How OTT Services Can Match the Quality of Broadcast

Thierry Fautier
Harmonic
San Jose California
Thierry.Fautier@harmonicinc.com

Abstract - *Recent events, such as the 2018 FIFA World Cup, demonstrated that live OTT experiences have some serious limitations, including low quality on TV sets and a delay of more than 30 seconds compared with broadcast services. If the industry does not solve those problems, they could halt the development of live OTT offerings. This paper will analyze the current challenges involved with delivering live OTT services and will describe solutions that are available today, such as content-aware encoding, artificial intelligence, and low latency CMAF. In addition, the paper will present new concepts that can be deployed as part of an end-to-end OTT solution to further optimize quality of experience for OTT services on existing infrastructure (i.e., headend, CDN, and devices).*

INTRODUCTION

OTT is becoming more and more mainstream and is even used now to watch live services on TVs. In the U.S., between Sling, DirecTV Now, Hulu, YouTube and Sony Vue, there were more than 9 million OTT subscribers at the end of 2018 [1]. Of course, OTT consumers expect to have an experience similar to broadcast TV, and the industry has to provide a solution to meet this expectation. This paper will first define what a broadcast delivery network is, along with what defines OTT today. Then it will look at the different ways to deliver OTT services and analyze each element of the network's contribution to the quality of experience (QoE). Next the paper will look at the technology being deployed, such as content-aware encoding (CAE) and MPEG CMAF and examine a more holistic approach that looks at the overall network analytics to deliver the best performance to the client.

BROADCAST DELIVERY

A broadcast delivery network is a dedicated network that sends one piece of video content to many viewers and can therefore scale by adding more capacity for more subscribers, without any impact on the experience. Table 1 lists the different broadcast network options to distribute content at scale.

Network	Standards	Delivery network	Client	Scaling with
DTT	ATSC/DVB-T/ ARIB/T-DMB	OTA	TVs	Spectrum capacity
Cable QAM	CableLabs/ DVB-C	HFC	STB	Spectrum capacity
DTH	DVB-S (2)	Satellite network	STB	Spectrum capacity
IPTV	DVB IPI	IP multicast	STB	Network capacity

TABLE. 1: BROADCAST DELIVERY NETWORKS.

These types of networks deliver a similar QoE, which is considered the gold standard by today's consumers.

IP DELIVERY

IP can be delivered in multiple ways. As is the case with broadcast delivery, IP multicast is very scalable, as only one stream is delivered to millions of subscribers. The other delivery mechanism used for all connected devices is HTTP, which relies on adaptive streaming or ABR (adaptive bitrate). OTT generally uses unicast delivery protocol over unmanaged networks. In IP deployments, there is also a way to deliver HTTP adaptive streaming over managed networks. The different ways to deliver unicast to connected devices are summarized in Table 2.

Delivery type	Service	Often called	QoS	Access network	Gateway	Device	App
Managed ABR	TV over IP	IPTV 2.0	VLAN dedicated to service	Operator network (DSL/Fiber/ DOCSIS)	Controlled by operator	Controlled by operator	Controlled by operator
Best effort ABR	TV over internet	OTT	Best effort	Operator network (DSL/Fiber/ DOCSIS)	Owned by consumer/operator	BYOD	Controlled by Content provider
						Controlled by operator (1)	Controlled by operator (1)

TABLE 2: UNICAST DELIVERY MECHANISMS.
(1) OPERATORS PROVIDE STBS TO CONSUMERS; THEREFORE, THEY HAVE MORE CONTROL OVER THE EXPERIENCE.

Figure 1 describes the different network topologies.

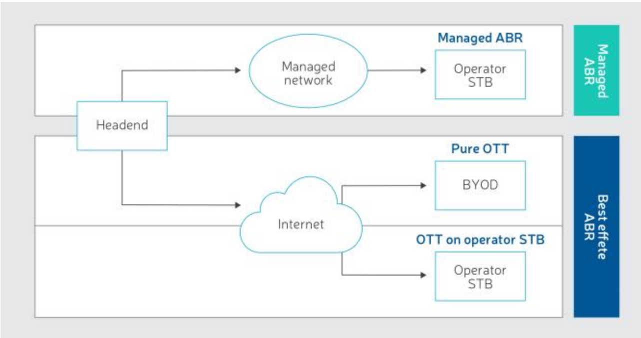

FIGURE. 1: UNICAST NETWORK TYPES.

