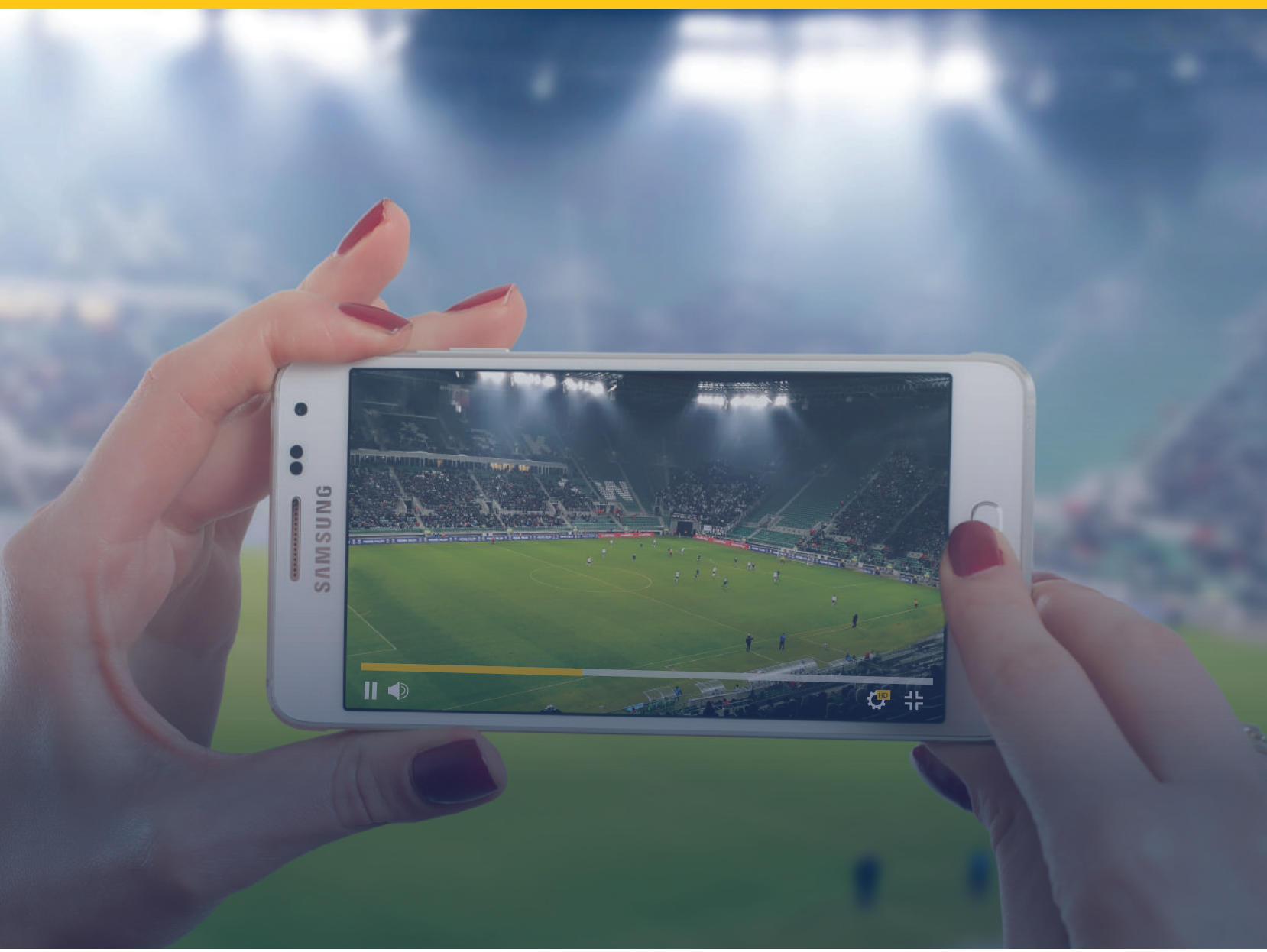
# Multicast ABR, Low-Latency CMAF, CTE, and Optimized Video Playback

