CDN Alliance

Low Latency Streaming

A First Step Towards Standardization

Table of contents

01 Introduction

This document provides an in-depth exploration by the CDN Alliance to address key challenges in the Content Delivery Network (CDN) industry, with a focus on Low Latency streaming. It outlines the goals, scope, and activities of the Low Latency Working Group (LL-WG), highlights the importance of Low Latency across various applications, and delves into the technologies and protocols that enable efficient content delivery. By standardizing definitions and fostering collaboration, this whitepaper aims to serve as a comprehensive resource for stakeholders in the CDN industry, from content providers to technology vendors to video platforms to end-users.

02 About the CDN Alliance

The CDN Alliance is an independent, nonprofit organization with the goal to connect, support, and represent the global CDN Industry and CDN Community. The Alliance is dedicated to defining best practices and raising awareness about the CDN Industry. By initiating and facilitating activities related to work on global industry challenges, the Alliance further bolsters the CDN Industry in relation to bit delivery, dynamic delivery, media delivery, security and edge across the industry, market, business, technology, and policy domains. It is an independent organization and therefore aims at creating a place where these industry challenges can be addressed in a collaborative, open, and safe environment.

The industry challenges that the CDN Alliance focuses on include, but are not limited to, availability, scalability, reliability, privacy, security, sustainability, interoperability, standardization, education, certification, and regulations. The CDN Alliance also aims at being the 'voice' and the 'face' of the industry both within the industry as well as for partner organizations, other industry associations, government bodies and government agencies, the press, and the general public. The goal is to create more awareness and to be the overall touchpoint for the CDN Industry, as well as to serve the interests of the CDN Industry and the CDN Community as a whole.

03 About the Low Latency Working Group

The Low Latency Working Group (LL-WG) is a working group composed of CDN Alliance members with expertise and interest in the subject. Low Latency is a dynamically evolving term used by the video streaming industry which describes a multitude of applications that enable video content to be delivered over the Internet with minimal latency (measured as the difference between time of occurrence and the time of viewing by the end user). This working group specifically focuses on the delivery of one-to-many scenarios of Low Latency (live) video over the Internet in order to solve the challenges that still exist. Therefore, the term 'Low Latency' used by this working group only refers to the delivery of live video for one-to-many scenarios over the Internet (e.g., not related to video-conferencing, which tends to represent one-to-one and one-to-few scenarios).

With this whitepaper, the LL-WG aims to create a comprehensive overview to provide clarity about the availability, possibilities, and specifications of the most common current and upcoming Low Latency delivery technologies with a focus on one-to-many scenarios.

The LL-WG has decided to categorize these technologies into four categories:

- Low Latency (LL)
- Ultra-Low Latency (ULL)
- Sub-Second Streaming (SSS)
- Real-Time Streaming (RTS)

The goal is to achieve consensus, normalization and – where possible - standardization regarding the language, terms, definitions, and specifications used in the market today and in the future around these four categories.
In our view, no clear overview with clear definitions and descriptions has yet been defined in relation to the delivery of LL, ULL, SSS, and RTS over the Internet, which we feel is hampering the adoption and optimization of Low Latency technologies and workflows at present and which, if not addressed, will likely add more complexity and friction over time as adoption grows.

The LL-WG is aware that maintaining a comprehensive and timely resource of definitions and descriptions of Low Latency technologies requires an ongoing process and therefore depends on input and collaboration from the industry at large. More information about the LL-WG and how to join can be found **here**.

04 What is Latency & Low Latency?

There are currently many different definitions and interpretations of what Latency and Low Latency mean. Within the context of this whitepaper about Low Latency, it is important to understand that our focus is on the delivery of live video for one-to-many scenarios over the Internet.

What is Latency?

Latency, in the context of this whitepaper, refers to the time taken for a video frame to be transmitted from the source location (point A) to the delivery location (point B) over the Internet. This is often described as 'end-to-end' or 'glass-to-glass' latency, encompassing the entire journey from the camera lens to the viewer's screen. The most common measurement methods include comparing timestamps between source and playback devices using synchronized clocks, or visual measurement with on-screen timers captured in both the source and displayed video.