I. Managed Network

In the case of ABR delivery over managed networks, the HTTP traffic is carried on a VLAN with a delivery priority (VLAN tagging) from the headend, down to the device. Even if it is delivered on Wi-Fi, operators can prioritize the Wi-Fi traffic on their own gateway. In this configuration, the QoE is close to IP multicast from a delivery point of view. Meanwhile, when traffic surges, if there is not enough network or streaming capacity, the servers will go to the lowest profile that enables smooth delivery. This is also called IPTV 2.0, and we see new telco deployments following this type of architecture.

II. Pure OTT

In the case of pure OTT, content providers use the public internet to deliver content to the end user, using HTTP protocols. Netflix and all of the OTT services are using this mechanism. Due to the popularity of Netflix traffic, there is congestion at the interchange level, and in order to avoid endless discussions with ISPs about the payment for traffic, Netflix is offering (for free) to put its own streaming server in the ISP PoP, through the Open Connect initiative [2]. Based on information shared by Netflix, the full catalog can fit on one server. With redundancy this equates to two servers per ISP PoP. There is no way to compare the difference in QoE with or without Open Connect, but many in the industry agree Netflix quality is the best compared with OTT services from other providers. Hence, we believe Open Connect caching is playing a big role in that.

III. OTT With Managed Devices

Another flavor of unmanaged ABR is where the operator brings its own device to consumers. This is the case for telcos that want to penetrate the customers of competitor ISPs or for OTT operators that want to better control the experience. This is a different case compared with the pure OTT one, as the operator fully controls the software stack running on the client and can therefore better control the user experience.

IV. Comparison of Different Architectures

There are very few operators that have deployed an ABR service on their network as well as OTT. We have gathered some data coming from a commercial deployment of one operator operating exactly the same service on its own network with its device and on OTT through its competitor's network. Table 3 provides the characteristics of the two deployments.

Service	Source	Network	Networking characteristics	Gateway	Gateway to STB	STB
Managed	ABR headend	IP private network	DSL, FTTH, DOCSIS	Controlled by operator	Wired	Managed by operator
Pure OTT	ABR headend	Public internet	DSL, FTTH, DOCSIS	Owned by consumer/operator	Wi-Fi	BYOD

TABLE. 3: UNICAST DEPLOYMENT COMPARISONS.

The difference in QoE (proprietary metrics defined by operator) between the managed and the OTT service is 300 percent. This is explained by the fact that on the managed side, the bandwidth can always be provisioned with the highest priority, there is no Wi-Fi interference and the client is fed with a wired Ethernet connection and is also managed by the operator. On the pure OTT side, the network is shared with internet traffic, the gateway does not give priority to the video service, Wi-Fi can create problems in a bad home network, and using a BYOD (bring your own device) leaves the operator at the mercy of the device's OS, even if the operator writes the app itself.

This example shows that a fully managed service has a much higher QoE, and this should give hope to IP operators (cable and telco) in their fight vs OTT operators that have to go through all of the hurdles of a pure OTT system. This could also open the path to use the ISP managed network to host OTT services in the future. To mitigate the pure OTT situation, operators are now developing their own (Android TV) STB. That way they can optimize the last part of the chain.

Needless to say, once the operator has access to analytics on the OTT device, it will do its best to migrate the customer on its own network, based on providing the "best consumer experience." This Trojan Horse approach is already being used by several operators and should seriously be considered by operators interested in grabbing more market share, not only for their OTT service but also for their broadband offering.

NETWORK DELIVERY ASPECTS FOR ABR DELIVERY

This section will look at how each element of the network can have an impact on the QoE. Let's first review the end-to-end delivery system in Figure 2.

