A FAIRNESS-AWARE SMOOTH RATE ADAPTATION APPROACH FOR DYNAMIC HTTP STREAMING

Li Liu [†], *Chao Zhou* [#], *Xinggong Zhang* [†⋆], *Zongming Guo* [†⋆]

[†]Institute of Computer Science & Technology, Peking University, Beijing, China
[#] Media Technology Lab, CRI, Huawei Technologies CO., LTD, Beijing, China
[⋆] Cooperative Medianet Innovation Center, Shanghai, China
Email: {liuli9203, zhangxg, guozongming}@pku.edu.cn, zhouchaoyf@gmail.com

ABSTRACT

Recently, Dynamic Adaptive Streaming over HTTP (DASH) has been widely deployed over the Internet. Under time-varying network conditions, it is, however, still a big challenge to provide smooth video bit-rate with high video quality, especially when multiple clients compete for the network resources where the fairness must be considered. In this paper, a fairness-aware smooth rate adaptation approach is designed for DASH under the scenario that multiple clients are competing for the network resources. To avoid the unfair bandwidth estimated by the client induced by the off-intervals during the downloading process, a probe-based bandwidth estimation method is designed which includes a logarithmic law based increase probing scheme and a conservative back-off based decrease probing scheme. Then, with the probed bandwidth, a dual-threshold based video bit-rate switching scheme is designed that buffer overflow/underflow is avoided, and smooth video bit-rate is also provided. The extensive experiments on our network testbed demonstrate that the proposed approach outperforms the existing schemes significantly.

Index Terms— Dynamic HTTP Streaming, Rate Adaptation, Fairness, Smoothness

1. INTRODUCTION

In recent years, Dynamic Adaptive Streaming over HTTP (DASH) has been widely used for video streaming service over the Internet[1, 2, 3, 4]. In DASH, a video content is encoded into multiple versions at various bit-rates. Each video version is further broken into small video chunks, which normally contains a few seconds worth of video. Using HTTP connections, a DASH client is able to dynamically request chunks from different versions. By dynamically monitoring the available bandwidth and client buffer occupancy, and throttling the video bit-rate to match the available bandwidth with version switching, DASH is able to achieve a continuous video playback at the best possible quality level[5]. However, due to the inherent bandwidth variations, it is still a big challenge to provide smooth video bit-rate with high video quality, especially when multiple DASH clients share a network bottleneck where the fairness must be considered[6, 7].

In DASH, video bit-rates are dynamically adjusted according to rate adaptation logic, which plays a critical role in guaranteeing high-quality video streaming service, thus attracting many research efforts, such as [4, 8, 9, 10, 11]. All these adaptation schemes aim to either adapt a video bit-rate to an available bandwidth so as to achieve a high bandwidth utilization, or ensure a buffer in the client to provide a continuous playback. However, due to the inherent bandwidth variations, there is a fundamental conflict between video bit-rate smoothness and bandwidth utilization, and existing schemes do not balance the needs for these two aspects well. Besides, all these schemes are designed with the underlying assumption that the TCP downloading throughput observed by a client is its fair share of the network bandwidth[12]. However, due to the off-intervals during the download processing[7], such schemes lead to video bit-rate oscillations and unfair bandwidth sharing when multiple clients compete over a common bottleneck link, as demonstrated in[12, 13, 14]. In[15], a rate-shaping approach is proposed to address the above problems by eliminating the off-intervals. However, this approach is conducted at the server-side, which has limitation in supporting the large-scale multimedia delivery since it will dramatically increase the burden on the web server or cache. In contrast to [15], to avoid the unfair bandwidth estimated by the client induced by the off-intervals, a probe-based method is proposed to perceive the fair-share bandwidth in [12]. However, it leads to a long convergence time and failure in properly tracking the time-varying bandwidth. Moreover, the tremulous probed bandwidth would lead to short-term rate oscillation and deteriorate user experience of streaming services.

