A growing number of video services use MPEG Dynamic Adaptive Streaming over HTTP (DASH) to deliver content. These flows are delivered using a Content Delivery Network (CDN), with a number of caches on the path. HTTP caches interact poorly with multiple flows of the same content: time and quality differences reduce cache hit-ratios, negatively impacting performance. Our techniques consolidate near-simultaneous flows based on time or quality, reducing the overall number of flows, and increasing the cache capacity for remaining flows. In large-scale streaming platforms, we estimate that there are a sufficient number of near-simultaneous flows for the same content for a non-trivial improvement in performance, and these improvements are only likely to increase as the use of DASH grows.

**STREAM CONSOLIDATION**

Our consolidating cache identifies near-simultaneous flows for the same content at potentially different rates, and selects (based on the methods below) one flow that could serve the clients of the other flows. The remaining flows are then eliminated from the cache, freeing up space for the surviving flow. HTTP redirects are used to serve clients with chunks from the surviving flow. To allow this, the cache intercepts and interprets [4] the media presentation description (MPD) for all DASH flows being requested. The MPD describes the URL structure of the content, and the different representations offered.

**Rate-based**

Simultaneous streams of the same content, but at different representations, can be consolidated. We select the best common representation available, based on the bandwidth capacity between the cache and each client, and the cache and the server.

Gouta et al. show that offering multiple bitrate representations decreases cache hit-ratios by 15% [3], indicating the performance improvements that this technique may yield. CF-DASH [2] allows clients and caches to agree on a representation. This agreed representation is not enforced by the cache, allowing a given client to provide a better quality of experience by requesting a higher representation than that agreed upon. This reduces the likelihood of deployment.

**Time-based**

Near-simultaneous streams of the same content, at the same bitrate, but at different time offsets, can be consolidated.

Near-simultaneous is defined as a time offset that, if the flows are consolidated, would not significantly impact the user. To reduce the likelihood of clients skipping to offset the impact of the technique, we prefer to repeat content that a client has seen, rather than skip ahead. Therefore, the surviving flow is that which is requesting content from the earliest time instant.

**IMPLEMENTATION CHALLENGES**

Degraded quality of experience Manipulating DASH flows in the way we propose will result in either in introducing skips or reducing the bitrate representation provided to users. This does not necessarily lead to a net reduction in quality of experience – our techniques allow for other metrics, such as start-up time, to be improved.

Widespread end-to-end encryption The techniques proposed would not work with DASH flows that have been encrypted end-to-end. This is a wider challenge for performance-enhancing middleboxes in general, but technical solutions exist [5] if the performance improvements prove worthwhile. Existing systems often only encrypt the payload, leaving the HTTP headers in the clear, and enabling caching.

**CONCLUSION**

Our techniques are designed to improve the performance of DASH applications by improving cache hit ratios. It remains for these techniques to be implemented and evaluated, including investigating interactions with other caching policies, such as prefetching algorithms. These techniques, and the performance improvements that they offer, will become increasingly beneficial as the use of DASH grows.

**REFERENCES**