## Reducing Latency in Live Online Video

Jonathan Tombes

A whitepaper powered by THEO Technologies & Broadpeak




THEO

broadpeak

# Multicast ABR, Low-Latency CMAF, CTE, and Optimized Video Playback

## Reducing Latency in Live Online Video

Jonathan Tombes


### Abstract

Demand for live streaming video is growing, along with consumer aversion to high latency. To enable the delivery of HTTP-based video within nearly 3 seconds, below that of standard IPTV, service providers can look to a proven set of technologies that include multicast ABR, the Common Media Application Format (CMAF) in its low-latency mode, HTTP 1.1's Chunked Transfer Encoding (CTE) mechanism, and video players that have been optimized for low latency. This solution involves tradeoffs and limits, yet fits predominant use cases in the live streaming video market.

## Table of Contents



## Introduction

The rise of live online video has made latency a hot topic in the streaming video world. Defined as the interval between live video being captured on a camera and displayed on a screen, latency results from various factors, depending upon video delivery scheme.

Real-time communication technologies, not surprisingly, deliver the lowest levels of latency. Among TV services, IPTV often has the least delay, with video arriving within about 4 seconds. Typical broadcast latency in the U.S. is slightly higher. For live online video, standard implementations of Adaptive Bit Rate (ABR) streaming, also known as HTTP Adaptive Streaming (HAS), can range between 30 and 45 seconds.

One reason latency has become such an issue today is that consumers are tapping into different technologies at once. As a result, they can compare performance. In a commonly invoked scenario, consumers are using an over-the-top (OTT) live streaming service to watch an athletic event while communicating on social media with spectators who are either at the game or watching it on a screen with much less delay. Being half a minute or more behind the live action is more than annoying; it leads viewers to question the value of their video service.

While high latency can lead subscribers to churn, low latency can boost usage. Global research presented by Limelight Networks indicates that 65 percent of respondents aged 26-45 would stream more sports if events were not delayed beyond the broadcast.[1]

Latency is not the only challenge facing streaming video. Buffering, start-up time and synchronization issues also impact quality of experience (QoE). Given how easy it is to lose online viewers, streaming video service providers should take these impairments seriously. The good news on latency is that a combination of existing technologies can enable live ABR streaming or HAS to perform extremely well, even better than IPTV.

Enabling live ABR to arrive in less than 4 seconds involves several steps. A service provider needs to replace unicast with multicast transmission; implement the low-latency (LL) or chunked flavor of the Common Media Application Format (CMAF); leverage Chunked Transfer Encoding (CTE); and optimize video playback. To see how this adds up, let's first look at the sources of latency.

## IPTV vs. Unicast ABR

Standard IPTV and unicast ABR video delivery differ in their components and latency budget. In a managed IPTV network that uses the MPEG Transport Stream (TS) digital container format, encoded video traverses the network and reaches the player, where it is decoded and buffered for playback and fast channel change (FCC). According to data and test measurements compiled by video delivery

[1] The State of Online Video 2018, Limelight Networks.

