Enhancing Over-the-Top Video Streaming Quality with DASH Assisting Network Elements

Jan Willem Kleinrouweler
Centrum Wiskunde & Informatica (CWI)
Science Park 123
1089 XG, Amsterdam
The Netherlands
j.w.m.kleinrouweler@cwi.nl

ABSTRACT

Dynamic Adaptive Streaming over HTTP (DASH) is the leading technology for delivering online video streaming content. However, DASH has performance problems on shared network links. This thesis investigates how DASH Assisting Network Elements (DANEs) can be used to optimize bottleneck links for DASH video traffic, with the goal to improve the viewers' Quality of Experience. DANEs are aware of active DASH players and divide the network bandwidth among the players and other traffic. In the first three years of the PhD, contributions have been made in the areas of multimedia systems and performance modeling: Two prototype implementations of DANEs have been developed and evaluated in both wired and Wi-Fi networks. Experiments with real DASH players show that DANEs significantly increase the video bitrate and reduce the number of changes in video quality. In addition, Markov models have been created to find out how network bandwidth should be divided, and what the effect of bandwidth sharing policies is on the resulting streaming performance. The model was thoroughly evaluated and has shown to be highly accurate. As such, it is a useful tool that can be used to configure and optimize bandwidth sharing in DANEs. In the remaining year of the PhD program, I would like to expand DANE technology and apply it to different use cases such as mobile networks.

ACM Classification Keywords

H.5.1 Multimedia Information Systems: C.2.1 Network Architecture and Design

Author Keywords

Dynamic Adaptive Streaming over HTTP (DASH); HTTP Adaptive Streaming; Video streaming; Network architectures; Performance modeling

TVX '17 Adjunct, June 14-16, 2017, Hilversum, Netherlands

ISBN 978-1-4503-5023-5/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3084289.3084295

INTRODUCTION

In the past few years, the media consumption model has undergone major changes. Content delivery is shifting from traditional television broadcasting to Internet based video on demand (VoD). Video content is distributed over-the-top. Streams are delivered over the Internet and not using dedicated and managed distribution infrastructures. The popularity of VoD services, such as YouTube and Netflix, is rising. Not only do they provide access to an ever growing catalog of video content, they also serve a large number of viewers. Video content requires high bandwidth and takes a significant share of internet traffic. For example, in North-America in 2016, over 50% of internet traffic during peak hours is attributed to online video streaming [14]. With an expected growth in this area, VoD will put a large on-even demand on our networks in the future.

From the viewers' point, video content should be always available on a variety of devices, ranging from smartphone to smart TV. Moreover, viewers are impatient: they want a stream to start immediately and navigate between different videos without delay. Streams cannot be interrupted, should be of consistent high quality, and video quality should look equally good on different devices. This rises the question how we can make our networks cope with these high bandwidth demand and satisfy the stringent requirements from viewers. Every step in the delivery pipeline must have sufficient resources to be able to deliver instant, stable, and high quality video.

Modern video streaming services are based on Dynamic Adaptive Streaming over HTTP (DASH). For instance, YouTube and Netflix implement this technology in their players [13]. A typical delivery pipeline for a DASH-based video streaming service is shown in Figure 1. High availability and scalability are accomplished by replicating the video content on Content Delivery Networks (CDNs). Video content is placed on replication servers close to the receiver to reduce delay, as illustrated in Figure 1. This approach ensures that there are sufficient resources at the content provider side. The bottleneck, therefore, is most often found in what is called the last mile: the users' DSL connections, Wi-Fi networks, or cellular networks. These networks have a limited capacity that is often shared among several users. Depending on the current load on the network, there is more or less bandwidth available for

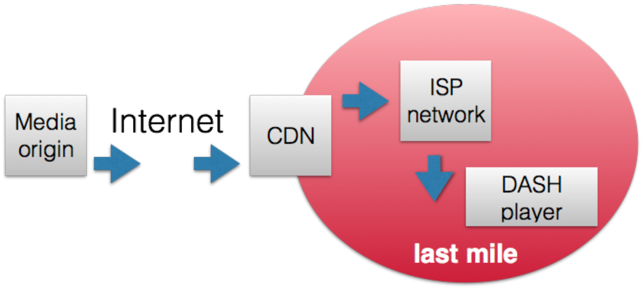

Figure 1. DASH-based video delivery pipeline

DASH. The adaptation algorithms in DASH players aim to adapt the video quality to changes in the network. However, this has proven to be difficult in shared networks with multiple DASH players and with background traffic, due to the greedy nature of DASH players [2, 1, 6]. Performance problems – resulting in stalling, lower video bitrate, frequent quality switches, and unfair bandwidth sharing – deteriorate the viewers' Quality of Experience (QoE). This eventually leads to user disengagement and abandonment [4, 16]. To realize a good QoE, it is important that last-mile network connections do not hinder DASH-based video streaming. The last-mile of the video delivery pipeline defines the domain of this thesis and is illustrated by the highlighted area in Figure 1.

