SCTE CABLE-TEC EXPO FAST FORWARD 2021
VIRTUAL EXPERIENCE
OCTOBER 11-14

2021 Fall Technical Forum
SCTE • NCTA • CABLELABS

How to Optimize TCO and QoE in a Cloud Environment Using a Context Adaptive Delivery Solution

A Technical Paper prepared for SCTE by

Patrick Gendron
Director, Innovation
Harmonic
2590 Orchard Parkway
San Jose, CA 95131
U.S
+1.800.828.5521
Patrick.Gendron@harmonicinc.com

Thierry Fautier,
Vice President of Video Strategy, Harmonic

Table of Contents

Title **Page Number**

1. Introduction ... 3
2. Current State of the Art for the Video Streaming Industry ... 3
3. Motivations to Move From Static Workflows to Dynamic Delivery of Content ... 4
4. Industry Trends and Research ... 5
5. Moving to More Dynamic Workflows ... 6
   - 5.1. Introduction ... 6
   - 5.2. AI-based Encoding/Content Aware Encoding ... 7
   - 5.3. Elastic Encoding ... 8
   - 5.4. Dynamic Resolution Encoding ... 8
   - 5.5. Dynamic Frame Rate Encoding ... 9
   - 5.6. Delivery Optimization ... 10
6. Conclusion ... 13

Abbreviations ... 14

Bibliography & References ... 15

List of Figures

Title **Page Number**

Figure 1 – AI-based Encoding ... 8
Figure 2 - Dynamic Resolution Encoding ... 9
Figure 3 - Dynamic Frame Rate Encoding ... 10
Figure 4 - Typical analytics collection architecture ... 11
Figure 5 - Context Adaptive Delivery (CAD) generic architecture ... 11

List of Tables

Title **Page Number**

Table 1 - OTT evolution from static to dynamic workflows ... 7
Table 2 - Delivery optimization summary ... 13

1. Introduction

OTT delivery is increasingly becoming a primary solution for the consumption of live video content. With OTT, the QoE provided to users should be at the same level as traditional broadcast TV.

OTT has become so mainstream that even live content is now available from video streaming providers. 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, according to a Fierce Video report. Yet, quality is sometimes an issue, and that's a problem because consumers expect the same video QoE for OTT as they've experienced with broadcast TV.

While the experience is expected to be the same or better, there are many technical differences between OTT and broadcast delivery. OTT targets a variety of devices (i.e., smartphones, tablets, desktop computers, connected TVs, game consoles) and delivers the content over a variety of networks (i.e., xDSL, fiber, radio 4G and now 5G).

To address these issues, many efforts have been made to define technical solutions for OTT streaming. One example is lowering the latency for the most popular streaming protocols HLS[1] and DASH[2]. This was presented in the 2019 SMPTE conference paper "How OTT Services Can Match the Quality of Broadcast"[3], but there are still some problems to be tackled when it comes to achieving massive at scale viewing of live events.

This paper will examine the different solutions that can be deployed in an OTT environment, comparing the technical merit, the integration aspects in an open ecosystem, the need for standardization and the overall impact on total cost of ownership (TCO) and QoE. It will provide suggestions on how scalability can be achieved to deliver high-quality live video to millions of subscribers on every device at any time, even during peak hours.

2. Current State of the Art for the Video Streaming Industry

There have been a lot of new streaming protocols and formats popping up over the past decade, but when you observe the current OTT delivery landscape for video on demand (VOD) and live content, it's clear that a vast majority of content is distributed using either HLS or DASH formats. Both use HTTP[4] as the underlying transport protocol and are based on adaptive bitrate (ABR) technology, which makes it possible to deliver video over unmanaged networks with variable available access bandwidth. The other formats have either reached obsolescence (e.g., Microsoft Smooth Streaming) or should be reserved for more specific use cases (e.g., ultra-low latency) as they come with some additional constraints. For example, WebRTC[5] can provide low latency but relies on peer-to-peer and thus has some serious scalability issues.

Since the focus of this paper is on large-scale content delivery, it is not a debate that the streaming industry will rely on two dominant streaming technologies for the next few years: HLS and DASH both using CMAF[6] as a common format for the delivery segments.

While there's been a growing demand for live content over the past few years, the vast majority of OTT consumption has always been and is still VOD content. This brings some additional constraints to the delivery workflow.

