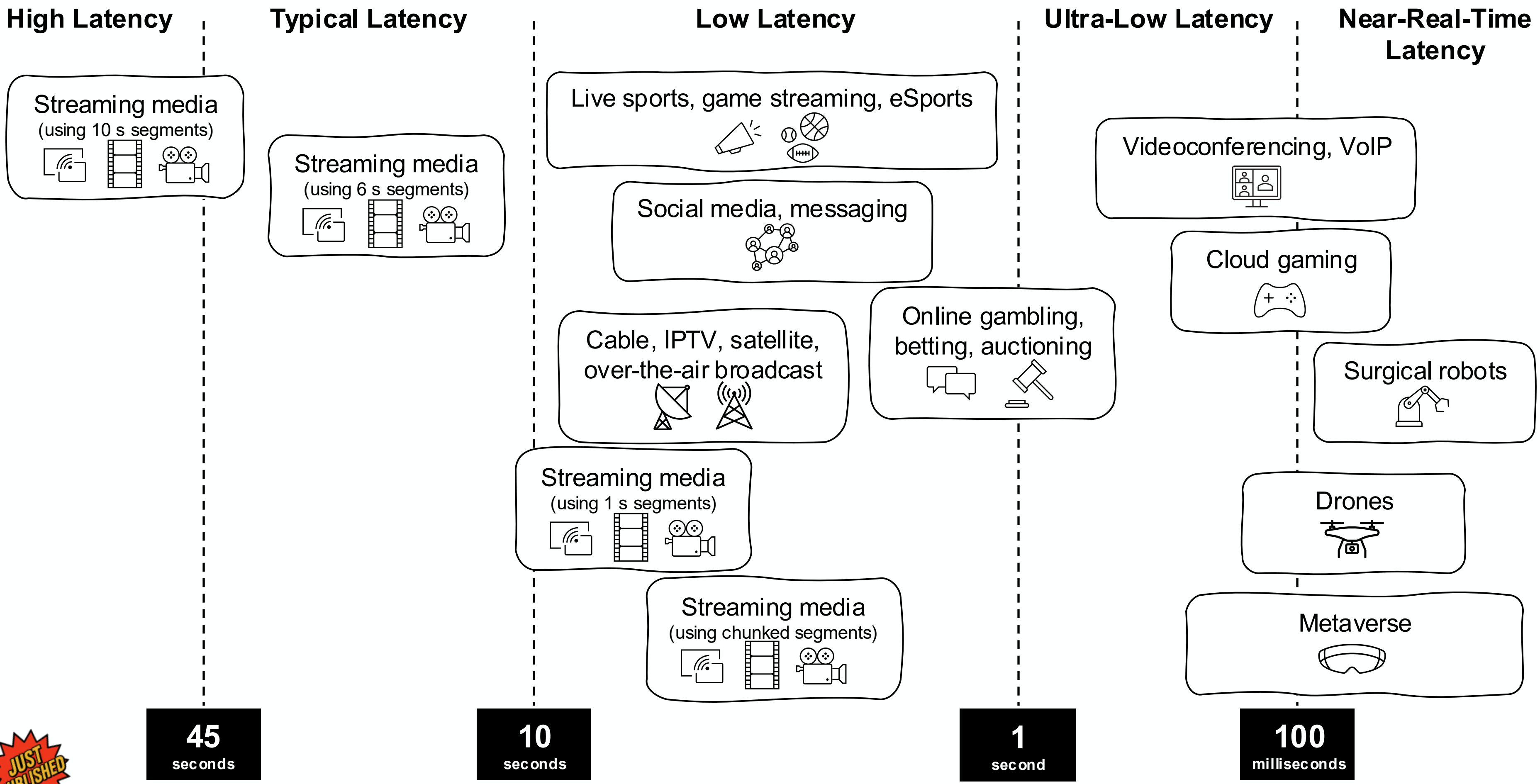
Live and Real-Time Video Delivery

Ali C. Begen

ali.begen.net



JUST PUBLISHED

Reading: Toward one-second latency: evolution of live media streaming – IEEE CS&T (2025) [link, preprint]

1969: The Greatest Single Broadcast in History (650M Viewers)

Image/sound signals transmitted via a lightweight antenna on the top of the lander

Three seconds over
250K miles

Source: BBC News, Science and Media Museum

50 Years Later: Super Bowl LIII

115 cameras including 8K ones, AR powered graphics and many other innovations

45 seconds over five miles

Source: CBS

Super Bowl 2025

Seconds behind on-field action – OTA broadcast average latency: 16-21 seconds

Source: https://blog.phenixrts.com/phenix-superbowl-latency-study-2025


Low-latency Guys
Streaming Providers

Broadcast

- One-to-many distribution
- Single (high-quality) version of a content
- Fixed capacity usage (cable, satellite, IPTV)
- Fixed schedule
- Fixed devices

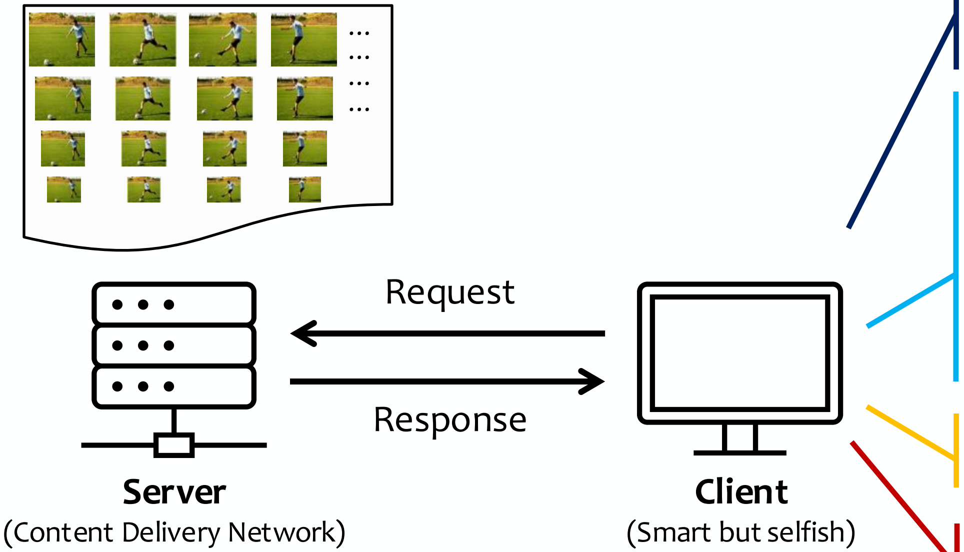
Streaming

- One-to-one distribution
- Multiple (varying-quality) versions
- Variable capacity usage
- Personalized schedule (any content, any time)
- Any device

How can we match or even exceed broadcast quality when streaming?

