TECH BRIEF

The Viability of Multicast ABR in Future Streaming Architectures

April 2019

The Streaming Video Alliance (the "Alliance") is an industry forum open to all companies from all sectors of the online video value chain. The Alliance focuses on the ecosystem, architecture and best practices needed to support the future of online video.

Membership is comprised of industry leaders from the entire online video ecosystem, including content providers, service providers, commercial CDNs and streaming video technology providers.

Notice:

This document has been created by the Streaming Video Alliance. It is offered to the Alliance Membership solely as a basis for agreement and is not a binding proposal on the companies listed as resources above. The Alliance reserves the rights to at any time add, amend or withdraw statements contained herein. Nothing in this document is in any way binding on the Alliance or any of its members. The user's attention is called to the possibility that implementation of the Alliance agreement contained herein may require the use of inventions covered by the patent rights held by third parties. By publication of this Alliance document, the Alliance makes no representation or warranty whatsoever, whether expressed or implied, that implementation of the specification will not infringe any third party rights, nor does the Alliance make any representation or warranty whatsoever, whether expressed or implied, with respect to any claim that has been or may be asserted by any third party, the validity of any patent rights related to any such claim, or the extent to which a license to use any such rights may or may not be available or the terms hereof.

streaming video alliance

TABLE OF CONTENTS

AUTHORS

Primary author:

- Brian Stevenson (EDGE GRAVITY by Ericsson)

Additionally, the following people contributed to this paper:

- Ali C. Begen (Comcast)
- Guillaume Bichot (Broadpeak)
- Glenn Deen (Comcast)
- Damien Lucas (Anevia)
- Yoav Gressel (Qwilt)

GLOSSARY

Adaptive Bit Rate (ABR)—ABR is a method of streaming multimedia over computer networks that works by detecting a user's bandwidth and CPU capacity in real time and adjusting the quality of the media stream accordingly[1]. ABR ready content gathers multiple bitrate representations of the same original content providing different qualities. Each of these representations sub-stream is composed into a series of segments having an equal duration allowing a receiver to switch from one bitrate representation to another according to criteria which is principally the available server bandwidth.

Cache—In computing, a cache is a hardware or software component that stores data so that future requests for that data can be served faster; the data stored in a cache might be the result of an earlier computation or a copy of data stored elsewhere. A cache hit occurs when the requested data can be found in a cache, while a cache miss occurs when it cannot[2].

Content Delivery Network (CDN)—a CDN is a geographically distributed network of proxy servers and data centers designed to more effectively deliver certain types of traffic (i.e., streaming video) over the Internet[3].

Data over Cable Service Interface Specification (DOCSIS)—DOCSIS is an international telecommunications standard that permits the addition of high-bandwidth data transfer to an existing cable TV (CATV) system[4].

Edge—the Edge is a distributed information technology (IT) architecture in which client data is processed at the periphery of the network, as close to the originating source as possible[5].

Ethernet Passive Optical Network (EPON)—generally delivers 1 Gbit/s symmetrical bandwidth. Employs a single Layer 2 network that uses Internet Protocol (IP) to carry data, voice, and video, generally delivering 1 Gbit/s symmetrical bandwidth. Costs of EPON equipment are approximately 10 percent of the costs of GPON equipment because it does not require multi-protocol conversions, and the result is a lower cost of silicon. In GPON, there are three management systems for the three Layer protocols. EPON equipment is rapidly becoming cost competitive with VDSL[6].

Forward Error Correction (FEC)—forward error correction (FEC) or channel coding is a technique used for controlling errors in data transmission over unreliable or noisy communication channels. The central idea is the sender encodes the message in a redundant way by using an error-correcting code (ECC)[7].

---

[1] https://en.wikipedia.org/wiki/Adaptive_bitrate_streaming
[2] https://en.wikipedia.org/wiki/Cache_(computing)
[3] https://en.wikipedia.org/wiki/Content_delivery_network
[4] https://en.wikipedia.org/wiki/DOCSIS
[5] https://searchdatacenter.techtarget.com/definition/edge-computing
[6] https://www.quora.com/What-are-the-differences-between-GPON-and-EPON
[7] https://en.wikipedia.org/wiki/Forward_error_correction

File Delivery over Unidirectional Transport (FLUTE)—a file transport protocol used to deliver files over IP networks, including the Internet and unidirectional systems, from a sender to one or more receivers. FLUTE uses UDP, an unreliable transport protocol, and so reliable delivery must be guaranteed by other means[8].

Gigabit Passive Optical Network (GPON)—provides three Layer 2 networks: ATM for voice, Ethernet for data and proprietary encapsulation for voice. Promises 1.25 Gbit/s or 2.5 Gbit/s downstream and upstream bandwidths scalable from 155 Mbit/s to 2.5 Gbit/s. GPON does not support multi-cast services. This makes support for IP video more bandwidth-consuming[9].

Internet Protocol (IP)—the principal communications protocol in the Internet protocol suite for relaying datagrams across network boundaries. Its routing function enables internetworking, and essentially establishes the Internet[10].

Multicast—In computer networking, multicast is group communication where data transmission is addressed to a group of destination computers simultaneously. Multicast can be one-to-many or many-to-many distribution[11].

NACK— NACK stands for Negative Acknowledgement. It is one of the error resiliency mechanisms in WebRTC. NACK is a way for the receiving end to indicate it hasn't received a specific packet[12].

Negative Acknowledgement Oriented Reliable Multicast (NORM)— This protocol can provide end-to-end reliable transport of bulk data objects or streams over generic IP multicast routing and forwarding services. NORM uses a selective, negative acknowledgment mechanism for transport reliability and offers additional protocol mechanisms to allow for operation with minimal a priori coordination among senders and receivers[13].

