HOW SUB-SECOND LATENCY STREAMING AND LIVE VIEWER INTERACTIVITY ARE CHANGING THE VIDEO LANDSCAPE

EXECUTIVE SUMMARY
Live streaming is attracting viewers online to watch major sports events, play games, participate remotely in educational opportunities, and bid at live auctions. But today, the latency of online video stream delivery is typically too long to provide the viewing experience users expect, resulting in unhappy viewers and lost revenue. Fortunately, new live streaming technology makes it possible to deliver live streams in less than a second, enabling exciting new experiences that engage viewers in multiple ways. For organizations that need to distribute live streams, it’s about increasing audience size and revenue. For viewers, watching streams in realtime with interactive data integrated with the live video enables new possibilities for how they can interact with you and each other. Read this brief to learn how sub-second latency streaming enables new business opportunities by making live viewing a more interactive social experience.

DEFINING LATENCY
Before talking about sub-second latency streaming, it’s important to define how latency is measured. Due to the complexity of live streaming workflows, a definition that comprises all the workflow elements will allow fair comparisons between different technologies. For the purposes of this brief, streaming latency is defined as the amount of time it takes for video to travel from the live source acquisition to an online viewer’s screen. The below diagram shows a complete live streaming workflow. Although the elements of a traditional television broadcast workflow differ from online streaming, latency is still measured as the time delay between live source acquisition and delivery to the viewer’s screen.
THE STREAMING LATENCY LANDSCAPE

There are many different live online streaming use cases, each with its own requirements for delivery latency. For reference, the below graph shows the typical latency of traditional chunked HTTP streaming formats and broadcast TV via cable along with the low latency delivered by other streaming technologies.

WHAT'S DRIVING THE QUEST FOR LOW LATENCY STREAMING?

Many of the streaming use cases that require very low latency involve some form of interactivity where online viewers need to react to live video in real-time. Some examples include:

- **Online gaming**—Online gambling is gaining traction, especially for 18-49 year olds, according to a study by YouGov. A barrier to greater adoption is the long and inconsistent delays in gameplay that gamblers experience online. This is because the casino must wait until all players are able to see the live action (such as the turn of a card or roll of dice) before players are able to determine whether they want to potentially raise their wager or hold. Delivering video streams in real-time to players anywhere in the world ensures everyone instantly sees the action on the table simultaneously, allowing them to place bets more quickly, increasing the number of rounds played per hour and the amount of revenue generated by the casino operator.

- **In-event sports betting**—This use case includes elements of online gambling along with sports viewing. In addition to receiving a real-time online stream of a live sporting event, viewers are provided with on-screen opportunities to wager on outcomes of specific in-game events that happen during the game or match. It is critical to ensure the streaming latency is as low as possible to ensure no online viewers have an unfair advantage by knowing the outcome of an in-event wager by being at the actual live sporting event or watching it on TV. It is also important to make sure all viewers have the same betting opportunities and experience, regardless of where they may be located or the device they are using to view the event.
• **Live auctions**—Remote bidders participating by phone are common at auctions, but now they can be online and have a “just like being there” experience with a live video stream to participate along with those on-site. The real-time video stream ensures the online participant can signal bids with confidence, knowing that they are viewing the action in the auction house at the same time as those on-site.

• **Trivia games**—The largest online trivia games can draw more than one million simultaneous players to a single event. Contestants use an app that displays a video stream of the game host, with multiple choice answers appearing on the screen a set time after the host asks a question. Like in-event sports betting, it is not only important that contestants’ video streams are delivered in real-time, but that the latency is consistent across players no matter where they are located to ensure everyone is able to respond at the same time.

**CHALLENGES DELIVERING SUB-SECOND LATENCY LIVE VIDEO**

The internet wasn’t originally designed for streaming live video. To address this limitation, HTTP-based live streaming formats such as HLS and MPEG-DASH were developed to allow live internet streaming using the TCP/IP protocol. Video streams are encoded in segments (or chunks) that are delivered to the receiving application and then buffered before being played. This allows the playback application to accommodate the inherent potential delay in the transmission of live video over the internet by buffering (or storing) the video before it is played. The typical stream latency of 30 seconds to over one minute is due to the amount of video that is typically buffered. While it’s possible to reduce the size of the buffer chunks to minimize the delay, making them too small increases the chance viewers will experience video rebuffering and other playback issues. To deliver live streams at the latencies required for the use cases discussed above, several new technologies are being tested and deployed by Content Delivery Networks (CDNs).

