Improving Adaptive Video Delivery Through Active Management

Santhana Chari
Arris
santhana.chari@arrisi.com

Mark Francisco
Comcast
mark_francisco@cable.comcast.com

Abstract

Adaptive delivery technologies which are in use to deliver services over unmanaged networks will allow for greater efficiency and capability of managed networks. A unified video processing workflow that can scale to large scale video delivery is described. Using this workflow from content ingest to distribution will provide a higher quality of experience over a wide range of devices. The concept of active management of adaptive delivery is proposed and described as a method of maintaining network capacity and incremental offering of new services and formats. A method of validating adaptive profiles is described and results presented comparing a number of different adaptive technology implementations.

OVERVIEW OF ADAPTIVE VIDEO DISTRIBUTION ARCHITECTURE

In this section, we present an overview of the architecture and basic video processing infrastructure used by an over-the-top video provider using Adaptive Bit Rate (ABR) video streaming. Delivery of video over an unmanaged network, like the public Internet, is based on best-effort delivery. Therefore the quality of user experience is impacted by various factors including the latency and congestion in the network, and uncontrolled variations of these network parameters over time. In spite of these limitations on-line video distribution has become prevalent recently and the quality of the video being distributed is ratcheting up slowly especially because of the increase in last-mile access bandwidth. Most on-line video distributors leverage some flavor of ABR streaming technology to adapt video delivery to vagaries of varying network conditions.

Most commercial over the top video delivery services (Netflix, Hulu, Amazon Prime, etc.) available today deliver on-demand content such as movies and TV shows, and not 24x7 live content. Exceptions to this are popular sporting events like the Olympics that are delivered over the Internet directly by the programmers. On the other hand, MVPDs (multichannel video programming distributor) have started to deliver live content using ABR (adaptive bitrate) streaming techniques to secondary viewing devices both inside and outside the home details of which will be presented in the subsequent sections.

Figure 1 shows a general content processing and delivery architecture employed by ABR streaming services. The first three elements, namely the Transcode, Package and Encrypt/DRM perform functions are related to content processing. The transcoder generates multiple *profiles* of video at different rates and resolutions, using H.264 for video and AAC (Advanced Audio Codec) for audio. Migration to more efficient codecs such as HEVC is expected to happen in the near future, but almost all streaming services currently use H.264 as the video format. Subsequent to transcoding, the packager

encapsulates the video and audio elementary streams into one of many different container formats. The container formats are MPEG-2 Transport Streams (TS) for HLS, and variants of fragmented MPEG-4 file for HSS and HDS formats. Recently announced ISO standard DASH container format can provide a non-proprietary alternative to the above formats depending on how quickly DASH sees adoption commercially. As ABR streaming has moved towards delivering HD and premium content, content encryption and digital rights management (DRM) have become integral part of a successful delivery service as shown in the figure.

Following the content processing modules are the Origin and Edge cache servers that are responsible for eventual delivery of the processed content to the end clients. Large scale content delivery is performed using a multi-level caching architecture containing one or more high storage capacity Origin servers and a large number of edge cache servers with high streaming capacities. Over the top services employ third-party CDN (Content Distribution Networks) to deliver video. The edge cache servers in the CDN help to move the content closer to the end subscriber as well as caching the most popular content for repeated consumption by other users in the same geographical area. Efficient delivery of content over the Internet with acceptable of quality of user experience continues to be a challenge as the display and processing capabilities of the devices consuming video continue to improve rapidly.

ADAPTIVE VIDEO DELIVERY IN A MANAGED NETWORK

One of the advantages of the ABR streaming discussed in the previous section is that it can be used deliver delay sensitive content such as video over the Internet using best-effort delivery. The flow-control mechanism in current ABR streaming is initiated and managed by client devices. Software in client devices dynamically request different profiles based on various factors, such as the estimate of available bandwidth, network latency, decoding buffer fullness or instantaneous CPU usage. This approach, although found to work reasonably well for unmanaged delivery, results in a non-optimal usage of network resources and end user quality of experience as described below.

Sub-optimal usage of network resources can be illustrated with a rather simple example. Assume that an ABR service has three profiles, a 5 Mbps profile for delivery to a STB-connected large screen TV, a 4 Mbps profile for delivery to PCs and a 3 Mbps profile for delivery to tablet devices. Individual clients are allowed to request any of the aforementioned profiles. Let us assume that there is an available bandwidth of 11 Mbps and two tablet devices initiate new sessions for the content. These two clients can progressively request higher bandwidth profiles and say they go up to the 4 Mbps profile consuming 8 Mbps of bandwidth. After this if a STB initiates a request, the available bandwidth is only 3 Mbps, thereby forcing the STB to deliver a less than optimal video to the large screen TV. Clearly individual clients in this scenario do not have knowledge of requests coming from other clients and therefore they make decisions in a rather greedy fashion. In an optimal delivery scenario, both tablet clients can deliver good quality user experience using the 3 Mbps profile leaving the 5 Mbps bandwidth to the STB.

In a managed network, optimal resource allocation can be accomplished by using a central controller that has information about all the adaptive clients that share the same pipe. Therefore the controller will have knowledge of all the requests coming from devices that are in the same node or service group. Based on the knowledge of the ABR

profiles and the capabilities of the client devices, the central controller can make optimal decisions on bandwidth allocation to individual clients.

The central controller can improve the quality of delivery by leveraging several factors in addition to the information on the type of clients and their requests as mentioned above. The central controller can monitor the instantaneous network utilization and proactively respond to any anticipated network congestion. Impact of any network congestion can be distributed over all the clients in the service group rather than adversely affecting a handful of clients. It can also implement business rules on content bitrates, priority of content and levels of subscription of end-user to decide on what content profile to deliver to individual clients. Figure 2 below shows one of the representative implementation scenarios of the central controller. Note that the central controller can reside in several points in the network, either in the cloud, a CMTS, one or more edge cache servers, or in a gateway inside the home.