Optical Network Terminal (ONT)—Equipment from the telephone company that terminates its optical fibers at the customer's premises. Using electricity from the customer's AC source, the optical network terminal (ONT) converts the incoming optical signals into electrical signals for telephone, TV and Internet[14].

Origin—a server or other storage location that contains files to deliver to a requesting user.

Reverse Proxy— A reverse proxy is a server that takes a client request and forwards it to the backend server. It is an intermediary server between the client and the origin server itself. A CDN reverse proxy

[8] https://onlinelibrary.wiley.com/doi/abs/10.1002/dac.835
[9] https://www.quora.com/What-are-the-differences-between-GPON-and-EPON
[10] https://en.wikipedia.org/wiki/Internet_Protocol
[11] https://en.wikipedia.org/wiki/Multicast
[12] https://webrtcglossary.com/nack/
[13] https://tools.ietf.org/html/rfc5740
[14] https://www.pcmag.com/encyclopedia/term/65027/optical-network-terminal

takes this concept a step further by caching responses from the origin server that are on their way back to the client[15].

Reliable Multicast Transport (RMT)—like the User Datagram Protocol, multicast does not guarantee the delivery of a message stream. Messages may be dropped, delivered multiple times, or delivered out of order. A reliable multicast protocol adds the ability for receivers to detect lost and/or out-of-order messages and take corrective action (similar in principle to TCP), resulting in a gap-free, in-order message stream[16].

Set-Top Box (STB)—A set-top box (STB) or set-top unit (STU) (one type also colloquially known as a cable box) is an information appliance device that generally contains a TV-tuner input and displays output to a television set and an external source of signal, turning the source signal into content in a form that then be displayed on the television screen or other display device[17].

Transmission Control Protocol (TCP)—TCP (Transmission Control Protocol) is a standard that defines how to establish and maintain a network conversation via which application programs can exchange data. TCP works with the Internet Protocol (IP), which defines how computers send packets of data to each other. Together, TCP and IP are the basic rules defining the Internet. TCP is defined by the Internet Engineering Task Force (IETF) in the Request for Comment (RFC) standards document number 793[18].

Residential Gateway (RGW)—a residential gateway (more commonly known as a home router or home gateway) is a device that allows a local area network (LAN) to connect to a wide area network (WAN) via a modem[19].

Unicast—Unicast is communication between a single sender and a single receiver over a network. The term exists in contradistinction to multicast, communication between a single sender and multiple receivers, and anycast, communication between any sender and the nearest of a group of receivers in a network. An earlier term, point-to-point communication, is similar in meaning to unicast. The new Internet Protocol version 6 (IPv6) supports unicast as well as anycast and multicast[20].

[15] https://www.keycdn.com/support/how-does-a-cdn-work
[16] https://en.wikipedia.org/wiki/Reliable_multicast
[17] https://en.wikipedia.org/wiki/Set-top_box
[18] https://searchnetworking.techtarget.com/definition/TCP
[19] https://en.wikipedia.org/wiki/Residential_gateway
[20] https://searchnetworking.techtarget.com/definition/unicast

ABSTRACT

This paper explores the technology behind Multicast ABR and its potential to improve how video content is streamed over IP, meeting the network operator's needs for a scalable way to deliver growing video traffic and end-users' demands for the best possible streaming experience.

Note: for the purposes of this paper and discussion, it is assumed that when considering mABR delivery, it is for live streaming content (which may or may not include time-shifting functionality). VOD content would still be delivered from a CDN.

Note: there are different terms often used to refer to chunked ABR delivery. In some cases, the word "segment" is used to describe a part of the video. In other cases, it is "fragments." And yet, in other cases, "chunks" is employed For the purposes of this paper, we consider these terms synonymous to represent chunked ABR delivery.

FOREWORD

Network operators are constantly on the lookout for technologies that will help them better manage and optimize the traffic on their networks. As streaming video viewership has grown dramatically, those same operators are struggling with the amount of HTTP traffic chunks flowing across their pipes.

Multicast Assisted ABR is a technology that could help alleviate some of that congestion for on-network delivery (i.e., across the operator transit and metro networks to the edge, and then possibly through the last mile to the home network). It's important to note that OTT service providers, like Hulu, and content owners, like Viacom, wouldn't implement this technology, at least not as an internet-based solution as it's primarily designed for deployment within managed networks, such as a cable operator.

This tech brief provides an overview of how Multicast Assisted ABR technology works, the potential benefits to a network operator, and some of the challenges in wide scale adoption.

EXECUTIVE OVERVIEW

As consumers have flocked to streaming video, the networks that transport it have come under increasing pressure. From the Internet to ISPs to cellular providers, infrastructure is choking on the volume of video traffic consumed through PCs, Smart TVs, connected devices, mobile phones, and more. It is an unfortunate situation for network operators who have little chance of being compensated to scale their network capacity even as their users demand a "near broadcast-like" streaming experience. The infrastructure problems are exacerbated further by the streaming format, which has moved from specialized protocols (like RTMP and RTSP) which could deliver video content more cost effectively, to chunked HTTP. Streaming billions of small objects in a Unicast model provides little opportunity to meet end-user quality expectations, especially as user streaming consumption continues to grow both in volume and geographic reach.

Network operators have long been looking for cost-effective and efficient solutions to the growing problem of streaming video scalability. A variety of industry associations, such as DVB and CableLabs, have put forth specifications and other documents that attempt to provide a better way to stream video content over Multicast Assisted ABR (mABR), as opposed to Unicast. But since these documents have been published[21], there have been few, if any, commercial deployments of the technology, leaving network operators to serve every single requesting user with an identical stream of content.