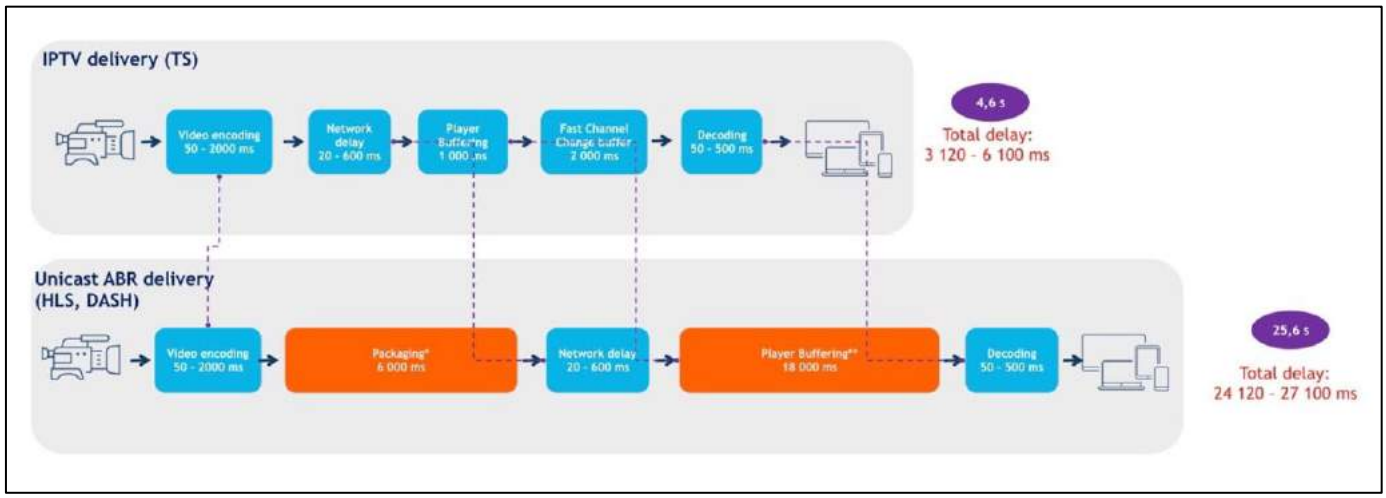



**Figure 1 – Latency Sources in Video Delivery Systems: IPTV vs. Unicast ABR**

solutions provider Broadpeak,[2] total latency for such systems averages 4.6 seconds. (See Figure 1.)

By contrast, in unicast ABR delivery, video is encoded and then packaged in HTTP Live Streaming (HLS) or Dynamic Adaptive Streaming over HTTP (DASH) formats. These packaged segments travel across the open internet and reach the video player, which decodes and adds them to a buffer.

Let's compare timing. The standard segment length recommended by Apple is 6 seconds (down from 10 previously). Buffering usually involves three segments, one of which is being encoded and packaged at any given time. Then there are a few seconds of static overhead. The result is that the average delay for HTTP-based adaptive streaming is more than five times as long as IPTV, or 25.6 seconds, as seen in Figure 1. (Longer segments will increase latency at the packaging and player levels.)

Packaging is the new element in unicast ABR delivery of HLS and DASH, adding 6 seconds to the timeline. While the lack of FCC buffering saves 2 seconds, player buffering increases from 1 to 18 seconds in ABR, accounting for 70 percent of total latency in this scenario.

## Reducing Latency

Packaging and buffering (the two large orange blocks in Figure 1) are obvious targets for latency reduction. Players are the first place one might want to start.

"It is crucial you upgrade your player to support low latency streaming, as more than 50 percent of latency is often due to buffers within the player." said Pieter-Jan Speelmans, CTO of THEO Technologies, developer of the THEOplayer video player. "However, low latency is an end-to-end story. If one component is not configured properly, the advantages won't be as big as they could be."

There is more to say about players and buffering. As for packaging, CMAF's low latency mode combined with CTE also delivers results, as we will describe below. But a root cause of high latency is the best-effort HTTP traffic of unmanaged networks.

### Multicast ABR

In unmanaged networks, absent sufficient buffering, video playout is consistently interrupted for re-buffering. This is where multicast ABR can change the picture: it

[2] Full disclosure: Broadpeak and THEO Technologies supported the production of this paper.

transforms a series of irregular, unicast bandwidth peaks into a relatively jitter-free, smoothed and prioritized traffic flow requiring no more buffering in the player than traditional IPTV.

Pioneered by Broadpeak, multicast ABR requires a transcasting device in a headend or central office to convert unicast into multicast, and a multicast-to-unicast agent embedded within a home gateway or set-top box. A protocol built on top of and adapted to simplify the NACK-Oriented Reliable Multicast (NORM) standard drives these conversions.