UNIFIED WORKFLOW

Modern video receivers can be serviced with a standard set of video formats using adaptive delivery methods. It is possible to support a wide range of device types and network conditions with a relatively small set of bitrate/resolution profiles. The maximum desired resolution for medium to large screens such as television, computer and tablet is nearly identical and all devices support the lower resolutions required by smaller screens such as smartphones and portable media players. To achieve greater network efficiency, it may be desired to limit smaller screens to a defined subset of profiles that exclude excessive bitrates. Using similar methods, it may be desired to limit large screens to a subset that exclude profiles providing too poor an experience. These limits can be implemented through player configuration or manifest conditioning.

Increasingly, a single version of an asset may be delivered as video on-demand over QAM and IP. In some cases, the same asset may also be downloaded using IP. Broadcast content may also deliver over both QAM and IP and may be captured as a file for delivery as streaming recordings or video on-demand. From a singular file or stream contribution content may be adapted to use cases including broadcast, IP-streaming, network recording, video on demand and download. A unified workflow allows a single asset to be converted for multiple use cases and device types. An important aspect is to limit the variations of content available for streaming or at rest to bitrate/resolution profiles in support of adaptive delivery methods. The primary function of unified workflow is to utilize a single contribution format and create a set of files or streams that satisfy the bitrate / resolution requirements of all device types and network capabilities.

A subsequent operation is performed as assets are delivered to package the content for the specifics of the device platform. Packaging operations may include file fragmentation, transport multiplexing, audio binding and identity binding of content security. Frequently referred to as just-in-time packaging, these operations are practical to implement in high stream capacity appliances that are a component of the content origin section of the CDN. An end to end illustration of unified workflow is included in Figure 3.

DYNAMIC SELECTION OF ABR PROFILES

As shown in Figure 1, one of the important design considerations in ABR

delivery is the selection of video profiles to be generated by the transcoder. Choosing a large number of profiles with small increments in bitrates and resolutions will enable client devices to smoothly switch between different profiles. However having a large number of profiles increases the complexity and cost associated with generating these profiles, maintaining the assets for future play-out and their ingest into origin and edge cache servers.

Table 1 shows a representative set of profiles used in an ABR service. These profiles, with their corresponding resolutions and bitrates, have been chosen *apriory* to match the capabilities of various display devices that are expected to have access to the content.

Profile	**Resolution (W x H)**	**Video Bitrate in mbps**
1	1280 x 720	3.0
2	1280 x 720	2.0
3	960 x 540	1.5
4	864 x 486	1.25
5	640 x 360	0.75
6	416 x 240	0.5
7	320 x 170	0.35

Table 1. Representative Set of Profiles

One of the challenges with hand-picking the profiles is that the selected bitrates and resolutions may be either too aggressive or conservative due to variability of the video content. We used a proprietary video quality tool to study the level of variability that can be encountered in practice. This tool estimates the perceived quality of ABR video segments and it was run on several different live video programs. Each of the video programs was coded at the profiles shown in Table 1 with segment duration of 6 seconds. The video quality tool that provides a no-reference scores for each segment in a scale of 1 to 100. A quality score of 75 or above is found to be visually acceptable based on our experiments. Figure 4 shows the plot of measured video quality of two different video content materials, "Home & Garden" and "CNBC" using Profile 1 encoding parameters. The "CNBC" video sequence exhibits a very high quality at Profile 1 encoding parameters, whereas the "Home & Garden" sequence exhibits larger variations and lower quality. This shows the difficulty associated with static selection of profiles.

We also investigated how the video quality scores change for a given live video program with different profiles used in Table 1. In some of the sequences, the video quality of different profiles is fairly evenly spaced apart, whereas for many sequences several profiles have similar video quality. Figure 5 shows the video quality score for the first three profiles of the CNBC sequence. As expected Profile 1 has a very high video score. Profiles 2 and 3 have a fairly similar video quality score indicating that Profile 3 could have been coded at a lower rate.

As the ABR video delivery technology matures, we believe that the profiles can be selected in a dynamic fashion in real time. A transcoder/encoder device that uses a two-pass video processing architecture can derive estimates on expected video quality for different profiles and then dynamically choose the combinations of profiles in a content adaptive fashion. In an alternate architecture, a transcoder can generate bitstreams corresponding to a large number of profiles. This can be followed by a downstream analysis device that inspects and compares the bitstreams corresponding to the individual profiles and then dynamically chooses a subset of profiles to be used. This can be accomplished in real-time with a small amount of processing delay.

<u>Profile selection - Adaptive performance comparisons across adaptive technologies</u>

The selection of encoding profiles is a key determiner of quality of experience for the customer and network capacity for the operator. Too few or improperly spaced adaptive variants may cause discontinuities in playback and marked video quality changes. Too many profiles increases the network storage requirements for a single asset and may cause unnecessary chattering between similar profiles as the adaptive player makes decisions based on available profiles and buffer condition. Encoding profiles are selected through a variety of rules of thumb and best practices. These include offering sufficient bits/pixel (typically 2-4), resolutions that are mod16 for efficient macroblock encoding, square pixel aspect ratio to ensure content is not anamorphically scaled on a variety of devices, and bitrate ratios between adaptive variants that are typically 1.5 – 2.

These rules of thumb can be tested by subjecting a player to dynamic network impairments while conducting subjective viewing of video quality and objective analysis of variant selection. Figure 6 illustrates an adaptive test system that was used to compare behavior of various adaptive technologies to a similar set of encoding profiles and network variations.