The problem, though, is not with the underlying technology. mABR has a lot of applications to help reduce the congestion associated with streaming video delivery within the operator's network. For example, deployment of mABR into the operator caching infrastructure could help alleviate some backhaul congestion. The problem of adoption ultimately, and long-term success, resides with the client device ecosystem. Unfortunately, there is not a very robust ecosystem in the market today to make operator investment in mABR technology and deployment a good decision. If they did choose to deploy, it might require new hardware at the client end-point (with software that supports mABR delivery). From STBs to smartphones to RGWs, the playback ecosystem must be robust enough to warrant deployment. Otherwise, it's a chicken-and-egg situation, with each side—the operators and the vendors—waiting for the other to make the first commitment.

The future promise of mABR is there, even as operators and CDNs continue to invest in edge cache capacity, even as 5G promises unheralded bandwidth. None of that ultimately solves the internal network congestion and backhaul issues that operators face when handling the exponentially-growing amount of video traffic. For that, mABR might be a valuable part of the solution.

[21] The first Multicast Assisted ABR specification was first published by CableLabs on 11/12/2014 (https://apps.cablelabs.com/specification/ip-multicast-adaptive-bit-rate-architecture-technical-report/). DVB released a mABR reference architecture (for comment) on March 9th, 2018 (https://www.dvb.org/news/dvb-releases-reference-architecture-for-ip-multicast) and is currently developing the specification related the referenced interfaces.

INTRODUCTION

The broadcast industry is undergoing a pronounced change, perhaps the most disruptive since the introduction of the Community Antenna Television (CATV) in the 1950s—the transition to IP. Where once broadcasters moved content around the globe via satellites and downlink facilities, IP is becoming the choice of transport. It's much cheaper, and faster, to backhaul content to multiple locations over an IP network than it is through satellite! But as broadcasters have done so, it's also provided them a new opportunity—with content encoded from MPEG-2 to MPEG-4, they can now deliver their content to a host of Internet-connected devices. From IPTV to OTT, broadcasters/content owners and network operators (especially as they partner for, or purchase, content to differentiate their offering) are recognizing the flexibility of IP and how it can enable new ways to reach viewers.

But this transition to IP isn't happening alone. It's converging with other macro-trends, most notably the rise of live streaming, OTT, and TV-Everywhere services along with the expanding popularity of connected devices, although it was traditionally broadcasters, with affiliate relationships to local television stations, that make video content available to consumers via their televisions. Big content brands like the NFL and Viacom, in conjunction with services like FOX GO, DirecTV Now, and XFinity [22], are now delivering traditional live broadcast television directly to consumers using IP technology. In the case of pure-play OTT providers, partnering with network operators (i.e., Comcast, Verizon, Liberty Global) to install local caches near the edge of the network can help ensure the best possible viewer experience with their content[23]. This collaboration has the potential of delivering even greater advantages by leveraging the multicast capabilities on the operator network - not normally available on the public internet – to scale live linear delivery. This shift in connectivity, combined with the exponential growth of connected devices capable of consuming video (such as smartphones and streaming boxes) is feeding the consumers "anytime, anywhere" attitude—allowing them to be able to watch whatever live video they want, at any time, from any device.

Together, these trends—the transition to IP, the growth of pure IP content delivery, and a myriad of connected devices that can digest and playback video—are causing the scalability problems experienced by many network operators. To take advantage of those trends, most video distributors who are delivering live streaming content are employing ABR (Adaptive Bitrate) delivery, enabling them to do two things. First, they can leverage the CAPEX already invested in their IP networks; and, second, they can meet consumer demand for consumption of video across a variety of devices. Under ABR instead of a single video stream with a constant bit-rate, content is distributed as a series of short-duration video files or segments that are requested via standard HTTP-based protocols and stitched together "just-in-time" in the video player to render a continuous video stream. This approach is highly suited to

[22] Note: when it comes to considerations for deploying an mABR solution, it is most likely the video distributors who deliver a lot of live content (i.e., NBCSports, Viacom, DAZN, and others) would benefit most. Of course, video distributors themselves wouldn't deploy mABR. It would be the network operator, perhaps in conjunction with a video distributor (or perhaps in response to significant traffic delivery for one or more live content video distributors) who would deploy this delivery technology.
[23] Although Netflix does not deliver live content, they have led the way with this approach: https://openconnect.netflix.com/en/

managing a fragmented, consumer owned client eco-system, particularly when having to traverse the complex firewall and NAT configurations found on many IP networks.

The technology does have one significant drawback, however—scale, which primarily manifests during live linear events such as sports and news. ABR delivery involves providing a Unicast stream for every viewer. That means the network must repeatedly send the same chunks to different users. Imagine one million people all watching a 3.5Mbps stream. That's a lot of bandwidth consumed for just one stream! It's not hard to imagine how difficult it will be to meet consumer demand for a "near broadcast-like" experience as more consumers stream higher bitrates. Even with next-generation encoding (which can feasibly save up to 50% on delivery bandwidth without quality degradation), it will be very difficult to maintain the highest quality as scale increases exponentially.

A potential solution involves the deployment of mABR (Multicast Assisted Adaptive Bitrate). This approach was initially defined in a CableLabs[24] initiative. The ability to leverage UDP in order to multicast content segments provides scalability for popular linear events such as sports and news, or to pre-populate content within the operator edge cache. mABR can reduce edge bandwidth requirements from Petabytes to Megabytes for extremely popular, live events, particularly on shared access network infrastructure such as DOCSIS.

[24] https://apps.cablelabs.com/specification/ip-multicast-adaptive-bit-rate-architecture-technical-report/

HOW MULTICAST ASSISTED ABR WORKS

mABR uses a fundamentally different delivery paradigm than conventional HTTP-based ABR distribution. While it utilizes the same source segments as HTTP delivery, it encapsulates those segments within UDP/IP multicast packets and sends them over a multicast/broadcast capable network such as PON, cable, or xDSL. There is no need for each video player to request a unique stream, and segments are readily available to every client that is "listening in."