In contrast with a unicast best-effort environment, this solution leverages two transport-layer technologies: the more persistent but loosely managed User Data Protocol (UDP) and the more tightly controlled Transport Control Protocol (TCP). "Multicast ABR also makes use of UDP but in a managed environment where there is no competition between UDP and TCP, and that ensures that there are no issues, for instance, with firewalls," said Nivedita Nouvel, Broadpeak VP Marketing.

The performance boost is substantial. Enabling 2-second segments and 1-second buffering, multicast ABR can reduce delay from 26 to 7 seconds, a 73 percent gain. The potential bandwidth savings are also a big draw. Instead of massive numbers of subscribers making individual unicast requests to origin servers, the embedded agent makes a request to join a multicast stream that can reach a host of endpoints.

Along with less buffering and bandwidth, multicast ABR can boost QoE because the home network, not the more contentious service provider network, determines which bit rate to use.

Multicast ABR has drawn the attention of industry organizations. In 2016, a year after Broadpeak launched its own solution, CableLabs issued a related Technical Report; and DVB released a reference architecture in 2018.[3] Transcasting and multicast technology is well out of the lab. According to analyst firm Rethink Technology Research, the market is poised for growth. Reporting that Comcast and two major French operators have been using multicast ABR "in stealth," the firm predicts revenues of $852 million by 2023.[4]

## CMAF and Low Latency

Even with a greatly reduced latency of about 7 seconds, multicast ABR combined with a small-buffer player is still 3 seconds over what traditional IPTV can deliver. The application of chunked CMAF can reduce that amount by half. Let's first review CMAF, then its low-latency mode.

The main driver behind CMAF was efficiency. In place of two separate media formats, the CMAF initiative achieved a significant, if incomplete, unification of DASH and HLS. Included in the specification is usage of the standard fragmented MPEG 4 (MP4) container; usage of Common Encryption (CENC); support of AAC, AVC and HEVC codecs; and more. What it does not specify is which of two manifest formats or block cipher modes to use.

On the plus side regarding manifests, both .mpd (DASH) and .m3u8 (HLS) are compliant with server-side ad insertion (SSAI) systems, thus safeguarding existing monetization models. As for counter mode (CM) and cipher block chaining (CBC) encryption, DRMs may align on their own. PlayReady and Widevine, for instance, are beginning to support CBC (in addition to CM) on more platforms.

What CMAF did unify is significant. The common format delivers a twofold increase in CDN efficiency and similar reduction in packaging/production costs. Of more particular interest here, however, is CMAF's low-latency mode.

At a high level, CMAF is a series of fragments, or segments. Simply making them smaller would not lead to reduced latency, but instead would incur a bandwidth tax – with no attendant boost to quality. That is because of

[3] IP Multicast Adaptive Bit Rate Architecture Technical Report, CableLabs, Oct 26, 2018; "DVB releases reference architecture for IP multicast," March 9 2018.

[4] "Transcasting and multicast-ABR market read to take off," John Moulding, Videonet, Sept 7, 2018.

the need for relatively large Instantaneous Decoder Refresh (IDR) frames at the start of each fragment. CMAF's low-latency mode, however, allows you to split these fragments into smaller chunks. These are the smallest encoded and addressable entities in CMAF, composed of a header (typically a movie fragment box or "moof") and media samples (media data box or mdat). (See figure 2.)

