Low Lag Solution

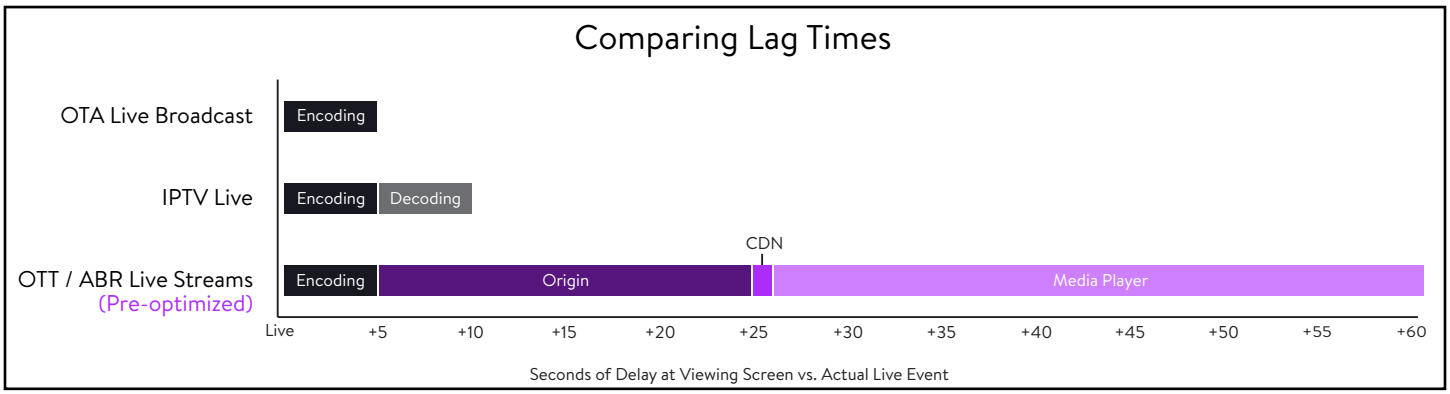
Lag Reduction for Live ABR Streaming

VELOCIX™

Velocix Origin Servers employ advanced techniques to reduce lag for live content delivered using HTTP adaptive bit rate protocols. A summary of the challenges and considerations associated with lag reduction are highlighted below.

The Problem

When viewing live television, consumers have a sense that the program they are enjoying is being watching by everyone at precisely the same time. In reality, while the original content comes from a live source, it traverses many different systems on its way to the consumer's viewing device, and each of those systems introduces a small to moderate delay. The near instant delivery we experienced in the days of analog TV when the workflows were simpler, turned into several seconds of delay with the advent of digital TV, and then roughly doubled when IPTV was launched, due in part to features like fast channel change. With the advent of HTTP adaptive bit rate (ABR) streaming and over the top (OTT) services the delay has only gotten worse. The gap between the time a program is viewed and the time the actual live event occurred can now be as long as a minute or more for streamed content.



Is this lag is a real problem? In most cases, it is not. When consumers watch live TV on a single screen, there is no frame of reference to suggest there is a time delay occurring. Consumers generally watch live programming unaware that they are viewing something that may have happened minutes earlier. However, there are certain circumstances under which lag becomes a real problem. Let's take the example of a consumer watching a football match on the TV using an OTT streaming service with a lag signature of one minute relative to the live source. If that consumer lives in an apartment building where one or more neighbors are accessing the same program using a different technology (ex. a digital satellite connection or over-the-air transmission), each with its own lag signature, the timing variance between these delivery methods will become obvious to everyone when the first point is scored. Noisy neighbors provide a frame of reference that exposes the intrinsic lag in ABR delivery and ruins the viewer's illusion that the content is being watched live. The effect of knowing someone scored before seeing the actual event on the screen can severely detract from the consumer's enjoyment. Therefore, time lag for live viewing can be a serious issue and considering the high viewership rates related to these instances (ex. live sports events) this is an important problem for video service providers to consider as they design and launch a new streaming service.

Analysis

When analyzing the lag in an ABR environment, one must first pinpoint how and where significant delays occur in the content delivery workflow. While there are delays introduced at many steps in the video distribution chain, the most significant ones can be found baked into the ABR protocol itself. ABR was originally designed with OTT services in mind, specifically streamed delivery of real time multimedia content over unmanaged networks (i.e. the Internet). The ABR protocol presumes:

- Video service providers have no control over the performance of the network.
- Network performance is highly variable and changes over time.
- Consumers will sacrifice video quality for stream continuity.