A typical implementation is comprised of three core components:

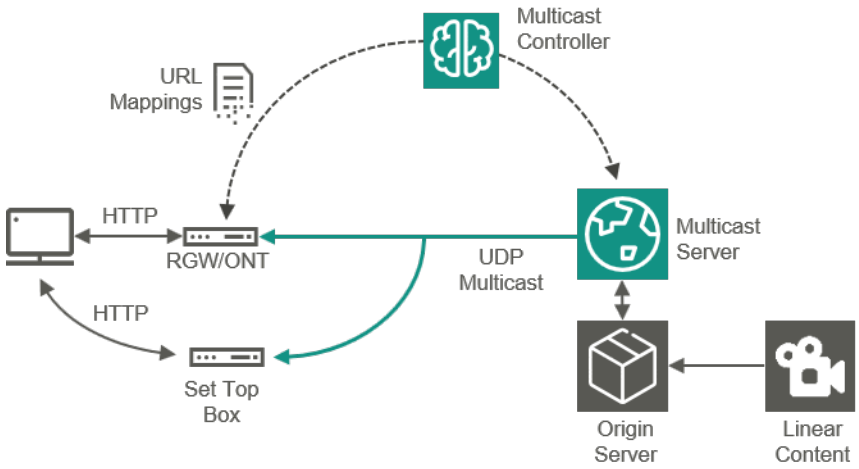


Figure 1: Typical Multicast ABR Implementation

- Multicast Server
- Multicast Client
- Multicast Controller

Multicast Server

The Multicast Server ingests content segments from source, normally a CDN origin or edge cache, and sends these over specific multicast address to Multicast Clients that reside downstream in the network. Depending on deployment considerations, one or more of the content representations for any given channel may be delivered[25]. Typically, only the highest bitrate representation is delivered, although multiple representations for the same channel may also be multicast to support low-capacity devices or

[25] Note: this is not only one bit-rate, but also one format (HLS or DASH), and one DRM. A video distributor or operator will often need to multicast multiple content adaptations to serve different devices which can significantly reduce the benefits of mABR and, in some cases, make the use of multicast technology only viable for delivery via the STB.

in-home variable-bandwidth networks which require lower bitrate content to maintain a suitable Quality of Experience for the viewer.

Multicast Client

The Multicast Client itself acts as a receiver and HTTP proxy. The deployment model is extremely flexible, supporting placements within multiple tiers of the network stack depending on network type, network topology, and the operator business case. Clients can be deployed within the home on either a STB or RGW, the ONT in the access network, or as part of the operator's edge caching infrastructure.

The Client is deployed downstream from the multicast server and fulfills the role of a proxy, supplying ABR content to any legacy ABR client (that supports only Unicast). The client joins multicast streams, acquires and caches the multicast segments and delivers those in response on a conventional HTTP Get request from any standards-based HTTP device (iPad, Roku, etc.). The fact that content is largely being delivered using multicast is completely transparent to the consumer device as the Multicast Client appears, for all intents and purposes, as a CDN edge.

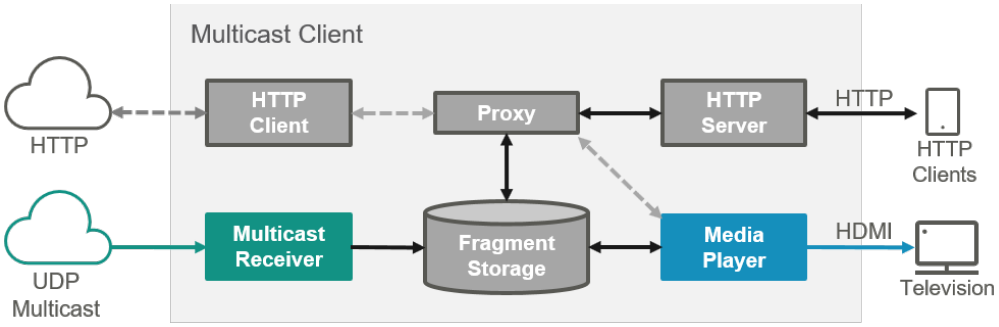


Figure 2: Multicast Client workflow

The Multicast Client exposes two interfaces on the operator network. The primary interface implements a multicast receiver that transparently caches content from the multicast streams in local storage to maintain a small DVR window, thus the nature of this storage is extremely transient. The client may optionally NACK all segments that contain uncorrectable errors for error handling and reporting. Failover and redundancy are handled seamlessly via a standard HTTP interface that, in the event of a multicast failure, may request content from the Service Provider CDN. If the Client receives an HTTP Get for content that, for any reason, is not in its local cache, the request for the missing segments is proxied to the operator CDN edge for fulfillment. This interface can also be utilized for "fast start" operations, retrieving the first chunk in a stream while the multicast receiver performs an IGMP/MLD Join to a channel's multicast address.

In order to service legacy ABR Devices, the client employs an HTTP Server interface on the consumer's network. When HTTP requests are received from any device on that network, the client fulfills them from local storage that contains segments cached from multicast streams.

Legacy ABR Devices are routed to the Multicast Client using standard DNS and 302 Redirects. These familiar approaches are shared with common CDN technologies and thus are transparent to the Legacy Device. As already outlined, the Multicast Client can proxy HTTP requests to the CDN edge on behalf of a Legacy Device to enable specific workloads, or the Legacy Client can be re-directed to the CDN Edge to support conventional HTTP delivery should the content not be available as a multicast stream. Depending on the CPU and memory footprint of the hardware hosting the Multicast Client, short trick-play operations – i.e. pause/rewind – could potentially be supported for Legacy ABR Devices, however a more characteristic workflow would be to simply route the client back to the CDN to consume more orthodox time-shift services as in general RWG's and ONT's are heavily resource constrained.

An alternate Client interface can be implemented within the context of a Set Top Box device[26]. This type of interface also consumes multicast segments from local storage in much the same way as described in the HTTP server interface scenario above. However, since the device interfaces directly with the screen, the Client can render the content via a standard interface such as HDMI, without the need to proxy it to a third-party.