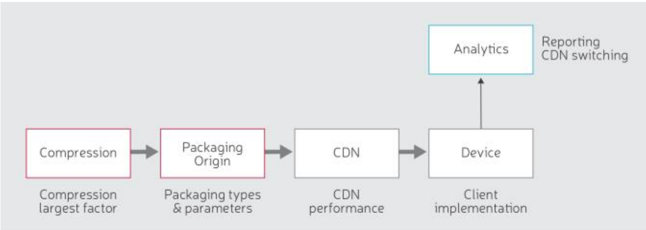

FIGURE. 2: END-TO-END QOE CONTRIBUTION.

I. Compression

Compression is the engine of ABR delivery. The lower the bitrate, the easier it is to deliver unicast to the device with the highest QoE. The problem is if the bitrate is too low, the quality will suffer and thus the QoE will be degraded. This can be resolved by using the most advanced codec, such as HEVC, but the current licensing terms have slowed down deployments. However, several recent reports [3], [4], [5] show that by 2019, 63 percent of operators will have an HEVC service deployed. With that being said, in June 2017, HEVC garnered extra attention after Apple announced that it will support the video compression standard across all of its platforms, including Mac, Safari, Apple TV and iOS11. This announcement was significant for HEVC, as iOS traffic counts for about half of all video streaming over the internet.

In order to work around the HEVC licensing problem, many companies have developed what Netflix calls "per tile encoding" for VOD and what the industry at large is now calling "content-aware encoding (CAE)," which supports both VOD and live applications. CAE received a huge blessing by Apple when it announced, in 2017, support of VBR for the VOD and live encoding in iOS11 [6]. The Ultra HD Forum has identified the use of CAE for Ultra HD delivery over the internet as critical and has included CAE in its Guidelines [7]. In addition, the Forum demonstrated the technology at NAB 2018 and IBC 2018, showing a consistent savings of 40 percent vs CBR for UHD ABR using CAE with multiple vendors (BeamR, Brightcove and Harmonic) across different applications (VOD and live). More details on how CAE works will be provided in the CAE section.

II. Packaging/Origin

The packaging technology used will have an impact on the QoE. Several reports have studied the differences between HLS and DASH. The most recent one [8] gives a good overview of the different techniques used to deliver the best QoE, including the client behavior.

One of the biggest problems facing OTT QoE is the end-to-end latency. With current HLS or DASH implementations, the end-to-end delay is between 30 and 90 seconds, as described in Figure 3.

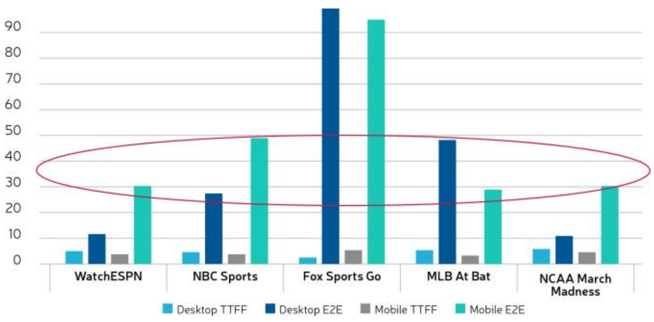

FIGURE. 3: U.S. OTT LATENCY (COURTESY WOWZA MAY 2017).

There are many ways to solve the latency problem: reduce the segment size, move to a different protocol like WebRTC or use MPEG CMAF. Reducing the segment size (down to one second) would increase the network traffic, may impact the video quality as it puts constraints on the compression schemes and has so far not been deployed in any commercial system. Using WebRTC, which is an old protocol, has the disadvantage of not being cacheable with off-the-shelf CDN servers. It requires dedicated infrastructure and has not been proven to scale for millions of concurrent sessions. MPEG CMAF will be presented in more detail in the low delay section.

III. CDN

CDN is a technology that is proven to work at scale for the most demanding events (i.e., the Super Bowl, Olympics, FIFA World Cup). The most popular event CDNs have been used for was in India with the IPL draining 4.8 million concurrent users [10]. To be efficient a CDN needs to have enough streaming capacity as well as servers located as close as possible to the clients. Operators have come to the conclusion that some CDNs perform better than others during certain times of the day and in certain locations. This is why we have seen a CDN selector technology market segment develop, with some success. The concept is described in [11].

