast clients on different segments of the HFC network may be experiencing differing delays. Hence it is important for the streamer to have an ability to pace the packets out smoothly.

Pacing packets out very smoothly such that there is little to no burstiness may help avoid packet loss in the network. However such an approach may increase the risk of cache miss as the last couple of packets for an ABR fragment may not yet be in the gateway's cache when the client requests it, resulting in fetching of the entire fragment via unicast. Hence a more prudent approach may be to pace the packets out of the streamer, yet pacing them out slightly (say 10-20%) faster than real-time. This may be a reasonable trade-off between reducing risk of packet loss in the network due to burstiness while simultaneously reducing the risk of cache miss on the home gateway.

Another aspect to keep in mind is the packet size used by the multicast server should be such that fragmentation in the network is avoided. Fragmentation in the network can lead to increase in bandwidth required due to additional overhead due to fragmented packets.

5.3. Cache Hit Optimizations

When a player's video fragment request is served out of the DOCSIS gateway's cache, we call that a cache hit. If a player's request is unable to be fulfilled from the cache, then a unicast request has to be made for that fragment, and it is delivered via unicast. Hence a cache miss is expensive because it is not able to leverage the multicast gains.

Operators can maximize the likelihood of a cache hit by properly planning their multicast delivery so that the most popular content and profiles are indeed delivered via multicast. Having done that, there are still a few more tools available for operators to further optimize their cache hit rate.

One such tool available to operators is manifest file manipulation. The manifest files made available to players could be manipulated to be one or more ABR fragments behind the latest one available via multicast. What this ensures is that the multicast client always stays one or more fragments ahead of the player. With such an approach, the fragment requested by the player is very likely to be already in the cache of the gateway. This approach eliminates the possibility of a race condition whereby the gateway has joined the multicast stream and is consuming bandwidth, while the player is requesting the same fragment that the gateway is currently receiving via multicast, resulting in a cache miss and a unicast request over the cable network. Such a cache miss would result in the stream consuming double the bandwidth, due to it being downloaded both via unicast and multicast. The above described manifest manipulation method eliminates such race conditions.

Additionally operators could choose to deploy gateways with sufficient cache such that even certain instances of time-delayed viewing can be addressed out of the cache. That is even if an end-user pauses/rewinds the linear video playback briefly, with a reasonable cache size, the gateway should still be able to serve the fragments requested by the player, even though it is a few fragments behind the fragment being delivered via multicast.

Furthermore gateways can have intelligent caching algorithms whereby they predict what video channels a player might switch to and hence start receiving and caching that multicast stream too, especially if it is already being streamed over the network to that service group. This would of course have to take into account the cache capacity available and the network capacity available. If done right, this can increase the chance of a cache hit thereby reducing the bandwidth required for the linear service.

5.4. Quality of Service (QoS)

A number of QoS choices are available to operators in delivering these streams. multicast QoS capabilities defined in DOCSIS 3.0 could be used to provide committed information rate (CIR) service to each of the multicast flows. The CIR rate provided to each stream must at least be equal to the bitrate of that ABR profile. In fact, as described above, it may be better to provide a CIR 10-20% higher than the profile bitrate to match the rate at which the streamer sends the fragment. Additionally the CMTS must have sufficient buffer to handle the burstiness of the streamer. The major advantage of per-flow QoS for the multicast streams is that each flow is protected from the other streams. That is, even if one multicast stream is misbehaving for some reason, it results only in delivery of that stream being impacted. Other streams in the same CMTS interface remain protected from such a misbehaving stream.

For a CIR-based QoS mechanism to function well, the CMTS needs to know what bandwidth to reserve for a given multicast stream. This could be either dynamically signaled to the CMTS or can be configured on the CMTS directly for all the multicast streams. Each multicast stream is identified via a combination of source and group address (S,G). With such a static configuration, the CMTS will reserve the associated bandwidth for that stream only when that (S,G) is requested by a multicast client.

Alternately, the operator could provision a pool of bandwidth across a number of ABR multicast streams. This aggregate pool of bandwidth approach will work well to absorb the bursts of packets from a plurality of streamers. Given a sufficient number of streams destined for the same DOCSIS bonding group, should minimize the risk of all streams bursting at the same time. However when multiple servers are streaming to multiple CMTSs or DOCSIS bonding groups, there is always a possibility that bursts from different streamers end up being aligned, resulting in a very large burst of traffic being received on the CMTS. So in any case it is important to deploy servers which can be configured to limit the burstiness of the packets as described in Section 5.2.

