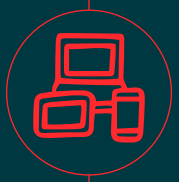
fastly

Fastly for Streaming Media

SOLUTIONS BRIEF



Maximizing Online Viewer Engagement and Interactivity with Low Latency Streaming

The promise of online media delivery with latency rivaling that of its cable counterpart is within reach as protocols and formats sees wider adoption among online broadcasters. Understand how to benefit from low latency delivery as well as the new business opportunities it can bring.

Overview

Sheltering at home opened the door to greater consumption of online live video, and the number of hours spent watching live content is not expected to shrink. For 2020 overall, live streaming doubled compared to the previous year[1]. The pandemic opened the door to greater consumption of online live video, and it is not expected to close. Live streaming accounts for three-quarters of the revenue of social apps, which are expected to grow to $17.7 billion in 2025, up 160% from $6.8 billion this year[2]. From Twitch to YouTube TV, and from Hulu + Live TV to Facebook Live, services that aim to bring the live experience home to consumers are predicted to continue their growth.

For providers, this instant growth poses significant challenges. The quality of any live streaming experience depends on driving latency out of the network. Latency — the lag between when the packet leaves the streaming source and when it arrives at the consumer's device — takes many forms, with the most common being lag, dropped frames, buffering, and reduced video quality. Businesses using virtual desktop infrastructure and consumers streaming games from cloud platforms each require ultra-low latency to offer the best user experience. One report predicts cloud gaming, for example, will grow nearly 60% every year until 2024, driven by such services as Google's Stadia, Microsoft's Xbox Cloud Gaming, Amazon's Luna, and a rumored service under wraps at Apple.[3]

Guaranteeing low latency for your content requires the right planning, the right technology, and the right partners. In this Solutions Brief, we discuss how to reduce latency for your content and achieve ultra-low latency for competitive advantage.

*1 Stephen, Bijan. "**The lockdown live-streaming numbers are out, and they're huge.**" Verge. News Article. 13 May 2020.*

*2 Sydow, Lexi. "**The Evolution of Social Media Apps: Live Streaming: The New Frontier for Social Media**." Market Data Blog. Blog Article. App Annie. 6 September 2021.*

*3 "**Cloud Gaming Market by Offering, Device Type, Solution, Gamer Type, Region - Global Forecast to 2024.**" MarketsAndMarkets. Analyst Report. December 2019.*

Livestreams grow, and so does importance of low latency

The market for live and low-latency content has continued to grow. Consumers have increased their demand for Internet-delivered media, and particularly live streams, since the beginning of 2020. The total quarterly hours viewed on the top-3 game-streaming platforms — Twitch, YouTube Live Gaming, and Facebook Gaming — grew to 17.8 billion hours in the first half of 2021, up from 7.3 billion hours in the first half of 2019.[4] By the end of 2021, video traffic — although, not just live-stream video — will account for 82% of the volume of data communicated over the Internet.[5]

The fast delivery of and ability to interact with content is now an expected part of consumers' experience, which means that content companies will have to compete on the basis of user experience.

Low latency enables interactivity, user engagement

Perhaps the most important aspect of that experience is eliminating latency to allow real-time applications. Users don't just want to stream content; they want to stream live and real-time content. They don't just want to hold video conferences for a small group; they want live presentations with hundreds of participants allowing real-time feedback for audience members. They don't just want to watch a live game; they want to engage with events in real time, taking quizzes, placing bets, and participating in the experience of the game.

If latency intrudes on those experiences, customers will turn to other providers, watch different events, or engage with other people and businesses. The real benefit of ultra-low latency, then, is that that content, user experiences, and data are delivered in near real-time, forming the basis of better user experiences and enabling new business models. Guaranteeing low latency can set your content apart from your competition, driving adoption, and minimizing customer churn.

Why now?

Prior to 2020, online users consumed a moderate and slowly growing amount of content, about 3.8 billion hours per quarter globally for the three top live-stream gaming services.[6] Stay-at-home orders and concerns about the pandemic boosted consumption of real-time streams, nearly doubling the number of hours consumed by users.

