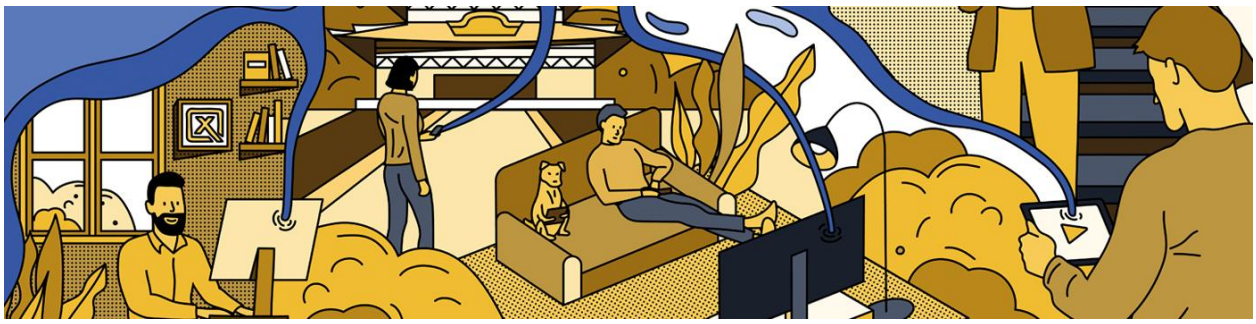

The Truth About Low Latency in OTT

Jérôme Viéron - September 24th, 2019

60 seconds … It took me 60 seconds to understand what happened, to figure out that the France team just scored against Croatia during the 2018 World Cup Final. 60 seconds between the clamor from my neighborhood watching TV and me watching the goal on my laptop via my favorite OTT streaming platform.

Believe me, it was a never-ending wait!

Why such a difference between OTT live streaming and broadcast TV services?

With the upcoming advent of OTT streaming as the de-facto standard for watching live events, being as much as 60 seconds behind the real action is not acceptable. Customers ask for broadcast TV like quality of experience and thus, do not want to be spoiled by viewers on other services. Before going into the possible solutions, let’s see where this latency comes from.

Latency? What are we talking about?

The latency represents the time it takes for “something” to propagate in a given system. In the context of HTTP Adaptive Streaming (HAS), the system and thus the definition of latency differs from people to people. At least two main definitions are commonly assumed:

- **Glass-to-glass latency** (or end-to-end): which represents the time it takes for the video to go from the glass of the camera to the glass of the viewer’s screen,
- **Encoder-to-screen latency:** which represents the latency between the input of the OTT video encoder and the viewer’s screen.

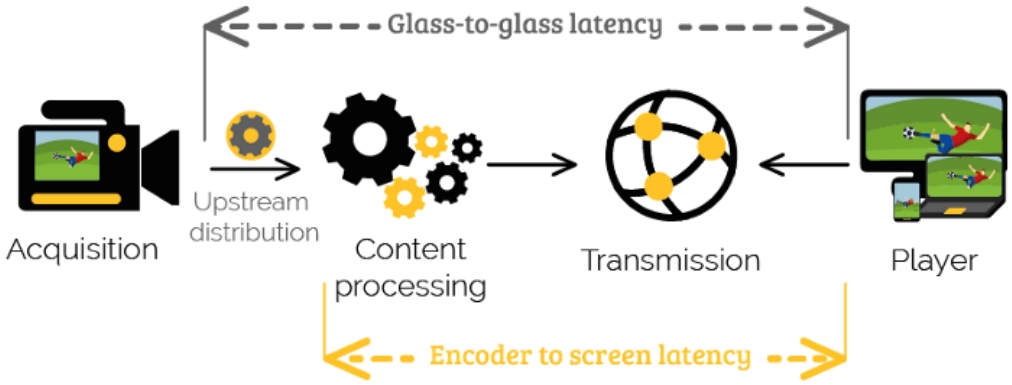

Figure 1: OTT live delivery chain

A live OTT delivery chain consists of multiple steps that all have an impact on the overall latency:

- **Acquisition** (i.e. Capture, Production, Post-production),
- **Upstream Distribution** (i.e. from the Production to the OTT Head-end)
- **Content Processing** (i.e. Transcoding, Encryption, Packaging, Ad-decision, Origin),
- **Player** (i.e. ABR logic, Decryption, Decoding, Ad-insertion, Rendering).