Several content delivery optimizations have already been deployed but most, though not all, are dedicated to VOD asset distribution or suffer huge limitations:

- Several years ago, Netflix introduced its per-title encoding [7], then per-scene encoding to provide a better video quality at a given bitrate. It also included a new paradigm of adaptive ladder, removing the profile when it doesn't bring any value to the end user from a video quality standpoint. These innovations improve the viewing experience, but they have been implemented for VOD assets with no real-time constraints on the processing.
- During the BEITC 2019 conference, Brightcove presented [7] some techniques to dynamically adapt the profile ladder based on the network conditions. These are a promising path but not yet available for live content distribution.
- Storage of recorded assets (e.g., cDVR applications) can be optimized by an offline profile curation removing the nonessential rendition in the profile ladder. The criteria to remove the nonessential rendition is based on a perceived QoE by the end user. This approach can work to optimize the storage volume and therefore the costs but again this is not yet something that can operate in real time for live content.
- Current multi-CDN strategies are based on static cache allocation. For large-scale events like the Olympics or FIFA World Cup, a major CDN would need to book physical resources up to 12 months in advance.
- Finally, as required for any closed-loop optimization client/CDN analytics may be collected in real time (which sometimes occurs at the end of the viewing session) but are generally not processed in real time to build actionable insights to optimize QoE.

When a popular event must be delivered live over a variety of networks to a variety of devices, it is a lot of work to make sure everything goes smoothly. For such an event or for regular peak audience, most popular services experience some QoE issues like rebuffering and long start time, when it is not a total impossibility to connect to the service.

The next section will discuss how the scalability issues and generally how a service operator can improve the delivered QoE by moving from a fixed, non-optimal workflow to a much more flexible workflow that is adaptative to the external context.

3. Motivations to Move From Static Workflows to Dynamic Delivery of Content

Depending on the business model for the service provider, the most critical metrics it needs to monitor and improve are:

- The acquisition of new subscribers, at a reasonable cost
- The churn ratio (or the ratio of new subscribers) for a subscription-based model or the total viewing time for an ad-subsidized model
- The cost of service (that can be approximated by the TCO value).

The first bullet will not be addressed in this paper, as the acquisition of new subscribers is mostly linked to the content proposed and service feature package, more than to the quality of the delivery.

The metrics reflected in the second bullet are the result of multiple factors, technical and nontechnical (typically based on the content offering and service features). However, for the technical side, which is what the platform can offer, the metric that is prominent is the QoE perceived by the end user. The QoE metric itself is a combination of multiple factors, including the video startup time (VST), video start failures (VSF), rebuffering ratio (CIRR), end-to-end latency and the perceived video quality. This metric can be estimated by collecting telemetry directly on the user devices or by deducing from other telemetry collected on the network.

The TCO includes any costs needed to run the service. This can encompass hardware costs and the cost of operation in the case of an appliance-based service or the cost of service invoiced when operated in SaaS mode. In any case, this total cost covers the headend and delivery (e.g., CDNs).

Any evolution of the current workflows should therefore be considered keeping these two goals in mind: improving the QoE and reducing or keeping the TCO under control.

Looking at the whole delivery workflow from the content capture to the end device, there are several areas where the processing needs to be flexible to get the optimum use from resources (i.e., computation resources, bandwidth, storage, etc.). In the video compression space, it has been well known for decades that the codec engine should enable the most appropriate mode depending on the content nature. This is a supported capability in any video codec since fixed QP approaches have been enriched by many other techniques driven by the content nature. Video codecs have dynamically adapted to the content characteristics for years, providing compression improvements. But things are now getting more complex with ABR distribution. Content is now made available in several bitrates, several resolutions and several frame rates. Even if one can anticipate general rules linking these parameters (for example, when the bitrate is reduced, at some point, it becomes more efficient to reduce the resolution), these thresholds are highly dependent on the nature of the content. The same applies for the frame rate. It is commonly agreed on that for sports content a high frame rate will provide a better result at a given bitrate, which is not the case for other types of content.

The delivery side, compared with the "good old simple broadcast era," is also much more complex. Delivery is made via multiple types of networks, all of which have limited resources that should nevertheless cope with the unicast paradigm used in OTT. Moreover, the QoS of the delivery network is highly variable over time and locations (on the open internet QoS is not guaranteed as it is on a cable network, for example). Dealing with these variable parameters on the network side can be achieved by overprovisioning resources. For example, putting more edge caching in the CDN. But this approach carries a significant cost that might impact the profitability of the service. If considering the worst case in a static configuration is not a viable option, then the alternative is to adapt the delivery to the current condition in order to find the sweet spot that will, at any time, give the best compromise and maximize the end-user satisfaction. All this is even more complex when consumption is on mobile networks, meaning there are even less predictable network conditions.

There is a clear motivation for the service provider to maximize the perceived QoE and keep TCO under control in a fast-moving environment. This multi-variable equation cannot be solved efficiently in a rigid, static delivery workflow. Moving from a static to a more dynamic approach can greatly impact the entire delivery workflow.

4. Industry Trends and Research

On the content preparation side, compression technologies have evolved and become more complex. In the past few years a new paradigm called content-aware encoding (CAE) has emerged. CAE embraces different technologies that are highly dependent on the encoder vendor but overall the codec decisions are more driven by an on-the-flow content analysis. More recently, artificial intelligence (AI) and machine learning (ML) technologies were added to cloud-based solutions, making these tools economically viable. Thanks to a big push by industry leaders like Netflix, AI entered the game to propose per-title and then per-scene encoding. Here the video content is not only analyzed in real time to extract the relevant feature but, then, a prediction model is created offline using a large database of content to select the best codec configuration. More details can be found on this topic in the SMPTE 2019 conference paper [9]. All these techniques, from CAE to more advanced AI-based processing, were first used in production for VOD

assets but now are starting to be deployed for live content, with the even greater challenge of matching real-time operation. This is the next step the industry should take, introducing some new dimensions and new flexibilities to create an optimal profile ladder at any time. These new areas will be described in the following sections.

