Llama: Low Latency Adaptive Media Algorithm

Tomasz Lyko

PhD Student at Lancaster University

Low Latency Live Streaming

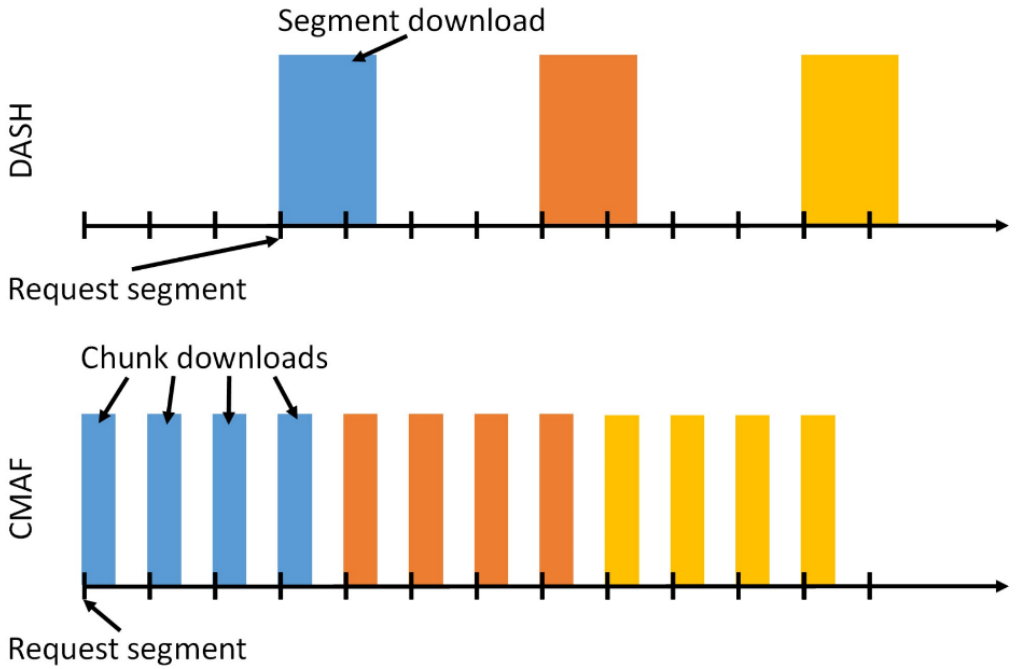
- Live video streaming over the internet suffers from high latency when compared to traditional broadcast methods such as satellite television

- The largest contributor to latency on the Internet is the client buffer
- The client buffer needs to be large enough for an ABR algorithm to be able to measure network conditions and adjust the quality bitrate in a timely manner

- Common Media Application Format (CMAF) has been recently introduced, which can aid ABR performance in live video streaming

Common Media Application Format

- Common Media Application Format (CMAF) introduces the concept of a CMAF chunk:
  - Segments are further divided into chunks
  - Chunks can be played out as soon as received by the client
  - Quality bitrate can be only changed at segment level
  - Reduces the minimum latency from one segment to one chunk
- CMAF chunks can be delivered using HTTP/1.1 Chunked Transfer to reduce the number of HTTP requests and the resulting overhead



ABR Algorithms

- Overall ABR goal: maximise the average quality bitrate while minimizing the negative QoE factors

