HOW TRUE, SYNCHRONIZED LIVE OTT CAN CHANGE THE SECOND SCREEN AND SOCIAL TV GAME

Per Lindgren, SVP Live OTT,
Ted Olsson, Product Manager Live OTT

Net Insight, Sweden

ABSTRACT

The shift to multiscreen TV “broke” the social and interactive elements of television viewing. It was thought that through companion and messaging apps the social aspect of TV viewing could be revived. Unfortunately, the delay in delivering Over The Top (OTT) content to different devices and lack of synchronization between the primary screen and secondary screens quickly became an issue.

The second screen has become a frustration rather than an enhanced social experience. For example, second screen users being informed of who scored a touchdown in a football game and how before it happens on their screen means viewers of live OTT events are forced to log out of social media and messaging platforms. They risk hearing about what is unfolding on someone else’s screen before it happens on theirs. This is only part of the problem. There is also little opportunity for real-time social messaging, viewer engagement and shared experiences when audiences are watching the same content on different screens and devices with a time delay ranging from tens of seconds to several minutes.

To bring the social element back to live TV viewing, OTT’s issue of synchronizing content delivery across all devices and harmonizing this with live linear broadcasts needs to be addressed.

This paper outlines the technical challenges in distributing true live OTT over today’s Content Delivery Network (CDN) platforms, and why their limitations in streaming live content breaks real-time social interactivity.

It further describes a software-based OTT distribution solution that is optimized for distribution of live content with low and synchronized OTT delivery, and further outlines the technical differences with today’s HTTP-based streaming solutions.

The last section of the paper provides examples of real-time social and audience engagement, and what this means for the entire media ecosystem.

INTRODUCTION

CDNs and OTT video distribution platforms today use technologies such as HTTP Live Streaming (HLS) and MPEG-DASH, which use segmentation of the video streams and HTTP for delivery. This provides a reliable, personalized delivery of the video streams to

consumers, optimized for on-demand content and catch-up TV. However, it also includes inherent caching and buffering of segments that in typical video OTT implementations result in a delay of around 30 to 40 seconds higher than linear TV broadcasts. Additionally, due to the use of TCP in HTTP and the way most players in the market are built, additional delays are accumulated during the playout of the video feeds. In long-tail content, such as live sports, this can result in a difference in delay of several minutes between devices, while HLS allows up to 15 minutes to be buffered.

This is not a major issue for video-on-demand (VoD) and catch-up TV, but for live content low delay and synchronized delivery is key to ensuring live TV experiences are not ruined, and social and interactive TV entertainment is maintained at the highest possible level.

Today, live and linear OTT is mainly used as a substitute when viewers are unable to watch live content on their largest screens and over standard broadcast TV. Contrary to this, research shows that a clear majority are willing to, and want to, watch major live events and sports content across OTT platforms, but ultimately actual live OTT viewing is still limited to a single-digit percentage compared with those watching live events on primary screens.

The OTT screen, with its strong computing power and graphics and its Internet connectivity, holds great promise in actually enriching and enhancing end-user experiences. However, because of the unsynchronized delay between TV screens, the OTT second screen promise has diminished to the second screen only holding data, and potentially graphics, but with no video content. This has, in general, resulted in a limited uptake for live OTT video distribution. An optimized true live OTT distribution in sync with regular broadcast television will change this by actually bringing a new, enhanced viewing experience that complements and harmonizes with the regular TV broadcasts, changing the way the second screen is used and how content is experienced.

In the coming sections we will look at the specific requirements for live OTT distribution and examine some use case examples in detail.

REQUIREMENTS FOR TRUE LIVE OTT DISTRIBUTION

The main areas that need to be focused on and addressed for optimized true live OTT distribution compared with regular CDNs are:

Low delay: For certain applications, such as stadium or on-site event solutions, a very low delay solution is of interest, typically within 1-2 seconds. This is difficult to achieve, but nevertheless possible with single bitrate, low-latency encoding and an optimized, more local, OTT delivery, preferably with an underlying multicast infrastructure. Betting is another area where very fast delivery is required, especially for so called in-game betting. A high delay will shorten betting windows, which means lost revenues.