Another way to look at CDNs is to build them inside of the operator's network, using either off-the-shelf technology from CDN vendors or technology developed by the operator itself like Comcast [12]. This approach offers the benefit of serving a certain type of traffic, overload can be off loaded to commercial CDNs, and in terms of depth in the network, the caching servers are located closest to the subscriber to ensure the best QoE. The drawback is that this is capex-intensive, as the operator needs to deploy the network and opex-intensive to manage the CDN 24/7.

IV. Device

The device is an important element of the delivery chain, and there is a wide diversity between the native clients (i.e., HLS, Android) and the custom clients developed by client companies. [13]

The OTT clients own their ABR selection rules using well known family of algorithms (based on buffer occupancy or bandwidth estimation). Advanced clients also embed smart processing to reduce the startup time and end-to-end latency to the minimum possible.

The unique streaming optimization technologies developed by companies, such as Giraffic [14], enable the client to fetch data based on rules, using TCP multi-connection strategies.

The unique client-side-only streaming optimization technologies developed by companies such as Giraffic, enable the client to fetch data based on rules and more robust TCP get strategies. This gets a much higher bitrate on average than the classical approach. Per Giraffic, the client can fetch the data sometimes two to three times faster than the classical approach, by using strategies such as TCP multi-connection and client-side congestion management based on variable byte range requests, as well as shaping the traffic ingested to the video player in order to facilitate better ABR switching decisions.

The downside of this technique is that potentially other devices in the home which are not enabled with such technologies might become more congested.

V. Analytics

Analytics can be of many sorts. They come with the CDN, and we have seen client analytics technology developed by several companies. Analytics can be used for business reports, which are of little use for QoE, but can be used for more important technical reporting of key parameters such as startup time, buffering rate, SD/HD profile ratio, and profile switch.

Out of those analytics, the operator has a good view into the user experience, but there is no indication on what to change in the network and what parameters should be modified in the video delivery network to increase QoE. This will be discussed in the network optimization section.

COMPRESSION IMPACT ON QOE

Content-aware encoding is one of the hottest technologies today in the encoding market, based on the fact that all encoding companies released in 2018 a similar form of CAE technology. The concept of CAE is that the encoding server looks in real time at the video complexity and in real time adjusts the encoding parameters to provide the best picture quality. It works similarly to VBR for statmux, except only one program is encoded and the video quality measurement is more refined since it is based on the Human Visual System Model compared with statistical multiplexing, which is based on basic rate control based on QP value. In order to have a more accurate video quality measurement, the CAE live system is trained offline using artificial intelligence technologies. For VOD, CAE can be used in one pass, as done for live. This provides the highest scalability (encoding speed) but not the lowest compression ratio. Alternatively, this can be done in several passes, where each encoding parameter set is encoded in parallel and the decision is made at the end of each encoding batch. This is the technique currently used by Netflix.

Table 4 provides the typical savings compared with CBR encoding on HD content using Harmonic's EyeQ CAE solution.

Resolution	Frame Rate	Profile	Max Bitrate (CBR)	Saving (%) Balanced
1920x1080p	50	AVC High	6	37
1280x720p	50	AVC High	3.5	34
960x540p	25	AVC Main	1.8	31
640x360p	25	AVC Main	1.25	31
320x180p	12.5	AVC Main	0.5	39

TABLE 4: THE SAVINGS PROVIDED BY HARMONIC'S EYEQ CAE TECHNOLOGY.

The overall savings compared with CBR averaged over all profiles is 34 percent, close to the 40 percent demonstrated by the Ultra HD Forum on UHD.

In addition to bandwidth savings, CAE also offers a better QoE to where when the video is better compressed, more HD profiles can be received, and when the bitrate is low, there are less buffering effects.

Figure 4 represents the analytics results when comparing CBR vs Harmonic's CAE EyeQ technology on a 4G network. The profiles used were the ones described in Table 3.

FIGURE.4: ANALYTICS SHOWING THE BENEFITS OF CAE.

For more details on CAE, Harmonic has published a white paper on the subject [15].