HTTP Adaptive (Multi-Bitrate) Streaming

ABR (adaptive bitrate) is misnomer



Client

- Uses the OS-provided HTTP stack
- Fetches and parses the manifest/playlist
- Uses decryption tools/keys for the protected content

Client monitors and measures

- Size of the playout buffer (in bytes and/or seconds)
- Chunk download times and bandwidth
- Local resources (CPU, memory, window size, etc.)
- Dropped frames

Client decides what to request next (rate adaptation)

Client measures and reports metrics for analytics

What Matters to Human Eye?

The tradeoffs in QoE modeling

Overall Media Quality

High resolution, frame rate, bitrate
HDR, wide color gamut
Multi-channel audio
No lip-sync issues

Quality Stability

Less switching across resolutions
Less audio switching

Robust Playback

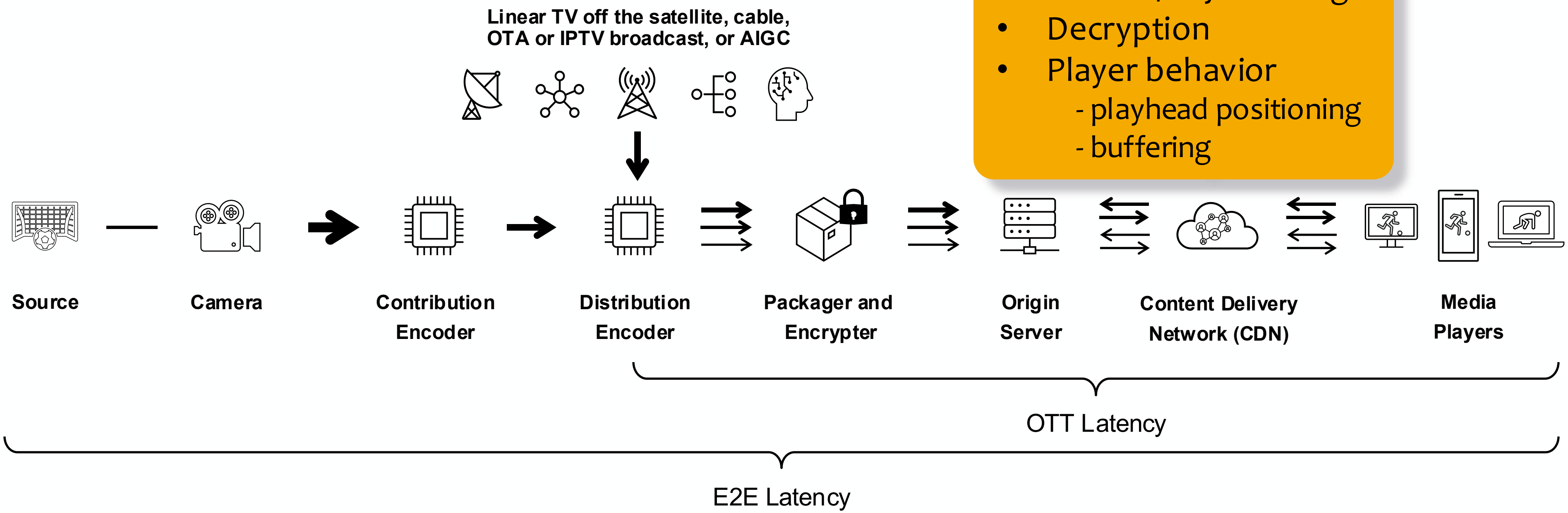
Short/no stalls
Less frequent stalls
Faster startup and seeking