There are several initiatives to reduce the delay of today's CDNs, however one of the most significant differences come when OTT delivery delay can be reduced to a point where it is on par with the delays in regular live TV broadcasts, i.e., harmonizing the delay of live OTT to live linear broadcast TV distribution. Such an OTT delivery thus offers a true broadcast TV experience over the Internet.

For IPTV/CATV operators it opens up the opportunity for harmonizing and unifying IPTV/CATV and OTT delivery. The upshot of this is that operators can now fully harmonize television and lower the delay to the same as today's standard TV broadcasts. This opens

up a huge opportunity to provide a new type of TV experience for live OTT audiences changing viewing behaviour. It also opens up the potential to engage with viewers in new and creative ways, while enabling them to create an environment for more interactive and social TV experiences.

For this to be possible, delay needs to be whittled down to around 8-9 seconds, including the contribution and encoding/transcoding aspects of video delivery. This poses a huge challenge when you consider the delay limitations inherent in standard CDN technology [1]. However, with a true live OTT solution in place (described later in this section), it is possible to provide national live OTT delivery at 5 to 6 seconds total delay. This has massive implications that will change the face of live television viewing, as we know it.

Synchronization: When watching live content, not only is delay important, but also that the delivery is synchronized to all viewers. Another problem with today's CDN technologies is that the delay in delivering live content to different screens increases over time. This is due to extra buffering primarily at the end device to cope with packet loss. In a VoD or catch-up TV context, this is not an issue, but for live content this increases the delay towards the first screen and hinders social interaction between OTT screens. With the significant increase in social interaction that takes place around large events, as mentioned above, this can ruin live TV experiences. At Super Bowl 49, for example, there were 265 million Facebook mentions [2] during the game and, in total, more than 500 million social mentions.

For live OTT, a low delay in combination with synchronized content delivery is key. Furthermore, a fixed delay to all consumers paves the way for new applications to be developed, aimed at social interaction and viewer interactivity such as voting, polling, as well as true live interactive gaming and betting.

Fast Channel Swapping: Channel swap time is another issue when watching live and linear content in commercial play services. It can take anywhere between 2 and 9 seconds to change between linear content. This does not bode well for a good end user experience when swapping between linear channels or when flipping through, for instance, player and driver cams. To experience a true live TV sensation, the channel swap times should be at least the same as in any other digital TV platform. Moreover, DRM solutions can further increase channel swap times, and also need to be optimized for live and linear content.

Scalability: Whether it's a large sporting event or the season finale of a popular linear television series, live events tend to attract large number of viewers. This means live OTT solutions need to scale to tens of millions of viewers. In addition, since all viewers are synchronized watching the same content at the same time, the systems also need be be able to cope with a massive amount of simultaneous transactions, such as channel swaps and back-end integrations, including the authentication and billing systems. The solutions must therefore be optimized for high performance and with minimum overhead. Today's CDN technologies typically show very high overhead when using low segment sizes.

Live Media Eco System: Increased use of live and linear OTT viewing will also affect other parts of the media ecosystem. This will not only include the back-end, such as DRM, authentication and billing systems, but also the ad model and value chain. Live content already represents some of the most lucrative ad revenues. A synchronized second (enhanced) screen with video content opens up a separate ad channel, either as a

separate personalized ad channel or a channel that can be used to enhance ads shown on the regular TV screen, using the interactive and online capabilities it inherently carries.

SOLUTION DESCRIPTION – TRUE LIVE OTT DELIVERY

This section describes the first commercial live OTT solution that provides all of the above requirements for true live OTT delivery. The presented solution, called Sye, is a piece of virtualized software that can be deployed over private, public or mixed clouds providing a low-delay virtualized OTT video delivery (CDN) solution over Internet with synchronized delivery.

There are several initiatives looking at low-delay OTT delivery e.g., within MPEG-DASH and Media MPEG Transport (MMT) groups, but these are typically not able to reach the reduction in delay required to harmonize with regular TV, which in turn reduces their value.