On the delivery side, the situation is a bit less mature, as this part of the global workflow has been a moving target in the past decade. Nevertheless, the global trend on the delivery network side is to transition to a more flexible software-based architecture. On the one side deploying new network elements (software based) according to demand is an option. On the other side, providing some hints to the end-user player so that it can make smarter requests to the network is also a possibility. There has been a significant amount of academic research on building some models of the various delivery networks. Much of the research is focused on the mobile network where data consumption (mainly driven by video content) explodes and will continue to grow in the coming years while QoS is still an issue, especially in crowded areas and at peak hours. Stanford University has done work on network optimization using deep reinforcement learning [10]. On the client side, MIT has developed research around an improved ABR algorithm using reinforcement learning to improve the player behavior in difficult network conditions [11].

As explained in Cassie Tolhurst's blog [12], deep learning algorithms can help secure highly demanding content like UHD delivered at a large scale. We are seeing AI spread more and more across the workflow from content preparation to network delivery to enhance the end-user experience. The different functions, part of these new dynamic workflows, are presented in the following sections.

5. Moving to More Dynamic Workflows

5.1. Introduction

Dynamicity of the workflows must be done in relation to external context, as explained above. Taking a holistic view of the global situation for live video streaming, we came to a conclusion about where it's important to build a new way to distribute video. This analysis leads, therefore, to the creation of adaptive workflows taking into account all possible contextual sets of information:

- **Content characteristics:** Live video is per nature changing over time. Encoding and packaging it with fixed configuration (i.e., bitrate, resolution, frame rate) as done today is not optimal to ensure the best QoE the delivery network can give at any time.
- **Content consumption:** Having the same encoding, same packaging (same profile ladder) for all live channels is sub-optimal. This should be dynamically adjusted over time using feedback from the network (i.e., player, CDN, access network).
- **Content importance:** Premium content with high value attached to it will have to be encoded with a higher quality than less valuable content. This should be dynamically adjusted over time according to the operator's preference.

Moving from the traditional static (set and forget) approach to a workflow where many configurations can change over the time, sometimes with a high dynamicity, is not a simple task. This covers many aspects summarized in Table 1.

Table 1 - OTT evolution from static to dynamic workflows

Static workflow	Dynamic workflow	QoE improvement	Cost improvement
All parameters are fixed	Variable parameters	✓	✓
Fixed resource allocation	Variable resource allocation	✓	✓
Fixed architecture/ maximum TCO	Usage-based architecture / Optimized TCO		✓
Siloed approach	End-to-end approach	✓	✓
Deterministic approach	AI-based approach	✓	✓

As illustrated in Table 1, the different areas moving from a static paradigm to a dynamic workflow may impact either the end-user QoE, the service provider costs or both.

The move to a dynamic approach implies that the decisions made should be driven by various criteria linked to the contextual sets of information mentioned above (i.e., content characteristics, content consumption and content importance).

The various dynamic actions across the delivery chain can be split into two categories:

- Actions and tools aimed at optimizing the production on the profile ladder, either for the purpose of improving the QoE or reducing the operation costs
- Actions or tools aimed at optimizing the delivery path to improve the QoE (mitigation of network congestion during peak audience)

The next sections give a description of the various tools, most being guided by AI, which are part of the Context Adaptive Delivery solution embracing the two categories of actions.

5.2. AI-based Encoding/Content Aware Encoding

One of the ways service providers are battling QoE issues for OTT is through advanced compression methods. Content Aware Encoding (CAE), a per-title encoding technique currently used by Netflix, is one such method that supports both VOD and live applications.

CAE assesses the video complexity in real time and adjusts the encoding parameters to provide the best picture quality. It works similarly to VBR for statistical multiplexing, except that only one program is encoded, and the video quality measurement is more refined since it is based on the Human Visual System (HVS) model. In order to have a more accurate video quality measurement, the CAE live system is trained offline using artificial intelligence technologies. For more details, see Harmonic's technical guide on EyeQ [13].

Over the past few years, CAE has made a real change in video compression and is now backed by Apple, Netflix, and the Ultra HD Forum, which has demonstrated a consistent savings of 40% vs. CBR for UHD ABR using CAE in 2018.