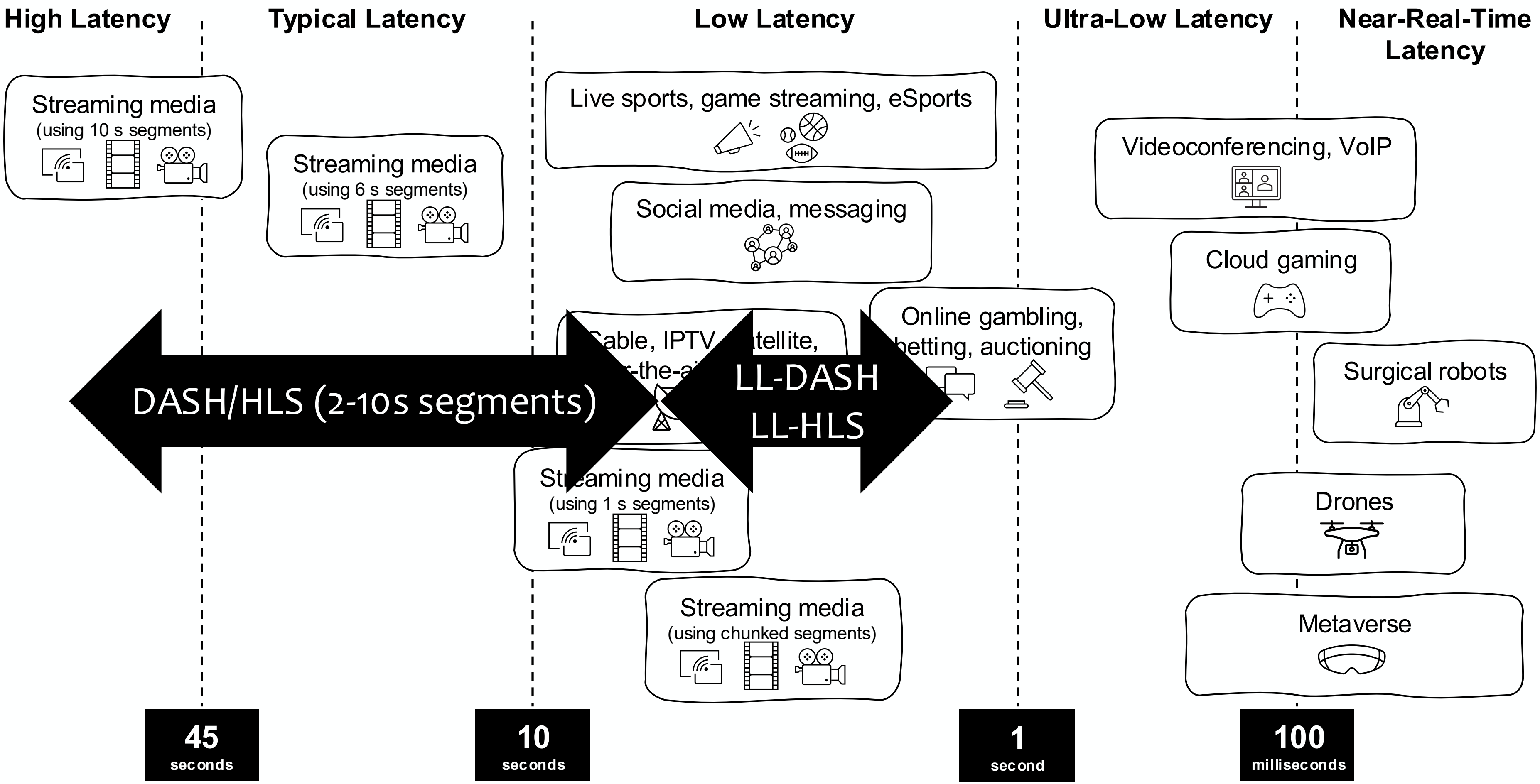
Proximity to Live Edge

Latency on par with traditional broadcast
Low latency to enjoy interaction/social media

End-to-End (Glass-to-Glass) Latency

- Encoding pipeline
- Packaging
- Ingest
- Network propagation
- Server I/O
- License/key fetching
- Decryption
- Player behavior
  - playhead positioning
  - buffering





Reading: Toward one-second latency: evolution of live media streaming – IEEE CS&T (2025) [link, preprint]

Twitch's Grand Challenge on

Adaptation Algorithms for Near-Second Latency

Blog post and testbed

https://tinyurl.com/2020llgc

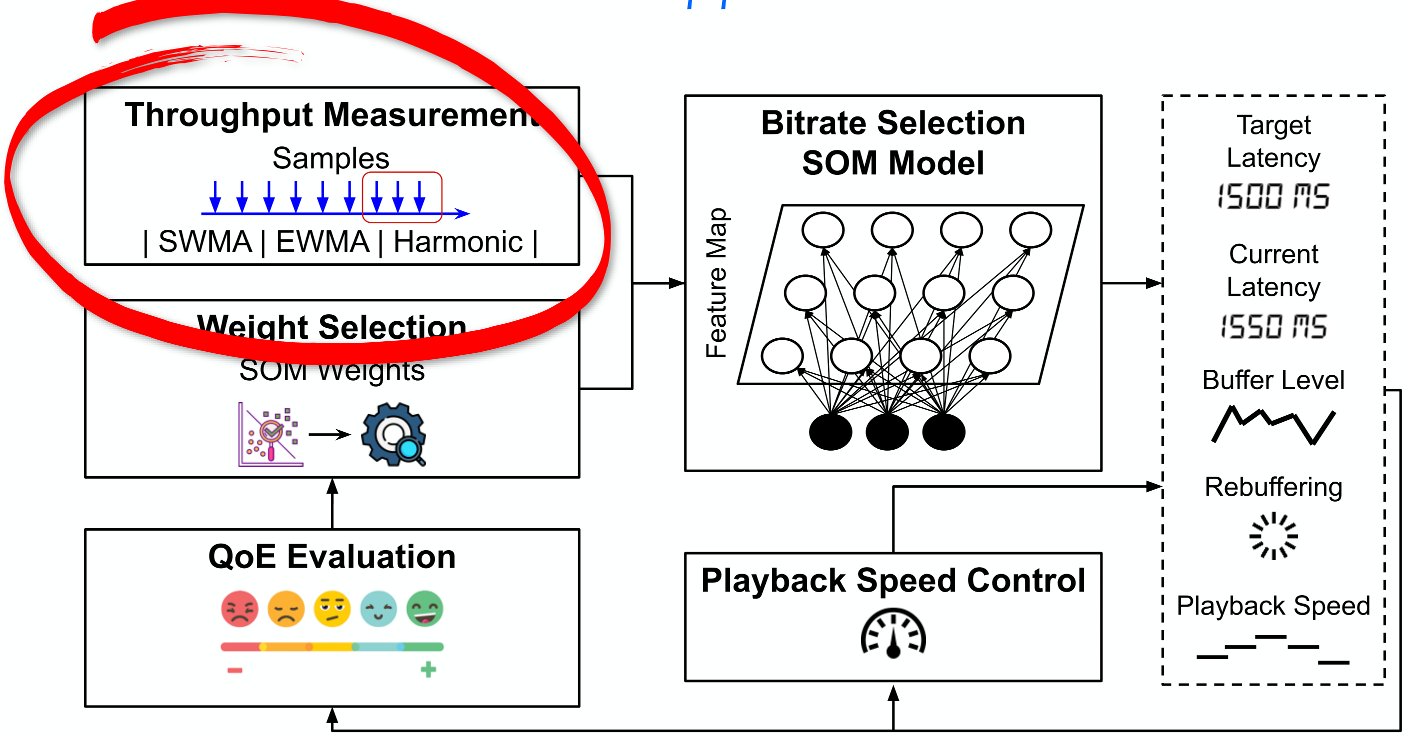
Twitch's findings

https://www.youtube.com/watch?v=rcXFVDotpy4

MMSys'20
ISTANBUL

The Low-on-Latency (LoL/LoL+) Algorithm

Streamline every component in the media pipeline



Reading: Catching the moment with LoL+ in Twitch-like low-latency live streaming platforms – IEEE TMM (2022) [link]
Code: https://github.com/NUStreaming/LoL-plus