In this paper, a fairness-aware smooth rate adaptation approach is designed for DASH under the scenario that multiple clients are competing for the network resources. First, to avoid the unfair bandwidth estimated by the client induced by the off-intervals during the downloading process, a probe-based bandwidth estimation method is designed to converge the probed bandwidth to the fair-share bandwidth by the probing mechanism. Specifically, when the probed bandwidth is

This work was supported by National Natural Science Foundation of China under contract No. 61471009 and National High-tech Technology R&D Program (863 Program) of China under Grant 2013AA013504.
⋆Corresponding author. Email:zhangxg@pku.edu.cn

smaller than the estimated bandwidth, which is equal to the chunk size divided by the time consumed to download that chunk, a logarithmic law based probing scheme is designed to increase the probed bandwidth to converge quickly whilst avoiding over-probing. On the other hand, when the probed bandwidth is higher than the estimated bandwidth, a conservative back-off scheme is designed to decrease probed bandwidth to avoid congestion. After obtaining the probed bandwidth, a dual-threshold based rate adaptation scheme is designed considering both the buffered video duration and the probed bandwidth. When the buffer occupancy stays between the two thresholds, the video bit-rate keeps unchanged and smooth video bit-rate is provided. Otherwise, when the buffer occupancy is too high or too low, an appropriate video bit-rate is selected to avoid buffer overflow/underflow.

The main contributions of this paper can be summarized in three-fold:

- A probe-based bandwidth estimation scheme is designed for DASH, which includes a logarithmic law based increase probing scheme and a conservative back-off based decrease probing scheme. It aims to converge the probed bandwidth to the fair-share bandwidth quickly whilst avoiding congestion.
- We propose a dual-threshold based smooth rate adaptation scheme combining the buffered video time and the probed bandwidth. In the scheme, the effect of bandwidth variation on video rate is eliminated and smooth video rate is provided. Moreover, the two thresholds are used as operation points to select the best video rate and a continuous video playback is obtained.
- Extensive experiments on our network testbed are conducted to investigate the performance of the proposed fairness-aware rate adaptation approach. And the experimental results demonstrate that the proposed approach outperforms the existing schemes significantly.

The rest of the paper is organized as follows. We present the probe-based bandwidth estimation method in section 2. A dual-threshold based smooth rate adaptation scheme is proposed in section 3. We show the experimental results in section 4, and conclude the paper in section 5.

2. FAIRNESS-AWARE BANDWIDTH PROBING

Due to the ON-OFF phenomenon in DASH, the bandwidth estimated by clients is discrepant. However, as demonstrated in [12], only when the bandwidth is oversubscribed, i.e., the congestion occurs, the bandwidth estimated by the clients is equal to the fair-share bandwidth (all the clients see the same available bandwidth). On the other hand, when congestion occurs, the requested video bit-rate cannot be supported by the bandwidth and playback freezing may happen. Thus, how to obtain the fair-share bandwidth without congestion is critical for improving the rate adaptation performance for DASH.

In this section, a probe-based bandwidth estimation method is designed that we try to converge the probed bandwidth to the fair-share bandwidth by probing mechanism. The intuition behind this method is that the bandwidth estimated by the clients is always the upper bound of the fair-share bandwidth due to the off intervals. Thus, the probed bandwidth will become closer to the fair-share bandwidth and equal to it during the probing process. On the other hand, the granularity of probing is very critical. Coarse granularity can converge the probed bandwidth quickly to the fair-share bandwidth, but it may over-probe and lead to congestion. While too fine granularity needs a long time to converge and cannot track the time-varying bandwidth well.