The next step of video compression improvement using AI is what Harmonic calls "Dynamic Encoding Style." We leveraged our first research in AI to embed prediction models in the encoding engines to feed the compression engines with the most appropriate set of parameters. As with many ML-based solutions, we train a prediction model offline using a large database of assets in order to find the best compromise among the huge set of encoding parameters that can be fine-tuned. Then, on the live system, the prediction model uses the video characteristics extracted in real time by the first blocks in the encoding

pipeline and matches it to the "optimum" set of encoding parameters. Dynamic Encoding Style is a natural complement to the CAE approach and shares a lot of common principles, including the video quality assessment technique used to build the prediction models in the Human Visual System Model.

Dynamic Encoding Style uses an AI based two-step approach, depicted in Figure 1.

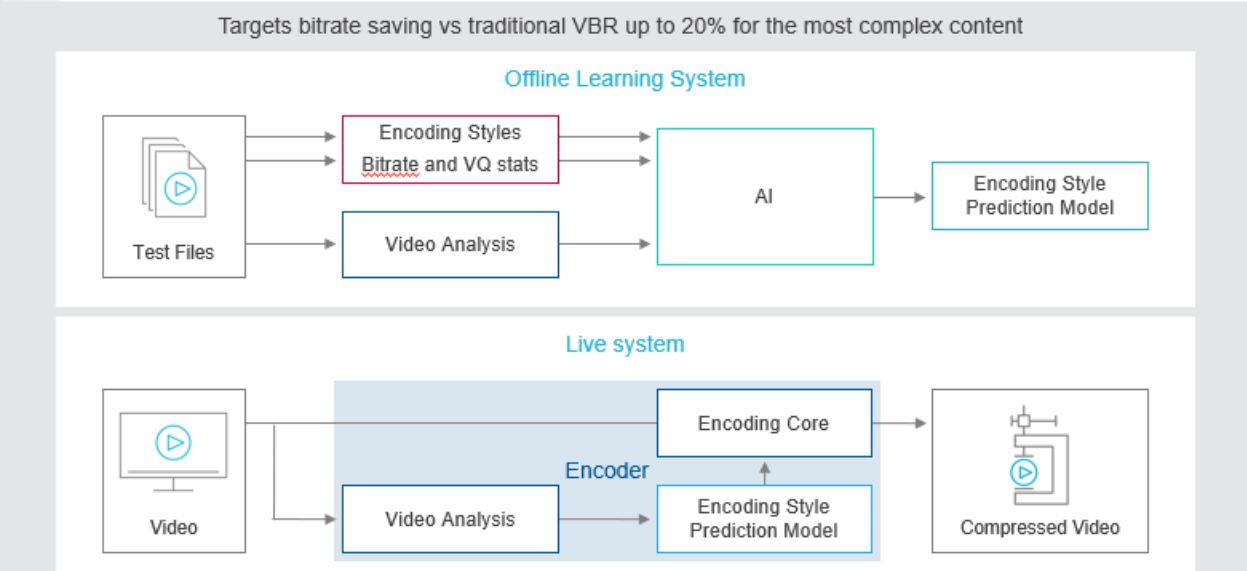

Figure 1 – AI-based Encoding

5.3. Elastic Encoding

Elastic encoding is the last tool mentioned in this paper that focuses on the compression core for live content distribution. The general approach is to use feedback from the network on the content popularity in order to allocate variable CPU resources for the transcoding. The new generation codecs have a very large toolbox that could, if they are all used for every content, have a huge impact on the solution density (number of transcoding instances that can run in parallel on a given cloud resource). All the live encoder vendors are therefore making compromises to find the sweet spot between quality, bitrate and CPU resources. This compromise can be different when the distributed content is very popular. Allocating more CPU cycles will lead to lowering the bitrate at a given video quality which, in turn, will reduce the CDN costs. This is very interesting when the CDN egress is high for popular content.

As it is sometimes difficult to predict which event will be very popular, having a flexible solution that can adapt dynamically when the live event is being distributed is very important.

5.4. Dynamic Resolution Encoding

This tool (and the next one) is different from the previous ones, as the AI is not used to modify the configuration of the compression algorithm but is used to select the optimal resolution of the encoded video. It is well known that there is a link between the representation bitrates and resolutions in an OTT profile ladder. For low bitrate representations, it is more efficient to reduce the content resolution before encoding to get the best QoE. As the threshold to change the resolution at a given bitrate is highly dependent on the content nature and evolves dynamically over time, a solution based on a ML prediction model is a good choice to estimate the best resolution before the encoding processes.

Like the previously mentioned tools, Dynamic Resolution Encoding uses an AI-based two-step approach, as depicted in Figure 2.

