RECLAIMING CONTROL OF THE NETWORK FROM ADAPTIVE BIT RATE VIDEO CLIENTS

John Ulm & Gerry White
Motorola Mobility

Abstract

This paper provides a brief introduction to adaptive bit rate (ABR) video and discusses why handling this class of traffic well is very important to the cable operator. It then examines the major differences between ABR and the current IP and MPEG video delivery mechanisms and looks at the impact these differences have on the network. Some interesting experimental results observed with real world ABR clients are presented. A number of problems which may develop in the network as ABR clients are deployed are discussed and possible solutions for these proposed. Finally, the paper looks at the cable modem termination system (CMTS) as a potential control point that could be used to mitigate the impact of the ABR clients and regain control of the access network for the operator.

INTRODUCTION

Adaptive bit rate is a delivery method for streaming video over IP. It is based on a series of short HTTP progressive downloads which is applicable to the delivery of both live and on demand content. It relies on HTTP as the transport protocol and performs the media download as a series of very small files. The content is cut into many small segments (chunks) and encoded into the desired formats. A chunk is a small file containing a short video segment (typically 2 to 10 seconds) along with associated audio and other data. Adaptive streaming uses HTTP as the transport for these video chunks. This enables the content to easily traverse firewalls, and the system scales exceptionally well as it leverages traditional HTTP caching mechanisms.

Adaptive streaming was developed for video distribution over the Internet. In order to deal with the unpredictable performance characteristics typical of this environment, ABR includes the ability to switch between different encodings of the same content. This is illustrated in Figure 1. Depending upon available bandwidth, an ABR client can choose the optimum encoding to maximize the user experience.

Each chunk or fragment is its own stand-alone video segment. Inside each chunk is what MPEG refers to as a group of pictures (GOP) or several GOPs. The beginning of each chunk meets the requirements of a random access point, including starting with an I-frame. This allows the player to easily switch between bit rates at each chunk boundary.

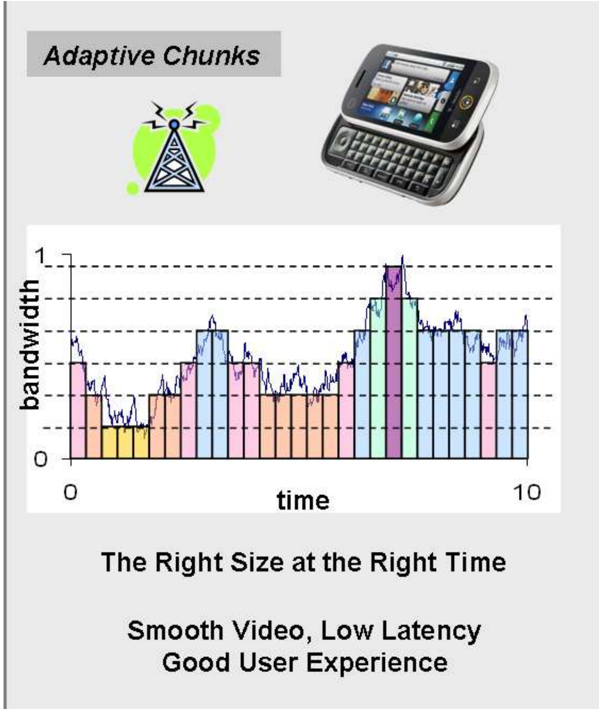

Figure 1 Adaptive Streaming Basics

Central to adaptive streaming is the mechanism for playing back multiple chunks to create a video asset. This is accomplished by creating a playlist that consists of a series of uniform resource identifiers (URIs). Each URI requests a single HTTP chunk. The server stores several chunk sizes for each segment in time. The client predicts the available bandwidth and requests the best chunk size using the appropriate URI. Since the client is controlling when the content is requested, this is seen as a client-pull mechanism, compared to traditional streaming where the server pushes the content. Using URIs to create the playlist enables very simple client devices using web browser-type interfaces. A more in-depth discussion of ABR video delivery can be found in [ADAPT]

IMPORTANCE OF ABR

Second and Third Screens

ABR based video streaming has become the de-facto standard for video delivery to IP devices such as PCs, tablets and smart-phones. ABR clients are typically shipped with (or are available for download to) these devices as soon as they are released. Given the short lifetime of this class of device this is a key enabler, especially compared to the time required to deploy software to traditional cable devices. As mentioned previously, ABR delivery simply requires an HTTP connection with sufficient bandwidth so that it is available both on net and off net. With these advantages, both over-the-top (OTT) and facilities based service providers are leveraging ABR so that essentially all video delivery to second and third screen devices uses this mechanism.