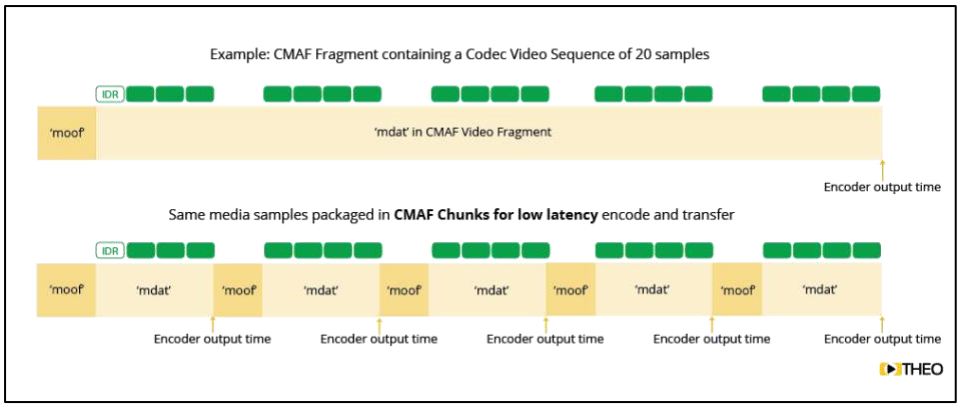



**Figure 2: CMAF Fragment and Low-Latency**

By itself, however, chunked CMAF does not move the latency dial. That requires the addition of CTE, the well-established streaming data transfer mechanism specific to HTTP 1.1.

## CTE, Players, Other Components

To maintain a HTTP persistent connection for dynamically generated content, CTE enables the transfer of an object or file, on the fly, piece by piece (i.e. chunk by chunk). It is the combination of the CMAF chunks with the CTE mechanism of transferring that positions the stream for lower latency. It also requires a player that can configure latency from the start, learn how to estimate bandwidth and maintain a correct minimum buffer size.

When playing media at low latency, players need to keep in mind network jitter. This is especially the case with unicast, rather than multicast over a managed network, as explained above. When there is no guarantee on delivery and frames can arrive late, the player's buffering algorithms remain essential for a smooth low-latency experience. Having a small buffer, and as a result a lower latency, could cause buffer underflow. Having a buffer which is too big, however, would introduce additional unwanted latency. If players fail to modify their buffering behavior, unstable connections will have a significant impact on the user's QoE.

One important qualification to chunked CMAF is that HLS does not yet have a real low-latency mode, while DASH does. There is the low-latency HLS solution that Twitter's Periscope implemented and discussed in a well-known article;[5] but so far, nothing has been standardized. Another point is that all components in the chunked CMAF video delivery chain must support HTTP CTE delivery. That said, this approach is gaining momentum.

"We see more and more vendors picking this up and making available the first pipelines that can handle chunked CMAF end-to-end, including encoders, packagers, CDNs and players," said Speelmans. "For some components, there are more alternatives than others. I expect new announcements around this going forward."

## How Low Can You Go?

The upshot is that multicast ABR, chunked CMAF and CTE, together with the right playback technology, can lower latency by another 4.5 seconds. According to Broadpeak, this combination further reduces packaging and multicast ABR transmission latency from 2 seconds to 250 milliseconds each. That's a combined reduction of 3.5 seconds. Chunked CMAF, incidentally, decouples latency from segment size, making segment duration less relevant; but the relatively jitter-free multicast link helps assure QoE. On the player side, algorithms cut buffering to 1 second, with end-to-end latency reaching 3.1 seconds, vs. 4.6 seconds for IPTV. (See Figure 3.)

[5] "Introducing LHLS Media Streaming," Mark Kalman et al., Periscope Code, Medium. July 21, 2017.