To solve the above challenges, a logarithmic law based probing scheme is designed to increase the probed bandwidth under the consideration that when the gap between the probed bandwidth and the fair-share bandwidth (we cannot obtain the fair-share bandwidth in practical, and its upper bound, i.e., the estimated bandwidth is used for approximation) is large, a coarse granularity probing scheme should be employed to increase the probed bandwidth quickly. On the other hand, when the gap is small, a fine granularity probing scheme should be adopted to avoid over-probing. When the probed bandwidth is higher than the fair-share bandwidth, congestion may occur leading to playback freezing which deteriorates the user experience heavily. In this case, a conservative back-off scheme is designed to decrease the probed bandwidth guaranteeing that no congestion happens.

Specifically, we denote $b^e(n)$ as the estimated bandwidth, i.e., the upper bound of the fair-share bandwidth for a client when downloading chunks, then we have

$$b^e(n) = \frac{r(n)T_0}{T(n)} \tag{1}$$

where $r(n)$ is the video bit-rate for chunk n, T_0 is the chunk length in seconds, and $T(n)$ is the time consumed to download chunk n. Let $b^p(n)$ denote the probed bandwidth initialized to zero, and $b^p(n)$ is used to select the video bit-rate for chunk n. Whenever a chunk is completely downloaded, the estimated bandwidth is update according to (1). Then, the probed bandwidth is updated as follows:

$$b^p(n) = \begin{cases} b^p(n-1) + \max(\frac{b^e(n-1)-b^p(n-1)}{2}, \triangle), \\ \qquad \text{if } b^p(n-1) < b^e(n-1) \\ b^p(n-1) + \alpha \cdot (b^e(n-1) - b^p(n-1)), \\ \qquad \text{if } b^p(n-1) \geq b^e(n-1) \end{cases} \tag{2}$$

where $\triangle$ is a constant to avoid too slow of the convergence and α is a positive constant satisfying that $\alpha \geq 1$.

3. SMOOTH RATE ADAPTATION

After obtaining the probed bandwidth in section 2, we now move to obtain the smooth video bit-rate with high band-

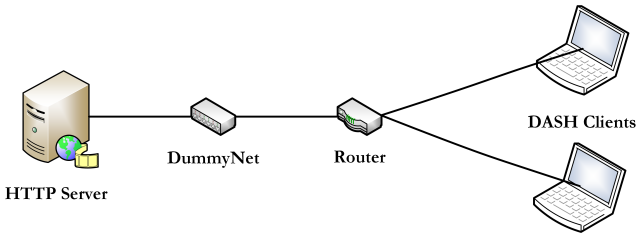


Fig. 1. Network topology in test bed

width utilization and a dual-threshold based rate adaptation scheme is designed combined with the probed bandwidth. In this work, the buffer occupancy is denoted by the buffered video duration considering that the client buffer may contain chunks from different versions, i.e., different video bit-rates and there is no longer a direct mapping between the buffered video size and the buffered video duration.

From the control system point of view, there is a fundamental conflict between maintaining smooth video bit-rate and stable buffer occupancy, due to the unavoidable network bandwidth variations. Nevertheless, from the end user point of view, video bit-rate fluctuations are more perceivable than buffer occupancy oscillations. The recent work in [16] has shown that switching back-and-forth between different bit-rates will significantly degrade user's viewing experience, whereas buffer occupancy variations do not have direct impact on video streaming quality as long as the video buffer does not deplete. Moreover, our prior work[17] has shown that the effect of short-term bandwidth variations on rate adaptation can be eliminated by using two thresholds. Thus, in this work, we also propose to use two thresholds, $q_{\min}$ and $q_{\max}$, to mitigate the effect of bandwidth variations on video rate adaption so as to provide smooth video bit-rate.

Combined with the probed bandwidth, a dual-threshold based smooth rate adaptation scheme is designed with the aim to provide a smooth video bit-rate whilst avoiding buffer overflow/underflow. When the buffer occupancy is lower than $q_{\min}$, to avoid buffer depleting and provide a continuous playback, the video bit-rate should be selected no higher than the probed bandwidth. Similarly, when the buffer occupancy is higher than $q_{\max}$, to ensure no buffer overflow happens while achieving high bandwidth utilization, a higher video bit-rate can be selected. At last, when the buffer occupancy falls in between the two thresholds, the risk of buffer overflow and underflow is low, and the video bit-rate should be kept unchanged so that smooth video quality is provided.