Primary Screen

ABR is also used to deliver a significant quantity of video to television screens in both standard and high definition formats. Over-the-top providers of video service leverage ABR clients installed in platforms such as gaming consoles, Blu-ray players, set-top box-like devices and smart TVs to provide video services to the primary screen. This content rides over the service providers' high speed data (HSD) service and, in many cases, constitutes the bulk of the HSD traffic.

ABR Traffic Load

Studies of Internet traffic patterns [SAND], [VNI] show that video has become the dominant traffic element in the Internet, consuming fifty to sixty percent of downstream bandwidth. Netflix alone constitutes almost thirty-three percent of peak hour downstream traffic in North America. Thus, how well the network supports ABR based IP video is obviously crucial to providing a satisfactory customer experience. In addition, delivery of Internet video to televisions is predicted to grow seventeen-fold by 2015 to represent over sixteen percent of consumer Internet video traffic (up from 7 percent in 2010) [VNI]. Thus, many of the customers will not only be viewing IP video, but will be doing so on a large screen device with expectations of high quality.

In addition to the Internet video explosion, significant amounts of managed service provider video will also migrate to an ABR mechanism, further increasing the percentage of ABR traffic on the network.

Having this much ABR traffic on the network means that it will be a key driver of network costs and with ABR delivering prime entertainment services, how well it is supported will be a key metric for customer satisfaction going forward. Therefore, understanding the issues around delivery of ABR over the DOCSIS network will be crucial for MSO's video service delivery, and for their ongoing profitability.

ABR vs. CURRENT VIDEO DELIVERY

ABR video delivery has a number of very significant differences to both MPEG video delivery and streamed IP video delivered over Real-time Transport Protocol/User Datagram Protocol (RTP/UDP) as used in a Telco TV system such as Microsoft Media Room [MMR]. A number of these differences are discussed below.

Client Control

ABR has been developed to operate over an unmanaged generic IP network in which bandwidth decisions (i.e. choosing the video bit rate to request) are made by the client device based on its interpretation of network conditions. This is fundamentally different from the approaches used for existing MPEG or conventional streamed UDP video delivery, where devices under the direct control of the network operator make the important decisions relating to bandwidth. Thus, in MPEG delivery, the encoding, statistical multiplexing and streaming devices determine the bit rate for a given video stream. These devices are under control of the service provider. Similarly for a UDP streaming solution, the video is encoded and streamed at a selected rate from devices owned by the service provider. In contrast, the behavior of ABR clients is specified by the developer which, in general, will be a third party outside the service provider's control.

Variable Bit Rate

As described previously, an ABR client will select a file chunk with a bit rate which it believes to be most appropriate according to a number of factors including network congestion (as perceived by the client) and the depth of its playout buffer. Thus the load presented to the network can fluctuate dramatically. This is in stark contrast to both MPEG and UDP video streams which are either constant bit rate (CBR) or are clamped variable bit rate (VBR) (i.e., bandwidth can vary up to a maximum bit rate but not beyond it).

A more detailed discussion on the impact on network loading of a number of factors is found in a later section of this paper.

Admission Control

ABR clients join and leave the network as users start and stop applications. From a network perspective, there is no concept of a session with reserved resources or admission control. Again this is the antithesis of MPEG or UDP video in which a control plane operates to request and reserve network resources and determines whether to admit a user. In a controlled network, adding a new user session can be guaranteed not to impact existing users. Once resources are exhausted, any additional session requests will be denied, introducing a probability of blocking into the system. In an ABR model under network congestion, each new session will reduce the bandwidth available to all existing sessions rather than be denied. Thus, users may see a variation in video quality as other ABR clients start and stop. This reduction in quality during peak times is analogous to statistical multiplexing in legacy MPEG video. During peak times, the statmux reduces bit rates across the various video streams to fit within its channel. The ABR system has an advantage in that it will be over a larger channel using DOCSIS bonding.