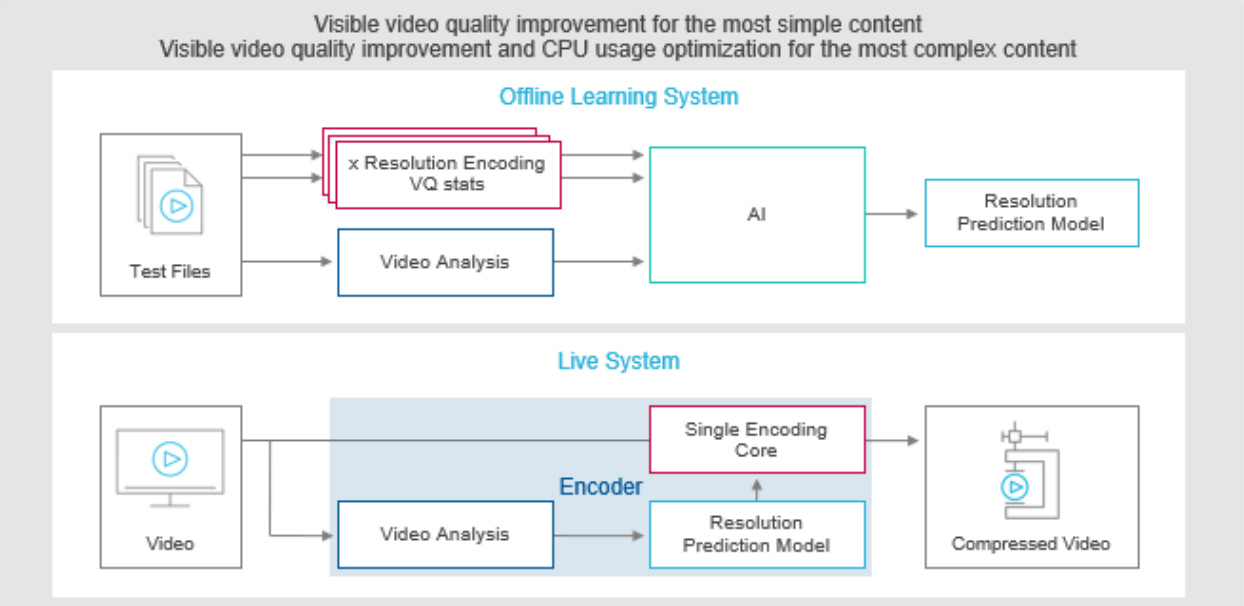

Figure 2 - Dynamic Resolution Encoding

Dynamic Resolution Encoding allows operators to improve the QoE by enhancing the picture quality perceived by the user. It also improves the density of the solution, thus the TCO, as less profiles need to be used for OTT.

5.5. Dynamic Frame Rate Encoding

Dynamic Frame Rate Encoding works on the temporal activity of the content. Relying on known properties in conjunction with what the user can perceive (all of this is described in the HVS model), this tool will determine the optimum frame rate for a given piece of content, making temporal decimation when a full frame rate is not required.

The value brought by this decimation is that the encoding core will not encode all the frames, therefore saving CPU cycles that can be used either to achieve a better bitrate or to improve the network reach (i.e., less stalling, less rebuffering when the content bitrate is lower). Another use for the CPU savings is to get a denser architecture and/or reduce power consumption, and this adds to the TCO for a service. Dynamic Frame Rate Encoding uses an AI-based two-step approach, as depicted in Figure 3.