A number of adaptive delivery technologies are available and differ in platform support, multiplexing and manifest syntax. The performance of adaptive delivery is not determined by these variations, but mostly by the implementation of adaptive heuristics in the player. Three different adaptive technologies were evaluated; HTTP Live Streaming (HLS), HTTP Dynamic Streaming (HDS), and Smooth Streaming (MS Smooth) Each player was subject to the same time-based network variation and was provided an identical set of content encoding profiles to select from. The content profiles used are shown in Table 2.

Profile	Resolution (W x H)	Video Bitrate in mbps
1	1280 x 720	6.1
2	1280 x 720	3.3
3	768 x 432	1.5
4	640 x 360	1.1
5	512 x 288	0.85

Table 2. Profiles used for adaptive performance evaluation

The results in Figures 7 through 9 show how the unique decision methods implemented in each adaptive technology player affect the ability to react to variations in network condition. During the testing, no video discontinuities due to buffering or decoder starving were noted. In general, a player which maintains the highest profile selection and switches less frequently is preferred to one that selects lower resolution profiles and/or switches profiles frequently. The results suggest under slowly changing network bandwidth conditions, HDS and HLS maintain higher profiles than MS Smooth while HLS switches profiles less frequently. Each player was subjected to a second scenario of quickly changing network conditions with results shown in Figures 10 through 12. The results provide similar conclusions that HDS and HLS maintain higher profiles than MS Smooth while HLS switches less frequently. Video artifacts were noted occasionally on all three players during the intervals of greatest network congestion, as noted on the charts. This suggests a lower bitrate profile might be desired to improve performance under quickly deteriorating bandwidth situations.

Adaptive performance comparisons across devices

Adaptive delivery is customarily applied to address variations in network conditions to maintain a best possible experience. Adaptive delivery can also address variations in device

performance. Figure 13 illustrates the response of devices with a variety of performance characteristics to an adaptive stream. The test environment shown in figure 6 was used to subject a number of devices to the same content profiles, adaptive technology and time-varying network variation. Each device was monitored to determine which content profile was selected during these variations to evaluate device specific behavior. The results suggest that most devices are more than suited for a wide range of content profiles and network variations. The least capable device in the test, the iPod Touch, was capable of selecting profiles normally targeted for large screen viewing, although the device switched away from the profile with less network congestion than higher powered devices.

THE ROLE OF ADAPTIVE DELIVERY WHEN OVERLAYING NEW FORMATS

A concern with unifying the formats among devices is the ability to introduce new profiles. Example applications may include advancing resolutions such as ultra high-definition (4K or 2160p), unique audio formats such as 7.1 or new video compression technologies such as HEVC. New video formats can be introduced incrementally by adding a variant to the adaptive encoding profile set. The set of profiles available to a device may be tailored to device capabilities through manifest conditioning or video player settings. A just-in-time packager is suited to introduce new audio formats by combining the appropriate audio and video formats for the device at time of delivery. This method can also be applied to introduce additional audio tracks such as alternate language or visually impaired commentary without burdening all streams with additive audio bandwidth.

SUMMARY

Traditional delivery of video over QAM infrastructure has been optimized over the last two decades. Adaptive IP video delivery, especially over a managed network, is still in its infancy. There are several challenges that exist in this new delivery paradigm and an array of new technologies that can be leveraged to improve IP video delivery. Using a unified work-flow for content ingestion will ensure that the content is generated at the highest possible quality and at resolutions that are consistent with various client devices. Experimental results presented in this paper show that clients respond somewhat differently to varying network conditions. Use of active management can ensure that the quality of experience can be judiciously and fairly maintained for all active users irrespective of client heuristics.