Congestion Control

With MPEG or UDP streaming video delivery, congestion control is not relevant as the control plane provides admission control to ensure it does not occur. When ABR is used for video delivery, congestion control is a potential issue. The situation is complex in that three levels of congestion control mechanisms are involved operating at different layers in the protocol stack. At the media access control (MAC) level, the CMTS is responsible for scheduling downstream DOCSIS traffic [MULPI]. Operating at the transport level is standard Transmission Control Protocol (TCP) flow control based on window sizes and ACKs, [TCP] and, finally, at the application level the client can select the video bit rate to request. The latter two levels of control (TCP and application) are the responsibility of the ABR clients and as such are outside the control of the network operator. Interaction between these three flow control mechanisms is not well understood at this time and may have unforeseen impacts.

Prisoners Dilemma

As noted above, ABR clients have the responsibility to select the quality (bit rate) of the video they request to download. The algorithms and parameters used by each client to make this decision are outside the control of the network operator. Each client is faced with a decision not unlike the classic "prisoner's dilemma" [PDIL] in that they can elect to optimize for their own benefit or they can optimize for the common good of all clients on the network (including their own). For example, a very selfish client may never request a lower quality file even during network congestion based on the assumption that other clients will do so, and thus resolve the congestion for them. Commercial pressures to create "better" clients may drive in this direction, but if all clients move to this mode the network will fail. This is not an issue with MPEG or UDP streaming delivery as the network operator has the incentive and necessary controls to offer a quality service to all customers.

Imperfect Knowledge

Clients base their decisions on what to request based on their local knowledge rather than on an overall view of the network conditions. This is in contrast to MPEG or UDP streaming where the network operator provisions the video bit rates based on knowledge of the end-to-end network and expected loads.

The following section on potential problems will address these issues in more depth and attempt to develop some potential solutions.

ABR CLIENT CHARACTERIZATION

As discussed previously, the ABR client plays a critical role in the operation of adaptive protocols. For an operator trying to provide a differentiated quality of experience, it is important to understand how different ABR clients behave under various circumstances.

Motorola research teams took multiple different types of clients into the lab to analyze their behavior. Previous work [Cloonan] discussed results from a simulator. Our goal was to capture live client interaction. Operation during steady state was relatively stable. The interesting observations occurred during startup and when video bit rates were forced to change.

At startup time, clients try to buffer multiple segments as fast as they can. This was particularly obvious for video on demand (VOD) assets where the entire content stream is accessible. Live content tends to have a limited playlist available to the client, preventing large buffer build up. During this startup period, the clients are also calculating the available bandwidth and may decide to switch bit rate. This action may cause some segments to be re-fetched with the new resolution. Overall, the differences between clients seemed fairly subtle for startup.

In our lab environment, the amount of bandwidth available to the ABR client was adjusted. In this manner, the client was induced to switch video bit rates. After reducing available bandwidth, the clients in general made a smooth transition to a lower bit rate. Some clients reacted more quickly than others in down shifting. When the available bandwidth is opened up again, clients started searching for new higher bit rates with the associated buffering of segments, similar to startup. It was in this phase where we saw the most differences between clients. In fact, we saw differences from the same device running different revisions of their protocol.

POTENTIAL PROBLEMS

Based on the above characterization, operators must be aware of some potential problems. As was discussed, there is a burst of additional traffic during startup and when switching to higher bit rates. The system must be capable of handling this additional traffic burst.

Actively managing ABR video traffic may be challenging given that every ABR client may be operating its own disjoint algorithm. This is also compounded since client behavior may change with the download of an updated revision. Bandwidth stability may become a concern if multiple clients become synchronized. For example, the network becomes congested causing a group of clients to lower bit rates. If these clients then sense that bandwidth is available (i.e. it is released due to downshifting by other clients), there may be a surge in traffic that causes congestion, and the cycle repeats.

In general, ABR clients are designed for general Internet usage, so they tend to back off quickly and may be slow to ratchet their bit rates back up. This will create some stability and should prevent the above oscillation, but this may make it challenging to fully utilize the network bandwidth.

There are several fairness concerns that must be taken into consideration. If the current bandwidth utilization is high, then new clients just starting their video may select a lower rate than other clients are currently using. Other forms of unfairness may be introduced when network congestion causes video bit rate changes. Some clients may decide to change while others remain at current bit rates, resulting in disparity between clients.

Another concern, especially for a managed video service, is maintaining a good Quality of Experience (QoE). The more that clients change bit rates, the more potential impact there is to QoE. The system should be designed to minimize unneeded bit rate changes.