The downside of an approach with an aggregate pool of bandwidth being dedicated to multicast traffic, with no per-flow CIR is that one misbehaving stream can impact the delivery of other well-behaved streams. One misbehaving stream can use up a disproportionate amount of bandwidth leading to loss of packets on other streams.

Over and beyond the server's packet emission characteristics, the encoder itself may encode the video at a variable bit rate. This may result in different video fragments being of different sizes, thereby leading to varying packet rate from one video fragment to another.

5.5. Bandwidth Sharing Across Services.

Many CMTSs include a connection admission control (CAC) feature. An operator can configure the CAC feature to control multiple services that are sharing the same DOCSIS bandwidth. For example, some operators use CAC to limit the aggregate amount of capacity on a given service group that can be allocated to QoS-enabled IP voice traffic. This provides a protection mechanism for best-effort traffic on the service group, should something go haywire with the amount of voice traffic. Operators may also

want to consider using CAC to limit the aggregate amount of capacity that can be allocated to multicast IP video traffic.

6. Analytics to Measure Efficiency of the Multicast Architecture

In this section, we will examine the key metrics operators can use to measure the efficacy of their multicast architecture and how to use analytics to further optimize network efficiency. We will touch on three important metrics that operators can monitor: CMTS interface multicast usage, multicast gateway cache hit ratio, and NORM forward error correction (FEC) statistics.

6.1. CMTS Interface Multicast Usage

Many operators already have an operational support system (OSS) infrastructure that monitors interface usage and percent utilization on all of their CMTS interfaces. Tracking utilization allows operators to schedule capacity augments as customer consumption grows over time. This existing infrastructure can be used to monitor multicast usage by making only minor modifications to infrastructure tools based on the simple network management protocol (SNMP).

Each CMTS vendor may instrument the DOCSIS interfaces on their chassis using the standard Internet Engineering Task Force (IETF) interface management information bases (MIBs) in a slightly different manner than another vendor. So the monitoring of multicast usage has to be slightly customized depending on the specific CMTS implementation.

For some CMTSs, operators may define a unique DOCSIS downstream bonding group for multicast within each high speed data (HSD) service group. This multicast bonding group may represent unique DOCSIS QAM resources and capacity or it may completely overlap with an existing unicast bonding group. In either case, measuring multicast use is trivial if a multicast bonding group is defined on the CMTS, as the CMTS will automatically instantiate a row in the IETF RFC 2863 ifXTable for this bonding group.

With a unique bonding group interface instantiated in the ifXTable, operators need to simply poll the ifHCOutOctets counter for each CMTS multicast bonding group interface. No doubt many operators are already polling the ifHCOutOctets counter for each existing (unicast) DOCSIS interface on their CMTSs today. Sometimes operators augment the use of the ifXTable counters with those provided by CMTS vendor proprietary MIBs. One interesting test is to compare usage on CMTS unicast DOCSIS interfaces before and after enabling Multicast. The drop in usage on the unicast interfaces is a direct indicator of the bandwidth savings provided by multicast.

For other CMTSs, operators may not have a dedicated multicast bonding group interface to monitor, so we need to get a bit more creative to work around the limitations of the ifXTable. In particular, the ifXTable lacks a high capacity (64 bit) output multicast octet counter. So as a workaround, we can make use of ifXTable's ifHCOutMulticastPkts counter combined with knowledge of the NORM multicast packet size to calculate multicast usage in units of bits per second.

6.2. Multicast Controller Analytics

Depending on the capabilities of the multicast controller infrastructure that an operator deploys, it can potentially provide a great deal of visibility into the health and efficiency of the end-to-end multicast IPTV delivery system. For example, an operator could build a management tool that collects several types of analytics from the multicast controller. We will touch on a few of these analytics next.

Since the multicast controller is responsible for runtime configuration of the multicast servers, it can be used to provide the list of multicast servers with which it is in current communication. For example, if the controller has lost connectivity to a particular multicast server, it can report that disruption. The controller can also provide a list of linear streams that are actively being transmitted by each multicast server and can indicate whether each stream is being multicasted due to near-real time channel popularity or because the channel has been statically configured to be multicasted, regardless of current popularity.