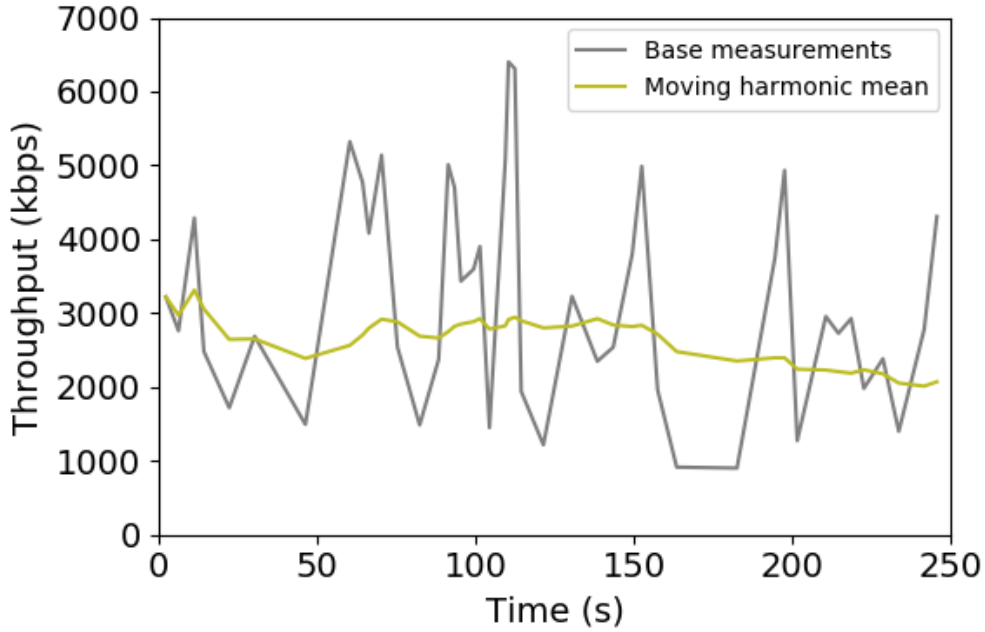
- Negative QoE factors include:
  - Rebuffering events
  - Quality Bitrate Variation

- Most of these QoE factors are in competition with each other, and therefore require a reasonable trade-off

Throughput Estimation

- Smoothing functions work well in environments with a large client buffer, however, in Low Latency Live Streaming, where the client buffer is small, it will lead to frequent rebuffering

- Fine grain measurement is great at detecting worsening network conditions quickly, however, it will lead to frequent temporary increases in video quality



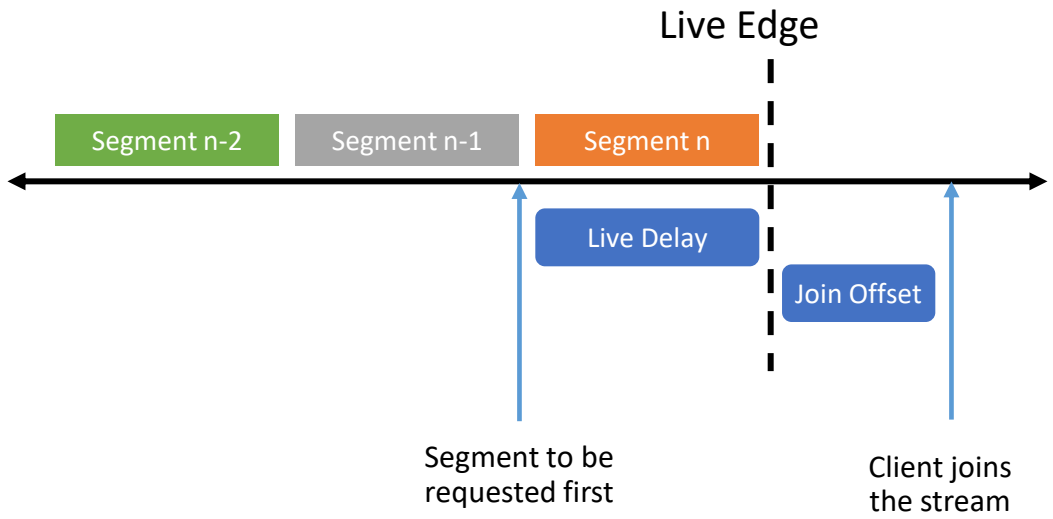
Llama: Low Latency Adaptive Media Algorithm

- Utilises the novel idea of using two independent throughput measurements made over different timescales
  - Short term measurement for decisions about reducing video quality
  - Long term measurement for decisions about increasing video quality

- Throughput of the most recent segment to provide quick reaction to worsening network conditions in order to avoid rebuffering
- Harmonic mean of past 20 segments to provide stable video quality