For future research, Motorola will expand its investigation to system-level behavior for a large number of disparate ABR clients. It is important that the industry grasps the system dynamics for adaptive protocols.

POTENTIAL SOLUTIONS

In a discussion of potential solutions to problems with ABR video delivery under network congestion, two types of ABR traffic must be considered: managed and best effort. Best effort video traffic is OTT types of service which, in general, would be indistinguishable from general Internet traffic.

Managed traffic would typically be video sourced by the service provider, or by a third party with whom the service provider has negotiated a carriage agreement. How well managed traffic is supported is a significant problem for a service provider as it is, in effect, a branded service for which customers will have a higher expectation.

In general, the following potential solutions apply to a managed IP video service. We will highlight where it also applies to OTT traffic.

Controlled Client

Managed ABR services may be made available only from a specific service provider application downloaded by the user. This removes the issues relating to client misbehavior and enables the operator to predict how the client will handle network congestion events.

It has the disadvantage that the operator must keep the application up to date both in terms of feature parity with other clients and with new devices and operating systems as they are released. It also makes it likely that the user must have multiple applications to access different video sources.

This is not applicable to OTT video from third parties, which will be typically be delivered to either a native client on the device or a client provided by the OTT service.

Session Control

One option to control ABR traffic is to implement a session mechanism similar to those used for more traditional video streaming. In this case a user (or possibly a proxy for the user such as a Fulfillment Manager) requesting a video asset would invoke resource checking and reservation mechanisms in the network control plane. The control plane would reserve access network bandwidth for the video session. Mechanisms such as PacketCable™ Multi-Media (PCMM) [PCMM] are in place today to enable quality of service (QoS) bandwidth reservation over DOCSIS. This is detailed later in the paper.

A problem with this approach is knowing when to start and terminate a session and specifically when to acquire and release the resources. For managed video this could be achieved by using a service provider application as described above. The application would invoke the session setup and teardown as part of the video selection and playing process. Even a controlled application implementation would need a back up mechanism to release resources as the user may simply power off a device or lose connectivity. At the minimum, a “no traffic timeout” would be needed (refer to CMTS section below for more details).

Network Override

In conventional ABR video distribution, the ABR client determines the bit rate of the next file to download from the options in the playlist and retrieves this directly from the content delivery network (CDN). This decision could potentially be overridden from the network in a number of ways.

The playlist file provides the bit rate options specified by the service provider. Normally this selection would be statically provisioned and implemented by the encoding and packaging processes as the video asset was processed. For example, each asset could have files created for 1, 2, 4 and 6 megabits per second (Mbps) and the client allowed to select between these. Modifying the selection options in the playlist file provides a potential mechanism for the network to influence the client operation. Thus in times of congestion, the high bandwidth option could be removed by providing a playlist with only 1 and 2Mbps options. This of course requires run time manipulation of the playlists. A potential problem is the lag from playlist manipulation to actual changes in bit rate selection. Even a short playlist file would probably need to represent video content lasting for a significant time so that this mechanism would have a very slow reaction time to network events. Thus, it would not respond to short term congestion events. However, if the network had well known congestion periods (e.g. 8:00 pm through 10:00 pm) it could be used to reduce congestion during these times. Alternatively, the Session Manager might provide notification when the system is congested. This mechanism would not be applicable to OTT traffic as detecting the playlist files would be problematic, and modifying the third party data is unlikely to be permitted.

CMTS AS CONTROL POINT

For users on an HFC network, IP traffic will always flow through the same CMTS port to reach a user at home. As the shared CMTS to CM link is normally the "narrow pipe" in the video distribution network, this is where congestion would be expected. Therefore the CMTS can potentially provide a useful control point to manage the ABR traffic.

Downstream Scheduling and Queuing

The DOCSIS standard provides very complete QoS functionality which may be useful for managing ABR traffic. DOCSIS QoS is based on the IntServ model of filter and flow specifications [INTS]. If a packet matches an installed filter (i.e. classification) it will be mapped to a specific service flow and then forwarded based on the parameters associated with that flow. Classification is based on matching fields in the packet header such as IP address and Differentiated Services Code Point (DSCP) fields. Thus it could be possible to recognize a managed ABR video packet from a well known source address (e.g. video server) or IP subnet. Alternatively all managed video traffic could use a DSCP marking indicating a preferential forwarding class [DSCP]. Inbound traffic to the network from non-trusted sources such as over-the-top (OTT) video would be subject to DSCP overwrite and set to a base priority such as best effort. The CMTS could then provide preferential treatment for the operator's managed video flows.