There are other partial latency measurements that can provide valuable diagnostic insights. However, these should not be confused with the total end-to-end latency that viewers experience. Here are some examples:

- **Processing Latency** Time required for encoding, transcoding, and packaging operations, isolating the computational overhead.
- **Replication and Delivery Latency** Time from when a video segment is ready at the origin server until it appears on the viewer's device, including CDN replication time.
- **Delivery Latency** Time from when a packaged segment is available on the CDN to when it's displayed on the viewer's screen.
- **Player Buffer Latency** Intentional delay introduced by video players that cache content before playback to prevent interruptions. This buffer creates a trade-off between playback stability and Real-Time delivery, significantly impacting the overall viewing experience in live streaming scenarios.

We propose in this whitepaper that latency should be measured end-to-end, and can be further categorized into three stages:

- Source Stage
- Transmission Stage
- Delivery Stage

Source Stage

The source stage includes the camera, encoder, upstream, and ingestion processes. In this phase, the camera stream is connected to a live software or hardware encoder. The encoded stream is then transmitted to the CDN or cloud infrastructure using protocols such as RTMP (Real-Time Messaging Protocol), SRT (Secure Reliable Transport), or WHIP (WebRTC HTTP Ingest Protocol). Although HLS (Apple HLS) or DASH (MPEG-DASH) ingestion is possible, it is rarely used in practice.

Transmission Stage

The transmission stage involves the CDN (Content Delivery Network), transcoding, and packaging processes. This stage is responsible for adapting the video stream to different formats and bitrates suitable for various devices and network conditions.

Delivery Stage

The delivery stage covers the final leg of the journey, from the CDN to the video player on the end-user's device. This phase requires compatibility with a wide range of end-user devices, including web browsers, smartphones, tablets, and connected TVs. Common technologies used in this stage are HLS and DASH.

The source and delivery stages have distinct requirements and challenges. The source stage typically involves lower data traffic and bandwidth utilization, as it deals with one or two concurrent streams being delivered to the transcoding and packaging layer hosted on CDNs or cloud infrastructure. In contrast, the delivery stage must handle large-scale deployments to accommodate thousands or even millions of concurrent viewers.

What is Low Latency?

The term 'Low Latency' is commonly used in an unspecific way when referring to latencies lower than those commonly achieved by delivery of HLS and DASH (according to the recommendations related to those standards). The term Low Latency can be further sub-categorized into Low Latency (LL), Ultra-Low Latency (ULL), Sub-Second Streaming (SSS), and Real-Time Streaming (RTS). These four categories are distinct because they involve different technologies, protocols, and specifications, making each appropriate for a different set of use cases. Despite these differences, confusion still arises as they are yet to be standardized and are often used interchangeably. Similar confusion can be found in several other key areas related to Low Latency, including (but not limited to): how Low Latency is measured, how secure the Low Latency video flow is, and the plethora of technologies, protocols, use cases, related language, specifications, terms, and lack of clarity in definitions.

One common misconception is that Low Latency is only defined based on technical definitions, when in fact the industry-wide usage of the term 'Low Latency' depends entirely on the marketing goal or industry perspective of whoever is using the term. A broadcaster delivering ad-based linear TV channels over the Internet will have a different opinion of what can be considered 'Low Latency' in comparison to a streaming service that offers premium live sports; live sports statistics and social media make the latency requirements of live sports far more demanding.

Auctions and iGaming, live casinos/betting scenarios, in which user interaction is the highest priority, offer different parameters for what latencies are acceptable and what technologies can be used.

Vendors of Low Latency services and solutions face similar challenges when refining precise technical deployments and latency measurement goals. As such, Low Latency not only provides a framework for a technical comparison between solutions and workflows, it also provides a framework for required latency measurements according to each particular use case.

05 Low Latency Use Cases

Low Latency can be applied to many different use cases, and the amount of use cases as well as the adoption of Low Latency across these and new use cases are growing. The following use-cases are those that we find to be most common.

Corporate Communications

For corporate communications, Real-Time or Ultra-Low Latency streaming ensures high-quality broadcasts and facilitates seamless Q&A sessions and other interactive experiences. This is particularly important for virtual meetings and webinars, where real-time interaction is critical for engagement and effectiveness.