Multicast Clients may also be embedded within a CDN edge. This allows for the edge cache to leverage multicast streams to populate content within a given cache and greatly reduces the number of requests for content from the edge to CDN mid-tier or origins, providing a significant efficiency when supporting multiple CDN edges or points-of-presence. Where the client is deployed in the CDN, it stores content directly in the edge cache.

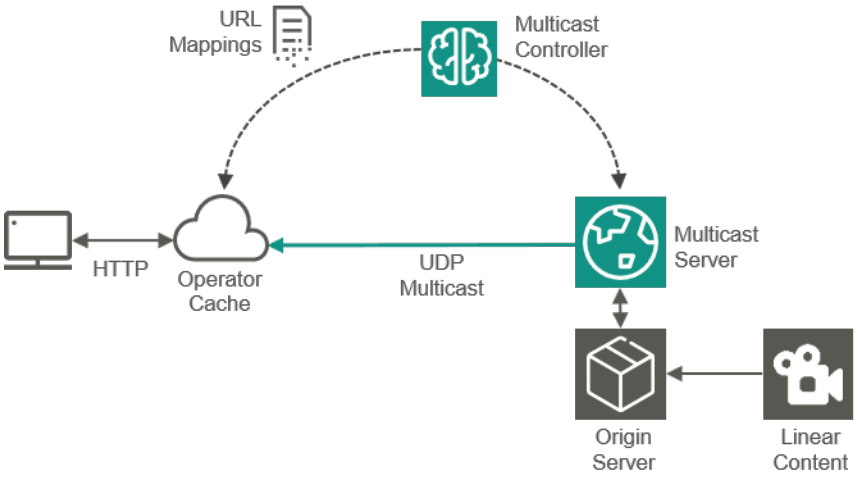


Figure 3: Multicast ABR Implementation with Inline Operator Cache

Multicast Controller

Delivery workloads are orchestrated by the Multicast Controller. The Controller is principally responsible for the management of all unicast to multicast URL mapping. This mapping is required by the Multicast Server when retrieving and encapsulating source content, allowing a specific URL on an origin to be

[26] Note that in Figure 2, the "Media Player" element would be software installed on the STB.

mapped to a downstream multicast address. It is also responsible for notifying downstream Clients of the "Channel Map," allowing the correct multicast address to be selected for a given channel or asset based on viewer selection.

Management of the map(s) can be either static or dynamic:

- Static map management—When implemented statically, the controller is configured with a list of all linear channel and asset URLs available on an Origin or Edge Cache with a corresponding output multicast address. The map, or a portion of the map, is sent to each Multicast Server, allowing it to source specific content URL's for multicast distribution. In Static mode, the Client is given the complete multicast map and joins a multicast group based on the channel selected by the viewer.

- Dynamic map management—Dynamic mode, as the name suggests, allows an optimal sub-set of multicast channels to be broadcast to a given Client based on the combined usage pattern of the access network. The Controller monitors content requests on the access network and when a set threshold for a channel is reached, it adds that channel to the map and informs the downstream Clients that the content is now available as a multicast stream. In networks where multicast resources are at a premium, this optimized set of channels allows a higher number of Clients to access the most popular content allowing viewers to achieve a higher Quality of Experience. An added benefit to dynamic mode is that the Controller can pre-join and pre-cache selected multicast streams to help reduce the unicast bandwidth traffic generated when many IGMP Joins occur in a short window of time[27].

Underlying Transport Protocols

NORM (NACK-Oriented Reliable Multicast) is the predominant, underlying transport protocol used in mABR deployments. It is standardized by the IETF in RFC 5740[28] and RFC 5401[29] and provides reliable end-to-end transport of bulk data objects or streams over common IP multicast routing and forwarding services.

NORM has the ability to leverage selective negative acknowledgment (an additional protocol mechanism to allow for operation with minimal coordination among senders and receivers) for added transport reliability. Ultimately, NORM is a congestion control solution to fairly share available network bandwidth with other transport protocols such as TCP and leverages the use of FEC-based (forward error

[27] In light of this, a question is exposed: "what happens when a channel becomes unpopular and removed from the multicast group? How is the client switched from multicast to unicast and vice versa?" The initial workflow would consist of the unicast<> multicast map being updated and communicated to the multicast server and multicast client. As a new multicast stream is instantiated or removed from the multicast group the consumer device can be redirected to the appropriate edge (proxy or CDN). A bigger challenge is that the multicast client must be aware of essential timing information expressed as a manifest delay parameter that will avoid race conditions when potentially switching between different delivery timelines and delays.
[28] https://tools.ietf.org/html/rfc5740
[29] https://tools.ietf.org/html/rfc5401

correction) repair and other IETF Reliable Multicast Transport (RMT) building blocks in its design. It can operate with both reciprocal multicast routing among senders and receivers and with asymmetric connectivity (possibly a unicast return path) between the senders and receivers.

FLUTE (File Delivery over Unidirectional Transport) as defined in RFC 6726[30] is another protocol for unidirectional delivery of arbitrary binary objects over the networks and could also be used as a basis for mABR. The specification builds on ALC (Asynchronous Layered Coding), RFC5775[31] as a base protocol that was designed for massively scalable multicast distribution.

ROUTE (Real-time Object delivery over Unidirectional Transport)[32] is another multicast protocol and has been selected as an ALC replacement for FLUTE that has been selected for the ATSC 3.0 initiative. It provides a single transport protocol, not only for live linear content, but also for non-real-time content and signaling metadata. It also reduces delivery latency by allowing early playout of segments leveraging bite range requests ("MDE" mode) and supports out-of-band and advanced delivery of file descriptors to enhance reliability of object recovery which can reduce signaling overhead. ROUTE is based on the 3GPP MBMS[33] download delivery protocol

[30] https://tools.ietf.org/html/rfc6726
[31] https://tools.ietf.org/html/rfc5775
[32] https://www.atsc.org/atsc-30-standard/3312017-signaling-delivery-synchronization-error-protection/
[33] https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=829

TECHNICAL CONSIDERATIONS

There are several technical considerations to keep in mind regarding the implementation of mABR. We have highlighted a couple of those below:

- Latency
- Targeted Advertising

Latency

Latency, particularly for content such as live sporting events, is an important consideration, especially when the ABR content is delayed when compared with the broadcast stream.