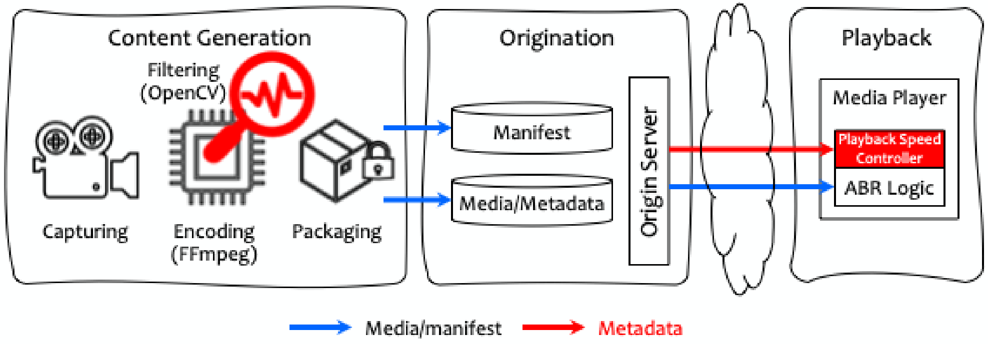
Low-latency sports

is a dish best served with ~~adaptive~~ playback

content aware

Reading: Content-aware playback speed control for low-latency live streaming of sports – ACM MMSys'21 [link]

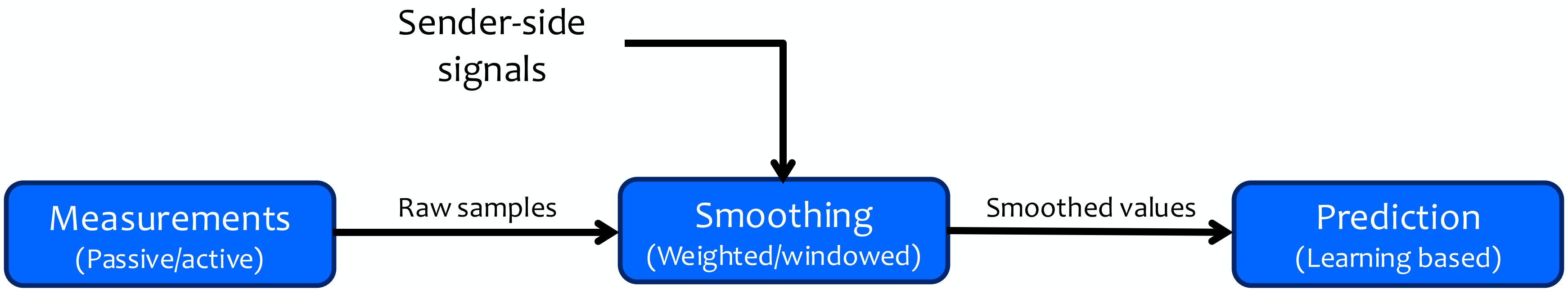
Automated Event Detection and Metadata Generation



Reading: Automated adaptive playback for encoder-adjudicated live sports – IEEE MIPR'22 [link]

(Client-Side) Bandwidth Measurement, Smoothing and Prediction

Assisted by sender-side signals

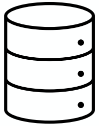


- Passive measurements have several advantages
  - No traffic overhead
  - No additional track and track management
  - No need to find a good value for probe size/frequency
- Active measurements have advantages, too
  - They can complement passive measurements
  - They work more reliably when there is not sufficient incoming data (e.g., when receiving very low-bitrate media)

Sender-side signals can assist client-side measurements but <u>**should not replace**</u> them

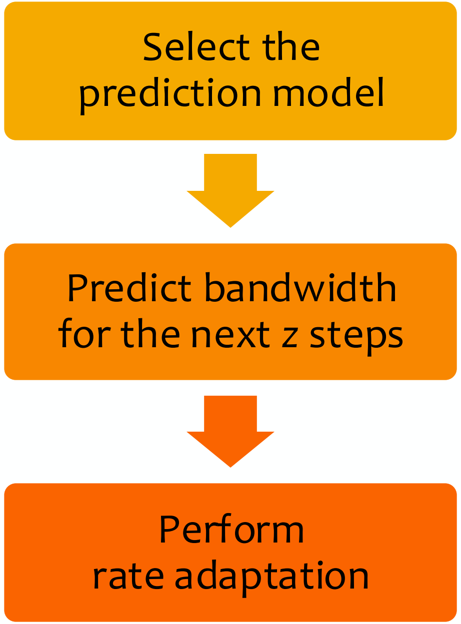
Automated (Prediction) Model Selection for Low Latency

Datasets



Training

Prediction Model



Statistical techniques

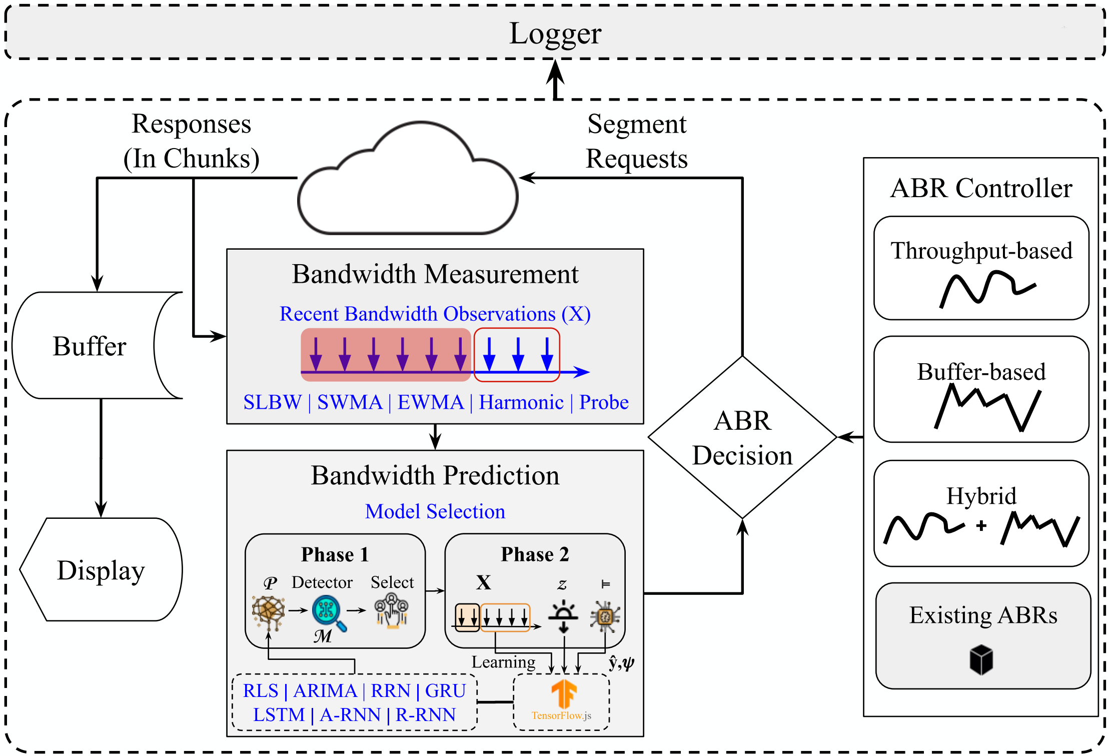
- adaptive RLS
- auto-regressive integrated moving average (ARIMA)