This thesis is centered around the key research question:

How can last-mile network connections be optimized for DASH video streaming resulting in a better quality of experience for the viewers?

It presents the concept of network elements (e.g. internet gateways or Wi-Fi routers) that actively manage DASH video streaming on a bottleneck link. Such elements – called DASH Assisting Network Elements (DANEs) – are aware of active DASH players and other network traffic. They have the task to divide the available bandwidth between DASH players and other traffic according to sharing policies. The following related research questions are answered:

- Does network assisted DASH improve the streaming performance in terms of bitrate, quality switches, stalling, and fairness?
- How can DASH assistance be realized, taking into account the agility of the solution and the privacy of the viewer?
- Which policies can be defined to divide network bandwidth among DASH players and other traffic?
- What is the effect of bandwidth sharing policies on DASH streaming performance?
- Can Markov modeling be used to efficiently evaluate bandwidth sharing policies in DANEs?

In the first three year of the PhD program, the research questions have been answered for DANEs in wired and Wi-Fi networks. The DANE prototypes successfully prevent freezes, increase the video bitrate, reduces quality switches, and improve fairness. Highly accurate Markov models have been created to describe bandwidth sharing in DANE. In the last year of the PhD, I would like to extend the DANE to mobile networks. The performance model can generally be applied, and will be used in the optimization strategy for mobile networks.

MANAGING DASH ON THE LAST MILE

To be able to provide stable and high quality video streaming, we have overcome two limitations of DASH players on shared network connections: errors in bandwidth estimation in DASH players' adaptation algorithms and reduced network throughput as a result of using HTTP over TCP.

In DASH, the player determines the video quality (i.e. bitrate) of the stream. A video file is split up into segments with a duration of a few seconds (usually between 2 and 10 seconds). Each segment is available in different bitrates and resolutions. The player downloads the segments one by one over HTTP. It selects segments in a bitrate/resolution that matches the current available bandwidth. However, obtaining reliable bandwidth estimations is difficult, especially since DASH players can only observe their own network transactions. Furthermore, estimations can only be made while a segment is being downloaded. If a segment download takes shorter than the duration of that segment, downloading the next segment will wait a bit. This is done to maintain a stable buffer level. However, players have a blind-spot in the interval between two consecutive segment downloads.

Network elements, in contrast to individual DASH players, have an overview of network capabilities and traffic on the shared link. They make well-informed decisions on how the available bandwidth must be shared. DASH Assisting Network elements become DASH-aware by DASH players signaling the DANE that they are active. DASH players report the representation in which a video stream is available to the DANE, as illustrated in Figure 2-1. The DANE takes all options for each player, combines this with statistics about other traffic on the network link, and assigns each DASH player a target representation. Target representations are communicated back to the DASH players, as shown in Figure 2-2. They use this target as a guide when selecting the final video bitrate. It depends on the download speeds that can be achieved in the network whether DASH players select the target bitrate or a lower quality representation.

DASH uses TCP as transport protocol, which may prohibit the high throughputs that are required for DASH. In crowded networks, TCP connections need time to increase the window size that enables the high throughput. However, for DASH, high

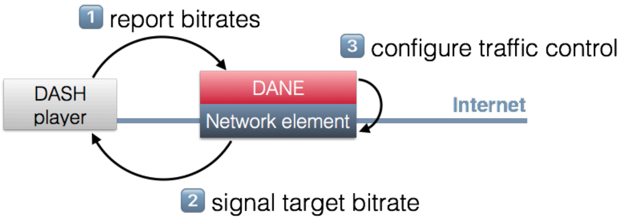

Figure 2. DASH Assisting Network Element – flow of events

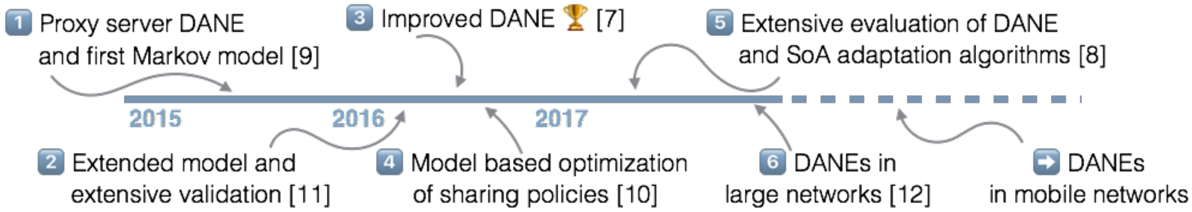

Figure 3. Timeline overview of publications during the PhD program (1-6). A contribution on DANEs in mobile networks can be expected by the end of 2017.

throughput must be available at the start of the stream. If bandwidth is not available, the stream can only be of low quality or will be frequently interrupted. Furthermore, the separated downloads of DASH segments prevent TCP windows sizes from growing and thus fails to reach high throughputs. To make sure that TCP will not be stuck at low speeds, the DANE allocates bandwidth for DASH in the form of traffic control, as depicted in Figure 2-3. Depending on the number of active DASH players, and when DASH players start and stop, the DANE dynamically changes the bandwidth allocation.