*4 Blees, Olivia. "**Q2 2021 Live Streaming Trends**." Stream Hatchet Blog. Blog Post. Stream Hatchet. 8 July 2021.*
*5 "**Annual Global Internet Report — Global 2021 Forecast Highlights.**" Cisco. PDF Report. n.d.*
6 Stream Hatchet Blog, ibid.

Following the pandemic, more consumers looked for online experiences that mimicked live events as closely as possible. As the impact of the pandemic on daily life slowly wanes, viewers will have in-person options and will decide whether to continue paying for content delivered over the Internet based on their experience. While some content may allow for latency without impacting user experience, the overall quality of the experience is always improved by reducing latency.

Attempting to achieve low latency is the top concern among content publishers, with 41% of professionals identifying latency as their primary challenge in a recent survey, with controlling costs taking second place at 33%.[7]

A primer on latency

The networks supporting content delivery have become more complex. Mobile devices are now the preferred viewing platforms for consumers, while businesses and advertisers have committed to livestream video for advertising, marketing, and customer engagement. As a result, latency can be caused at a growing number of points along the content supply chain. Reducing and eliminating latency is no longer a discussion with a single provider, but requires visibility into the performance of a chain of third-party network and content providers.

Defining latency and low latency

The pipeline of content from creation to transmission to eventual reception by the consumer device requires processing and bandwidth, and so requires time. It can often take tens of seconds for a live event to display on a consumer device. While the average HD cable broadcaster experiences 4 to 5 seconds of latency, about a quarter of content networks are challenged by anywhere from 10 to 45 seconds of latency.[8]

While there are no standards today, typically, “low latency” is defined as content that is delivered between 1 second and 5 seconds after it was first captured. For live events, such as concerts and sports, low-latency delivery means that the video is delivered to a consumer’s screen less than 5 seconds after the live action. So-called “glass-to-glass” latency — from the camera lens to the viewer’s screen — is often around 20 seconds, while high-definition cable TV content is the benchmark for low latency, at about 5 seconds of latency.[9]

*7 “**Video Developer Report 2021.**” BitMovin. PDF Report. N.D.*

*8 “**What Is Low Latency and Who Needs It? (Update).**” Wowza Resource Center. Wowza. 8 September 2021.*

9 “What Is Low Latency and Who Needs It? (Update).” Wowza.

What causes latency?

There are many causes of latency in broadcast and delivery networks. The mere act of encoding a live video into packets to be sent over a network introduces delays into the video stream. Add to that the delivery through a variety of third-party networks to the end user's device, and the latency grows longer. In addition, different protocols have different strengths and weaknesses, and the primary consideration is not always reducing latency.

Apple's HTTP Live Streaming (HLS), for example, has a default of 30 seconds of latency, because it prioritizes using existing infrastructure (HTTP) for delivery as opposed to more efficient, but less widespread, protocols. Other protocols might optimize the network route to the destination, minimize delays from encryption and other secondary network functions, and choose optimized encoding and decoding techniques.

Network latency is a key factor in achieving low latency delivery, and poses challenges that the content publisher is ill-suited to solve. Content delivery networks (CDNs) have the expertise and network architecture to solve these issues and deliver low-latency bandwidth.

How does the media format affect the need and deployment of low latency?

The industry has typically taken a two-pronged approach to developing protocols to deliver broadband media while benefiting from the widespread adoption of existing protocols:

1. retrofitting widespread protocols, such as HTTP, to streaming while attempting to reduce its latency, and
2. using tailor-made protocols whose primary focus is low-latency, or real-time, applications.

Different media formats are used depending on the application and the focus on what needs to be accomplished — often not just fast delivery, but delivery to heterogenous technology platforms and to users with a variety of bandwidth capabilities. WebRTC, for example, is an open-source effort that allows real-time communications over the web, achieving sub 500 millisecond latency. The protocol works with all major browsers and can send information to laptops, smart phones, and smart TVs. HTTP Live Streaming (HLS), on the other hand, is more focused on using existing infrastructure, scalability, and easy integration into web applications.

3. Achieving low latency

Reducing the latency in the content delivery network requires planning and engineering, and an acceptance of tradeoffs between latency and cost. While building an efficient method of recording, encoding, and initially transmitting content can help remove inefficiencies and latency from early in the process, much of the latency occurs in the delivery network.