Computational intelligence techniques

- recurrent neural network (RNN)
- long short-term memory (LSTM)
- gated recurrent unit (GRU)
- attention RNN (A-RNN)
- residual RNN (R-RNN)

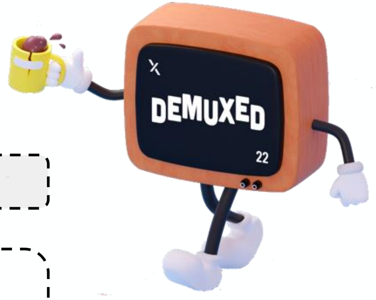
Reading: Data-driven bandwidth prediction models and automated model selection for low latency – IEEE TMM (2021) [link]

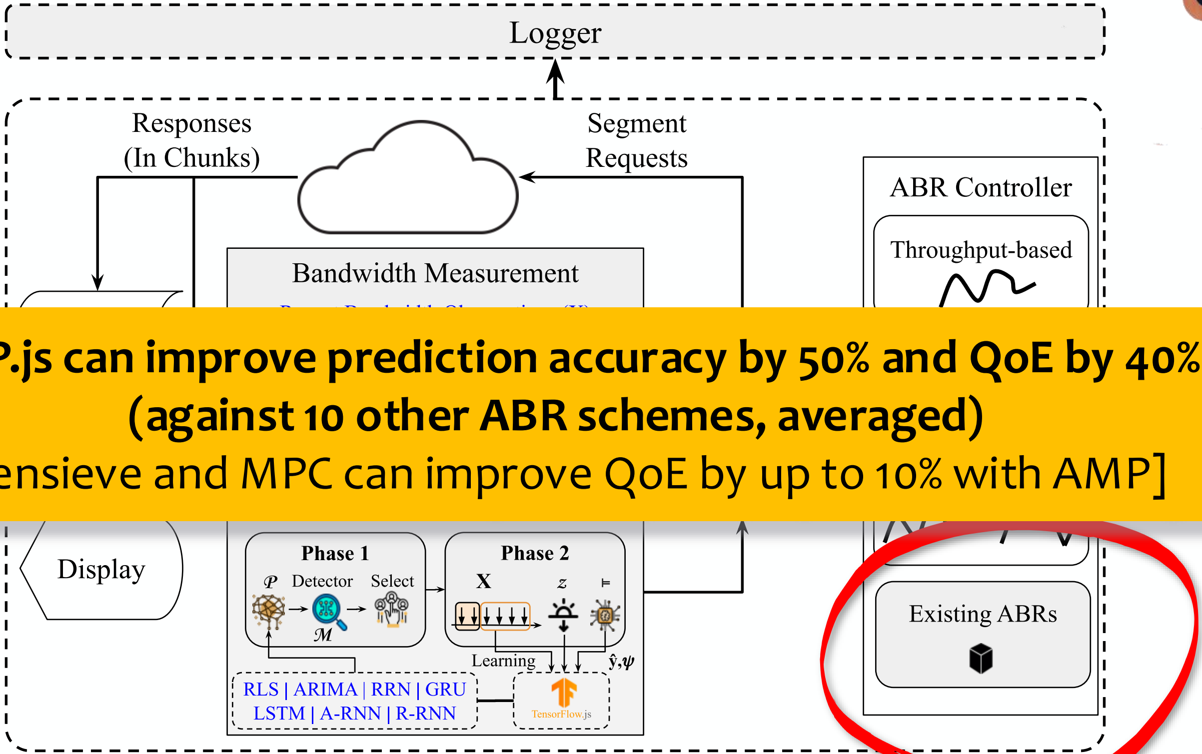
Open-Source Implementation in dash.js



Reading: Data-driven bandwidth prediction models and automated model selection for low latency – IEEE TMM (2021) [link]
Video: https://youtu.be/8jSRFvx70co