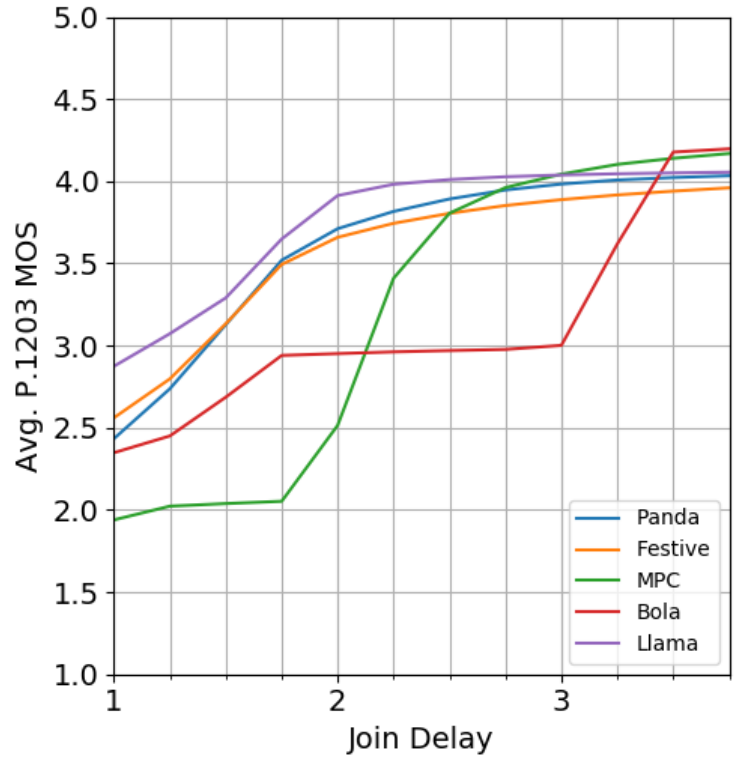
Evaluation of ABR Performance in Low Latency DASH & CMAF

- Utilising a simulation model developed in NS-3:
  - Supports Live DASH, CMAF Chunks, Traffic Shaping, and Latency Configuration
  - Verified against the DASH.JS player
- Applied realistic network scenarios to the simulations based on throughput traces extracted from CDN logs of a commercial live TV service (BT Sport 1)
- Tested four popular ABRs across multiple latency configurations using two parameters
  - Live Delay - Specifies which segment the client requests first
  - Join Offset - Determines the time at which the client first requests a segment relative to the time at which segments are made available on the server



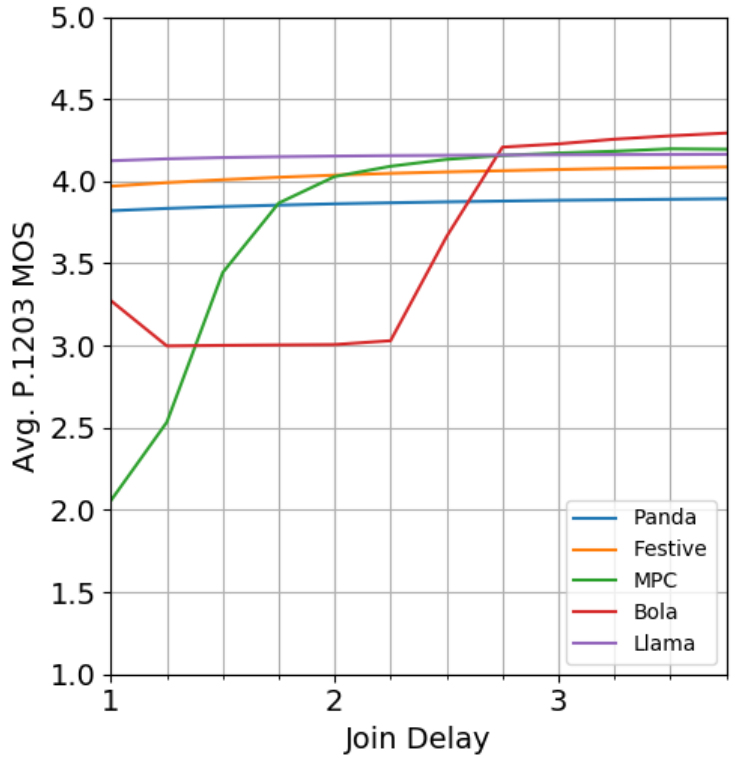
Llama Offers Better Performance in Low Latency DASH