For large-scale events such as town hall meetings or product launches, immediate feedback and interaction can significantly enhance the effectiveness of the communication. By providing a seamless and interactive experience, Real-Time streaming helps organizations maintain a high level of engagement and efficiency in their virtual communications.

iGaming with Live Dealers

For iGaming, Low Latency improves the quality of experience (QoE) by providing smooth, real-time interactions. This rapid interaction capability increases player engagement and satisfaction, leading to higher revenues as faster betting cycles allow more bets to be placed within the same time frame.

The interactive nature of live dealer games, such as blackjack or roulette, relies heavily on real-time communication between the dealer and the players. Any delay can disrupt the flow of the game and negatively impact the player experience. Low Latency streaming ensures that players can make decisions and place bets without noticeable delays, maintaining the excitement and engagement of the game. Additionally, it allows for more dynamic and responsive game mechanics, which can further enhance player satisfaction and retention.

In-Stadium Sports

Real-Time technologies are a must-have requirement for luxury boxes so that the 'crack of the bat' is heard on screen simultaneously to the actual sound from across the stadium. This premium, on-prem service demands the highest level of service in terms of latency, reliability, and image quality.

There are several approaches that have proven successful for this use case, including private 5G networks and distributed on-prem encoders and decoders.

Linear Broadcasting

Low Latency streaming benefits linear broadcasting by significantly enhancing viewer engagement and accessibility. Unlike traditional broadcast methods, Real-Time streaming allows viewers to access live content on various devices, including smartphones, tablets, and smart TVs, regardless of their location. This flexibility ensures that audiences can watch their favorite shows, news, and events as they happen, fostering a more immediate and connected viewing experience. The interactive capabilities of Real-Time streaming, such as live chats and social media integration, further enhance viewer participation and create a more dynamic and engaging broadcast environment.

Live Auctions

In live auctions, every millisecond counts. Real-Time streaming ensures that bids can be placed in real-time, preventing delays that could result in missed bidding opportunities. This leads to higher revenue and a better overall QoE, eliminating issues related to buffering, latency, or synchronization.

The fast-paced nature of live auctions demands that bidders receive information and place bids with minimal delay. Real-Time streaming allows auctioneers to manage the auction efficiently, keeping all participants on the same timeline. This capability is crucial for high-stakes and low-stakes auctions alike, as the slightest delay can result in lost bids and reduced revenue. By ensuring real-time interactions, Low Latency streaming enhances the competitiveness and fairness of the auction process

Live Concerts/Performances

Low Latency streaming significantly enhances the user experience of broadcasts of live concerts and performances by allowing audiences to engage with events as they unfold. This immediacy creates a shared experience, enabling viewers to feel the excitement and energy of the performance in sync with the live audience. The ability to stream concerts in real-time breaks geographical barriers, giving fans worldwide access to events they would otherwise miss. This democratizes the concert experience, fostering

a global community of fans who can interact through live chats and social media, enhancing the sense of connection and participation.

Additionally, Low Latency streaming opens new revenue streams for artists and organizers by reaching a broader audience beyond the physical venue. Virtual tickets, exclusive online content, and interactive features can be monetized, providing additional financial support to the performing arts industry. It also offers valuable data insights into viewer preferences and behaviors, allowing for targeted marketing and personalized experiences. Real-Time streaming thus not only amplifies the reach and impact of live performances but also supports the sustainability and growth of the performing arts sector in an increasingly digital world.

Live eSports

Low Latency streaming has revolutionized the eSports industry by providing immediate access to live events and fostering a global community of viewers. Platforms like Twitch, YouTube Live, and Facebook Gaming have become essential for broadcasting eSports competitions, allowing fans to watch and interact with live matches from anywhere in the world. This accessibility has been crucial in increasing viewership and engagement, with hours watched soaring by 75% since 2020[1]. The ability to stream live has made eSports more inclusive, connecting fans who might not have the opportunity to attend events in person and creating a more interactive viewing experience through live chats and social media integration.