Technology Comparison

All CDN streaming technologies today, i.e. MPEG-DASH, HLS and Smooth Streaming, create small files, or segments, out of a stream, which is then distributed through cache servers using Hypertext Transfer Protocol (HTTP). HTTP is a TCP (Transmission Control Protocol) based protocol, a protocol typically used for best effort file transfer using a conservative retransmission scheme to recover packet loss and ensure reliable delivery. TCP scales well for on-demand types of applications, but is not optimized for a premium live streaming user experience. Instead most real-time IP video applications like IPTV, Skype, Lync and FaceTime use User Datagram Protocol (UDP). UDP in contrast to TCP does not have a built-in retransmission mechanism.

Today's CDNs thus treat the segments as small file transfers using TCP-based store and forward mechanisms to deliver the content to the receiving device. The duration of each segment is typically 2 to 10 seconds.

Additionally, if the receiver does not receive or cannot recover enough packets, it uses an Adaptive Bitrate (ABR) mechanism instead to ask for a lesser bitrate, i.e., lower resolution.

Due to the inherent buffering mechanisms and TCP's inefficiency, Sye instead uses UDP as the basic transport protocol, but with enhanced selective retransmission capabilities, an optimized ABR scheme for streaming, and a synchronization protocol for synchronized delivery. The more optimized selective retransmission protocol carries lower overheads compared to TCP retransmission and much higher packet recovery probability. This directly translates to a better viewing experience by providing and maintaining higher profile viewing for a longer period of time, compared to legacy HTTP streaming, before reverting to a lower resolution bitrate.

In addition, Sye does not segment the stream into small file segments since this further increases delay. This means it uses the same encoders and transcoders as used in current CDNs, but ingests the ABR profile streams as IP MPEG streams before it enters into the segmenter/packager in normal CDN origins.

It uses Egress servers that distribute the content to end devices, similar to today's edge caches, but without adding any extra cache buffers. Instead it uses a small fixed-size buffer at the receiving device end, which is also used as a retransmission buffer.

Figure 1 shows the difference in buffering mechanisms between the two protocol types. HLS and other HTTP-based streaming technologies typically buffer three segments at the receiver. However, the delay is not fixed to 3x[segment time], but will accumulate over time due to retransmissions and extra buffering within the players. HLS allows up to 15 minutes of buffering at the receiver.

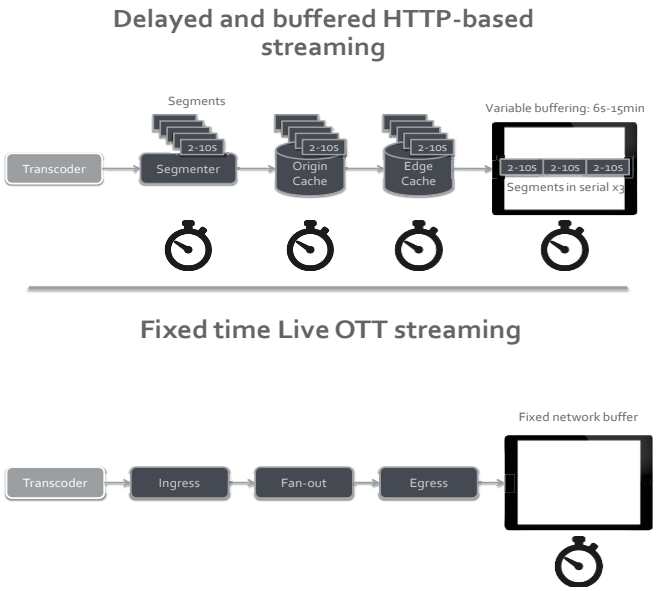

Figure 1 – Comparison between HTTP-based streaming and optimized live OTT streaming

Sye true live OTT uses robust UDP transport as the first defense toward packet loss, and ABR enforcement as the second defense.

For a typical HTTP based streaming solution, “bandwidth training” kicks in to probe if the next level of quality could be requested. It proceeds in this fashion until the highest quality is achieved. This takes an unnecessary amount of time and typically provides the end user with a significantly degraded viewing experience.