- For Join Delays between 1 and 2.75, Llama outperformed all other ABRs in terms of P.1203 MOS
- At Join Delay of 3, it still performed the best, along with MPC which achieved the same P.1203 MOS
- Beyond Join Delay of 3, Llama was outperformed by MPC and Bola which achieved slightly higher P.1203 MOS



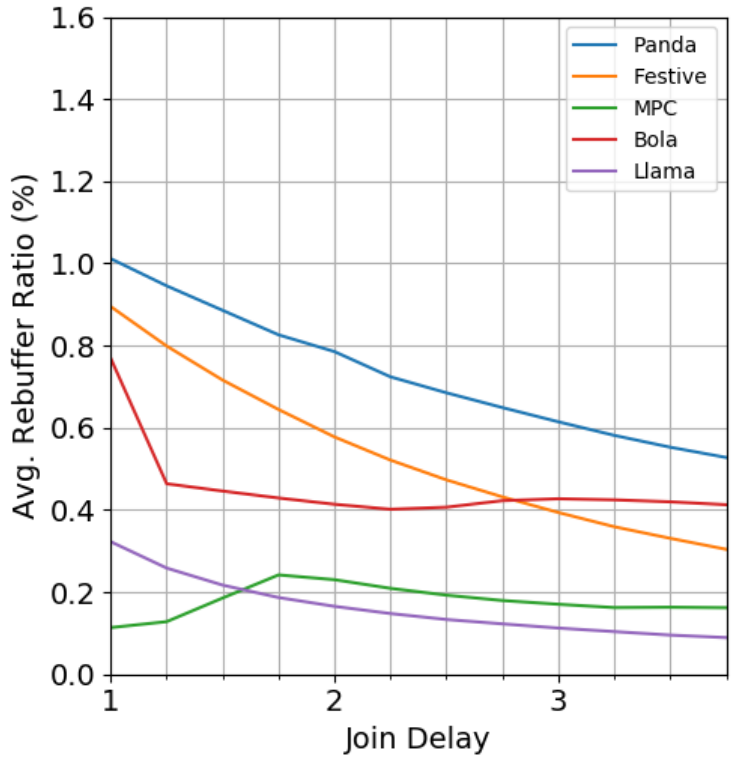
Llama Offers Better Performance in Low Latency CMAF

- Llama performed the best for Join Delays of 1-2.5
- For Join Delays of 2.75-3.75 Llama performed slightly worse than MPC, however both of the ABRs were outperformed by Bola