EVALUATING BANDWIDTH SHARING POLICIES

The DANE (partially) takes over the task from the DASH players to select a video bitrate. It oversees the active DASH players and assigns each player a target bitrate. A simple policy would equally divide the available bandwidth among the players. However, different devices may have different capabilities and requirements (e.g. different screen size and resolution). Which sharing policy should be applied depends how many simultaneously active players are active in the network, and how often users start and stop video streams. The sharing policy, combined with the dynamics in the network, determine the resulting streaming quality.

In small networks with a few players (e.g. ≤ 5 players), the streaming quality can be assessed with reasoning. Larger networks are more dynamic (i.e. more players that start and stop). As a consequence, new methods for evaluating sharing policies are required. The previously applied methodology with experimental evaluations is no longer applicable. Experimental evaluations in testbeds and simulation environments are a time consuming and error prone process. Changing parameters in a sharing policy requires experiments to be repeated. This is also impractical for network administrators who would like to try multiple sharing policies before deploying a DANE. Therefore, this thesis investigates if performance modeling can be applied to speed up this process.

From experiments with the DANE, we learned that the achieved streaming bitrates match the assigned bitrates. This means that for every number of active DASH players, the streaming bitrate is predictable given a sharing policy. Starting and stopping DASH players can be modeled using a birth-death Markov process. For each state in that process (i.e. an active number of DASH players), the sharing policy is applied. By observing how much time the Markov process spends in each state, and how often the process transitions between states, the mean video bitrate and the number of quality switches are obtained. These two factors play a major role in the viewers' QoE [3, 5, 15].

STATUS AND OUTLOOK

At the time of writing, the last year of the PhD program starts. In the first three years of the program, contribution have been made in the areas of multimedia systems and performance modeling. A timeline overview of the publications is given in Figure 3.

Multimedia systems: A proxy server based implementation of a DANE is presented in [9]. HTTP traffic is routed through this proxy server which inspects the traffic to detect DASH flows and their characteristics. Adaptation assistance is provided by altering the contents of the HTTP request header. The design and implementation of an improved version of a DANE is presented in [7]. Compared to the proxy server, this DANE works out-of-band. DASH players connect to the DANE over a WebSocket. This DANE is more lightweight, because it no longer requires to inspect all HTTP traffic. This means that the DANE can be deployed on low-powered hardware such as Wi-Fi routers and DSL modems. The out-of-band approach also significantly increases user privacy. DASH players only have to report the available representations. The DANE cannot see which video will be streamed. Furthermore, this implementation is compatible with encrypted streams over HTTPS. In extensive evaluations in a Wi-Fi testbed it was demonstrated that our DANE outperforms state-of-the-art DASH adaptation algorithms [7, 8]. The performance of DANEs in large networks (up to 600 players) was investigated in [12].

Performance modeling: The first version of the Markov model for network assisted DASH is described in [9]. This model describes the expected video bitrate and expected number of changes in video quality. It allows to define different types of DASH players where each type can be treated differently in the sharing policy. For example, policies for devices with different screen sizes, or for premium users, can be defined. An extended version of the model that includes sensitivity for buffer sizes is presented in [11]. Moreover, thorough evaluations are performed to validate the model and show its high accuracy. In [10], the Markov model is applied to compare three bandwidth sharing policies: DASH priority, Background traffic priority, and an in-between policy. The parameters for the in-between policy are determined via a weighted-sum optimization.

This thesis covers technical aspects of DANE implementations as well as highly abstract analytical performance models. However, both topics are centered around the concept of DANEs and have shown to be a good example where both fields meet. In future research efforts, including those to finish this PhD, we would like to further exploit this relationship by extending the use of DANEs to mobile networks. Up to this point we deployed the DANE in wired and Wi-Fi networks. In the light of 5G, we would like to generalize the technology mobile networks. The network resources needed for certain download speeds depend on the signal quality in mobile networks . An optimization strategy that specifies not only which bitrate has to be selected, but also how the buffer must be filled, has to be defined. The Markov model described in this thesis is general applicable and will build the core of the optimization strategy.