Open-Source Implementation in dash.js



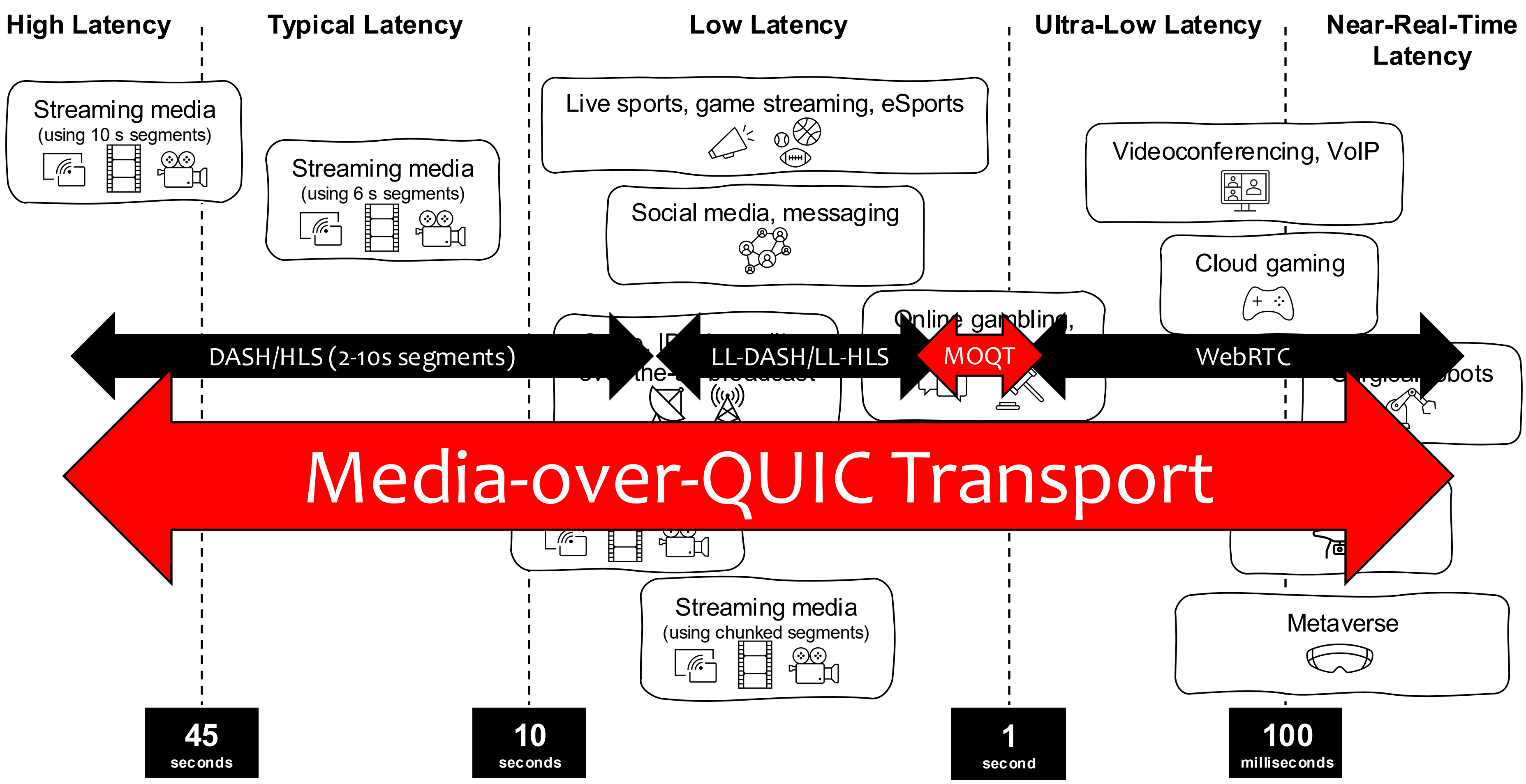


AMP.js can improve prediction accuracy by 50% and QoE by 40% (against 10 other ABR schemes, averaged)

[Pensieve and MPC can improve QoE by up to 10% with AMP]

Reading: Data-driven bandwidth prediction models and automated model selection for low latency – IEEE TMM (2021) [link]

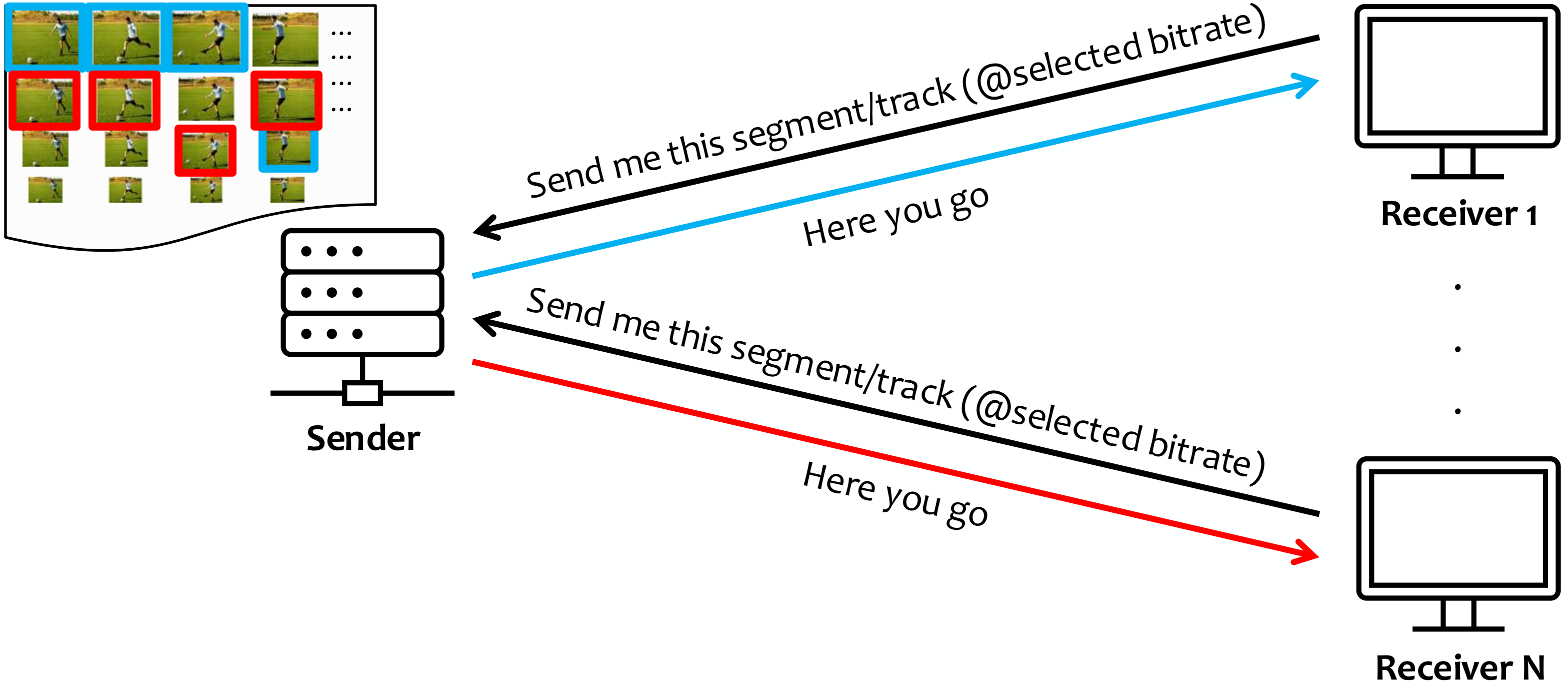
Video: https://youtu.be/8jSRFvx70co



Reading: Toward one-second latency: evolution of live media streaming – IEEE CS&T (2025) [link, preprint]