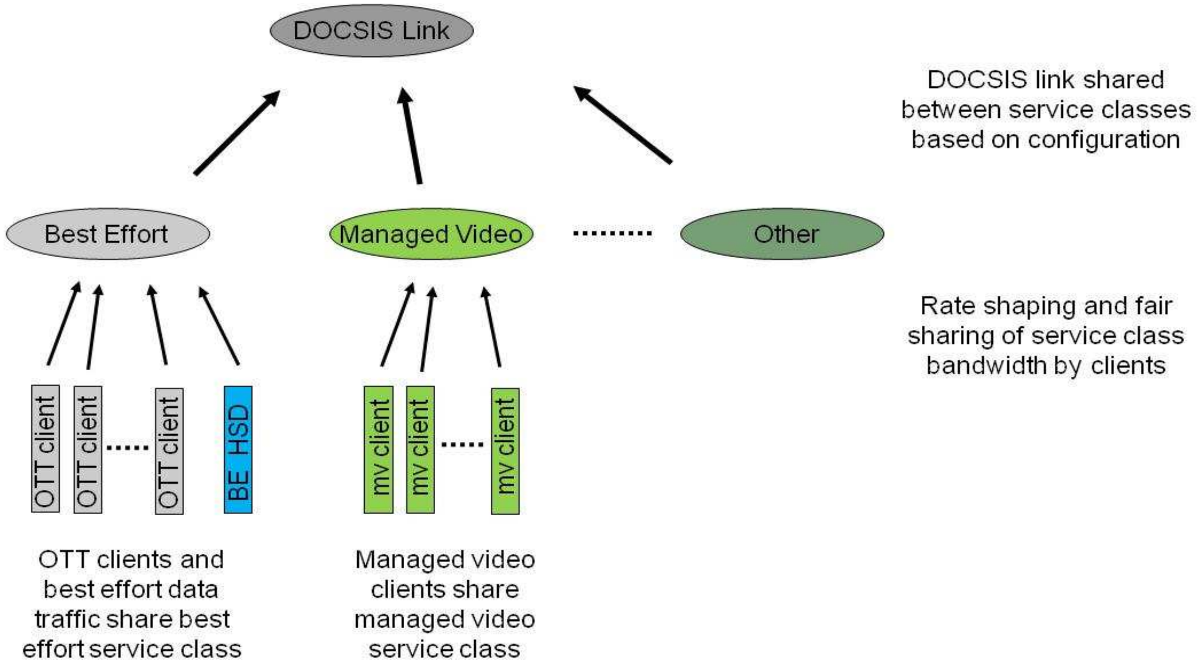
DOCSIS Link
DOCSIS link shared between service classes based on configuration
Best Effort
Managed Video
Other
Rate shaping and fair sharing of service class bandwidth by clients
OTT client
OTT client
OTT client
BE HSD
mv client
mv client
mv client
OTT clients and best effort data traffic share best effort service class
Managed video clients share managed video service class

Figure 2 DOCSIS Link Sharing

If the CMTS supports multi-level scheduling and per-flow queuing as shown in Figure 2, then it can provide fairness between video flows. In this case, each video packet would be mapped to an individual queue (based on the header fields in the packet) within a particular scheduling class such as managed video or best effort traffic. All queues within the same scheduling class share the bandwidth assigned to the class equally so that a single user receives only their fair share and cannot disrupt other video sessions. This mechanism applies to both managed video and OTT ABR video. OTT traffic will be put into the best effort class but will still receive a fair share of the assigned bandwidth for this class. It will, of course, share this with all general Internet traffic. Each scheduling class would be assigned a percentage of the available bandwidth proportionate to its expected load.

Session Control at CMTS

The DOCSIS infrastructure has a mechanism to reserve bandwidth for a flow based on the PacketCable™ Multimedia specification [PCMM]. This provides a potential mechanism to implement resource reservation at the session level. It requires a session establishment and teardown mechanism. In the PCMM model, client applications communicate with an application server (AS) that initiates the QoS requests to the policy server and CMTS. The ABR client application server might be co-located with a session/fulfillment manager, edge server, or user interface (UI) server depending on an operator's control plane infrastructure. Therefore, it would be suited to a managed video service but not OTT. The PCMM / CMTS mechanisms are well understood and include error recovery functions such as the timeout of orphaned sessions.