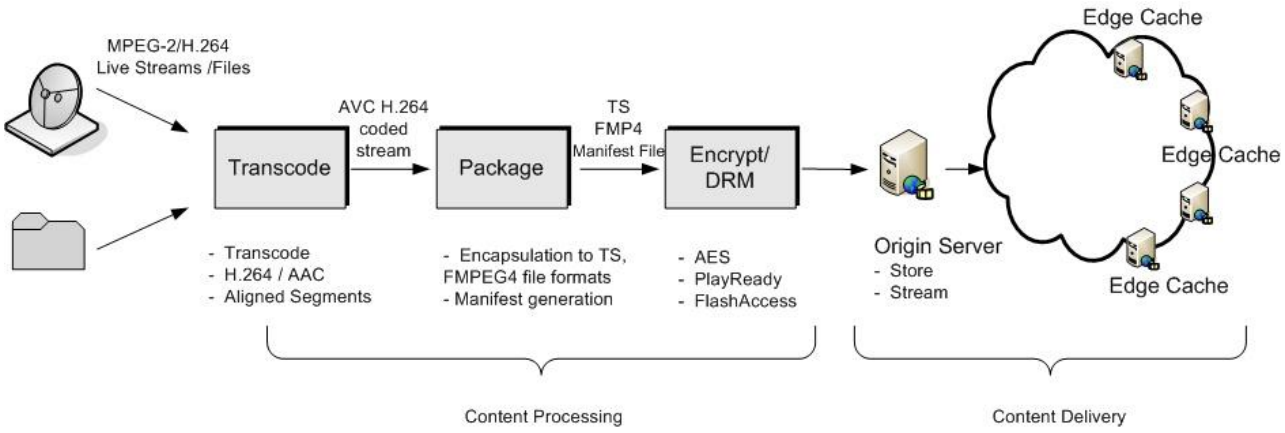

Figure 1. General Content Processing and Delivery Architecture

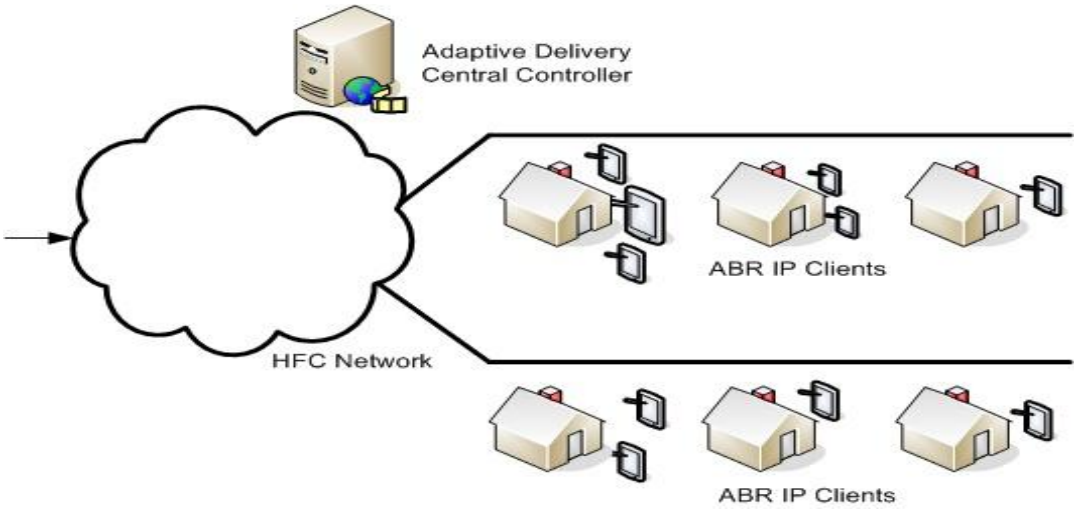

Figure 2. Central Control Representative Implementation

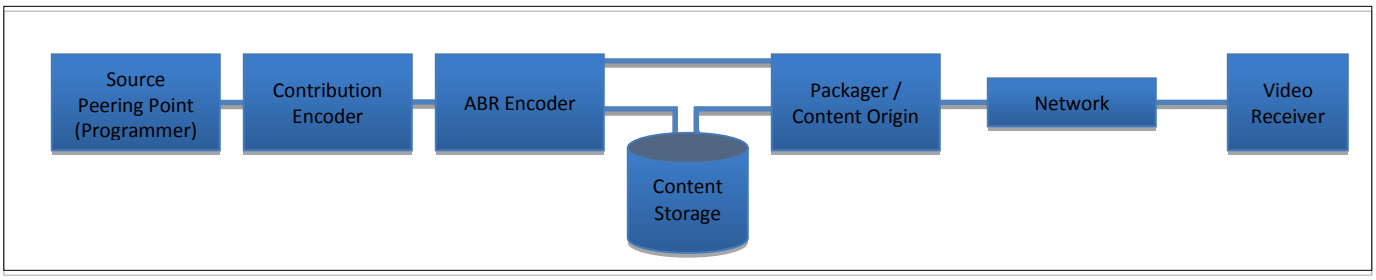

Figure 3. Unified Workflow