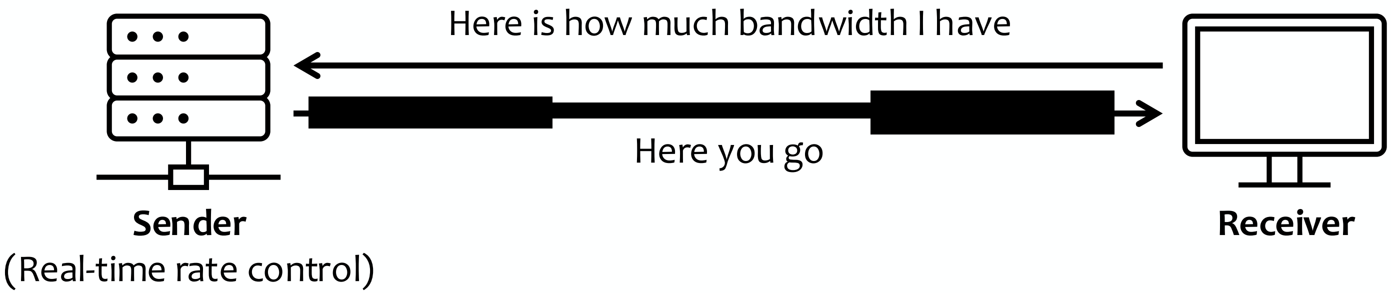
Multi-Bitrate Encoding and Streaming

No receiver-specific encoding at the sender



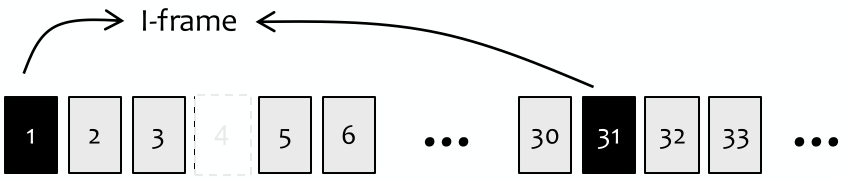
Single-Bitrate Encoding and Streaming

Receiver-specific encoding at the sender



QUIC’s Flexible Loss Recovery

Latency-aware operations (especially useful when the encoder’s rate control is slow)



TCP

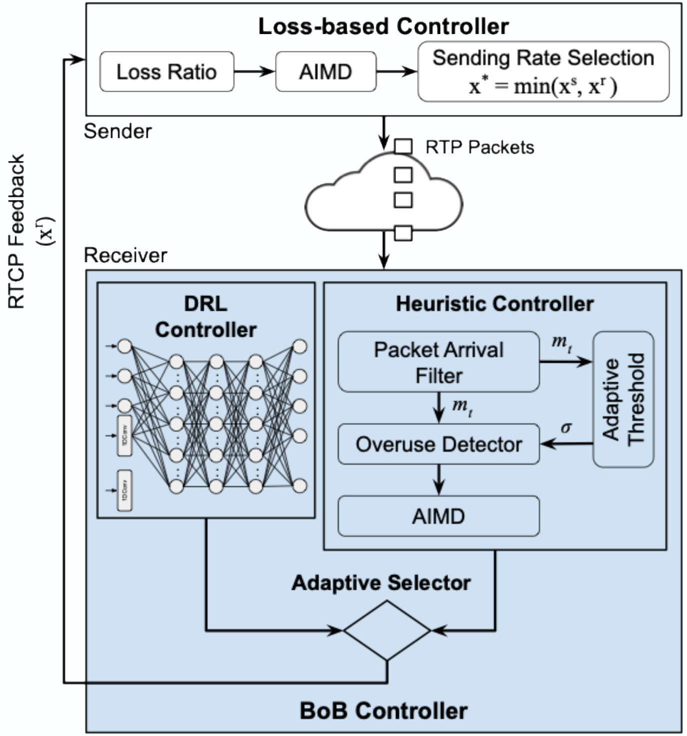
- must recover #4 (HOL blocking)

QUIC

- can do what TCP does
- can continue with #31 - #60 and skip #4 - #30
- can continue with #31 - #60 and then try recovering #4
- can recover #4 using FEC (in the app layer)

BoB (Bang-on-Bandwidth): Hybrid Bandwidth Prediction for RTC

BoB combines a heuristic-based controller (inspired by the GCC algorithm) with a DRL controller



Sender: Loss-Based Controller

Compute target sending rate x^*

Receiver: BoB Controller

Compute x^r using the BoB prediction

Reading: BoB: bandwidth prediction for real-time communications using heuristic and reinforcement learning – IEEE TMM (2023) [link]

Heuristic Controller

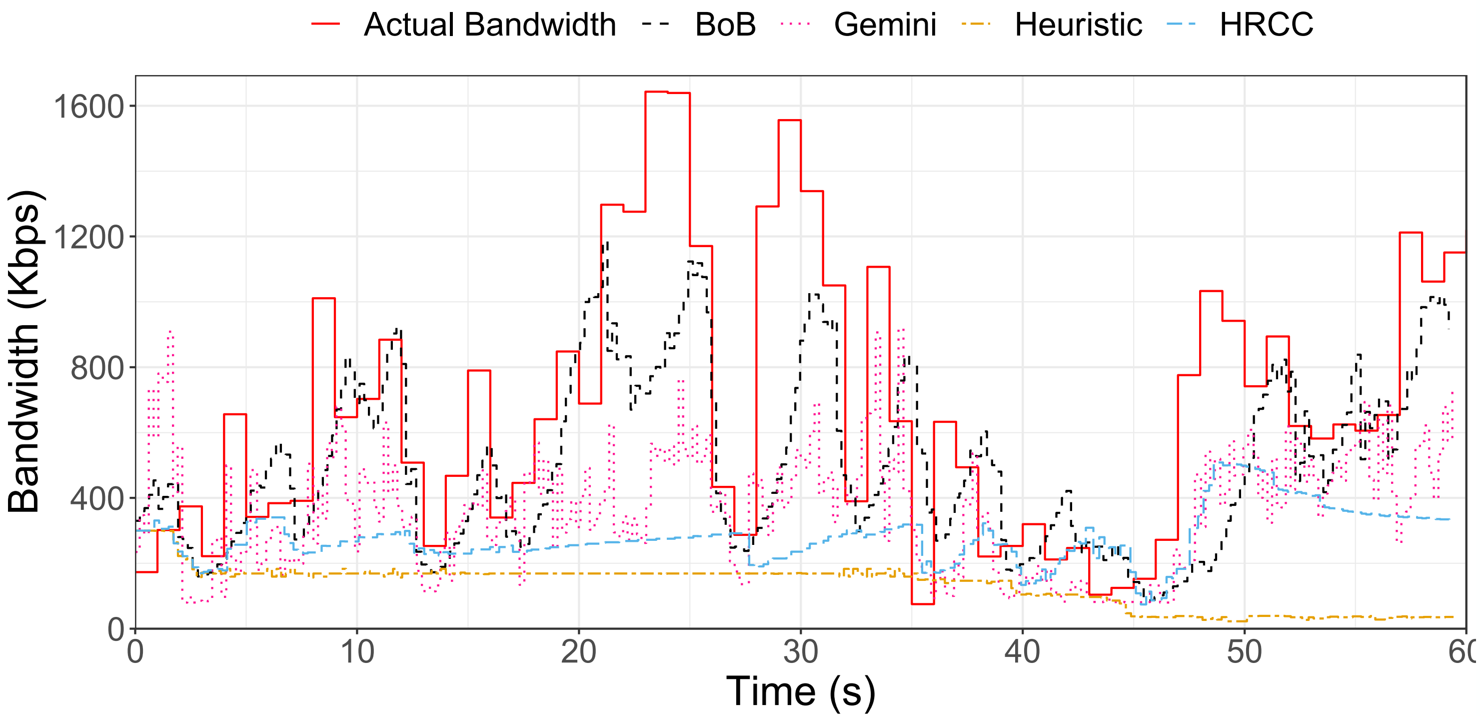
- Key idea
  - Estimate the bandwidth based on aggregated per-packet RTP feedback
- Strength
  - Good convergence in some scenarios
  - Easy to implement
  - Low overhead
- Weakness
  - Fail to achieve consistent high performance in diverse network environments

