WHITE PAPER
REDUCING LATENCY FOR LIVE OVER THE TOP VIDEO DELIVERY
TABLE OF CONTENTS

Introduction .................................................................................................................................................. 3
Live Streaming With TCP/IP ............................................................................................................................ 3
Tuning Chunk Sizes ......................................................................................................................................... 4
Limelight Multi-Device Media Delivery Live (MMD Live) ............................................................................. 5
Common Media Application Format (CMAF) .................................................................................................. 5
CMAF Encoding .............................................................................................................................................. 6
CMAF Requirements ....................................................................................................................................... 7
WebRTC and Limelight Real-Time Streaming ................................................................................................ 8
Conclusions .................................................................................................................................................... 9
About Limelight Networks ................................................................................................................................ 9
Figures
  Figure 1: Overview of Live Streaming Using TCP/IP Protocol ........................................................................ 3
  Figure 2: Video Players Require 3 Video Chunks to Buffer Before Playback Begins ...................................... 4
  Figure 3: MMD Live Small Chunk Size Streaming ....................................................................................... 5
  Figure 4: Low Latency CMAF Delivers Smaller Chunks of Video More Often. ................................................. 6
  Figure 5: Requirements for Low Latency CMAF .............................................................................................. 7
  Figure 6: Limelight Realtime Streaming ....................................................................................................... 8
INTRODUCTION

Live over-the-top video delivery is becoming increasingly more popular and competitive. There are more major media companies adding live streaming services to compete in the OTT space. One way for media distributors to differentiate their OTT platform is by providing high-quality video with lower latency. In this whitepaper, you’ll learn about different low latency streaming options that can be used to reduce latency for live OTT delivery.

It takes time for live video to process and propagate through the network to the viewer’s device. The time that it takes for video from a live camera to be encoded and delivered to an online player or OTT device is what we refer to as live streaming latency.

Why is latency important? OTT latency requirements vary by content. For example: With live events like the Super Bowl, World Cup or the Kentucky Derby, having ultra-low latency may add value, while other VOD content like movies or TV shows may not require low latency.

Have you ever been watching a live sporting event online while also monitoring the score on social media? Because of the difference in latency between platforms, spoilers can occur, where we see the score update on social media before the live video is seen. Lower latency can reduce the occurrence of spoilers. Viewers want to enjoy the same viewing experience regardless of device or location.

USE CASES FOR LOW LATENCY AND REAL-TIME STREAMING

Low Latency: Live Sporting Events, Webinars, Town Halls, Meetings and more.
Low Latency with Interactivity: Voting, Auctions, Wagering, Gaming, and more.
The integration of live data and social media requires low latency.

LIVE STREAMING WITH TCP/IP

The Internet was not originally designed to deliver video. HTTP live streaming protocols emerged as a way to deliver video files over the internet by breaking the stream into a sequence of small http-based video chunks. By using http transactions, HTTP live streaming can traverse firewalls and use standard web servers to deliver video streams and take advantage of Content Delivery Networks to scale delivery. Figure 1 is a simplified illustration of the process of live streaming with TCP/IP.
TUNING CHUNK SIZES

What is chunk size tuning? Chunk size tuning reduces the video segment length for encoded media.

When our encoder creates video chunks (or segments) they are typically created in 10 seconds chunks. The player will need to buffer 3 video segments before playback begins. If we have 3 chunks that are 10 seconds each, we have 30 seconds of video that must be buffered by the player before playback begins.

Now let’s consider the same 3 video segments, but the length has been reduced to 6 seconds each. What that allows is instead of 30 seconds of video to be buffered, we only have to buffer 18 seconds of video before our player begins playback. Figure 2 below demonstrates that concept.

There is no universal setting for chunk sizes. Overly reducing chunk sizes can introduce rebuffering from the player if the stream is disrupted and the buffer becomes empty. Each workflow should be tested accordingly.
LIMELIGHT MULTI-DEVICE MEDIA DELIVERY LIVE (MMD LIVE)

With Limelight MMD Live, chunk sizes are configurable. With chunk size tuning, Limelight MMD Live can achieve end-to-end latency as low as 6 seconds by reducing the chunk sizes as small as one second. MMD live supports all major HTTP chunked streaming formats, including HLS, HDS, MSS and MPEG-DASH, simplifying the process of delivering live video to a myriad of different devices.

MMD Live is a great choice for low latency delivery, especially when there is not the need for real-time latency or interactivity.