A potential problem arises in that a video asset may be delivered from one of multiple sources within the CDN. Thus, the filter specification used to identify the packets associated with the session would need to be capable of handling this. This may be as simple as using a known sub-network for the video sources. A more complicated problem is that within the single session, multiple file chunks at different bit rates may be requested due to local events in the client device. The resource reservation for the session could be selected to provide the maximum data rate expected from the client. However, if the client downshifted, this reserved bandwidth would not be used for the managed video but released for use by best effort traffic.

The lab investigations showed that the ABR clients tend to require additional bandwidth during startup and following bit rate increases. The PCMM mechanism can be used to provide a "turbo" mode in which additional bandwidth bursts are allowed for these periods.

CONCLUSION

The impact of ABR traffic on the network is already considerable and is likely to grow significantly as more video is distributed using this mechanism. ABR traffic operates very differently from existing video delivery mechanisms, and in the conventional use case, control over access network bandwidth is essentially abrogated to the device clients. Motorola experiments indicate that these clients vary from device to device and are not necessarily well behaved. Given that they have an incentive to be greedy rather than cooperate for the common good, it seems imperative that the operator finds other mechanisms to control ABR traffic impacts.

A number of options are discussed and the CMTS appears to be a promising location to implement this control. For OTT ABR traffic, the CMTS can provide rate limiting and fair sharing of bandwidth between both ABR clients and other best effort users. This is implemented using existing DOCSIS QoS and CMTS downstream scheduling. For managed ABR traffic, these QoS and scheduling mechanisms may also be used and can also provide segregation of the managed traffic from best effort traffic. With the addition of a session management function in the network, additional control is possible. This enables PCMM control mechanisms to be used to establish service flows for the video streams with defined QoS and reserved bandwidth.

The existing functions provided by the CMTS appear to provide the operator with an excellent control point to impose order on the access network despite the potential for aberrant client behavior.

REFERENCES

[ADAPT]	Adaptive Streaming – New Approaches for Cable IP Video Delivery J. Ulm, T. du Breuil, G. Hughes, S. McCarthy, The Cable Show NCTA/SCTE Technical Sessions spring 2010
[SAND]	Global Internet Phenomena Report Fall 2011; Sandvine
[VNI]	Cisco® Visual Networking Index (VNI) 2011
[MMR]	Microsoft Media Room -www.microsoft.com/mediaroom/
[MULPI]	DOCSIS 3.0 MAC and Upper Layer Protocols Interface Specification www.cablelabs.com
[TCP]	RFC 2581 TCP Congestion Control M. Allman, V. Paxson, W. Stevens
[PDIL]	Kuhn, Steven, "Prisoner's Dilemma", The Stanford Encyclopedia of Philosophy (Spring 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2009/entries/prisoner-dilemma/>.
[INTS]	RFC 1633 Integrated Services in the Internet Architecture: an Overview R. Braden, D. Clark, S. Shenker
[PCMM]	PacketCable™ Multimedia Specification www.cablelabs.com

ABBREVIATIONS AND ACRONYMS

CCAP	Converged Cable Access Platform
CDN	Content Delivery Network
CMTS	DOCSIS Cable Modem Termination System
COTS	Commercial Off The Shelf
CPE	Customer Premise Equipment
DOCSIS	Data over Cable Service Interface Specification
DRM	Digital Rights Management
DVR	Digital Video Recorder
DWDM	Dense Wave Division Multiplexing
EAS	Emergency Alert System
EQAM	Edge QAM device
Gbps	Gigabit per second
HFC	Hybrid Fiber Coaxial system
HSD	High Speed Data; broadband data service
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
MAC	Media Access Control (layer)
Mbps	Megabit per second
MPEG	Moving Picture Experts Group
MPEG-TS	MPEG Transport Stream
nDVR	network (based) Digital Video Recorder
OTT	Over The Top (video)
PHY	Physical (layer)
PMD	Physical Medium Dependent (layer)
PON	Passive Optical Network
RF	Radio Frequency
STB	Set Top Box
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VOD	Video On-Demand
WDM	Wave Division Multiplexing