The protocol was therefore designed to accommodate significant fluctuations in the performance of the network and to prioritize stream continuity over video quality. Notably, lag time was not a primary consideration, as reducing lag tends to work counter to the goals of stream continuity.

If we dig deeper into the ABR content formats and protocol we can begin to understand the source of increased lag durations for live streaming. Video content prepared for ABR delivery is first time-sliced into discrete segments that are 2 to 10 seconds in duration. Segments are then encoded at different bit rates to create several versions of the content ranging from low to high quality. When playback of ABR content is initiated by a consumer, the consumer's viewing device is responsible for selecting the right video segment to request from the CDN based on current network performance. When network bandwidth is measured to be low, the client will automatically request low quality segments. As network bandwidth improves, the client will begin to request higher quality versions. Since segments are time-sliced, the visual quality of a stream can change every 2 to 10 seconds (the duration of one segment) during playback in response to changes in network performance during the viewing session. In order to make this negotiation process as smooth as possible, viewing devices also take advantage of client-side buffering. Most consumer devices contain some level of solid state memory (RAM), which can be used to buffer video segments in advance of playback. The buffer serves as a margin of safety in the event the network connection is temporarily interrupted, giving the device a little time to re-establish the connection and avoid presenting a "please wait" or spinning hourglass symbol to the consumer.

While segment size and client-side buffering are the most significant contributors to ABR lag, additional delays are generated during the content delivery process. Encoders, the first step in the delivery chain, can add up to 5 seconds of lag to live streams, however this delay is not unique to ABR delivery. Encoding lag is common across all TV delivery methods, so it doesn't contribute to the timing variance consumers might notice between one screen and another. Moving on to the next step in the delivery chain, Origin servers are a common source of lag due to embedded processing tasks such as ABR packaging, digital rights management encapsulation, and the writing of content to persistent storage in support of time-shifted TV services. Depending on the software implementation, origin processing tasks can add up to 2 full segment durations of delay or more. Further downstream in the workflow is the CDN. Unlike origins, CDNs are not a major source of lag. Processing within CDN caches is minimal and while some delay is introduced due to application (HTTP) and transport (TCP) level protocols, added delays from the CDN typically amount to less than a second in aggregate.

In addition to the buffering delays, players can introduce about a half a segment size of delay on average during live edge detection as a result of time synchronization between the creation of the segment and the reading of the manifest.

Armed with this information, we can calculate the expected time delay relative to the live source by summing up the common contributors to lag as follows:

Lag Calculation for Live ABR Streams (Before Optimization)

Sources of Lag	Added Delay (10 second segment size)
Encoder*	~5 seconds*
Origin (Processing)	~20 seconds
CDN	<1 second
Media Player (Client Side Buffering + Edge Detection)	~35 seconds
Total	**~61 seconds (~56 seconds after encoder)**

*Common to all video delivery methods

In summary, if we remove the effects of the encoder, which apply equally to all video delivery mechanisms, the time gap between ABR delivery and legacy broadcast delivery of live content can be as high as 56 seconds. This degree of delay would be very noticeable to a consumer if they are within earshot of another screen that is being served using more traditional delivery mechanisms. Such lag could prove to be very disruptive to the consumer's viewing experience and might be sufficient to cause customer dissatisfaction. Fortunately, there are several ways this delay can be reduced.