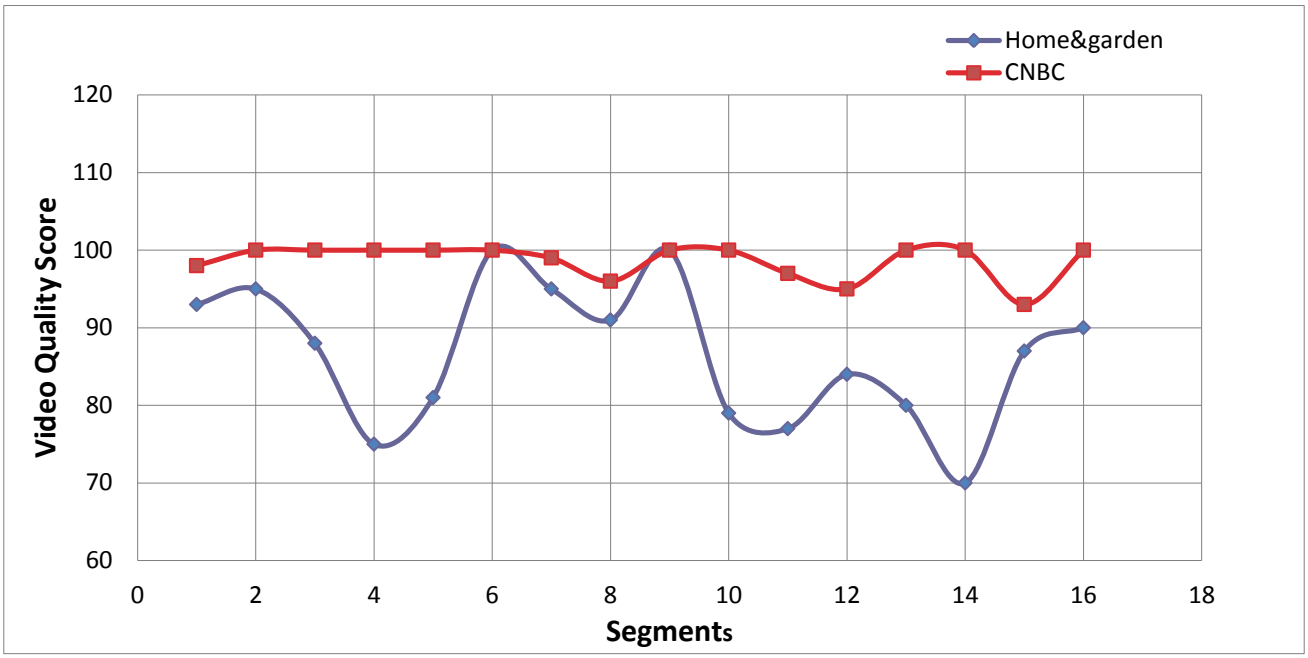

Figure 4. Video Quality with Table 1 Encoding Profiles

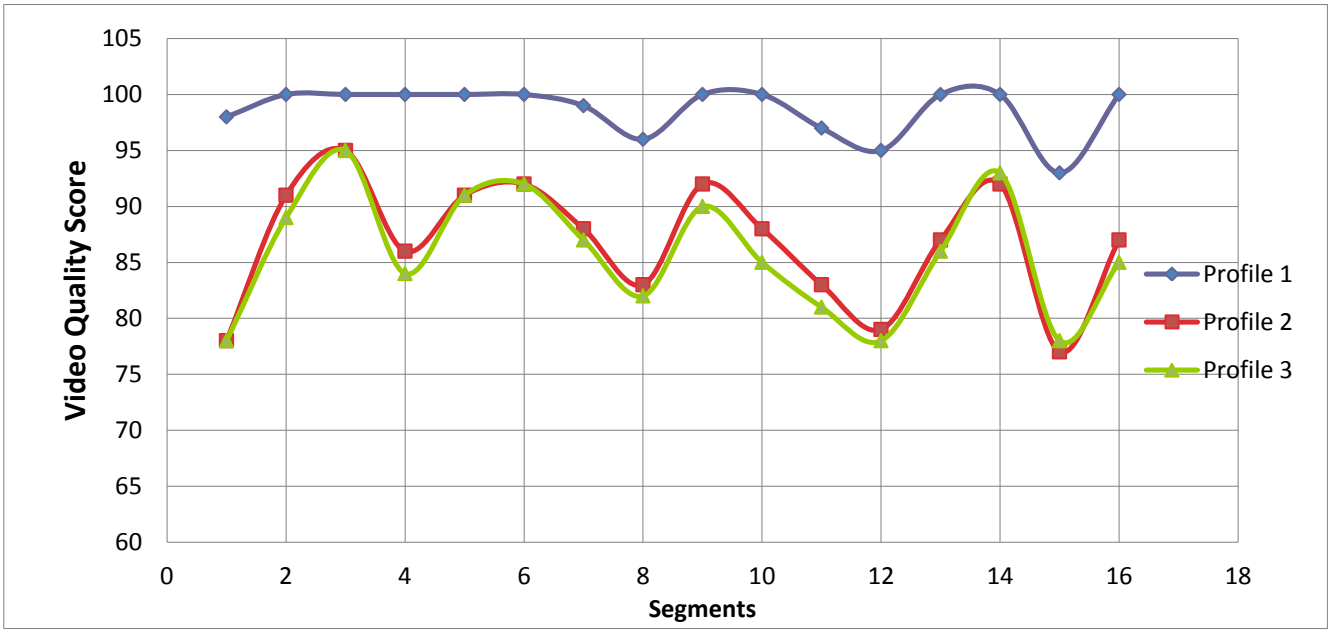

Figure 5. Video Quality Score for First Three Profiles

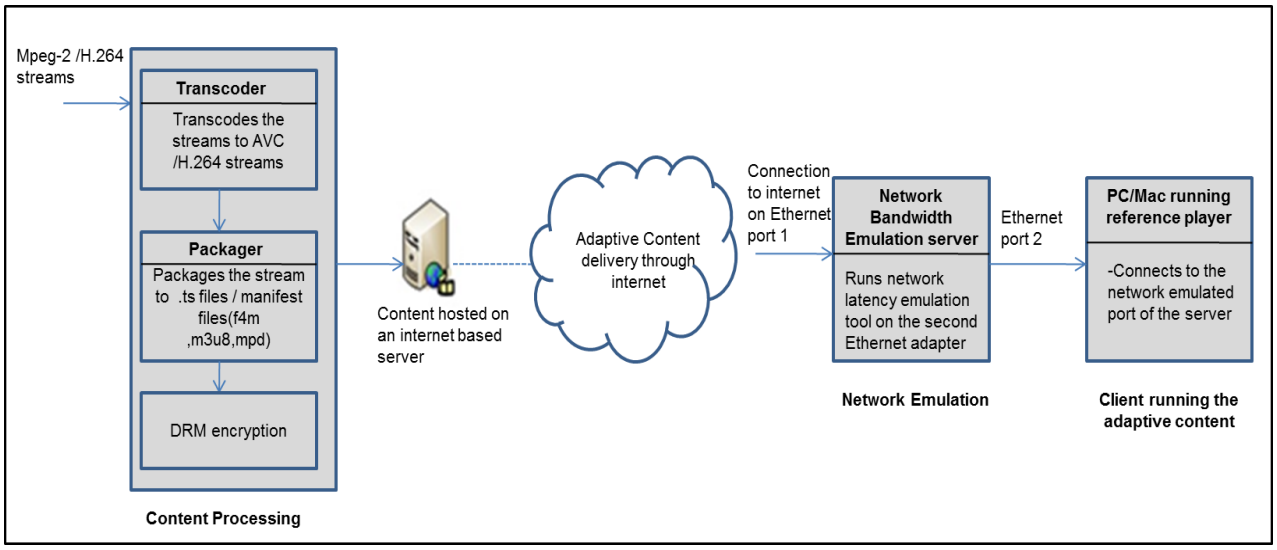