DELAY IMPACT ON QOE

Latency is an issue during ABR delivery, whether over managed or unmanaged networks, and more recently this problem was noticed during the FIFA World Cup 2018, when viewers watching the event via OTT services discovered that the platforms were not very stable (see the network delivery aspects for ABR delivery section). The OTT services were way behind the broadcast delivery. Figure 5 explains the difference between a classical ABR delivery and CMAF LLC delivery.

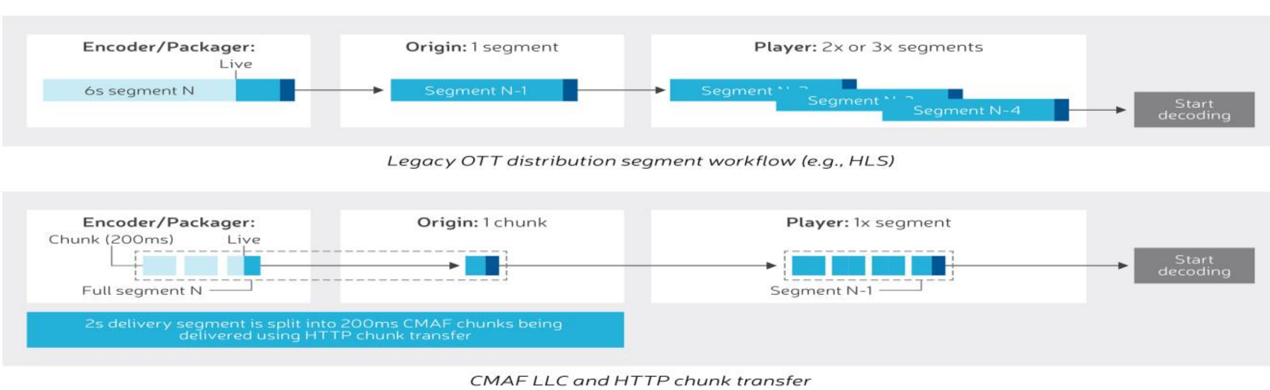

FIGURE. 5: COMPARING LEGACY OTT DISTRIBUTION WITH CMAF LLC AND HTTP CHUNK TRANSFER.

The MPEG Common Media Application Format (MPEG-CMAF) is a media container standard that was recently introduced to simplify the distribution of OTT at scale. MPEG-CMAF is based on fMP4 (ISOBMFF) and can be used by both MPEG-DASH and HLS delivery formats with a common encryption scheme. Indeed, CBC scrambling is now supported by Apple, Microsoft, Google, and Adobe, allowing for a single scrambled media segment, even though different DRM are used by end-user devices. Different MPEG DASH manifest and HLS playlists are still required, but with CMAF, the media segment becomes common across both formats, which greatly simplifies the heavy load of the content distribution workflow.

The second major benefit of MPEG-CMAF is the LLC (Low Latency Chunk) option to support low latency. MPEG-CMAF LLC introduces the possibility to deliver a segment by small chunks (e.g., 200 ms) before the full XX-second or six-second segment is calculated. With CMAF LLC, data transmission is accelerated across the whole workflow, including in the decoder, which can potentially start decoding/display before a complete segment is encoded/received. MPEG CMAF LLC performs very well and can deliver end-to-end latency of three seconds or less with a short latency encoder and an optimized DASH OTT player (as iOS 11 supports CMAF but not LLC). Of course, the HLS player can still decode the stream but may introduce an additional latency compared with a MPEG DASH CMAF LLC player.

Harmonic has performed a variety of field tests to compare broadcast vs ABR delivery, and full reports of the tests are available [16]. Table 5 summarizes the different delays measured during the field tests.

Test conditions	Network	CDN	Network	Device	Measured delay
On-premises	Unmanaged	Akamai	Wired	Wired	5.5s
			Wireless	Wireless	7.5s
Public cloud	Unmanaged	Akamai	Wired	Wired	7.0s
			Wireless	Wireless	9.5s

TABLE 5: FIELD TRIAL RESULTS ON LOW LATENCY.