Low Lag Solutions

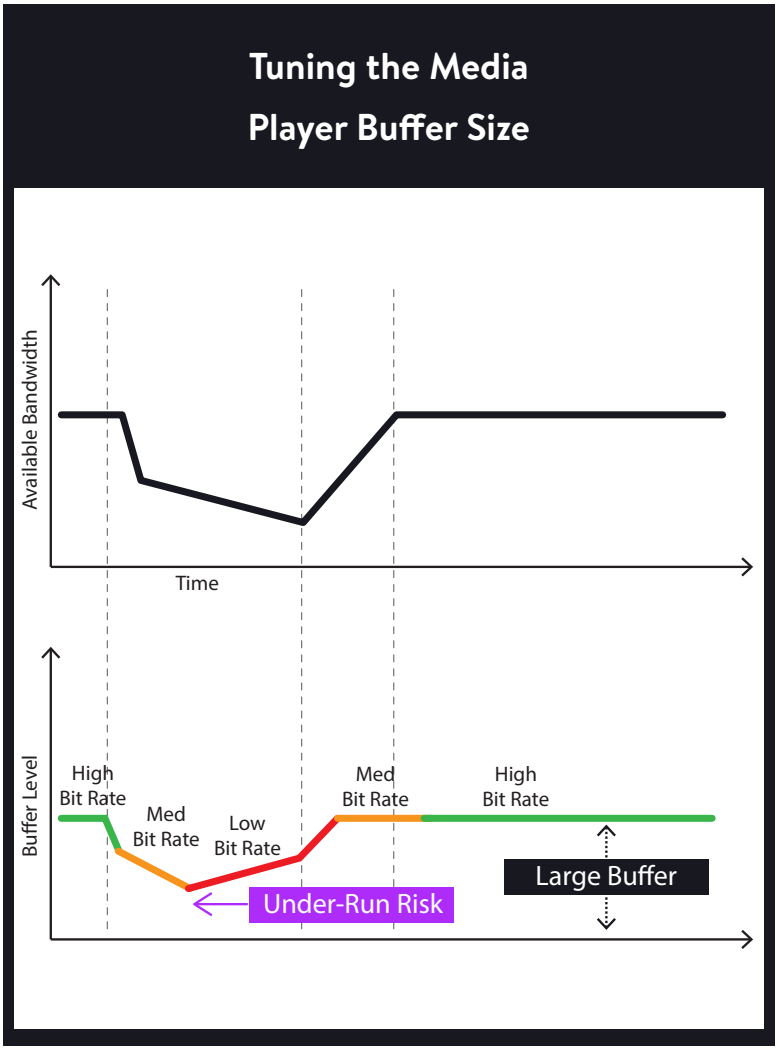
To achieve significant reductions in lag times for live ABR streaming, several approaches can be taken to minimize the delays at each step in the workflow. These approaches can be employed individually or used in concert with each other to progressively improve performance.

- **Reducing Segment Duration**

  One quick and easy way to reduce lag is to decrease the segment duration for ABR content. Shortening the segment size can improve origin processing, client-side buffering, and edge detection delays by orders of magnitude. For example, reducing the segment size from 10 seconds to 2 seconds will generate a 5x improvement at points where lag is tied to segment duration. There are trade-offs to this approach, however. Most importantly, using smaller segment sizes will diminish the protections designed into the ABR protocol to account for network fluctuations. In other words, when network bandwidth changes, the likelihood of stream interruptions will be higher because the margin of safety is lower with small segments. Service operators therefore need to carefully consider the network performance characteristics before taking this approach. For services delivered over public Internet or mobile networks, dramatically lowering segment duration may not be a viable option. However, for service providers that deliver ABR streams over their own managed network, the more stable performance characteristics of the network make reducing segment size a good option for significantly reducing lag. Service providers that use their own CDN should also account for the fact that reducing segment size will increase the number of files managed within the CDN. Care must be taken to choose a segment size that strikes the right balance between lag improvements and service reliability.

- **Media Player Modifications**

  Given a large proportion of lag is introduced by the media player on the consumer's viewing device, improvements are easier to achieve when the software on the device can be modified. When this is the case, Media player software can be enhanced to change the client-side buffering logic. By attaching to the very latest segment in the manifest rather than buffering several segments before playback, lag can be reduced by the duration of as many as 2 segments (e.g. ~20 seconds using our example). Player software can be further modified to start playback at an advanced point within the last segment, instead of at its start. This is only possible if the segment has multiple random access points (e.g. intra frames) within it, and it should be done taking into account the creation time of the segment. The player must be careful not to start from a point in the segment that will produce a buffer under-run due to the next segment not being available in time.



- **Origin Processing Enhancements**

  Reducing the processing delay in the origin/packager can speed up both the creation and delivery of ABR manifests and segments. Specialized tuning can yield lag reductions as much a 2 segment durations (~20 seconds). Potential side effects must be considered in the tuning implementation, however, as errors resulting from very fast segment requests must be handled properly by downstream system components. Advanced techniques such as delivering a segment from an origin as it is still being built, can further reduce the origin server component of lag to seconds or less. The server can make available pieces of the current chunk (sub-segments) as they become available, and signal them appropriately. As with the other lag reduction options, care must be taken to test enhancements under real-world conditions to ensure that stream quality and continuity are not negatively affected as a result of overly aggressive tuning.