Do we need Low Latency? Ultra Low Latency ?

Traditional broadcast latencies are in the range of 3–12 seconds, whether we talk about Satellite (DTH), IPTV or Cable TV. A typical latency for live OTT Streaming today is in the 30–45 seconds range.

What is the real latency target? My belief is that for the vast majority of the live events (e.g. Sports, Concerts, Breaking news) reaching the latency (or slightly better) of traditional broadcast TV is the real need. We are talking here about Low Latency (i.e. 1–5 seconds) which is achievable now. Targeting Ultra Low Latency (i.e. sub-second) with HAS in a short-term frame is not realistic and may even be useless.

Every step in a live OTT delivery chain introduces some latency. In order to close the gap, all the components of the chain will have to be tuned.

Where does the latency come from? Encoders?

For years, I heard people complaining about the video encoder/transcoder latency. The encoder may be the ideal culprit, but it would be naïve to think it's the only lever.

Assuming a common workflow between broadcast and OTT, the first added latency is due to the fact that the Upstream Distribution includes one or several encoding stages to deliver the mezzanine feed. Those stages, based on traditional video encoder/transcoder add around 3–5 seconds latency per stage. Such latency is not negligible and has to be considered in the whole OTT Latency problem.

Once in the headend, the video transcoder in charge of producing the multiple ABR representations in the desired codec (e.g. H.264 or HEVC) adds another 3–5 seconds depending on the targeted tradeoff video quality/bitrate. Indeed, increasing for instance the look-ahead, the number of B-frames and/or the GOP (Group of Pictures) size will increase the video quality at a given bitrate, but comes with a penalty overhead.

It is clear at this point that the intricate relationship between latency and quality will be key in the latency reduction quest.

Then come the segments!

In the rest of the OTT chain, the live content is divided into a series of contiguous files called segments that will be packaged, encrypted, sent and downloaded.

With HLS (Apple — HTTP Live Streaming) and DASH (MPEG — Dynamic Adaptive Streaming over HTTP), traditional packager/origin adds at least one segment duration, so 2–6 seconds additional latency to the chain. Indeed, Apple still requires a 6 seconds segment duration for an HLS compliant stream, and the packager/origin needs to wait for a full segment to be available before starting its job.

It is also not uncommon to see buffer-based poor implementations introducing additional latency for the ad-decision and DRM/encryption process.

Regarding the transmission, going through a CDN can take up to a few seconds for a cold cache (i.e. caches are empty and segment has to be sourced from the customer origin), down to a few milliseconds in case of cache-hits. The CDN latency is estimated as a function of the video segment duration. Typically, Akamai provides the number of half the duration of a segment.

The player is the largest source of latency

For historical reasons, when Apple introduced the HLS format, they recommended a three-segment initial buffer in the ABR player, which became then the de facto standard for the industry. Such buffer has been introduced in order to deal with network drop outs and thus, to reduce the chances of a stall or re-buffering. But the storage of HLS and DASH segments, prior to being decrypted and decoded, is also used by the ABR logic to manage the selection of the appropriate representation. Of course, it is possible to operate with a one-segment storage but with higher risk of not being able to hide network issues.

When adding the segment Decryption latency (theoretically negligible but usually constrained by slow key retrieval with DRM Servers) and the MPEG video Decoder latency of a few seconds, the overall latency for the player is typically around 20 seconds. In fine, the end-to-end latency for most of the live OTT deployments can be computed as a function of the segment duration D as follows.

	Upstream	Encoding Packaging	CDN (s)	Player	Total
D	1*D	D+5	0,5*D	3,5*D	
6s	6	11	4	21	41s
2s	2	7	1	6	15s

Reducing the overall streaming latency is not about taking a new standard and hope it makes magic. It's about taking best in breed components all along the chain, combined with the latest standards and innovative architectures such as our Just In Time Workflow. Stay tuned for the next article!