Furthermore, Low Latency streaming has opened new revenue streams for the eSports industry. By reaching a wider audience, esports organizations can monetize through various methods such as virtual tickets, advertisements, and exclusive online content. The integration of advanced analytics allows for better understanding of audience preferences, enabling targeted marketing and personalized experiences. Co-streaming, where individual streamers broadcast official events, has also contributed to higher viewership numbers, leveraging the influence of popular content creators to draw in larger audiences. This blend of accessibility, engagement, and monetization through Real-Time streaming has been pivotal in propelling the esports industry to new heights.

1 www.demandsage.com/esports-statistics/

Live Sports

Low Latency streaming is a goal for sports OTT broadcasting as it provides a competitive edge to licence holders over other streaming services that use outdated videoflows, eliminating the typical 30-45 second end-to-end latency disadvantage that HLS and DASH suffer in comparison to faster terrestrial or satellite broadcasts (such as Digital Video Broadcasting DVB-S or DVB-T). This subject is of increasing interest for sports OTT broadcasters as the importance of preventing spoilers and enhancing viewer engagement on second screens and over social media continues to grow.

Sports Betting

Real-Time streaming transforms the user experience by enabling play-by-play betting on events such as tennis (per serve) and boxing (per round). It also allows for the betting window to stay open longer and closer to the start of the event for increased betting activity and revenue as seen for the opening of the gates in horse racing.

This immediacy significantly increases per-event revenue and helps decrease betting advantages by ensuring that all bettors receive the same information at the same time.

The use of Real-Time technologies creates immersive fan experiences, instant access to micro-betting opportunities, and data-driven insights. Social interactivity is also enhanced, creating new market opportunities for betting platforms. The ability to place bets on live events as they unfold - such as predicting the next goal or the outcome of a specific play - depends on minimizing latency to ensure all participants receive information simultaneously. Real-Time streaming supports these dynamic betting opportunities, extending betting windows and increasing per-event revenue. When combined with real-time analytics, it further enhances decision-making and engagement, improving the overall user experience.

Virtual Live Events

Low Latency is crucial for virtual live events because it ensures real-time interaction and engagement, which significantly enhances the experience of the virtual attendees. This immediacy is vital for maintaining the flow and engagement of the event, allowing participants to react and interact in real-time without noticeable delays.

The benefits of Low Latency extend to various types of virtual events, including corporate webinars and virtual congresses. Low Latency facilitates real-time participation through live chats, networking exercises, polls, and collaborative activities, making the experience more immersive and the connections made more significant.

06 The Latency Definitions

The lack of standardized terminology in the industry has caused considerable confusion. One of the most common misconceptions is the agreement on the latency related to a specific term. The LL-WG aims to clarify and standardize the use of the below terms in relation to Low Latency. This specifically relates to delivery only and not in relation to glass-to-glass. This is partly based on research done by the LL-WG amongst 50 vendors of Low Latency services and solutions in 2024. Although it is expected that no classification can capture every specific use case and not every organization/individual will agree, it is believed this is seen as a common guideline to use.

Real-Time Streaming (RTS) (0 - 0.5 seconds glass-to-glass)	Typically used for real-time (interaction) applications like sports betting, live casino gaming, live meetings with live audience participation, live shopping, and auctions. Common technologies used include **WebRTC** and **WebRTS** and there is much discussion for **MoQ** when it can be put to production.
Sub-Second Streaming (SSS) (0.5 - 1 seconds)	Typically used for both real-time (interaction) applications like RTS but also used for live events related to social media such as live sports, live esports and non-interactive, data-related applications and participating audiences such as town hall meetings or trivia-type setups. Common technologies used include **HESP**, **Websockets**, **WebRTC**, and **WebRTS** as there is much discussion for **MoQ** when it can be put to production.
Ultra-Low Latency (ULL) (1 - 3 seconds)	Typically used for non-interactive live events related to social media such as live sports, live esports and data-related applications such as OTT Live Sports, OTT Broadcasting and Live Events. Common technologies used include **HESP**, **LL-DASH**, **LL-HLS**, **Websockets**, **WebRTC**, and **WebRTS** as there is much discussion for **MoQ** when it can be put to production.
Low Latency (LL) (3 - 12 seconds)	Typically used for any live events and linear channels that relate to regular broadcasting and less commonly live sports events. Common technologies used are **LL-HLS** and **LL-DASH** as well as **optimized HLS** and **optimized DASH**.