Content companies need to find solutions for both the front end of the system and the network delivery components to achieve the lowest latencies possible. While 42% of respondents in a 2021 survey considered high-quality video to be the No. 1 priority for user experience, low latency came in a close second, with 32% of respondents considering responsive video to be the top priority.[10]

Consider the impact of workflow components

Content publishers should benchmark their pipelines to find out which stages are causing the most significant delays. The first place to look is the workflow used to capture, encode, and package content — each stage of which can dramatically impact latency. The codec and frame rate used during encoding, for example, can significantly slow down the processing of the video stream. Using multi-second video segments requires that the encoder wait for the length of the segment before processing the video, resulting in a delay of equal time.

The network propagation techniques and the data transport protocols are also common sources of latency in a media distribution network. Content delivery networks (CDNs) that are not designed for low latency will also contribute to delays. Finally, the policies and settings of the player on the device can have a significant impact, especially if the player is configured for significant error correction or extensive buffering.

Major places to reduce latency

Latency can be reduced by tuning the encoding workflow for faster processing. However, doing so will cause inefficiencies — and higher costs — elsewhere. Smaller network packets and video segments amount to more overhead and less bandwidth, but will reduce latency, while larger segments increase the overall bandwidth and efficiency at the cost of a real-time experience.

*10 "**2021 Video Streaming Latency Report.**" Wowza Media Systems. PDF Report. 21 September 2021.*

The workflow of capturing and encoding media is a good place to look for opportunities to reduce latency. A well-tuned workflow can quickly produce encoded video segments, but focusing on minimizing processing time is not the only goal. Spending more time processing can often produce more compact data streams, reducing the overall network latency. Thus, there is a dial between processing efficiency and network-transport efficiency, and content publishers need to find the right balance.

Limits of latency reduction

Theoretically, sub-1 second latencies from the camera lens to the device screen — so-called "glass to glass" — are quite possible. Network latencies depend on the network, but vary from 10 milliseconds for modern carrier networks to 60 milliseconds for 4G cellular networks to 800 milliseconds for satellite networks.[11] Add to that the capture, encoding, decoding, and delays for out-of-order packets, and the overall latency can quickly climb.

"Real-time" cloud processing actually does not exist. Human perception of real-time activity is typically argued to be 6 milliseconds to 20 millisecond by academics,[12] while the time it takes light to travel between San Francisco and New York City is 14 milliseconds, or 21 millisecond over optical fiber one way and 42 millisecond round trip.[13] Despite this impossibility, real-time-like interactivity is possible. Acceptable latency, or lag, in online gaming is considered to be 40 millisecond to 60 milliseconds, with anything greater than 100 milliseconds becoming noticeable to players.[14]

Yet, only a few applications need ultra-low latency. Media publishers should assess their needs and determine what their target should be for latency and then architect their process and select the right partners.

*11 "**What is network latency (and how do you use a latency calculator to calculate throughput)?**" SAS Blog. Blog Article. N.D.*

*12 S. Kudrle, M. Proulx, P. Carrières and M. Lopez, "**Fingerprinting for Solving A/V Synchronization Issues within Broadcast Environments**," in SMPTE Motion Imaging Journal, vol. 120, no. 5, pp. 36-46, July 2011,*

*13 Grigorik, Ilya. "**Primer on Latency and Bandwidth.**" High Performance Browser Networking. O'Reilly. Blog Post. 2013.*

*14 "**How to improve your gaming latency.**" CenturyLink Support Article. Web Blog. N.D.*

Conclusion

Reducing latency is a competitive advantage, and companies should always strive to decrease the latency from glass-to-glass. However, latency reduction has a number of tradeoffs, the most important being the tradeoff between low-latency hardware and networks versus the cost of that equipment and service.

Companies should first determine what their latency goals are and then determine the path to get there, taking into account the necessary hardware, architecture, network services, and cost. Partners who have a focus on driving down latency are critical for any effort to improve the user experience.

Getting started

Find out why customers like
Shopify, Stripe, and LaunchDarkly choose Fastly.

To learn more, please contact us at sales@fastly.com, and visit our Software-as-a-Service webpage.

Contact Sales