If you consider that the broadcast delay today for pay TV is around five seconds for satellite networks, you can conclude that only the most favorable conditions (on premise/wired) can provide a delay approaching that of broadcast. Other delivery mechanisms will be between 40 percent and 90 percent additional delay. Note that if the encoder delay can be reduced, then the overall delay will be reduced with an impact on the consumed bandwidth, which today operators are not ready to sacrifice. In the same way, a more aggressive client could save a fraction of a second with the impact being more buffering risks, which again the operator is not ready to sacrifice.

We have meanwhile reduced the delay from 30 to 90 seconds to five to 10 seconds, which is a gigantic 6-9x step, or more, especially without any impact on encoding, a minor impact on the packager, CDN and client that will all have to be upgraded with CMAF LLC. Note that if a client is only CMAF compliant without LLC support, the delay will have to be increased by a few seconds (double buffering of a segment). This could be the lazy option for operators to take until Apple releases a native CMAF LLC client.

NETWORK OPTIMIZATION

In the analytics description, this paper noted that analytics have mostly been used to monitor the client behavior today. Adding analytics to the CDN and the network enables operators to gather a lot of information that will help to better understand what the chocking points in the network are.

Figure 6 describes the state of the art today in terms of network monitoring.

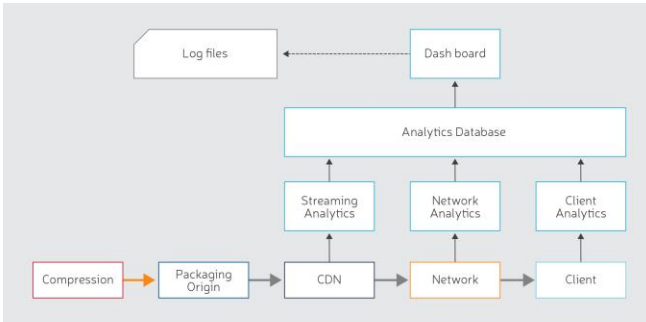

FIGURE. 6: GENERIC NETWORK MONITORING SOLUTION.

Various publications cited in this paper [8] [13] make a point that not all clients are equal in their way to intelligently exploit ABR technology, and some of them are not even fair (always trying to get the highest bitrate on the manifest). We therefore believe that in severe congestion situations, it is best to guide the client and to offer only a limited manifest during the congestion period. This guidance provided to the OTT client can also be driven by business rules, giving, for example, access to top representation for premium customers and not for others during these severe network conditions.

After building a strong monitoring solution like described in Figure 6, the next step is to build an orchestration system, which we will call the “network optimizer” that will in real time, at the segment pace, collect all of the network analytics and will define key parameters to be applied to the network elements, including the edge cache, origin server and encoder.

Figure 7 describes a high-level architecture for implementing a network optimizer function.