Standardizing these terms will help eliminate misconceptions and provide a clear framework for the industry. It is essential for all stakeholders, including content providers, technology vendors, and end-users, to have a common understanding of what each latency tier represents and the corresponding use cases. This clarity will facilitate better decision-making and more effective implementation of Low Latency streaming solutions and services.

Furthermore, the LL-WG emphasizes that these technologies are scalable and capable of supporting large-scale deployments in risk-averse broadcast environments. This counters the belief that Low Latency streaming solutions and services are niche, immature technologies requiring bespoke deployments. By highlighting successful large-scale implementations, the LL-WG aims to demonstrate the robustness and reliability of these technologies, encouraging broader adoption across the industry.

Low Latency streaming is pivotal across various sectors, enhancing user engagement and driving revenue. Understanding the differences between the technologies and standardizing terminology will pave the way for broader adoption and a more effective implementation of Low Latency technologies. With the advancements and clarifications provided by the CDN Alliance LL-WG, stakeholders can confidently deploy scalable and resilient Low Latency solutions and services to meet the demands of modern broadcasting and interactive applications.

07 Technologies that support Low Latency

There is a lot of confusion on the differences between the terms Real-Time Streaming (RTS) Sub-Second Streaming (SSS), Ultra-Low Latency (ULL) and Low Latency (LL). Especially on how this relates to the available technologies and what typical use cases are supported with such technologies and terms. The below table will give the overview on how to relate all those to determine what technologies are useful to use with your use case and as such what terms are relevant to use as part of it.

Technology	Low Latency (LL)	Ultra-Low Latency (ULL)	Sub-Second Streaming (SSS)	Real-Time Streaming (RTS)
HESP	Yes	Yes	Yes	No
HTTP(S) 1.x / 2.x (LL-HLS/LL-DASH)	Yes	Yes	No	No
HTTP(S) 3.x (LL-HLS/LL-DASH)	Yes	Yes	Yes	No
Media over QUIC (MoQ)	•	•	•	•
WebRTC	Yes	Yes	Yes	Yes
WebRTS	Yes	Yes	Yes	Yes
WebSockets	Yes	Yes	Yes	Yes
Use-cases				
iGaming w. Live Dealers	No	No	Yes	Yes
Live Auctions	No	No	No	Yes
Linear Broadcasting	Yes	Yes	Yes	Yes
Live Concerts/ Performances	Yes	Yes	Yes	Yes
Live eSports	Yes	Yes	Yes	Yes
Live Shopping/ E-business	Yes	Yes	Yes	Yes
Live Quiz and Trivia Games	No	No	Yes	Yes
Sports Betting	No	No	Yes	Yes
(Interactive) Live Events	No	No	No	Yes
(Interactive) Virtual Live Events	No	No	No	Yes

• Not in production for delivery as of the date of this publication

08 Comparing Streaming Technologies

As Low Latency (LL) and Ultra-Low Latency (ULL) are commonly achieved by using well-known technologies like HLS, DASH, LL-HLS and LL-DASH, Real-Time and Sub-Second technologies are often less familiar. A description of the most important of these technologies can be found below.

HESP

HESP (High Efficiency Streaming Protocol) is a proprietary TCP-based solution protocol developed by THEO Technologies (since acquired by Dolby in 2024) which includes a royalties schedule and was launched in conjunction with the HESP Alliance members in 2020 aimed at achieving real-time latencies. While it may not match WebRTC's speed in many instances, HESP is an HTTP-based TCP protocol that offers better compatibility with network security measures and can work seamlessly with traditional CDNs.

HESP is particularly suited for large-scale broadcast environments where traditional CDNs are used to distribute content. Its ability to work efficiently over HTTP enables broadcasters to leverage existing infrastructure, reducing the complexity and cost of deployment.

Additional reference:
Wikipedia

Media over QUIC

Media over QUIC (MoQ) is an innovative protocol under active research and development which has not been put to production for delivery to viewers as of the date of this publication and is meant to enhance Low Latency media delivery, leveraging the capabilities of the QUIC transport protocol. MoQ aims to bridge the gap between high-latency, scalable streaming services and Low Latency, real-time communication tools. By building on QUIC's efficient handling of streams and datagrams, MoQ ensures minimal delay in media transmission, making it ideal for live streaming, gaming, and video conferencing applications.