In addition, it may be useful for an operator's multicast gateways to report performance statistics up to a central collection point. An existing command and control plane between the multicast controller and the multicast client in the gateways may be utilized for this purpose. If the control plane supports analytics reporting from the client, one of the most important statistics to report is cache hit ratio. The cache hit ratio is a measure of how many ABR fragments were served to in-home IPTV players out of the gateway's cache, versus how many required a unicast HTTP Get Request and Response across the access network. Ideally, cache hit ratio statistics can be aggregated and presented by the controller on a per-gateway and on a per-stream basis. A low cache hit ratio value would indicate there is a problem in the end-to-end multicast IPTV delivery network which needs to be corrected.

6.3. NORM Forward Error Correction Statistics

If the operator has a control plane between the multicast controller and the multicast client that supports analytics reporting from the client, it may be useful to the have the client report NORM layer statistics up to the controller. For example, each client could report the number of NORM objects that were corrected by the NORM FEC for a given stream, and the controller could present aggregated NORM statistics to the operator. Ideally, the controller could present these NORM stats on a per-gateway and per-stream basis.

The operator can use the NORM statistics to determine how "clean" a given portion of their network is. For example, if the clients are reporting that a high number of NORM objects were not correctable, the operator may want to increase the number of NORM FEC bits used within the multicast streams. If there are little to no uncorrectable NORM objects, the operator may be able to reduce the number of NORM FEC bits, thereby reduce NORM network overhead, and increase access network efficiency.

7. Summary

Cable operators are launching more and more IP video services. Although the trend started with on-demand services to unmanaged devices, it has grown to linear services to unmanaged devices. Furthermore, operators are starting to test the delivery of linear services to managed devices such as IP-STBs. Linear services tend to drive high concurrency for the most popular content, which leads to

significant bandwidth requirements. It is imperative that the industry has a bandwidth efficient method to deliver these linear services. In this paper, we have outlined an ABR-based multicast delivery method for linear services. Leveraging ABR to deliver linear services over multicast enables operators to address their subscribers' needs while leveraging their investment in ABR technologies and yet using their network resources efficiently.

In this paper we have discussed what a multicast ABR delivery architecture might look like and various choices available to operators in this respect. It is important to recognize and understand the implications when using multicast for delivering ABR, because originally ABR was designed for unicast delivery over TCP. We have discussed various factors that influence the efficiency of the system and their trade-offs and included recommendations on how operators can optimize their deployments. We have also outlined analytics that can help operators understand how well their network is performing and can help measure efficiencies achieved and identify problem areas. With careful planning and consideration, operators can successfully deliver linear services via multicast ABR.

Acknowledgements

The authors would like to thank Matt Flowers and Harish Venkataramudu of Comcast Cable for their contributions to this paper.

Abbreviations

ABR	Adaptive Bit Rate
CAC	Connection Admission Control
CIR	Committed Information Rate
CMTS	Cable Modem Termination System
DOCSIS	Data-Over-Cable Service Interface Specifications
FEC	Forward Error Correction
HFC	Hybrid Fiber-Coax
HSD	High Speed Data
HTTP	Hypertext Transfer Protocol
IGMP	Internet Group Management Protocol
IP-CDN	Internet Protocol - Content Delivery Network
MCR	Multicast Controller Router
MIB	Management Information Base
NORM	NACK-Oriented Reliable Multicast
OSS	Operations Support Systems
QoS	Quality of Service
SDV	Switched Digital Video
STB	Set-Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol

Bibliography & References

1. Cisco White Paper, "*Cisco Visual Networking Index: Forecast and methodology, 2014-2019,* "http://newsroom.cisco.com/dlls/index.html
2. Comcast's Stream Service, http://corporate.comcast.com/comcast-voices/a-new-streaming-tv-service-from-comcast
3. *" DOCSIS 3.0 MAC and Upper Layer Protocols Interface Specification"*, Cable Television Laboratories
4. *"IP Adaptive Bit Rate Architectural Technical Report"*, OC-TR-IP-MULTI-ARCH-V01-141112, Cable Television Laboratories
5. *NACK-Oriented Reliable Multicast (NORM) Transport Protocol*, RFC 5740, B. Adamson, C. Bormann, M. Handley, J. Macker, November 2009
6. *"Pioneering IPTV in Cable Networks"*, by John Horrobin and Gitesh Shah, Cisco Systems, SCTE 2013
7. "HFC Capacity Planning for IP Video", Sangeeta Ramakrishnan, Cisco Systems, SCTE 2011
8. "*The Interfaces Group MIB*", RFC 2863, K. McCloghrie, F. Kastenholz, June 2000, http://www.rfc-editor.org/rfc/rfc2863.txt