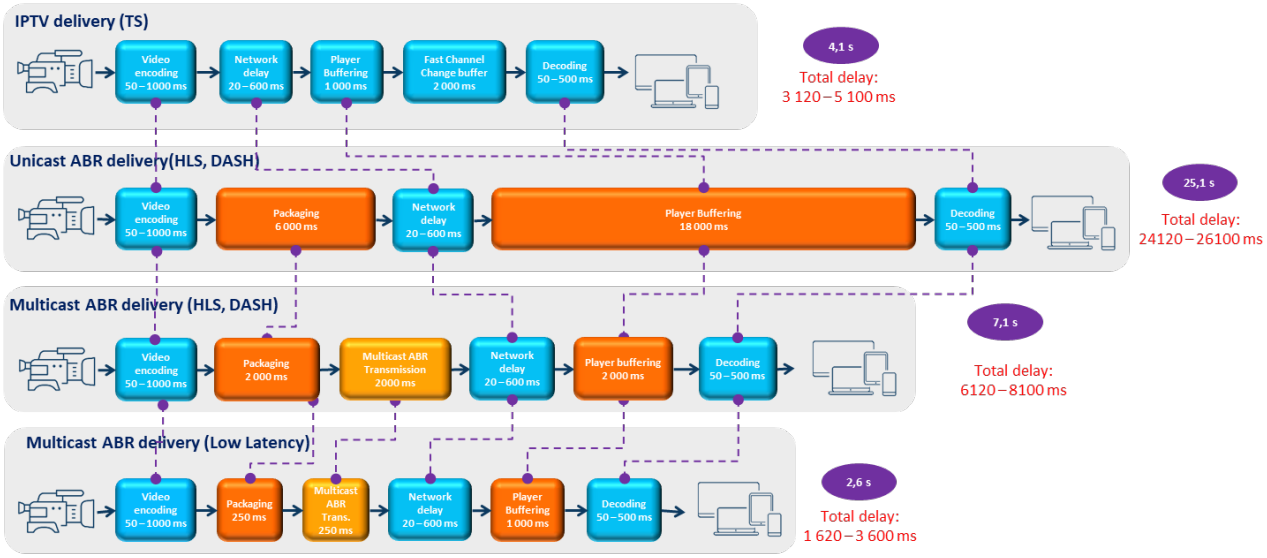


Figure 4: Latency in video delivery (unicast and multicast)

Because mABR needs additional processing after segments are generated, additional delays can be inserted into the workflow. In a typical standalone deployment model, content is retrieved by making standard HTTP Get requests to an origin and assumes that the mABR server must wait until a complete segment is created prior to retrieval.

An innovative approach implemented by several encoding vendors is to by-pass the origin server completely. A multicast stream is created as a primary or secondary output of the encoder/packager and as the NORM packets are generated, it pushes them directly to the downstream multicast address. Thus, the workflow does not need to wait until a complete segment is created prior to sending it.

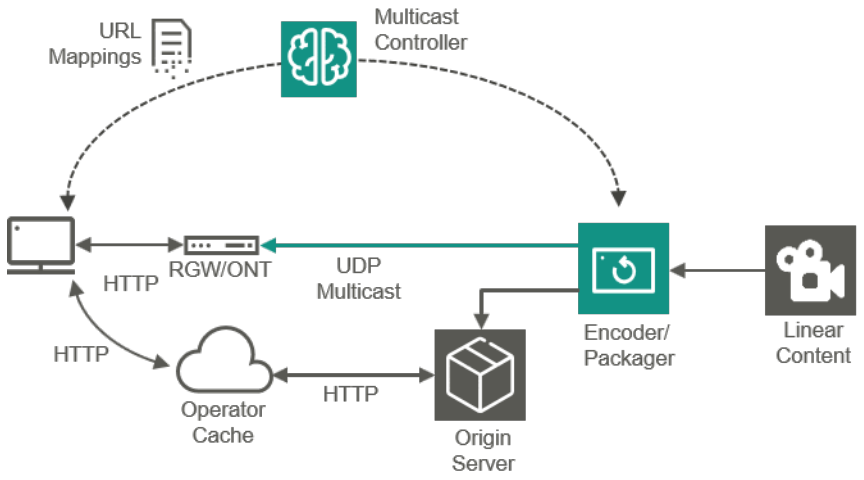


Figure 5: Combined Encoding and NORM Encapsulation

Chunk transfer, available in a number of standard protocols, can also help to improve latencies. This allows an ABR segment to be further split into smaller chunks and delivered to the player without the need to wait until the entire segment is assembled on the mABR client thereby enabling frames to be rendered almost immediately within the constraints of bi-directional prediction and propagation of a given GOP, or contiguous GOP structure.

It is common to use an IPTV delivery infrastructure as the baseline (for which latency is typically between 3 to 5 seconds). A unicast ABR case illustrates that Apple HLS can impose severe constraints to avoiding the player's buffer underflow and for limiting the server/network request overflooding. Apple recommends a segment length of six seconds and a buffering of 3 segments. In multicast ABR, as shown above, these constraints are relaxed and a two second segment duration is a good compromise that ensures quality without imposing too much latency. Note that the multicast ABR infrastructure imposes an additional delay due to the multicast client that must reform the segment before caching and forwarding.