By intelligently implementing a combination of the aforementioned lag improvement techniques, delay times associated with the viewing of live ABR content can be brought within range of legacy TV delivery mechanisms. Running a calculation of lag contributors after optimization, showcases the magnitude of the difference that can be realized by combining the various techniques.

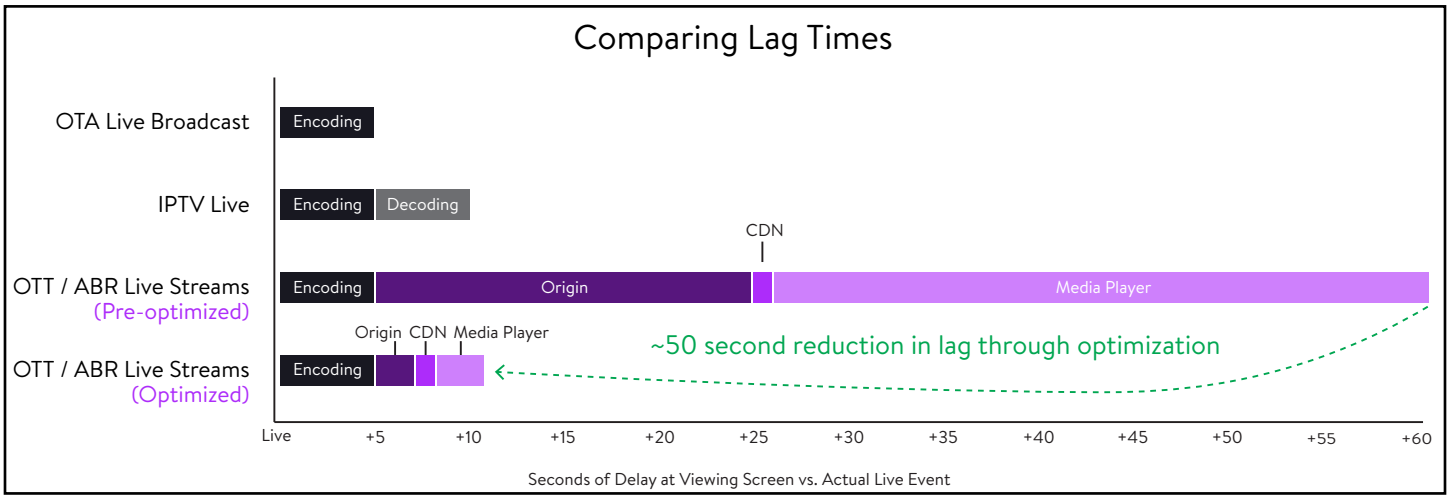
Lag Calculation for Live ABR Streams (After Optimization)

Sources of Lag	Added Delay (2 second segment size)
Encoder*	~5 seconds*
Origin (Processing)	~2 seconds
CDN	<1 second
Media Player (Client Side Buffering + Edge Detection*)	~3 seconds
Total	**~11 seconds (~6 seconds after encoder)**

*Common to all video delivery methods

The Future of Low Latency Delivery

There are several emerging standards for low-latency ABR delivery, such as CMAF and Low Latency HLS (LHLS) that strive to formalize some of the methodologies discussed in this paper, as well as introduce other advanced techniques. These standards hold a lot of promise, however it is unclear how long it will take for them to be broadly adopted. What is clear is that video service providers can take steps today to improve latency for live ABR streaming and the benefits of doing so can be fairly dramatic.



In Conclusion

In conclusion, reducing lag times for live content delivered using HTTP ABR protocols is important to improving the quality of experience for consumers. There are a variety of techniques that can be used to accomplish this, each carrying benefits and trade-offs that need to be considered when architecting a live streaming service. Velocix has tuned the performance of its origin server to minimize lag and proven the benefits in real-world scenarios. By working closely with service providers to optimize origin and client configurations, Velocix can make Live TV viewing on connected screens more enjoyable for consumers.

Contact a Velocix salesperson to learn more.