**SOLVING LOW LATENCY DELIVERY CHALLENGES**

One approach for delivering lower latency live HLS and DASH video streaming is to reduce the chunk sizes to minimize the amount of video that is buffered before playback. If chunk sizes are reduced to one or two seconds, the total resulting latency will be in the 6-10 second range, which is close to the latency of broadcast cable delivery. Deploying this solution for a major sports event with a large traditional TV broadcast and online viewing audience would mitigate the potential spoiler issue where a viewer watching online receives a text from a friend viewing the TV broadcast about an amazing play they haven’t seen yet due to online latency.
Another current industry approach for reducing latency with chunked streaming is the use of the Common Media Application Format (CMAF). The original goal for the development of CMAF was the ability to create a single encoding that could be used to deliver both HLS and DASH video. However, CMAF also provides a low latency mode where media segments can be divided into small chunks which can be streamed, downloaded, and played even though the entire segment has not been delivered yet. This can reduce live streaming latency to as low as 1 to 2 seconds. For more detailed information on CMAF, see this article.

TECHNOLOGIES FOR SUB-SECOND LATENCY STREAMING DELIVERY

One technology that is gaining significant traction for sub-second streaming is Web Real-Time Communication (WebRTC) https://webrtc.org/, originally released as open-source by Google in 2011 to enable sub-second latency communications between browsers. Many of the initial implementations of WebRTC were peer to peer video chat applications, such as the Google Hangout capability. There are WebRTC-based solutions being developed for distribution of low latency live video. Some use Peer to Peer connections between clusters of viewers to enhance content delivery. Content is automatically accessed from other end-user devices using the WebRTC protocol. The idea is to offload bandwidth and the associated cost of content delivery by shifting some of the content distribution to end-user devices. Potential challenges with this solution are its critical dependency on having the right mix of peers who are viewing the same video across regions and ISPs, and the fact that no guarantees can be offered about performance and video quality. Other WebRTC streaming solutions are based on a cloud model where live source video streams are converted to WebRTC at the edge of the network.

To leverage WebRTC’s sub-second latency to support large global audiences, Limelight Realtime Streaming deploys an edge compute model that distributes ingest streams to edge servers in CDN Points of Presence (PoPs) around the world where they are duplicated and scaled to support large numbers of viewers. By leveraging the CDN's global capacity, this architecture easily scales to provide sub-second realtime streaming to viewers everywhere in the world.

Another significant capability available in Limelight Realtime Streaming is the ability to share bidirectional data. This sharing of integrated data along with video allows content distributors to develop their own creative ways to use data as part of live video workflows. For example, viewers watching live sporting events could receive statistics about their favorite players and even vote on their favorite plays or choose which camera angle they would like to view. Gamers can have an integrated chat channel with their video, and online auctions can be streamed along with the ability for viewers to bid on items in realtime by hitting a button. These interactive capabilities open many new business opportunities.
IT’S NOT JUST ABOUT LOW LATENCY

Viewers have come to expect high-quality online viewing experiences. This means a sub-second latency solution must be as robust as today’s global HLS and DASH chunked delivery infrastructure, providing the same levels of resiliency, redundancy, security, and global scale. Audiences expect a broadcast quality experience from online video. Adaptive Bitrate Streaming (ABS) enables delivery of the highest possible picture quality to each viewer, even over changing network conditions, ensuring the best possible online experience. With adaptive streaming, the source video is encoded in multiple bitrates and resolutions. When a viewer plays the video, the highest possible picture quality that is supported by the viewer’s network connection will be streamed. If conditions change and video is streamed at a slower rate, the adaptive player will request a lower bitrate encoding to avoid rebuffering. Similarly, if network conditions improve, a higher quality encoding will be requested. Adaptive bitrate streaming ensures viewers receive the highest picture quality and the lowest rebuffer rates by adapting to the speed of the user’s internet connection.

SUMMARY AND RECOMMENDATIONS

Consumer demand for low-latency live video streaming is growing. With multiple solutions being developed to deliver low latency streams, it’s important to understand the characteristics of the different technologies and how they can support your specific use case. In addition to reducing the delay in live video streams, new capabilities are being developed that change the live viewing experience. Probably the most significant, based on potential new revenue opportunities, is the ability to integrate realtime data with live video to create interactive experiences. As with every commercial streaming video service, monetization is critically important, and the new business models enabled by interactivity will provide additional revenue opportunities.

If you are currently distributing live video or planning to do so in the future, you should consider how low-latency streaming solutions such as Limelight Realtime Streaming could help you increase the monetization of your live content. By integrating data with your video streams, you can offer new interactive experiences that provide new revenue opportunities and increase viewer engagement.

Limelight offers a complete range of live video streaming services with the experience delivering and supporting the largest live events in the world. No matter which technology fits your use case requirements, Limelight is committed to helping you deliver the best online experiences to your viewers and maximizing the value of your online content.

ABOUT LIMELIGHT NETWORKS

Limelight Networks Inc., (NASDAQ: LLNW), a leading provider of digital content delivery, video, cloud security, and edge computing services, empowers customers to provide exceptional digital experiences. Limelight’s edge services platform includes a unique combination of global private infrastructure, intelligent software, and expert support services that enable current and future workflows. For more information, visit www.limelight.com, follow us on Twitter, Facebook, and LinkedIn.