Thanks to the MPEG CMAF (Common Media Application Format)[34], it is possible to split an ABR segment into pieces or chunks. Depending on the encoding structure (i.e. compression mode), a chunk may transport at a minimum one video frame. This format combined with HTTP Chunk Transfer Encoding (HTTP CTE) allows the ABR delivery chain (e.g. packager) to propagate segment chunks on the fly and allows the ABR player to start decoding a segment's chunk without waiting for the complete reception of the segment. Depending on the CMAF chunk duration (which must be at least one picture/frame as e.g. 40ms for 25fps), the impact on latency is tremendous[35].

[36] IP over Coaxial
[36] IP over Coaxial

Targeted Advertising

To the casual observer, the concept of providing targeted advertisements within multicast video streams may seem to be contradictory. However, the concept is well supported in mABR and a few Tier 1 and Tier 2 operators have already been monetizing targeted ads as part of their legacy multicast delivery services for many years. In theory, everybody could have an individual ad break, but a more common scenario is to segment viewers into a small number of groups and provide the same ad-pod to everyone in that group. This could be based on location, demographic, etc.

To substitute ads on a legacy device requesting HTTP fragments from a multicast ABR client, the device must receive an appropriate personalized manifest file and the device must download the appropriate fragments. This workflow is transparent to the device. Personalized manifests can easily be obtained by proxying the request to any standard manifest manipulation technology, either using Pre-CDN or Post-Cache models.

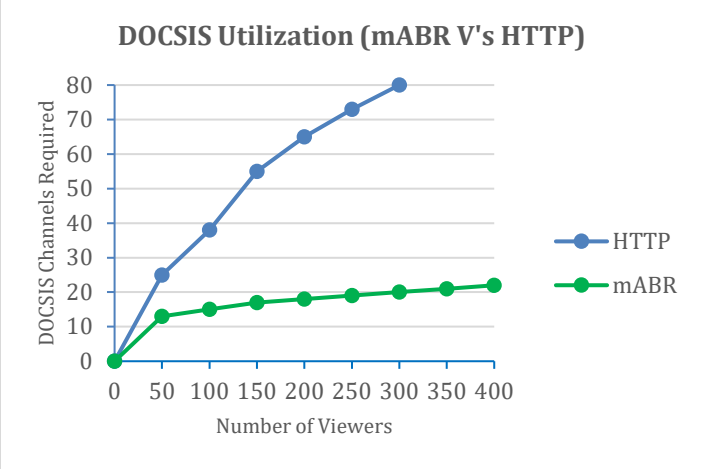
The simplest method to retrieve personalized fragments is to have the Multicast Client fetch Ad Avail chunks over unicast http download. While this means each viewer could theoretically receive a unique ad break, it also means that enough CDN capacity is required to serve all concurrent users, negating the benefits of mABR.

Ad Prepositioning is an elegant solution to this problem. A POIS notifies the solution that an avail is approaching and provides the ESAM identification of the channel (acquisitionPointIdentity) and the Ad signal information. This allows one or more ad decisions—depending on granularity—to be retrieved from an ADS or Ad Service and the alternate content is sent ahead of time over multicast. The Multicast Client caches these and so will serve cache hits during ad breaks. This is multicast and scales well. Another common method is to leverage Dynamic Ad Streams. These are multicast streams containing each Ad Avail rendition. The client simply joins appropriate stream, and these may be parallelized or concurrent. Again, this is multicast but has the added benefit that no Multicast Client storage is required.

THE MULTICAST ASSISTED ABR ADVANTAGE

mABR delivers two primary advantages over unicast delivery. First, it mitigates playback jitter as a result of being a managed, rather than "best effort" service. Second, it can provide significant bandwidth savings on both DOCSIS[36] and G/EPON[37] networks, although the actual deployment model within the network and relative position within the network hierarchy may vary slightly.

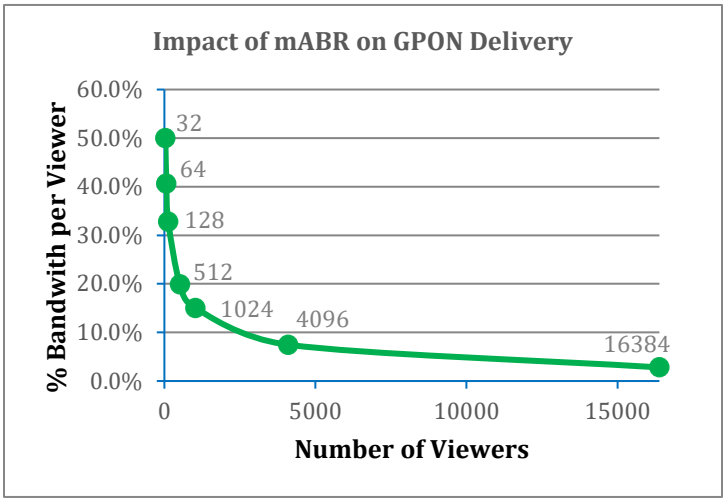
DOCSIS networks delivering unicast ABR require enormous resources in terms of bandwidth and DOCSIS channels which grow in a linear fashion as more and more viewers begin to consume content.



As illustrated, mABR is highly effective in reducing these requirements by allowing viewers to share multicast segments.

As viewers begin to consume content on the multicast network, there is an initial spike in demand for resources, but this quickly stabilizes as more viewers join and share multicast streams. As illustrated, the overall resource requirement is minimal compared to a full unicast solution, and the savings are realized throughout both the core and access networks as the mABR stream is terminated on the RWG or STB within the home.

As with DOCSIS, delivering unicast ABR in a GPON network requires enormous bandwidth, the greatest impact being in the core network to the ONT. Since the viewer has dedicated access from the home to the ONT, there is relatively little benefit delivering multicast past the ONT and into the home in this scenario.