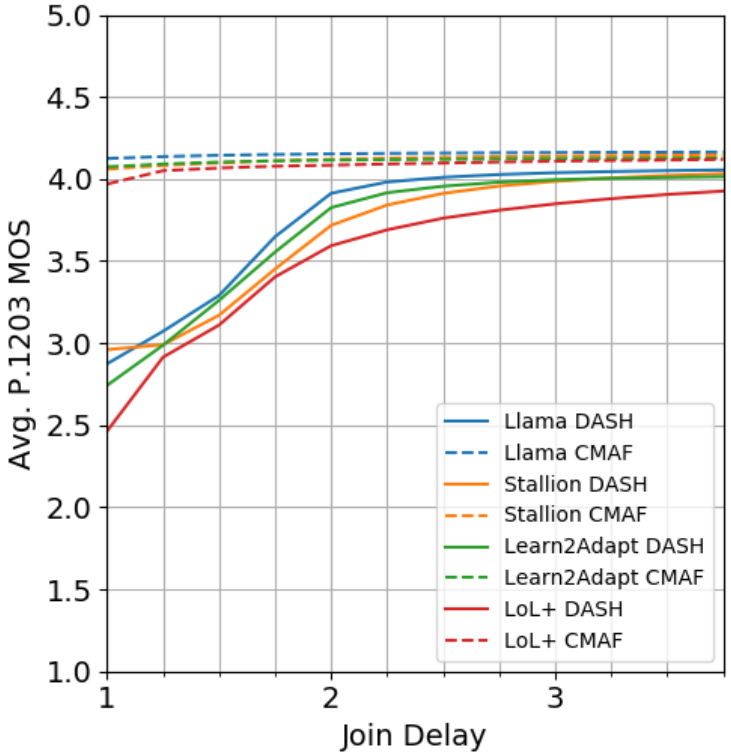


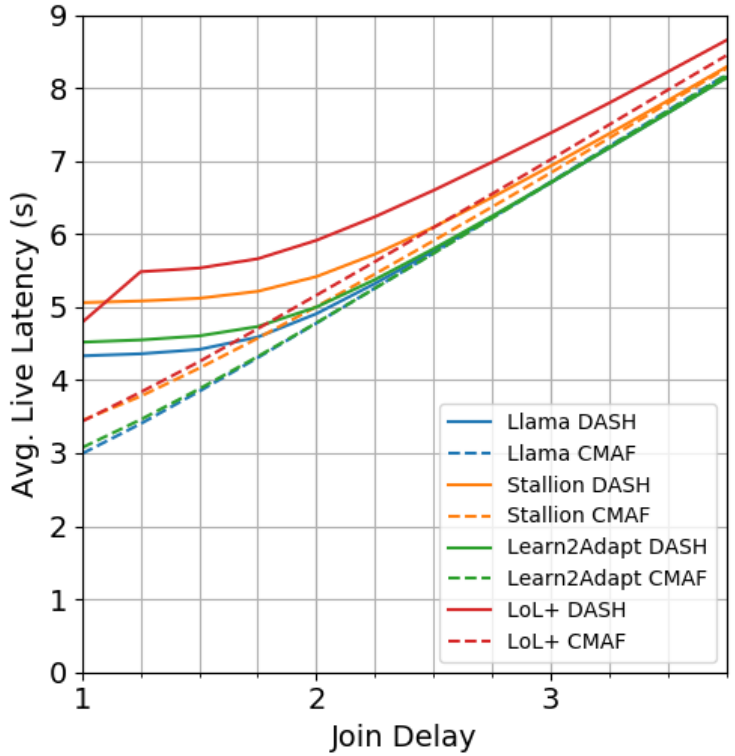
Llama Reduces Rebuffering

- For Join Delays between 1 and 1.5, Llama achieved the second lowest average Rebuffer Ratio
- For higher values of Join Delay, Llama achieved the lowest average Rebuffer Ratio
- At Join Delay of 1 and 2, Llama reduced rebuffering by up to 68% and 79% respectively



Llama Outperforms Other Low Latency ABR Algorithms





Summary

- Llama introduces a new approach to throughput estimation, which improves performance in Low Latency Live Streaming
  - Improved P.1203 MOS by up to 2.06
  - Reduced rebuffering by up to 75%

- This approach consists of using two independent throughput measurements made over different timescales
  - Each designed with a different goal in mind

References

- Simulation model: https://github.com/tomlyko/ns3-dash-cmaf-model
- Throughput traces: https://github.com/lancs-net/ABR-Throughput-Traces
- Lyko, T., Broadbent, M., Race, N., Nilsson, M., Farrow, P., & Appleby, S. (2020, June). Evaluation of cmaf in live streaming scenarios. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video* (pp. 21-26).
- Lyko, T., Broadbent, M., Race, N., Nilsson, M., Farrow, P., & Appleby, S. (2020, December). Llama-Low Latency Adaptive Media Algorithm. In *2020 IEEE International Symposium on Multimedia (ISM)* (pp. 113-121). IEEE.