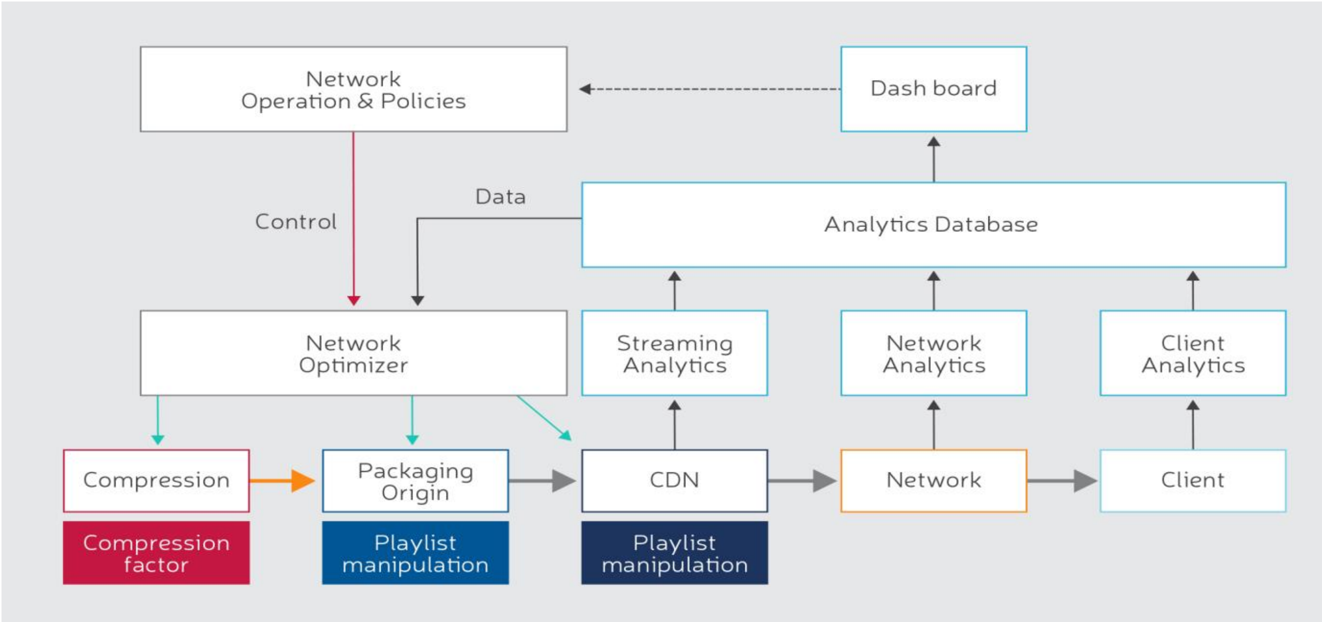

FIGURE. 7: NETWORK OPTIMIZATION SOLUTION.

The challenge to implement such a system is that there are no standards for the analytics collection and in order to implement this solution, the solution provider has to work closely with the operator.

BENCHMARKING WITH NETFLIX

Beyond the fact Netflix is considered today to have the best user experience for pure OTT delivery, we wanted to compare OTT service delivery for the following reasons. If a consumer has a good experience with Netflix and if an OTT provider on the same network, using similar technical parameters such a resolution, cannot match the Netflix QoE, we predict the service will not be successful. This section will look at all of the weapons Netflix has at its disposal.

I. Encoding

Netflix was the first to deploy "per tile" then "per chunk encoding" in 2015, which is the equivalent of CAE. Netflix is using HEVC only for UHD and is preparing to deploy AV1 for all of its devices in the 2019-2020 timeframe. But all of this is for VOD asset encoding, not for live.

II. Edge Caching

Netflix, via its Open Connect program deployed in 2012, has deployed its own caches in ISP networks. Based on this, we can say that Netflix, when deploying Open Connect, is not an OTT service provider anymore and has a network that is close to a managed network for distribution to the home.

Footnotes

III. Traffic Prioritization

As Netflix is a pure OTT service provider, it cannot influence the QoS inside the home.

IV. Client Optimization

Netflix has not made any publication on its client optimization technology, but Giraffic has published, in January 2018 [17], a benchmark showing Netflix is fetching faster than regular clients.

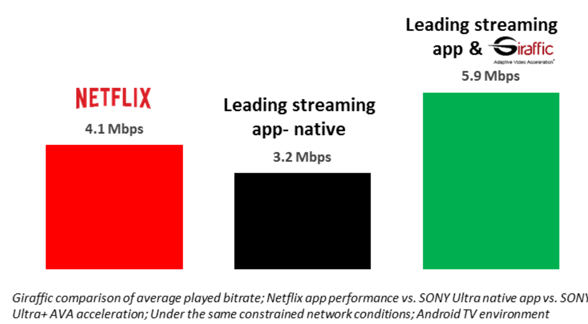

FIGURE. 8: CLIENT DOWNLOAD COMPARISON.

V. Network Monitoring

Netflix has not published much on how it monitors its network, but we can expect this to be a major subject of research, as in the U.S. only it serves 58 million subscribers.

VI. Summary

Table 6 provides a comparison between legacy OTT services, Netflix's service and what we believe can be an ABR delivery service that can match broadcast quality.