If an operator is delivering 500 linear channels and 80% of viewers are watching 10% of the channel lineup, core bandwidth savings based on an initial 32 customers being served from the PON is approximately 50% of that required for unicast.

As the number of viewers increase past 64 and 128 to 512, the bandwidth

[36] IP over Coaxial
[37] IP over Optical

required is only 20% that of unicast while at 16384 viewers the bandwidth per viewer requirement drops to 2.8% on the core network.

In both scenarios the Bitrate consumed for a high number of viewers is considerably smaller for shared popular linear content resulting in greater savings on transit and last mile networks.

A point worth noting is that the infrastructure cost to implement mABR may require less network CAPEX spend that implementing other solutions, as most of the necessary routers and network elements that would be employed in mABR delivery are already present.

BARRIERS TO ADOPTION

The primary barrier to wide spread mABR adoption has been the lack of a deployed Client ecosystem.

Legacy RGW's and ONT's typically do not have the capacity to process Multicast ABR streams, and the majority do not meet the minimum specifications in terms of CPU cycles and memory that is required to install and run an mABR Client. A secondary problem, specific to NORM-based deployments, is that third-party clients can be bloated and overweight, and typically designed for PC deployments as the primary requirement. While new CPE and Networking equipment generally has enough processing power and memory to run an embedded mABR Client, a concerted effort is required by the vendor community to build an ecosystem of devices and embedded clients that provide standards-based client applications that expose an interoperable platform with ubiquitous connectivity in every device.

In general, mABR is usually complementary to CDN caching strategies, and can even act as the content source for most edge cache deployments. Depending on the network topology, it may drive tremendous scale in the mid-tier if many points-of-presences are deployed[38]. However, it can occasionally conflict with some caching strategies. For Operators that deliver mostly niche, on-demand traffic and have relatively low cache hit ratio requirements, adding mABR to drive scale is simply not cost effective.

Many operators still have IPv4 as their primary network addressing scheme, and while IPv4 has some multicast capabilities, these are somewhat limited compared with IPv6 and limited to a broadcast metaphor in order to transmit packets to the destination clients. In addition, not every router or host supports broadcast with the IPv4 protocol. Multicast was added onto IPv4 late in IPv4's development and as a result IPv4 multicast works best in networks that are end to end managed; in contrast, multicast was designed in to the architecture from the beginning of IPv6 development with high-bandwidth and fault tolerant content delivery in mind. Unlike IPv4, which is limited to a single Broadcast address for a subnet, IPv6 has Multicast addresses that can be used to specify groups of nodes.

[38] Note: in a star-based topology,

CONCLUSION

Multicast Assisted ABR (mABR) represents an interesting solution to the trends happening within the broadcast and media/entertainment space. As more broadcasters and video distributors utilize IP to deliver streaming video to a growing audience (using ABR), the traditional unicast method over TCP of sending HTTP chunked video segments will continue to weigh on the infrastructure. It's almost become a battle between capacity and demand—can network operators keep adding infrastructure to meet viewer demands for not only more streaming video, but higher quality? With mABR over a DOCSIS or E/GPON network, significant bandwidth can be recovered ultimately enabling network operators to better manage traffic over their networks while meeting subscriber demands.

But this begs the question, "is mABR really needed anymore?" Since the initial establishment of the mABR standards, network operators, video distributors, and others have made substantial investments in CDN technology and capacity. One could argue that the "edge" of the network has grown significantly, perhaps more than is needed to handle the demand for high-quality streaming video. Of course, that position is predicated on network operators, video distributors, and CDNs continuing to expand the edge caching infrastructure. A secondary challenge is that mABR is not particularly suited to OTT delivery given the lack of multicast support on the public internet. Combined with advancements in video compression such as HEVC, AV1, and VVT, and new transport protocols such as QUIC, SRT and WebRTC, the streaming video industry may well stay ahead of the exponential growth curve as consumers continue to supplant traditional television for online streams. And, what about 5G? Will the coming of the new spectrum, and its copious bandwidth, obviate the need for delivery savings within the operator network?

Although the long-term impact of mABR on the streaming video experience is still up-in-the-air, there is some low-hanging fruit—the operator cache population. mABR is fairly easy to deploy on the cache and, when combined with the available processing power on those caching boxes, could offload a lot of traffic from the backbone, depending on the network topology.

But after that? It's unclear how mABR will fair in assisting the delivery of the live streaming video experience. As identified in the Executive Summary, the problem is a chicken-and-egg situation. The operators aren't going to expend a lot of CAPEX to install and support mABR technologies within their networks. While newer CPE may scale to support the deployment of an embedded mABR client, older devices will almost certainly need an upgrade in the field to provide the functionality. It's possible that, over time, as operators upgrade client devices, multicast clients will be included, or the box will support remote updating to provide a client to support mABR network deployment. But, it's probably the case that not a lot of client device vendors are going to expend money to build, install, test, and support multicast client software on their devices when there is no infrastructure to utilize it (just a matter of priorities).

Ultimately, this situation will continue to hold back mABR implementations. And so long as consumer devices can consume HTTP-based video, however inefficient it may be to deliver, network operators and

video distributors may be reticent to expend significant capital in deploying mABR technology for an unknown return.

FIGURES

ABOUT THE STREAMING VIDEO ALLIANCE

Comprised of members from across the video ecosystem, the Streaming Video Alliance is a global association that works to solve critical streaming video challenges to improve end-user experience and adoption. The organization focuses on three main activities: first is to educate the industry on challenges, technologies, and trends through informative, publicly-available resources such as whitepapers, articles, and e-books; second is to foster collaboration among different video ecosystem players through working groups, quarterly meetings, and conferences; third is to define solutions for streaming video challenges by producing specifications, best practices, and other technical documentation. For more information, please visit www.streamingvideoalliance.org.

Streaming Video Alliance
5177 Brandin Court
Fremont, CA 94538 USA
(510) 492-4000
streamingvideoalliance.org