DRL Controller

- Key idea
  - Predict the bandwidth based on NN model trained using various network conditions
- Strength
  - Strong adaptability to diverse network environments
- Weakness
  - Not easy to deploy
  - Sometimes it may not converge

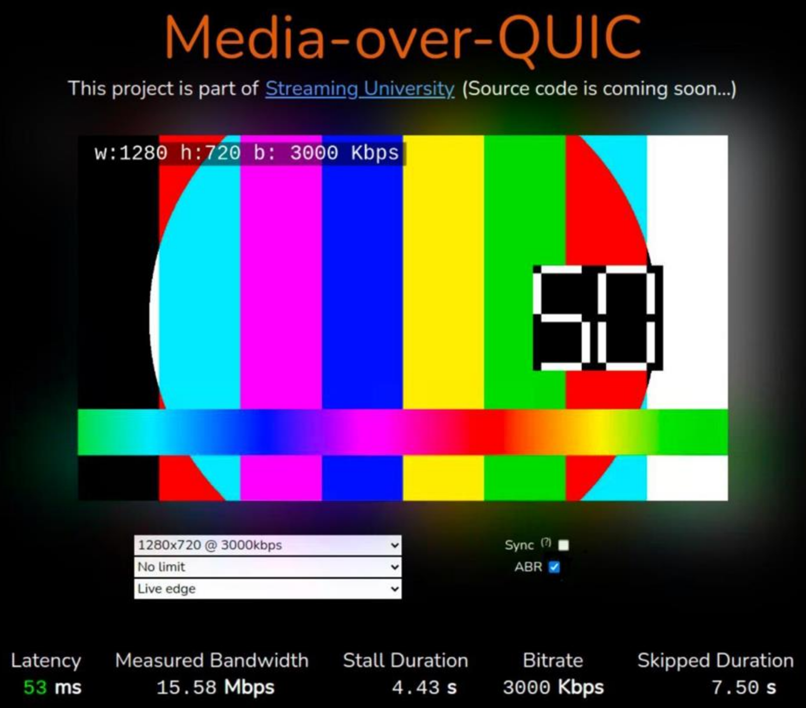
Actual and Predicted Bandwidth

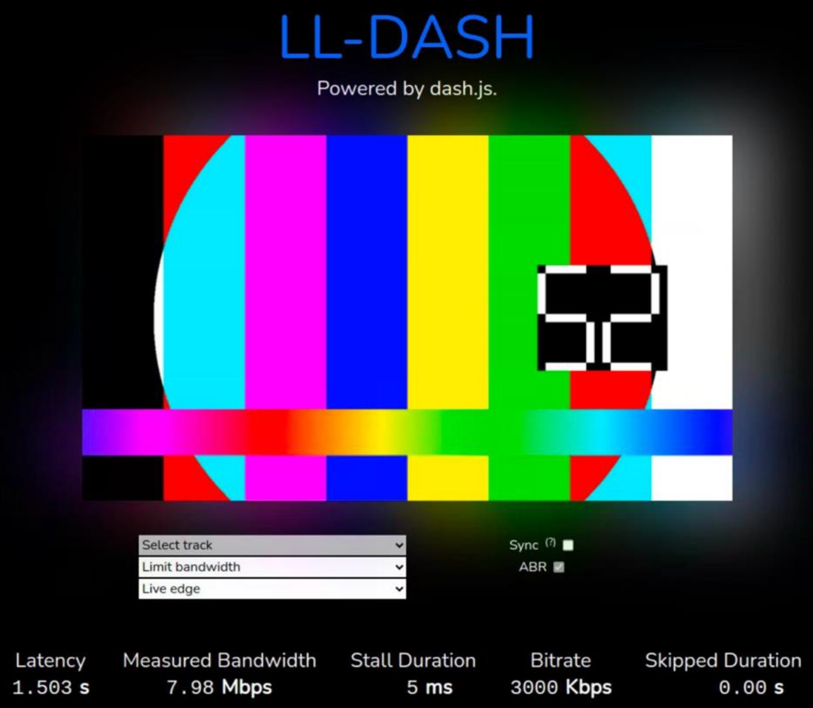
Example network trace



MOQT (with Client-Side ABR) vs. LL-DASH

https://moq.streaming.university





Reading: Media-over-QUIC transport vs. low-latency DASH: a deathmatch testbed – ACM MMSys'24 [link]

- 🚦 Congestion response policies
- 🚫 Selective drop policies
- 🎚️ Subscriber-side track (bitrate) selection lo
- ⚙️ Publisher-side rate adaptation
- 🚨 Prioritization schemes
- ⏱️ Scheduling strategies