Technology	Category	Standard OTT services	Netflix	Advanced OTT services
CAE	Encoding	No	Yes	Yes
New codec (HEVC/AV1)	Encoding	Some	Yes	Yes
CMAF LLC	Packaging (Live)	Being deployed	NA	Yes
	CDN			
	Client			
Deep caching	CDN	Some	Yes	Yes
CDN switching	CDN	Yes	Not necessary	Not necessary on managed network
Traffic prioritization	Home network	Managed networks only	No	Managed networks only
Client acceleration	Client	Some	Yes	Yes
Network optimization	Monitoring	No	Unknown	Yes

TABLE 6: COMPARISON OF DIFFERENT OTT SERVICES.

As one can see in the table, legacy OTT services are far from having all of the tools to match the QoE provided by Netflix, while the advanced OTT services can not only compete with Netflix but also provide an experience that can be close to the broadcast experience for live services.

CONCLUSION

This paper has reviewed the different network architectures that provide a better QoE for ABR delivery. Encoding and especially CAE are key to provide the best QoE. CMAF is an essential technology to reduce the ABR delivery end-to-end latency to a level close to what broadcast offers. CDN caching is an important factor, and client acceleration should also be considered. Lastly, to optimize the network resources and avoid unfair client behavior, we advise operators to deploy a network optimization technology. All of those elements make us confident that in the very near near future an ABR delivery mechanism can be close to a broadcast delivery network on a managed network. On an unmanaged network it would be more of a challenge, as a lot of parameters are out of the control of the OTT service provider.

ACKNOWLEDGEMENTS

The work presented in this paper would not be possible without the support of many people at Harmonic, especially from Noam Koren and Raz Nitzan, who have respectively pioneered the content-aware encoding and network optimization technologies before they became a commercial reality.

REFERENCES

[1] Virtual MVPD subscribers will reach 9.2M by end of year, analyst predicts: https://www.fiercevideo.com/cable/virtual-mvpd-subs-will-reach-9-2m-by-end-year-analyst-predicts

[2] Netflix Open Connect: https://openconnect.netflix.com/en/

[3] The impact of Apple's HEVC adoption: A survey-based report; http://www.streamingmedia.com/Research/7694-The-Impact-of-Apple's-HEVC-Adoption-A-Survey-Based-Report.htm

[4] Bitmovin 2018 video developer survey reveals shifting technology landscape; https://bitmovin.com/bitmovin-2018-video-developer-survey-reveals-shifting-technology-landscape/

[5] HEVC adoption slow but steady says Beamr Report: https://streaminglearningcenter.com/codecs/hevc-adoption-slow-steady-says-beamr-report.html

[6] HLS authoring specification for Apple devices: https://developer.apple.com/documentation/http_live_streaming/hls_authoring_specification_for_apple_devices

[7] Guidelines for UHD Phase B implementation: https://ultrahdforum.org/resources/phaseb-guidelines-description/

[8] A survey on bitrate adaptation schemes for streaming media over HTTP: https://ieeexplore.ieee.org/document/8424813/

[9] Per-title encode optimization: https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2

[10] Hotstar and Akamai set global streaming record during VIVO IPL 2018: https://www.akamai.com/us/en/about/news/press/2018-press/hotstar-and-akamai-set-global-streaming-record-during-vivo-ipl-2018.jsp

[11] Pluralize your CDN for peak performance: https://www.comcasttechnologysolutions.com/blog/pluralize-your-cdn-peak-performance

[12] Building a large scale CDN with Apache Trafficserver: https://events.static.linuxfound.org/sites/events/files/slides/apachecon_jvd_2014_v2_16x9.pdf

[13] Comparing adaptive HTTP streaming technologies; https://www.nctatechnicalpapers.com/Paper/2014/2014-an...of-http.../download

[14] Giraffic adaptive video accélération white paper: http://www.giraffic.com/wp-content/uploads/2014/08/Giraffic-White-Paper-AVA.pdf

[15] EyeQ - Achieving superior viewing experience: http://info.harmonicinc.com/Tech-Guide-Harmonic-EyeQ%20

[16] DASH CMAF LLC to play pivotal role in enabling low latency video streaming: https://info.harmonicinc.com/low-latency-white-paper

[17] Giraffic goes after streaming platforms with SDK release: http://www.giraffic.com/in-the-news/giraffic-on-streaming-media-magazine-ces-18-giraffic-goes-after-streaming-platforms-with-sdk-release/