![Figure 3: MMD Live Small Chunk Size Streaming](image)

COMMON MEDIA APPLICATION FORMAT (CMAF)

CMAF Common Media Application Format is an effort to move towards a common low latency file format. HLS and MPEG DASH are the two dominant HTTP streaming formats in use today. HLS and DASH have traditionally used different methods of containerizing video and audio. By using a common framework to store the files, content owners wanting to deliver video to the various tablets, computers, phones and OTT devices in use can store and package files once, instead of having two separate sets with each holding exactly the same audio and video data.
CMAF ENCODING

Let's take a look at CMAF encoding in more detail. At the top of figure 4, we have a non-CMAF chunked video fragment. It starts with a Moov (Movie Fragment Box) that contains timing and duration for each video sample. The MDAT (Movie Data Box) contains the described video samples. After the complete segment is created, the video samples are output from the encoder, transferred over the network, and the decoder is able to start the decode.

At the bottom of figure 4, we have a CMAF chunked encoded segment where additional ‘moofs’ are injected to describe smaller chunks of video. We refer to these smaller video chunks as video fragments. There is a separate header that is required to recognize smaller video fragments and initialize the CMAF chunk streaming playback. The video fragments can be transferred before the entire segment has been encoded. This allows the decoder to start playing back video before the entire video segment is received. As you can see, CMAF chunked encoding allows delivery of smaller chunks of video more often, resulting in reduced end-to-end latency.
CMAF REQUIREMENTS

There are several requirements that must be met in order to achieve lowered latency with CMAF:

**Encoding** – The media must be CMAF chunked encoded and the encoder must update the manifest to properly describe fragments of video and signal the availability of smaller video fragments.

**Network** – The network must support Http 1.1 chunked transfer or similar method that allows persistent http connections. Chunked transfer encoding must be used through the entire distribution chain, from the encoder to the video player. With chunked transfer encoding, the header can be written and read before the full media content is generated.

**The Video Player** – The video player must be able to read and react to the updated manifest. The player must be able to decode the video fragments as they are received. Also, the player must be able to dynamically adjust its buffer.

![Figure 5: Requirements for Low Latency CMAF](image)
WEBRTC AND LIMELIGHT REAL-TIME STREAMING

What is WebRTC? Web Real-Time Communications is an open source project for Browser and mobile application communications. WebRTC is supported by Google, Mozilla, Opera and other major industry players. Unlike HLS and MPEG-DASH, WebRTC uses UDP instead of TCP/IP to broadcast streams. Streams do need to be chunked by the encoder before broadcasting. WebRTC was initially used for browser to browser conferencing. Although WebRTC was originally designed for peer-to-peer communications, Limelight Networks has implemented it in a way that easily scales to support global audiences with Limelight Realtime Streaming.

Limelight Realtime Streaming uses WebRTC technology to stream live video with less than one second of latency through the fast and efficient UDP data transfer protocol. It allows playback and delivery on standard web browsers so video playback can happen without the need for plug-ins or special video players. Limelight Realtime Streaming also supports adaptive bitrate streaming to deliver the highest possible picture quality to each viewer, even over changing network conditions. Limelight Realtime Streaming leverages Limelight’s global private network which has the capacity, reach, and connectivity to ensure a high-quality, real-time viewing experience for viewers wherever they are.

Limelight Realtime Streaming also features the ability to share bi-directional live data. Using this 2-way data-channel, content providers can build interactive experiences alongside sub-second video streaming.

Figure 6: Limelight Realtime Streaming
CONCLUSIONS

There are many low-latency streaming options and no single best solution for delivering low-latency video for every workflow. The best solution for your needs will depend on the type of content you stream and your specific video workflow requirements.

Here are some pros and cons for some of the different low latency streaming technologies:

HTTP live streaming protocols like HLS and MPEG-DASH have a latency penalty when the encoder has to create video chunks. Chunk-size tuning can be used with the popular HLS and MPEG-DASH protocols to reduce latency caused by the player buffer. However, over-reducing these chunk sizes can cause rebuffering. Limelight MMD Live supports all http chunked formats and can provide a source to viewer latency as low as 6 seconds.

However, when ultra-low live streaming latency is required for workflows that require interactivity, Web-RTC streaming technologies may be the best option. WebRTC does not require video chunking and buffering required by HLS and MPEG-DASH.

Limelight Realtime Streaming utilizes WebRTC technology to provide sub-second latency for live video streams. Integrated live data sharing is also available to provide the ultimate interactive live viewing experience.

Talk to Limelight about which streaming solution is best for your workflow!

ABOUT LIMELIGHT NETWORKS

Limelight Networks, Inc. (NASDAQ: LLNW) is a global leader in delivering the highest quality online video experiences and edge-enabled workflows. Limelight’s edge services platform makes knowledge, information, and entertainment instantly accessible anywhere in the world and accelerates the next generation of real-time, interactive and immersive content. For more information, visit www.limelight.com, follow us on Twitter, Facebook, and LinkedIn.