MoQ operates by creating a single protocol for both the ingestion and distribution of media, eliminating the need for intermediary repackaging. This approach allows for efficient error recovery and scalability. Media relays are used to cache and distribute content, reducing the distance data must travel and thereby decreasing latency. Additionally, MoQ supports end-to-end encryption and flexible rate adaptation strategies to maintain high-quality media transmission under varying network conditions.

MoQ is still in development, and is the only technology on this list that is not yet truly production-ready (for example, industry-standard features, such as multibitrate rendition ladders, are not yet supported). However, MoQ has been well-received by the industry at large due to the backing of many established companies.

WebRTC

WebRTC (Web Real-Time Communication) is an open-sourced UDP-based solution released by Google in 2011 designed for real-time communication. It excels in scenarios requiring sub-500 ms latency, such as video conferencing and live streaming. WebRTC can be deployed in such a way as to handle spotty network conditions effectively, providing Low Latency and resilient video even under less-than-ideal circumstances. Most WebRTC providers use adaptive bitrate technologies to prevent drops and interruptions, making it a preferred choice for real-time applications.

Industries such as telehealth, remote learning and live customer support industries have adopted WebRTC to facilitate long-distance interactivity. Additionally, WebRTC's native integration with web browsers without the need for plugins simplifies its deployment and broadens its accessibility.

Additional reference:
Wikipedia

WebRTS

WebRTS (Web Real-Time Streaming) is an open-sourced Real-Time streaming framework for both TCP and UDP developed by Ceeblue and publicly announced in 2024 which is designed to deliver sub-500 millisecond end-to-end latency with high stability and minimal artifacts. It addresses network congestion through advanced frame-skipping techniques, thereby optimizing Quality of Service (QoS) in real-time scenarios. WebRTS supports both adaptive as well as fully reliable modes.

In terms of compatibility, WebRTS is transport, protocol, and codec agnostic to ensure broad support across various workflows. The framework introduces a new containerless format that reduces network load by 5% compared to CMAF, resulting in significant bandwidth savings. Its efficient demuxer and playback engine further contribute to reduced origin load, making WebRTS a cost-effective solution for Real-Time streaming. WebRTS is compatible with traditional CDNs and supports DRM.

WebSockets

WebSockets are highly effective for Low Latency video streaming due to their ability to maintain a continuous, bi-directional connection between client and server. This persistent connection minimizes the overhead of repeatedly establishing new connections, thereby reducing latency significantly.

In video streaming, WebSockets enable real-time transmission of video and audio data with minimal delay, making them suitable for interactive applications such as live auctions, online betting, and virtual events. The technology supports adaptive bitrate streaming, ensuring a high-quality viewing experience.

09 Balancing Quality and Latency

The particular challenge of Low Latency always supposes the balancing of latency and bitrate; reducing latency often means compromising on bitrate, power consumption (codecs), and video quality. It's important to note that achieving both Low Latency and high bitrates is still a challenge, especially at scale, and there is no one-size-fits-all solution as the choice of technology depends on the specific use case.

QoE gauges viewer satisfaction with a streaming experience, based on factors like latency, video quality, buffering, and network metrics. While standard QoE measurements might offer general insights, specific technologies may necessitate a bespoke or proprietary QoE system for a more precise evaluation.

The table below shows the standard prioritization of the different last-mile delivery technologies:

	HESP (TCP)	**LL-DASH** (TCP)	**LL-HLS** (TCP)	**MoQ** • (UDP)	**WebRTC** (UDP)	**WebRTS** (TCP&UDP)	**WebSockets** (TCP)
Priority	Quality	Quality	Quality	Latency	Latency	Quality	Latency

• Not in production for delivery to viewers as of the date of this publication

10 How about DRM and Low Latency?

Security when streaming video is important to ensure content cannot be accessed, redistributed, pirated, or altered by unauthorized parties. A lot of use cases also relate to either monetization such as live streaming of live sports or esports events, concerts, auctions and betting or are confidential in nature, such as closed (interactive) webinars for companies, education, faith, and so on. These characteristics require a good line of defence. While it is beyond the scope of this document to highlight and compare all the possible security capabilities of all technologies available, the security measure most demanded by commercial low-latency applications is generally Digital Rights Management (DRM).