Figure 6. Adaptive Performance Test Environment

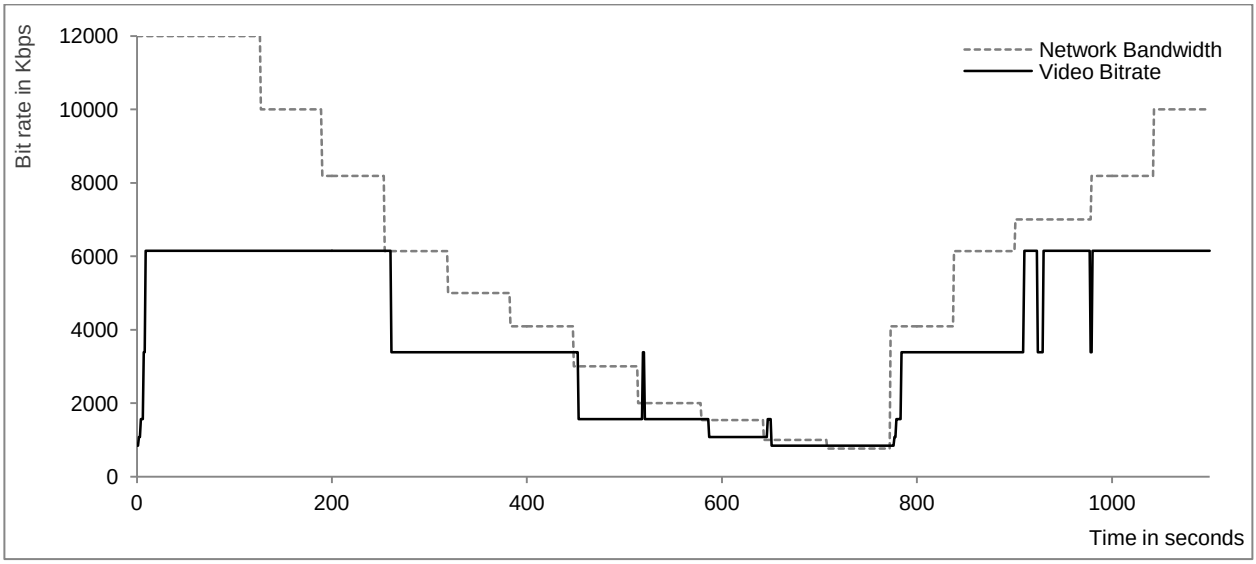

Figure 7. Adaptive Content Performance to Slow Variation - HDS

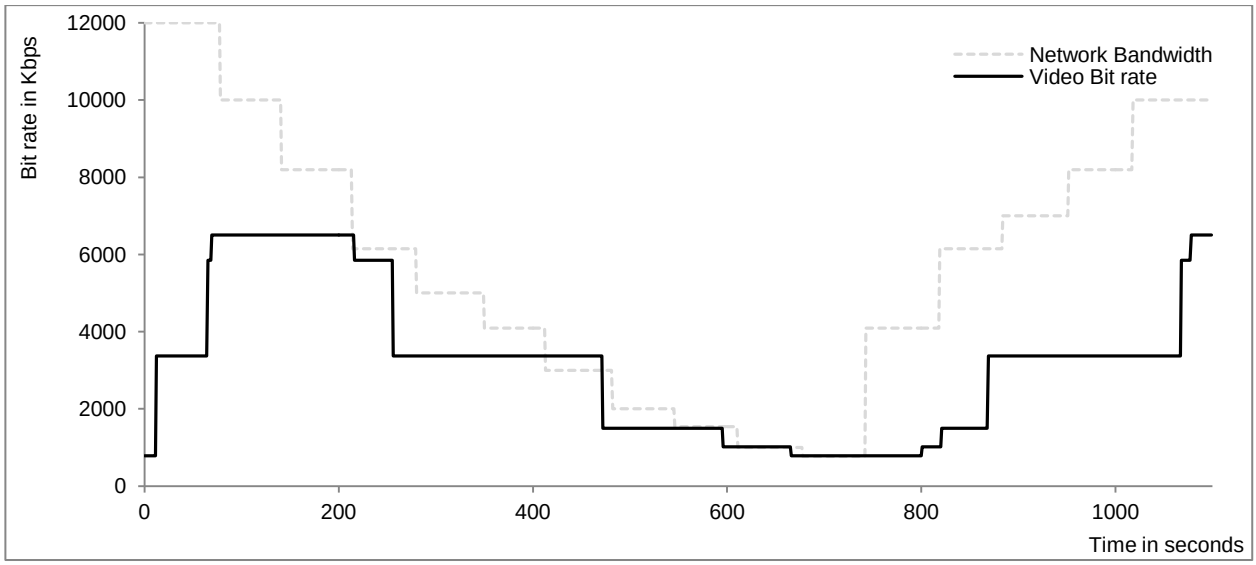

Figure 8. Adaptive Content Performance to Slow Variation - HLS