While today’s CDN platforms and ABR implementations are fully “stateless” and client-centric, Sye in contrast uses “stateful” ABR and a combined client and server centric approach. It is continuously aware of the *Client* network environment, i.e. available bandwidth, round trip delay, packet loss, connection type and more. During normal operation this is primarily used to optimize end-user experience during changes, such as channel swap changes, fast restoring of higher ABR levels and server load balancing. However, during a period of congestion, more intelligent network and traffic engineering decisions can be performed. Additionally, the constant monitoring of network and user behaviour can be used for both network performance analytics, as well as end user analytics tools.

Channel change is another area where the “stateful” ABR solution delivers a better end-user experience by using the knowledge of the current ABR level when changing to a new channel. The server-side, network-aware function comes into play for the channel change case and only reduces ABR level one step and can quickly restore the highest quality possible, given the available bandwidth. “Stateless” ABR, as used in today’s CDNs, begins with the lowest profile and works its way up. Typical profile change times range from 2 to

10 seconds and can thus take along time until full quality is re-obtained, despite there being enough capacity in the network.

The difference between “stateful” and “stateless” ABR is outlined in Figure 2, below.

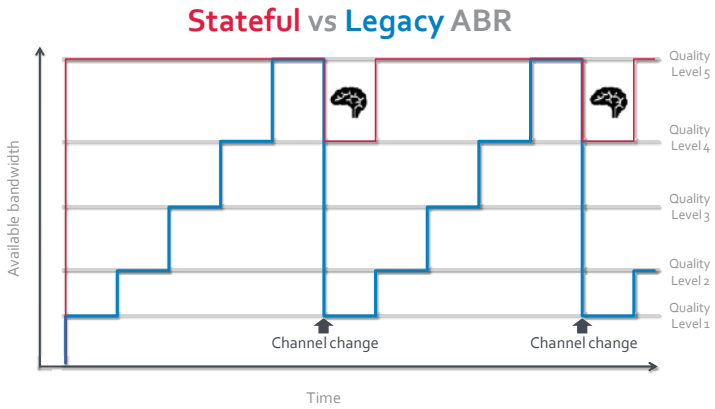

Figure 2 – Difference between “stateful” and “stateless” ABR

Another crucial aspect of live content delivery is that larger live events are typically mission critical and have a huge audience base. Because of this, it is essential to continuously monitor the performance of both network and video quality, in case of network or equipment failure, and to be able to analyze where the problem lies and quickly recover the service. In normal CDNs performance monitoring and metrics gathering is performed separately and in parallel to the streaming service and often requires a separate SDK to be implemented at the end-user devices.

Sye offers built-in monitoring and metrics, including video quality, round trip time, packet loss, and video profile.

Architecture of Sye True Live OTT Solution

The back-end server components are split up into two different functions: *Data Plane* and *Control Plane* as shown in Figure 3.

The *Data Plane* functions are *Ingress*, *Fan-Out* and *Egress*. These functions are responsible for distributing streaming media as an overlay network on top of any type of underlying network infrastructure. The Data Plane functions are described in some more detail below.

The *Control Plane* functions consist of *Controller*, *Front-End* and *Front-End Balancer*. These functions are responsible for *Client* resource requests, load balancing of data and media, configuration, provisioning, alarming, monitoring and metrics. The system is a pure software solution. It is hyper scalable and therefore deployable on bare metal, a fully virtualized environment or in a hybrid physical/virtual approach.

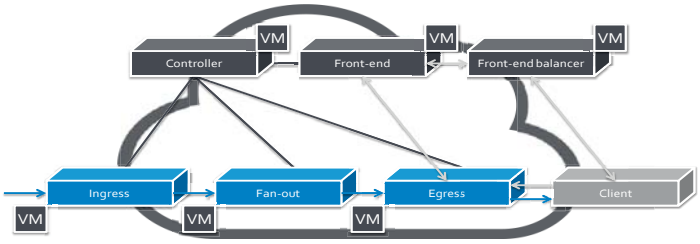

Figure 3 – Virtualized CDN architecture for Live OTT delivery

Data Plane

Ingress

The main functionality of the *Ingress* function is to process and prepare the incoming streams for live distribution where every packet is enriched with synchronization and latency information.