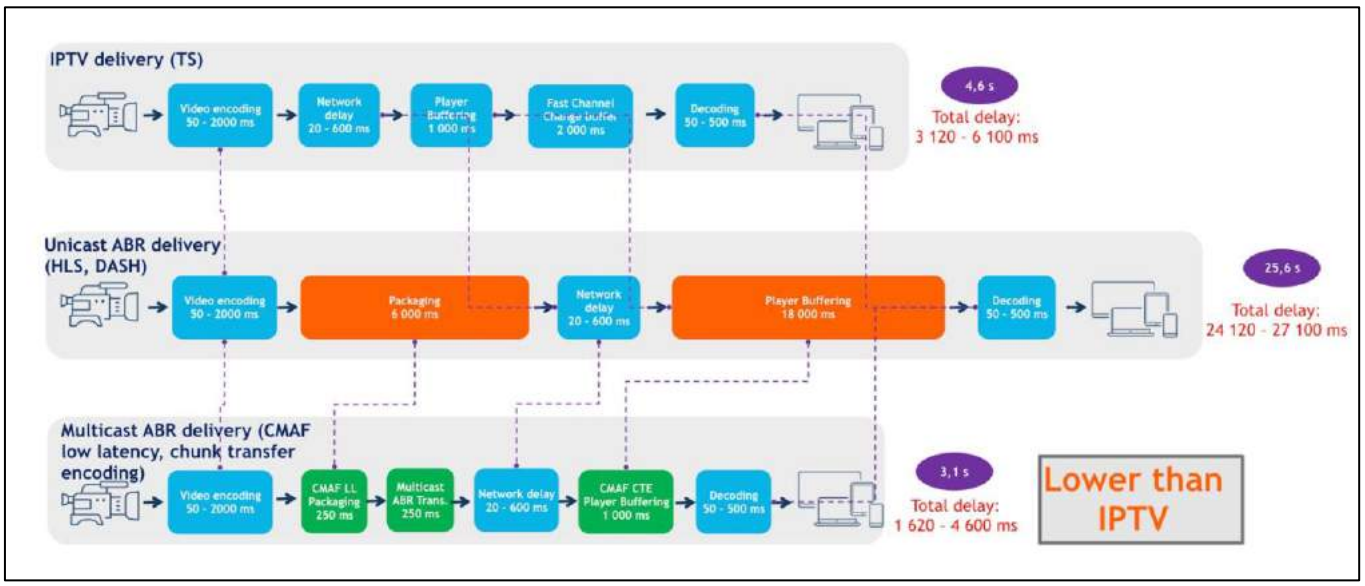



**Figure 3 – Latency Comparison: IPTV TS vs. Unicast ABR vs. Multicast ABR (Chunked CMAF)**

Can it go further? It can, at least another second, according to Limelight Networks, which tested chunked CMAF over its CDN. "We have seen CMAF work really well at two-seconds latency," said Limelight VP Product Strategy Steve Miller Jones, in a Videonet webinar.[6] Below 1 second, however, he said that chunked CMAF's client-and-server chatter leads into "file not found" errors, with QoE beginning to deteriorate.

# Tradeoffs and Options

## Startup Time, Synchronization

As noted with respect to buffer size, latency involves tradeoffs. Low latency and startup time are also in competition. Lowering latency requires waiting for a segment that has not yet been received, which increases startup time. A short startup time requires using what is already available in the buffer; with such content being older, latency increases.

Device synchronization is another area of contention. In a standard OTT system, two users switching to the same channel at a different moment may experience desynchronization up to the size of a segment. This could also arise from different services of different operators, or broadcast vs. OTT from the same operator. Sound desynchronization can cause an annoying echo effect if several people are watching the same content on different devices next to each other.

Multicast ABR with low latency can counter desynchronization, but as noted, it will have a negative impact on startup time. One best practice is to configure individual channels for low latency (such as sports) and others according to what needs to be optimized.

## Other Technologies

Apple's deprecation of Flash in 2010 accelerated interest in alternatives to the related Real-time Media Protocol (RTMP). Other approaches soon emerged. Underlying transport protocols have also evolved.

WebSocket, while not a streaming protocol, is worth mentioning in this context. A TCP-oriented socket (i.e. IP address and port number combo) WebSocket was standardized in 2011 and provides full-duplex communication. It was designed to support video calling, but several companies have added proprietary layers and leveraged it for some broadcasting use cases.

[6] "The Power of Now: Enriching TV with sub-second latency for live streaming," Videonet, Jan 2019.

Open-source WebRTC is another option. Promoted by Google and first implemented in 2011, it uses APIs to enable real-time communication on Web browsers and mobile apps. Apple lent official support in 2017, and Limelight Networks is one CDN that has embraced it. Supported by most browsers, WebRTC operations yet face interoperability challenges due to multiple stacks.