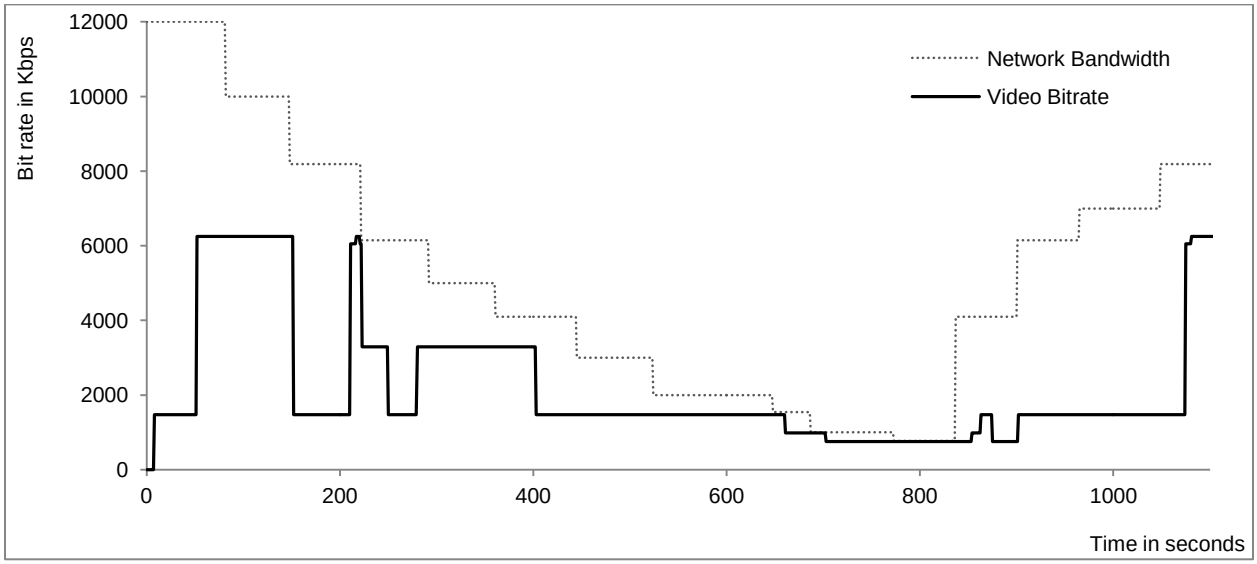

Figure 9. Adaptive Content Performance to Slow Variation -MS Smooth

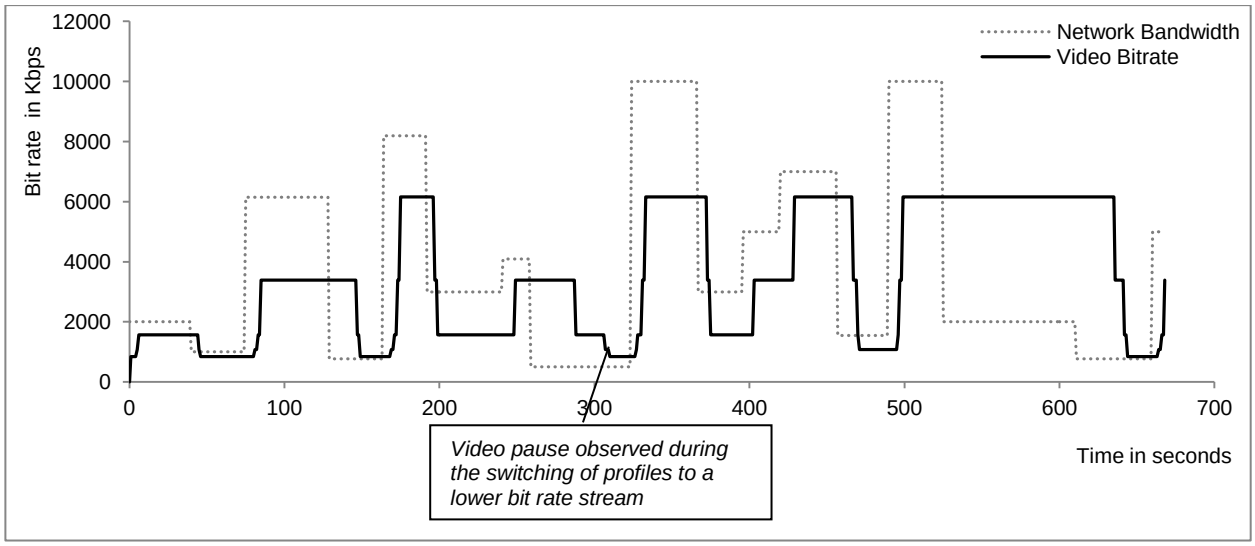

Figure 10. Adaptive response to rapid changes in network bandwidth - HDS

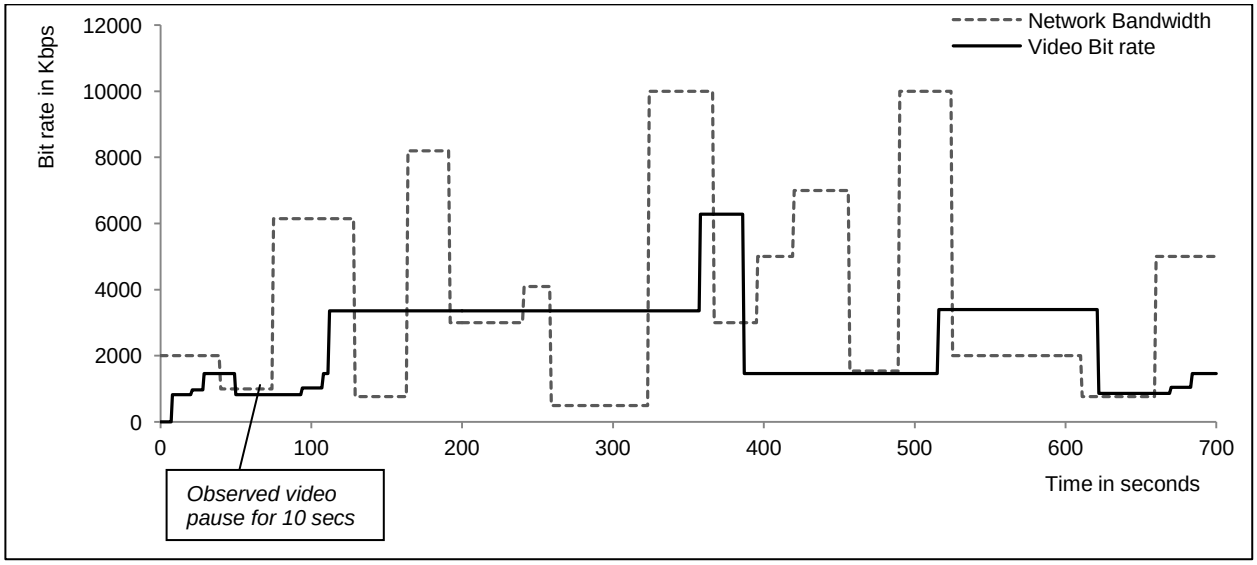