Looking further ahead, DANE technology may be an enabler for other applications: Virtual Reality (VR), 360 degree video, and immersive TV. These applications share the need for instant availability of high bandwidth in the network with VoD. However, the requirements for the system are tighter because of user interactivity, to which the system needs to respond without delay. Network assisting technology based on the DANE could be further developed to support new application areas. At the TVX Doctoral Consortium, I would like to seek the advise of the iTV/TVX community and explore how network assistance can be integrated to enhance the experience of new applications.

ACKNOWLEDGEMENTS

Jan Willem is a PhD student at Centrum Wiskunde & Informatica, supervised by Pablo Cesar, Prof. Rob van der Mei, and Prof. Dick Bulterman.

REFERENCES

1. Saamer Akhshabi, Lakshmi Anantakrishnan, Ali C Begen, and Constantine Dovrolis. 2012. What happens when HTTP adaptive streaming players compete for bandwidth?. In *NOSSDAV '12: Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video*. New York, New York, USA, 9–14.

2. Saamer Akhshabi, Ali C Begen, and Constantine Dovrolis. 2011. An Experimental Evaluation of Rate-adaptation Algorithms in Adaptive Streaming over HTTP. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems*. ACM, New York, NY, USA, 157–168.

3. Nicola Cranley, Philip Perry, and Liam Murphy. 2006. User perception of adapting video quality. *International Journal of Human-Computer Studies* 64, 8 (2006), 637–647.

4. Florin Dobrian, Asad Awan, Dilip Joseph, Aditya Ganjam, Jibin Zhan, Vyas Sekar, Ion Stoica, and Hui Zhang. 2013. Understanding the Impact of Video Quality on User Engagement. *Commun. ACM* 56, 3 (March 2013), 91–99.

5. Tobias Hoßfeld, Michael Seufert, Christian Sieber, Thomas Zinner, and Phuoc Tran-Gia. 2015. Identifying QoE optimal adaptation of HTTP adaptive streaming based on subjective studies. *Computer Networks* 81 (2015), 320–332.

6. Te-Yuan Huang, Nikhil Handigol, Brandon Heller, Nick McKeown, and Ramesh Johari. 2012. Confused, timid, and unstable: picking a video streaming rate is hard. In *IMC '12: Proceedings of the 2012 ACM conference on Internet measurement conference*. New York, New York, USA, 225–238.

7. Jan Willem Kleinrouweler, Sergio Cabrero, and Pablo Cesar. 2016. Delivering Stable High-quality Video: An SDN Architecture with DASH Assisting Network Elements. In *Proceedings of the 7th International Conference on Multimedia Systems (MMSys '16)*. ACM, New York, NY, USA, Article 4, 10 pages.

8. Jan Willem Kleinrouweler, Sergio Cabrero, and Pablo Cesar. 2017. An SDN Architecture for Privacy-friendly Network Assisted DASH. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2017).

9. Jan Willem Kleinrouweler, Sergio Cabrero, Rob van der Mei, and Pablo Cesar. 2015. Modeling Stability and Bitrate of Network-Assisted HTTP Adaptive Streaming Players. In *27th International Teletraffic Congress (ITC 27)*. Ghent, Belgium.

10. Jan Willem Kleinrouweler, Sergio Cabrero, Rob van der Mei, and Pablo Cesar. 2016a. A Markov Model for Evaluating Resource Sharing Policies for DASH Assisting Network Elements. In *28th International Teletraffic Congress (ITC 28)*. Ghent, Belgium.

11. Jan Willem Kleinrouweler, Sergio Cabrero, Rob van der Mei, and Pablo Cesar. 2016b. A model for evaluating sharing policies for network-assisted HTTP adaptive streaming. *Computer Networks* 109, Part 2 (2016), 234 – 245. Traffic and Performance in the Big Data Era.

12. Jan Willem Kleinrouwer, Britta Meixner, and Pablo Cesar. 2017. Improving Video Quality in Crowded Network Using a DANE. In *Proceedings of the 27th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV '16)*. ACM, New York, NY, USA, 6 pages.

13. Stefan Lederer. 2015. Why YouTube & Netflix use MPEG-DASH in HTML5. Availble online https://bitmovin.com/status-mpeg-dash-today-youtube-netflix-use-html5-beyond/ (accessed February 8, 2017). (Februari 2015).

14. Sandvine, Inc. 2016. Global internet phenomena report. Available online https://www.sandvine.com/trends/global-internet-phenomena/ (accessed February 8, 2017). (2016).

15. Michael Seufert, Tobias Hosfeld, and Christian Sieber. 2015. Impact of intermediate layer on quality of experience of HTTP adaptive streaming. In *2015 11th International Conference on Network and Service Management (CNSM)*. IEEE, 256–260.

16. R. K. Sitaraman. 2013. Network performance: Does it really matter to users and by how much?. In *2013 Fifth International Conference on Communication Systems and Networks (COMSNETS)*. 1–10.