DRM is a widely recognized method of content encryption. It uses third-party keys, delivered through a separate mechanism, to decrypt content for live events only for users that are eligible to watch. It is mostly in relation to content that is monetized, such as live (e)sports and concerts, and is also very commonly used for video-on-demand streaming, including downloads as well. With the shift to new Low Latency technologies as an alternative for live streaming, it becomes less clear if and how DRM is supported. The table below provides an overview of the currently available technologies, indicating whether DRM is supported and which technologies can be used with DRM.

	HESP	LL-DASH	LL-HLS	Media over QUIC (MoQ)	WebRTC	WebRTS	WebSockets
Full DRM support ●	Yes	Yes	Yes	No?	Yes	Yes	No
Encryption Support ●	Yes	Yes	Yes	Yes?	Yes	Yes	Yes
Native DRM-support ●	No	Yes	Yes	No?	No	Yes	No
Vendor DRM-support ●	Yes	Yes	Yes	Yes/No?	Yes	Yes	Yes
3rd-party DRM -support ●	Yes	Yes	Yes	No?	Yes	Yes	No
Open player/ client support ●	No	Yes	Yes	No?	No	Yes	No

- ● Does the technology support a mechanism of full DRM with encryption of the content with key for decryption to be delivered separately to a player/client based on a DRM-license?
- ● Does the technology support a mechanism of (a minimum of 128-bit) encryption of the content with a key to be delivered to a player/client based separately?
- ● Does the technology support native DRM-support with a separate license server that can be used in conjunction with the encryption?
- ● Is the technology supported by vendors that offer proprietary DRM-support (in combination with their own player/client) in order to support DRM for the given technology?
- ● Is the technology supported by 3rd-party DRM-vendors that offer DRM-support (as a service or solution) for the given technology?
- ● Does the technology support the use of open clients/players (such as open source) to use DRM with?

11 Video Players, CDNs & Low Latency

Traditional CDNs are vital for large-scale internet broadcasting and have been successful for Low Latency and Ultra Low Latency applications. Real-Time and Sub-Second latencies are currently delivered by specialty CDNs for the most part. For Real-Time Streaming, small broadcasts are often supported by cloud-based software solutions, while large broadcasts require traditional CDNs that utilize networks supporting Edge Computing, HTTP/3, Sockets, etc. Regular Low Latency streaming is widely supported by most modern CDNs using technologies like HTTP/1.1 CTE, and HTTP/2 PUSH.

For HTTP-based protocols, Low and Ultra-Low Latency delivery requires special options such as 'Chunked Transfer Encoding' or 'Continuous Transfer,' and playback is facilitated by numerous players such as dash.js, hls.js, video.js, or native players for protocols such as HESP. WebRTC-based delivery involves additional media servers and possibly native or WHEP-enabled players, while other protocols may require their own specific transmission and playback logic, typically provided by vendor solutions. The iOS Safari browser has been known to often cause issues due to its numerous streaming technology limitations. However, with iOS 17.1 and the introduction of 'Managed Media Source' support, playing videos in iOS Safari is now much simpler. The table below briefly shows the possibilities for each technology, the more advantages (+) the better.

	HESP	LL-DASH	LL-HLS	MoQ	WebRTC	WebRTS	Web Sockets
Scalability	+++	++	++	•	++	+++	+
ABR	+++	+++	+++	•	++	+++	+
Cross-platform playback	++	+++	+++	•	+++	+++	+

• Not in production for delivery as of the date of this publication

12 Conclusion

In conclusion, the work of the CDN Alliance's LL-WG is vital to fostering clarity, standardization, and innovation in Low Latency streaming technologies. By addressing ambiguities in terminology, defining use cases, and comparing protocols, the working group aims to create a shared framework that empowers stakeholders to make informed decisions. The insights provided in this document highlight the transformative potential of Low Latency solutions and services across industries and underline the importance of collaboration in overcoming technical and operational challenges. As the working group continues to promote consensus and best practices, it plays a crucial role in driving adoption, ensuring scalability, and enabling resilient, high-quality streaming experiences that meet the evolving demands of modern audiences.

