A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP — Public Review

Keith Winstein
Stanford University
Stanford, CA
keithw@cs.stanford.edu

Video streaming accounts for two thirds of global consumer Internet traffic, a figure predicted to rise to 80% by the year 2019 [2]. The SIGCOMM community has worked in this area for decades [5], but a time-traveling visitor from SIGCOMM 1995 might be shocked to see how things have turned out.

"Where is the bespoke congestion control?" the visitor will ask. "Or multicast transport? What video-specific error-concealment techniques are used to recover from datagram (or cell!) losses?"

No, we'll say, the industry didn't end up using those approaches. The clients for YouTube, Netflix, and HBO just grab multi-second video chunks from a Web server with HTTP-over-TCP-over-best-effort-IP. That was the easiest thing to deploy and to scale up.

Of course, the questions of *which* video chunks to grab, and *which* servers to serve them from, are big business. In a new wave of research, the community has created and critiqued schemes that consider available throughput, video-buffer occupancy, and other signals [e.g., 1, 3, 6, 7]. Netflix Inc. recently switched to a "buffer-based" algorithm and reduced the rate of rebuffering pauses by 10–20% [4].

In the current paper, Yin et. al. unify and place this recent work in a principled control-theoretic framework: the authors formulate video streaming as an instance of an optimization problem in sequential decision-making under uncertainty. Their algorithm, a practical model-predictive controller called FastMPC, depends on a quality-of-experience metric—an externally-specified compromise between the desires for high- and consistent-quality video and for fast startup with infrequent rebuffering pauses—and an estimator that forecasts future throughput on the connection.

Given those inputs, and some approximation and pre-calculation to make the algorithm run in real time, FastMPC's behavior is a product of whatever preferences and throughput predictor are fed in. FastMPC is implemented as a patch to the `dash.js` video client.

In trace-based emulation experiments representing broadband and cellular network paths, FastMPC outperformed the authors' implementations of prior "rate-based" and "buffer-based" algorithms, and the unpatched `dash.js`, on multiple quality-of-experience metrics.

The authors found a 10–15% improvement in quality-of-experience over prior work. It's hard to know viscerally what this would feel like to the viewer, and it may be that only large video operators and the service providers that tend to them are in a position to assess precisely the practical magnitude of FastMPC's improvement across a broad and diverse user base.

Irrespective of the size of the gains shown so far, the elegance of uniting prior work on common axes and demonstrating that this problem can be solved in real time by a computer—once we tell the computer how to compromise between multiple goals—makes this an exciting contribution.

What other areas of the SIGCOMM program will be reduced to an optimization problem and practically solved by a computer? If a visitor from today traveled to SIGCOMM 2035, will computers be doing most of the research?

REFERENCES

[1] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience for internet video. SIGCOMM '13, pages 339–350, Hong Kong, China.

[2] Cisco visual networking index: Forecast and methodology, 2014–2019. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf.

[3] T.-Y. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari. Confused, timid, and unstable: Picking a video streaming rate is hard. IMC '12, pages 225–238, Boston, Massachusetts.

[4] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson. A buffer-based approach to rate adaptation: Evidence from a large video streaming service. SIGCOMM '14, pages 187–198, Chicago, Illinois.

[5] H. Kanakia, P. P. Mishra, and A. Reibman. An adaptive congestion control scheme for real-time packet video transport. SIGCOMM '93, pages 20–31, San Francisco, California.

[6] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. Begen, and D. Oran. Probe and adapt: Rate adaptation for HTTP video streaming at scale. *Selected Areas in Communications, IEEE Journal on*, 32(4):719–733, April 2014.

[7] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang. Qdash: A QoE-aware DASH system. In *Proceedings of the 3rd Multimedia Systems Conference*, MMSys '12, pages 11–22, Chapel Hill, North Carolina, 2012.