Figure 11. Adaptive response to rapid changes in network bandwidth – HLS

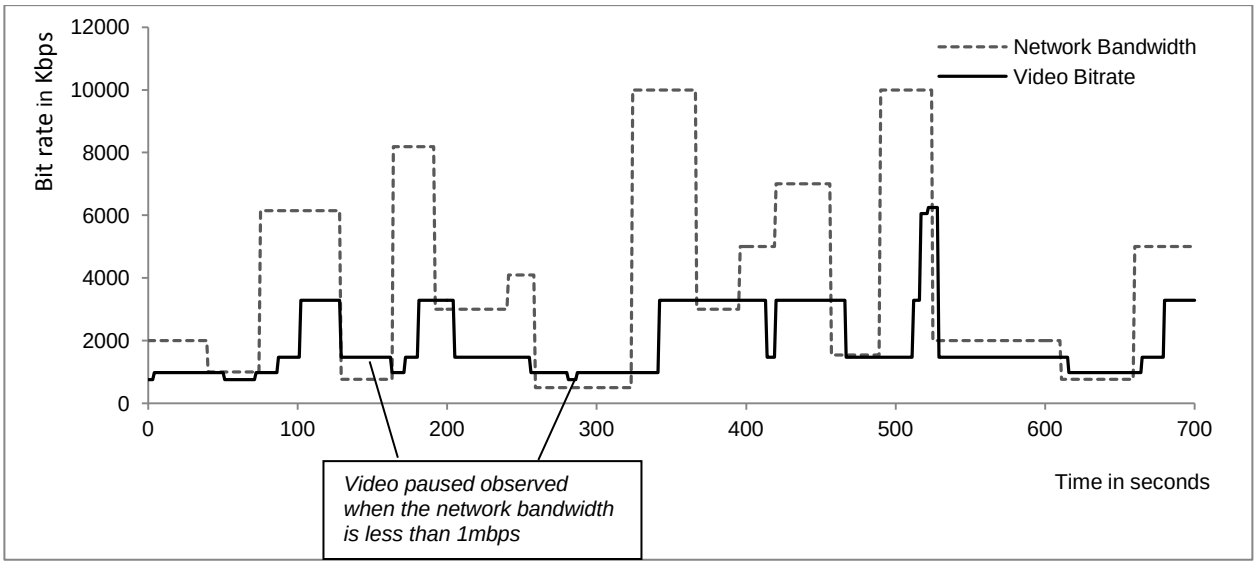

Figure 12. Adaptive response to rapid changes in network bandwidth – MS Smooth

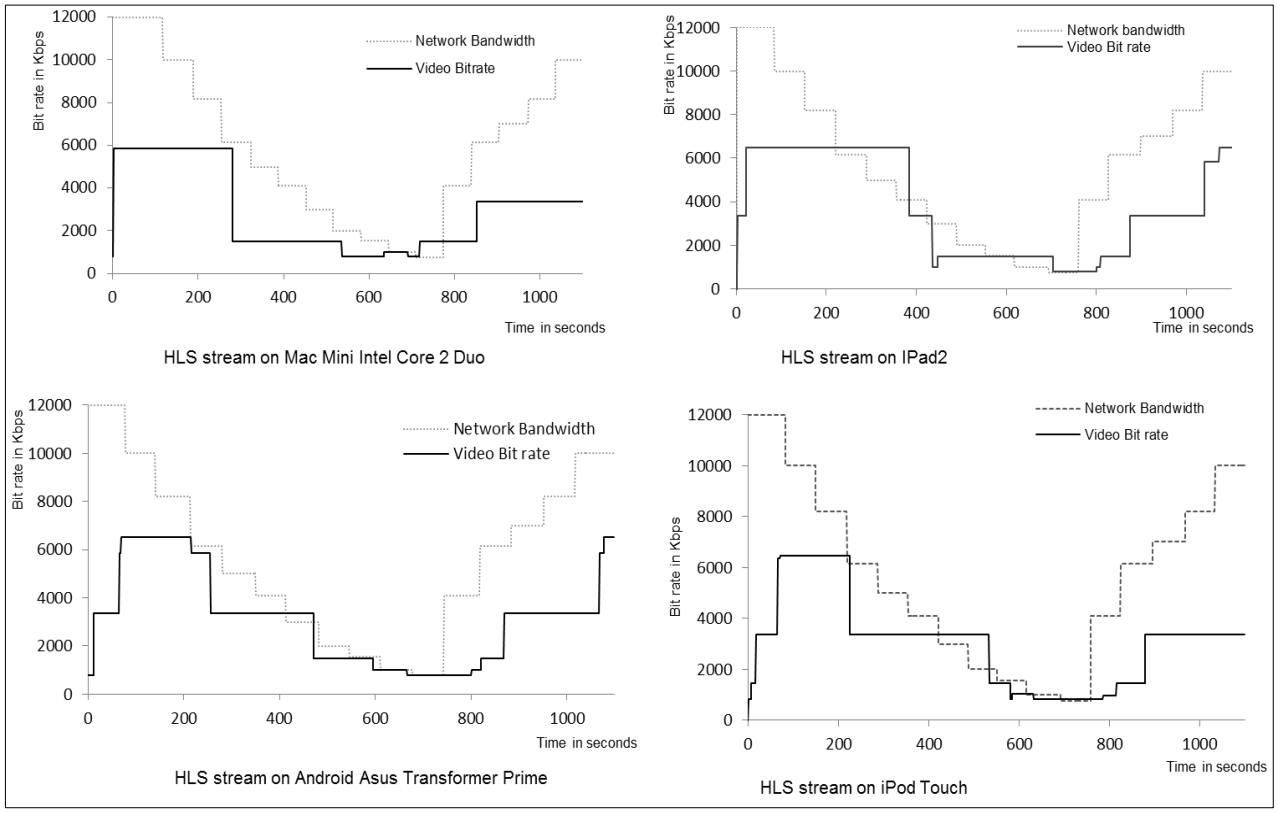

Figure 13. Adaptation to Device Capability