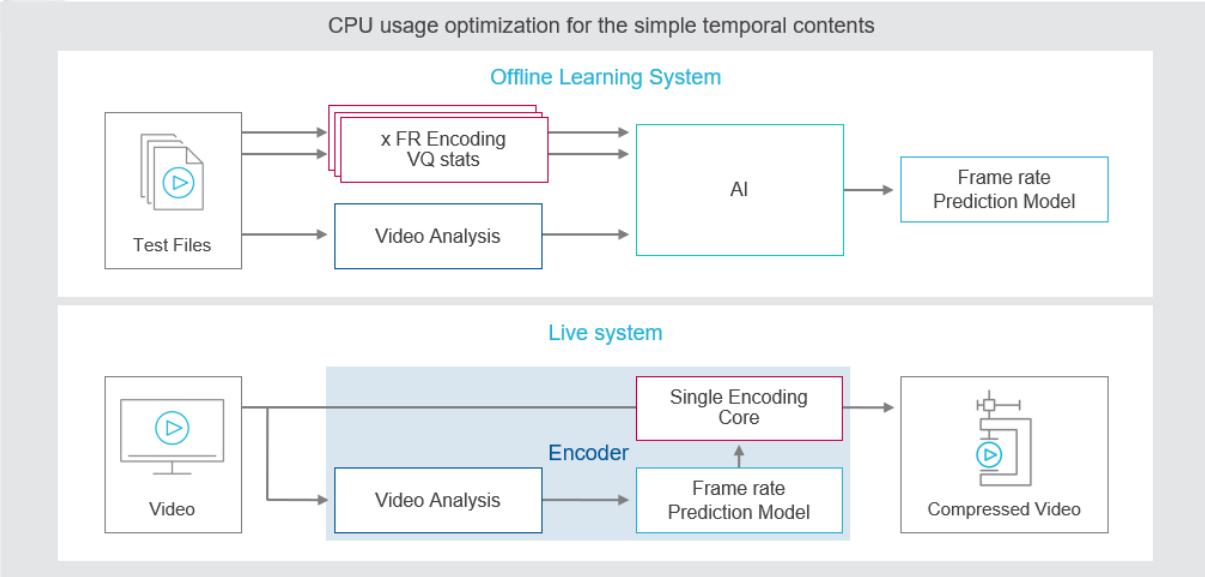

Figure 3 - Dynamic Frame Rate Encoding

5.6. Delivery Optimization

The different tools presented above are aimed at selecting the best encoding configuration depending on the content characteristics. Using these approaches, or a combination, one can create an adaptive content preparation workflow that should optimize the profile ladder based on the content itself or its popularity. But, at this point, another important aspect needs to be considered. The content will be sent over various networks from a core network to the edge and then to the delivery network with a lot of different situations depending on whether the user is on a fixed or radio network, and depending on if it is in a geographic area where this content is very popular compared with other areas (think about a sports match between team A and team B where the audience will be higher in regions A and B compared with the rest of the eligible territory).

With a traditional broadcast paradigm, the service provider delivers one single stream to all the users, making sure that the signal-to-noise ratio will be good enough to ensure a reliable reception on the covered geographic area. On the other hand, a modern OTT distribution platform needs to cope with multiple CDNs, multiple devices, and adapt to much different situations. It seems very ambitious or may be suboptimal to define one single strategy to dynamically adapt the delivery workflows to all these situations. Therefore, flexible architectures will be easily tunable to find the best configuration at a given time.

Whatever the strategy for this delivery optimization is, there are some basics that need to be met on the network. Collecting information from the different elements in the networks is necessary to understand, in real time, how the network is behaving and what the actual QoE is for end users.

A typical analytics collection architecture is illustrated in Figure 4.

In many deployed systems, analytics are collected for the purpose of offline marketing dashboarding but not for real-time usage on a feedback loop.

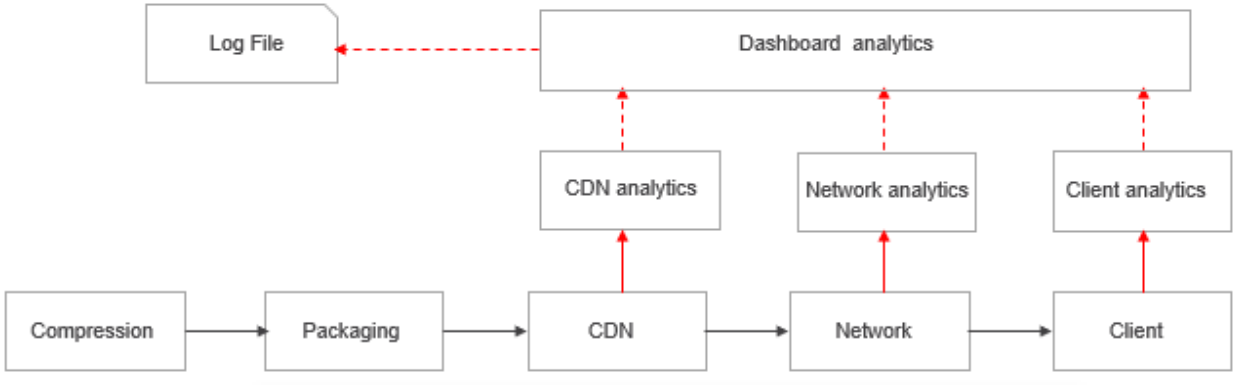

Figure 4 - Typical analytics collection architecture

As presented in the first category of tools that will dynamically adjust the profile ladder, all these decisions may be influenced by the current situation in the delivery network.

The generic architecture for a full Context Adaptive Delivery (CAD) workflow, including the profile ladder optimization and the network path optimization, is illustrated in Figure 5.

Many variants can exist, but the high-level idea is to use the raw data collected on the network to feed a decision engine that will trigger some actions on:

- The compression engine (tools mentioned in the previous sections)
- The packager/origin
- The network path controller

