Low-latency Live Streaming at Scale

Yueshi Shen | August 2018, Denver
About Speaker

Yueshi Shen
Principal Research Engineer, Twitch

- In charge of Twitch’s core video technologies: codec, HLS, ABR playback, video quality, low latency
- Inventor of AV1's Switch-Frame and has filed 15 patents
- Oversees Twitch’s transcoder team
- Principal Engineer at OnLive, worked on ultra low latency (<100ms) video streaming
- Earlier career in the linear TV and semiconductor industries, worked on MPEG-2/H.264 ad splicing, ASIC H.264/SVC codec

yshen@twitch.tv
What is Twitch.tv?
World’s largest live-streaming platform
Twitch In a Nutshell:
Easy video broadcasting, Interactive live streaming
Content on Twitch.tv

Interactive Live & VOD
Games, eSports, original content
Community for gamers & creative arts
Social video platform
Twitch’s Scale Numbers, 2017

15+ MILLION unique daily visitors
2+ MILLION unique monthly broadcasters
355 BILLION minutes watched
Xbox E3 2018 briefing sets Twitch record of 1.7 million concurrent viewers

Microsoft pulls ahead again.

E3 2018 was an absolute blast for gamers because a lot of exciting titles were announced. Publishers like Microsoft and Sony showcased incredible experiences like Gears 5 and a Resident Evil 2 remake.
Twitch’s Growth

Alexa Rank History:

<table>
<thead>
<tr>
<th>Year</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011/7</td>
<td>411,522</td>
</tr>
<tr>
<td>2012/1</td>
<td>2,344</td>
</tr>
<tr>
<td>2014/4</td>
<td>267</td>
</tr>
<tr>
<td>2017/1</td>
<td>73</td>
</tr>
<tr>
<td>Now</td>
<td>32</td>
</tr>
</tbody>
</table>

We are the 3rd most popular video website behind YouTube and Netflix.
Interactivity:
Twitch’s differentiation from linear TV
Live Demo & Broadcaster’s Feedback

Piebo VOD recording: https://www.twitch.tv/videos/279506342

Takarita clip: https://clips.twitch.tv/TemperedIntelligentCheddarRiPepperonis
Traditional Live Broadcast vs. Interactive Live Streaming
Twitch Extension:
Let viewers help streamers who are playing the game as Show Directors
Low Latency Facilitates Communication

- **_absolute** latency matters
- **relative** latency matters

<table>
<thead>
<tr>
<th>Milliseconds</th>
<th>Absolute Latency</th>
<th>Relative Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4000</td>
<td>Gaming</td>
<td>Consumption</td>
</tr>
<tr>
<td>2000</td>
<td>Conversation</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>Interaction</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

www.twitch.tv
P2P vs. Broadcast:
Trade-off among latency, scalability, and QoS
Trade-off Based on Applications

Linear TV, OTT Live Streaming

- Low Latency
- High Scalability
- Good QoS
  (High Video Quality, Smooth Playback)
Trade-off Based on Applications

- Linear TV, OTT Live Streaming
- Teleconferencing

Diagram showing the trade-offs between Low Latency, High Scalability, and Good QoS (High Video Quality, Smooth Playback).
Trade-off Based on Applications

- Linear TV, OTT Live Streaming
- Teleconferencing
- Twitch’s Interactive Live Streaming

Low Latency

High Scalability

Good QoS
(High Video Quality, Smooth Playback)
Backend

Video distribution pipeline, still HTTP streaming
HTTP-Streaming-Based Video Contribution/Distribution Pipeline

1. Broadcast tool (output: RTMP)
2. Broadcaster’s ISP
3. Ingest Server
4. Transcoder Server (input: RTMP, output: HLS)
5. Distribution network
6. Viewer’s ISP
7. Cross-platform player
Live Streaming: Pull vs. Push

RTMP: Low scalability & Low latency

HLS: High scalability & High latency
HTTP Chunked Transfer Encoding
## Where Are My Milliseconds?

<table>
<thead>
<tr>
<th>Step</th>
<th>Median Latency without HTTP Chunked Transfer Encoding</th>
<th>Median Latency with HTTP Chunked Transfer Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Broadcast Tool</td>
<td>1s</td>
<td>1s</td>
</tr>
<tr>
<td>2. Broadcaster’s ISP</td>
<td>100ms</td>
<td>100ms</td>
</tr>
<tr>
<td>3. Ingest Server</td>
<td>50ms</td>
<td>50ms</td>
</tr>
<tr>
<td>4. Transcoder</td>
<td>2.5s</td>
<td>0.5s</td>
</tr>
<tr>
<td>5. Distribution Network</td>
<td>200ms</td>
<td>200ms</td>
</tr>
<tr>
<td>6. Viewer’s ISP</td>
<td>100ms</td>
<td>100ms</td>
</tr>
<tr>
<td>7. Playlist</td>
<td>1s</td>
<td>0s</td>
</tr>
<tr>
<td>8. Player buffer</td>
<td>5s</td>
<td>2.5s</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>9.95s</strong></td>
<td><strong>4.45s</strong></td>
</tr>
</tbody>
</table>
Backend Stability Example:
*HTTP request time of a replication site*
Frontend

HTTP streaming, ABR, buffer control
Segment Streaming

Time

Past

Future

Transcoder

Chunk

Chunk

Chunk

Content delivery network

In-progress segment (1 HTTP request)

Player: HTTP receiver, demuxer

Chunk

Player: Buffer

Sample

Sample

Sample

Sample

Transmuxed video data

www.twitch.tv
Adaptive Bitrate (ABR) Playback

playback bitrate = f(download bandwidth, player buffer level, ...)

www.twitch.tv
HTTP Chunked Transfer Encoding = NOISY Input

What is my download bandwidth???
Streams API Support

Summary

Twitch is world’s largest live streaming platform, with 15+ million daily unique visitors and 2+ million monthly unique broadcasters.

Twitch’s live streaming model creates strong interaction between broadcasters and their audiences. **Low latency streaming for scalable broadcast is a different technical problem from low latency streaming for point-to-point communication.**

Achieving low latency along with scalability and QoS requires full control of the entire video ingest/distribution pipeline, as well as innovations in both backend and frontend.

HTTP Chunked Transfer Encoding is utilized across transcoder, origin, replication/edge and player to realize median 4.5s glass-to-glass latency, including <0.5s latency in transcoder and <2.5s latency in player.
Thank You and Questions

Yueshi Shen

yshen@twitch.tv