The *Ingress* function is the entry point for media into Sye's true live OTT system. The *Ingress* ingests unencrypted transport stream (TS) based linear content over UDP. Multiple profiles per content are used to provide the ABR functionality.

Fan-Out

The main functionality of the *Fan-Out* function is to save bandwidth in the core distribution.

The *Fan-Out* function is an intermediate geographical scaling function when using unicast in the core distribution. It provides point-to-multipoint replication in the live distribution network for geographical distribution spread.

Egress

The *Egress* function is the *Client* facing streaming component responsible for delivering robust and error free linear media to the *Clients*, with fixed delay and frame synchronization. The *Egress* function also handles adaptive profile change and instant channel change. This is based on the per-individual continuously available *Client* network bandwidth. It communicates with a client SDK integrated into existing apps.

IMPROVED LIVE TV EXPERIENCE

With a low-delay OTT solution where all viewers see content at the same time, the current issues of getting information about events and results before seen on one's own screen is solved and it opens up for a more social and interactive viewing experience. People can watch content on different places and socially interact with one another. You can invite your friends with a link to watch the same content together, in sync, and socially interact about the content. It also enables direct interactivity within the programs whether that is voting, participating real-time in games or use in-game betting.

However, it will also change how people watch and experience live content. Today, linear OTT typically distributes and show content in the same way as regular TV. However, an optimized true live OTT distribution that streams in sync with regular broadcast television will change this by actually bringing a new, enhanced viewing experience that

complements and harmonizes with the regular TV broadcasts, changing the way the second screen is used and how content is experienced.

New Enhanced TV Experience Combining First and Second Screens

An optimized true live OTT distribution in sync with regular broadcast television will enable a new, enhanced viewing experience that complements and harmonizes with the regular TV broadcasts, changing the way the second screen is used and how content is experienced. The first live example where Sye was used for this was a trial with Formula 1 and TATA Communications [3], where TATA used Sye to distribute driver cams as individual OTT streams. The fixed delay of the Sye OTT streams was set to harmonize with the normal SKY broadcast in the UK. This provided a game-like experience where you could view the drivers getting into the first curve in perfect synchrony with what was seen on the big regular TV screen. The fast channel swapping allowed to quickly swipe through the camera feeds to select your favourite driver.

In many sports, there is a lot of parallel action occurring in parallel during an event, which regular broadcast television does not follow. In an OTT environment this can be produced and easily distributed so viewers can quickly change between feeds to always follow the most relevant content. Such feeds will not replace existing broadcast, but instead complement and enhance the experience for the viewer, making it a richer and more interactive experience. This signals profound changes to the uptake of live OTT, live content production and the whole OTT value chain.

SUMMARY

This paper has described a new CDN streaming architecture called Sye, which is optimized for Live OTT streaming. Where current CDN technologies offer end-to-end live OTT delays of 40-60 seconds, Sye provides sub 10 seconds delays, including encoding/transcoding delays. In addition, Sye uses an in-band synchronization mechanism to ensure synchronous play-out to all devices. This offers a broadcast TV experience over the Internet that can be harmonized with regular TV broadcast.

For OTT audiences, the harmonization of all screens opens up massive potential for real-time social interaction, as well as interactive entertainment opportunities that current CDN's and OTT platforms cannot deliver. And, for broadcasters, content owners and advertisers, true live OTT enables new and exciting ways to engage with television audiences. It revitalizes existing revenue models and helps to create completely new ones, while at the same time enhancing user experiences, allowing innovative multi-screen content to be produced and consumed in new ways. An optimized live OTT solution in sync with regular TV broadcast will change the mobile device from a TV substitute, or a data-oriented second screen, to be part of a seamless, enhanced viewing experience that includes complementary video, data and interactive tools.

REFERENCES

[1] M. Dbrowski, R. Kolodynski, W. Zielinski, "Analysis of video delay in Internet TV service over adaptive HTTP streaming", federated conference on Computer Science abd Insformation Systems, DOI: 10.15439/2015F91
[2] http://www.cnbc.com/2015/02/02/super-bowl-xlix-and-social-media-most-tweeted-nfl-game-ever.html
[3] http://advanced-television.com/2015/11/26/tata-powers-live-ott-f1/