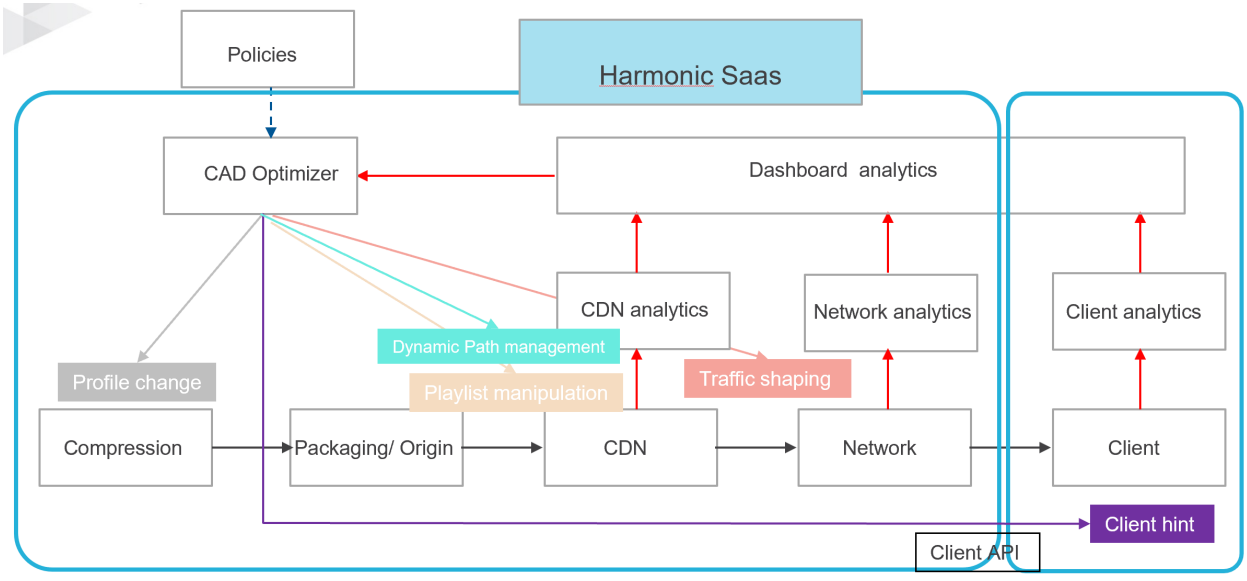

Figure 5 - Context Adaptive Delivery (CAD) generic architecture

The decision engine located in the "CAD optimizer" box above can be as simple as logical switches triggered on fixed thresholds of any of the raw data. But a more forward-looking architecture is to bring AI into this system to enable action before the network crash occurs. Building an accurate and reliable

model of the complete delivery network, from the core network to the last mile, is still a research topic but modern AI approaches look very promising. Because the network conditions are evolving over the time, the AI should adapt automatically based on the current network situation. To achieve this goal, reinforcement learning[14], an area of machine learning, provides some interesting tools. Through a rewarding mechanism, the decision engine gets instantaneous and continuous feedback from the network on the actions taken. It can then adjust the decisions to find, at any time, and under any conditions, the optimum configuration.

As mentioned above, different strategies or scenarios to optimize the delivery at scale for live events can be imagined using the raw information collected in the network. The aggressiveness of the scenario depends on the policy the service provider wants to use to prevent or reduce the breakdown in case of peak audience for a given event.

The global optimization scenario will therefore combine, with a holistic view, the optimization that can be made on the profile ladder and the class of actions to:

- propose a dedicated manifest to some category of player (through manifest manipulation approach),
- dynamically change the delivery path by selecting the most appropriate CDN or delivery nodes

As the second part of the contextual information that the CAD should use, the content consumption is translated into the network load that can be heterogenous, typically based on geographical distribution. To estimate this network load in real time, the system should collect information telemetry from different points in the delivery network. This includes client-side telemetry as well as CDN analytics and network traffic measurements. The collection and capability to perform real-time processing on this information is critical in order to have a timely answer (feedback loop) to any significant change in the content consumption.

Below are some possible scenarios that the CAD optimizer can implement using the network telemetries combined with the service provider policies:

- Based on the reported network load, either global or in some geographic areas, the Customer Management System (CMS) can decide, when a given threshold is reached, whether to prevent any new subscribers from connecting to the service.
- Based on reported network load, either global or in some geographic areas, the traffic can be routed to one CDN or another (when the problem doesn't come from the last mile).
- Based on reported network load, either global or in some geographic areas, the manifest generator can be instructed, when a given threshold is reached, to remove one (the top one) or several high demanding representations in the manifest either to all or a subset of subscribers. This decision can also be influenced by some business rules to give higher privileges to premium customers.
- Some geographic areas can be isolated and treated with a particular scheme if the network indicates that something is wrong in this area.

These scenarios can be seen as a reaction to a given situation but should bring more value if, thanks to an accurate prediction model, the action anticipates and therefore avoids a future crash.

As depicted in Figure 5, the loop-back action can be directed to different elements in the delivery workflow:

- This can be on the encoder where an action can be decided based on content consumption information. This is the elastic encoding tool presented above.
- This can be at CMS level where new subscribers to the service are rejected when network capacity is exceeded.

- This can be on the packager/origin where playlist manipulation can be done to present the best profile ladder to all or a subset of end-user players.
- This can be on the path management system that can dynamically move the delivery from one CDN to another based on reported consumption or risk of overload on the delivery path.
- This can be on the player itself where some instruction or guidance can be delivered in real time to make sure it will request the most appropriate resource (this may include some hints to help the ABR decision algorithm, for example)

In summary, depending on the contextual set of information on the content characteristics, its consumption and its importance and on scenario choices, Table 2 gives an overview of the different actions triggered by the CAD optimizer. This comes together with the content characteristics-only related tools mentioned in the previous sections.