13 Glossary

ABR | Adaptive Bitrate A streaming technique that adjusts video quality in real-time based on network conditions.

CDN | Content Delivery Network A network of distributed servers that deliver web content to users based on their geographic location.

CMAF | Common Media Application Format A standard designed to simplify the packaging and delivery of HTTP-based streaming media.

CTE | Chunked Transfer Encoding A data transfer mechanism that splits content into chunks. It allows streaming of dynamic content ranges without requiring knowledge of the total size or duration.

DASH | Dynamic Adaptive Streaming over HTTP Also known as MPEG-DASH, is an adaptive bitrate streaming standard which became an International Standard in 2011.

Downstream Refers to components of the videoflow closer to the end-user (viewing audience) of the video.

DRM | Digital Rights Management Encryption technologies mostly referring to the application of FairPlay, PlayReady, and Widevine to HLS and DASH streams to help prevent pirating and other unauthorized access to video content.

DVB | Digital Video Broadcasting A standard for TV broadcasting. DVB-S for satellite, DVB-C for cable, DVB-T for terrestrial.

HESP | High Efficiency Streaming Protocol A proprietary protocol designed for efficient streaming of high-quality video content in low latency launched by THEO Technologies in 2020 in a partnership with Synamedia (Theo was acquired by Dolby in 2024).

HLS | HTTP Live Streaming An adaptive bitrate streaming standard developed by Apple Inc. and released in 2009.

HTTP | Hypertext Transfer Protocol The foundation protocol of data transfer on the web.

LATENCY The measurement in time for a video frame to be delivered from point A (the source location) to point B (the delivery location) over the Internet.

LL | Low Latency Used to describe a service, protocol, technology, experience, etc., that has lower latency than standard HLS and DASH. Used to refer to streaming with less than 12 seconds of latency.

LL-DASH | Low Latency MPEG Dynamic Adaptive Streaming over HTTP An version of DASH for reduced latency using smaller and partial segment sizes. See also: DASH.

LL-HLS | Low Latency Apple HTTP Live Streaming A version of HLS for reduced latency using smaller and partial segment sizes . See also: HLS.

LOW LATENCY Commonly used in an unspecific way when referring to latencies lower than those commonly achieved by delivery of HLS and DASH.

MoQ | Media over QUIC A streaming protocol designed to enhance low latency media delivery using QUIC transport protocol. See also: QUIC.

OTT | Over-The-Top The delivery of media content over the Internet directly to viewers.

QoE | Quality of Experience A measure of user satisfaction, usually through user-centric metrics (e.g. video quality, startup delay, buffering, etc.).

QoS | Quality of Service A measure of service performance (e.g. bandwidth, latency, jitter, packet loss, etc.).

QUIC | Quick UDP Internet Connections A transport layer network protocol designed to improve the performance of web applications. See also: UDP.

RTS | Real-Time Streaming Streaming with less than 500 milliseconds of latency.

SRT | Secure Reliable Transport A protocol for the transport of secure, low-latency video streams.

SSS | Sub-Second Streaming Streaming with less than 1 second of latency.

TCP | Transmission Control Protocol A transport layer network protocol.

UDP | User Datagram Protocol A transport layer network protocol.

ULL | Ultra-Low Latency Streaming with less than 3 seconds of latency.

Upstream Refers to components of the videoflow closer to the originating source (broadcast) of the video.

WebRTC | Web Real-Time Communication An open-sourced UDP technology enabling real-time communication and streaming to end-users through web browsers and applications released by Google in 2011.

WebRTS | Web Real-Time Streaming An open-sourced framework and machine-learning algorithm for real-time TCP and UDP streaming announced by Ceeblue in 2024.

WHEP | WebRTC HTTP Egress Protocol A protocol for streaming WebRTC content over HTTP.

WHIP | WebRTC HTTP Ingest Protocol A protocol for ingesting WebRTC content over HTTP.

Colophon

Title Low Latency Streaming | A First Step Towards Standardization
Version 1.0
Publication Date July 2025
Published by CDN Alliance

Editing & Final Review CDN Alliance

Contact

www.cdnalliance.org

info@cdnalliance.org

CDN Alliance

Version History Version 1.0 - July 2025 (initial release)