At the transport layer, buffering delays inherent to TCP have fueled interest in UDP. WebRTC defaults to UDP, although it can also use TCP when faced with corporate firewalls. Secure Reliable Transport (SRT) developed by video streaming solutions company Haivision and supported by a 170-member alliance also takes advantage of UDP.

A similar evolution is underway with HTTP. Early on, HTTP 1.1's CTE enabled a persistent server connection; HTTP/2 implements several other latency-decreasing techniques; and the proposed HTTP/3 may simply become the new label for Quick UPD Internet Connections (QUIC), a Google-led protocol focused on correcting the inefficiencies of TCP's congestion control and boosting the performance of web applications.

Whatever upper-layer technology is involved needs to address UDP's lack of flow control, consumption of shared bandwidth when coexisting with TCP, and need for special techniques to traverse Network Address Translation (NAT)-based firewalls. As noted, multicast ABR leverages UDP, but within a contention-free managed framework.

### Fitting Tech and Use Case

Which technology works best, and where? Some low-latency solutions, such as SRT, may be suited for ingest or contribution, but not elsewhere. "WebRTC is great technology, but it was designed for P2P communication," said Oliver Lietz, Founder and CEO of nanocosmos, in a Streaming Media webinar.[7] "It is hard to scale well and get it to all devices."

The solution that Lietz himself developed relies in part upon WebSocket, which raises a similar question. "Yes, you can scale WebRTC and WebSocket," said Speelmans. "But there is a cost associated with doing it." Another objection is vendor lock-in. Application developers (gambling, auction, quiz, etc.) may be happy working with any vendor option, provided latency is low enough, but video service providers tend to avoid solutions controlled by a single company.

What matters, according to Speelmans, is the use case, scale and target audience. "Are you trying to achieve something like broadcast latency, 5-8 seconds ballpark?" he asked. "Or are you going for sub-second real interactivity where people are interacting with each other and need close feedback?" In the first case, or even where latencies as low as 2-3 seconds are required, the standards-based, multi-vendor approach discussed in this paper is a strong candidate.

## The Live Streaming Future

The vast majority of HTTP-based adaptive streaming involves on-demand video, but consumer appetite for live events, especially sports, is strong. According to Cisco's Visual Networking Index, live streaming will grow 15-fold from 2017 to 2022.[8]

To support this growth, minimize viewer frustration and drive usage, video service providers are exploring low-latency technologies. The test results for multicast ABR, chunked CMAF and optimized video playback are impressive. Replacing unicast with multicast ABR leads to a 73 percent reduction in latency. Using chunked CMAF with CTE can further reduce that amount by more than half, below what IPTV can deliver today. This overall approach can lower latency from 26 to 3 seconds.

As always, there are limits and tradeoffs. But industry support for multicast ABR and chunked CMAF is growing, smart playback technology supports it, and demonstrated results make this combination a promising catalyst for the future of live streaming. **- JT**

[7] "Latency Still Sucks – So What Can You Do About It Today?" Streaming media, Dec 6, 2018..

[8] Cisco Visual Networking Index. Forecast and Trends, 2017-2022.

## About the author

Jonathan Tombes is a technology and business writer who has served as a consultant and freelance writer for numerous companies, organizations and publications. He has extensive experience covering the cable, telecommunications and IT industries. For more information, visit www.jtombes.com.

This paper was sponsored by Broadpeak, a designer and manufacturer of video delivery components for content providers and network service providers; and THEO Technologies, developer of THEOplayer, a cross-platform, universal video player.

THEO

# INTERESTED IN MULTICAST ABR, LOW-LATENCY CMAF, CTE, OR OPTIMIZED VIDEO PLAYBACK?

Get in contact with one of our THEO experts.

**Pieter-Jan Speelmans**

**Joeri Devisch**

**Willem De Saegher**

**Ask an expert**

THEO Technologies NV
www.theoplayer.com
contact@theoplayer.com