Table 2 - Delivery optimization summary

Tool category	Usage	QoE improvement	Cost improvement
Profile change	Adjustment of the profile ladder based on delivery network status	✓	
Playlist manipulation	Provide different playlist/manifest to groups of users based on network congestion, user category, device groups	✓	
Dynamic path management	Optimize the distribution between several CDNs or private delivery nodes based on reported consumptions and business rules	✓	✓
Traffic shaping	Set business rules to limit the traffic (enhanced zero rating approach)		✓

6. Conclusion

Context Adaptive Delivery is a new paradigm that takes the end-to-end video delivery workflow to the next level in order to address the two most important aspects for an OTT service provider: delivering better QoE to end users while reducing or keeping the TCO under control. Taking into consideration the dynamicity of the content's consumption over a variety of delivery networks is the next step now that OTT technologies are ready for the main screen. Moving from today's rigid configuration to much more adaptive workflows should be seen the same as the transition from hardware- to software-based solutions. With Content Adaptive Encoding, Dynamic Resolution Encoding, and Dynamic Frame Rate Encoding, the content preparation can be much more flexible and adapt to the content type itself as well at its popularity. All these tools should be used to prepare the optimum profile ladder at any time. Then, once the profile ladder is optimized, using network optimization in an adequate scenario allows one to provide the best delivery to users, no matter what their location, device or subscription is. This will quickly create opportunities to improve the user experience, leveraging the value that AI-based processing and cloud-native deployments bring into this landscape.

With more adaptive workflows for the delivery of large-scale live events, the end-user experience can be dramatically improved to reach the expected level and match broadcast services. In addition, service operator profitability can be improved. Even better, we can imagine the introduction of new services and new user experiences that are not possible today.

Abbreviations

ABR	Adaptive Bit Rate
AI	Artificial Intelligence
CAD	Context Adaptive Delivery
CAE	Content Aware Encoding
CDN	Content Delivery Network
cDVR	Cloud Digital Video Recorder
CIRR	Connection Induced Rebuffering Ratio
CMAF	Common Media Application Format
DASH	Dynamic Adaptive Streaming over HTTP
HLS	HTTP Live Streaming
HTTP	HyperText Transport Protocol
MIT	Massachusetts Institute of Technology
ML	Machine Learning
OTT	Over The Top
QoE	Quality of Experience
QoS	Quality of Service
QP	Quantization parameter
SMPTE	Society of Motion Picture and Television Engineers
TCO	Total Cost of Ownership
UHD	Ultra High Definition
VOD	Video on Demand
VSF	Video Start Failure
VST	Video Start Time
WebRTC	Web Real-Time Communication

Bibliography & References

[1] HTTP Live Streaming 2nd Edition draft-pantos-hls-rfc8216bis-07. https://datatracker.ietf.org/doc/draft-pantos-hls-rfc8216bis/

[2] ISO/IEC 23009-1:2019, Information Technology — Dynamic Adaptive Streaming Over HTTP (DASH) — Part 1: Media Presentation Description and Segment Formats (4th Edition), https://www.iso.org/standard/79329.html

[3] T. Fautier, "How OTT Services Can Match the Quality of Broadcast," SMPTE 2019 Annual Technical Conference & Exhibition.

[4] Wikipedia, "HTTP," https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol

[5] WebRTC, https://webrtc.org/

[6] ISO/IEC 23000-19:2020, Information Technology — Multimedia application format (MPEG-A) — Part 19: Common media application format (CMAF) for segmented media, https://www.iso.org/standard/79106.html

[7] "Per-Title Encode Optimization," Netflix, https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2

[8] Y. Reznik, et al, "Optimizing Mass-Scale Multi-Screen Video Delivery," Brightcove, 2019, http://reznik.org/papers/ReznikY_BEITC2019.pdf

[9] J.L. Diascorn, "How AI Technology is Dramatically Improving Video Compression for Broadcast and OTT Content Delivery," SMPTE 2019 Annual Technical Conference & Exhibition.

[10] S. Chinchali, et al, "Cellular Network Traffic Scheduling With Deep Reinforcement Learning," Stanford University, 2018, http://asl.stanford.edu/wp-content/papercite-data/pdf/Chinchali.ea.AAAI18.pdf

[11] H. Mao, R. Netravali, M. Alizadeh, "Neural Adaptive Video Streaming With Pensieve," MIT Computer Science and Artificial Intelligence Laboratory, http://web.mit.edu/pensieve/content/pensieve-sigcomm17.pptx and http://web.mit.edu/pensieve/content/pensieve-sigcomm17.pdf

[12] C. Tolhurst, "Deep Learning Algorithms Could Secure the Future of 4K Streaming," Venture Beat, 2017, https://venturebeat.com/2017/10/26/deep-learning-algorithms-could-secure-the-future-of-4k-streaming/amp/?__twitter_impression=true

[13] Harmonic, "EyeQ Achieving Superior Viewing Experience," https://info.harmonicinc.com/technical-guide/achieving-superior-viewing-experience/?hsLang=en

[14] Wikipedia, "Reinforcement Learning" https://en.wikipedia.org/wiki/Reinforcement_learning