Specifically, consider that a video content is encoded into k versions with different video bit-rate. The set of video bit-rates is denoted by $R = \{r_i | 1 \leq i \leq k\}$, where r_i is the bit-rate of i-th video version and it satisfies that $0 < r_i < r_j$, $\forall i < j$. We define $q(n)$ as the buffer occupancy when starting to download chunk n. With the buffer occupancy $q(n)$ and the probed bandwidth $b^p(n)$ according to (2), the video bit-rate for chunk n is determined as

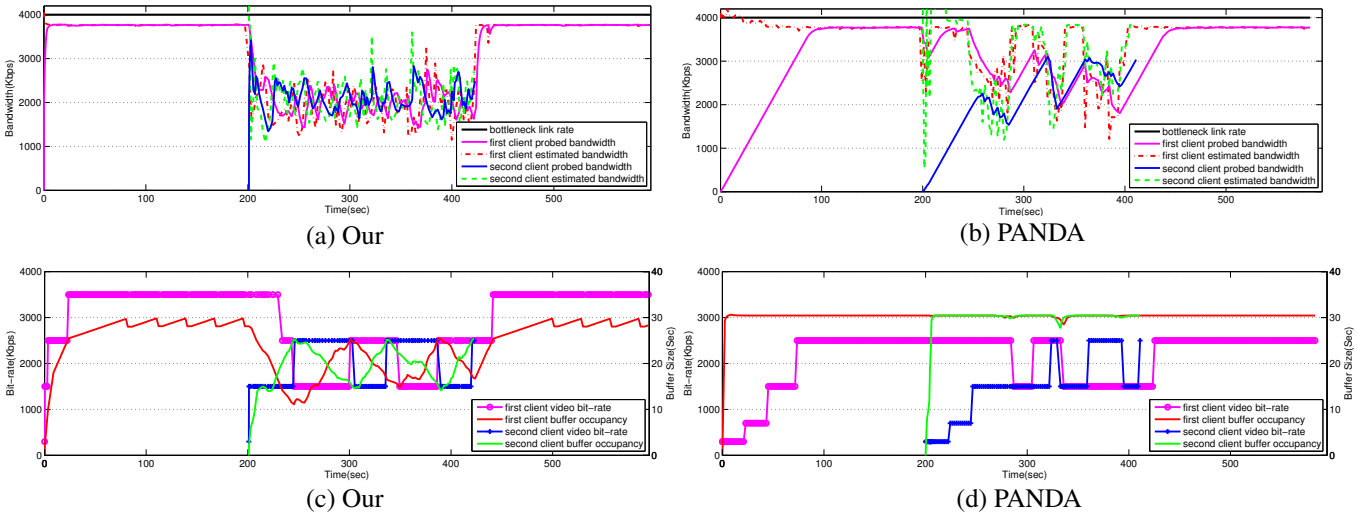
$$r(n) = \begin{cases} \max\limits_{1 \leq i \leq k} \{r_i | r_i \leq b^p(n)\} & \text{if } q(n) < q_{\min} \\ \min\limits_{1 \leq i \leq k} \{r_i | r_i \geq b^p(n)\} & \text{if } q(n) > q_{\max} \\ r(n-1) & \text{else} \end{cases} \quad (3)$$

where when the buffer occupancy is smaller than the $q_{\min}$, the maximal video bit-rate which is no higher than the probed bandwidth is selected so as to guarantee a continuous video playback; when the buffer occupancy is higher than the $q_{\max}$, the minimal video bit-rate which is no lower than the probed bandwidth is selected so as to improve video quality; otherwise, the video bit-rate keeps unchanged and smooth video bit-rate is provided.

4. PERFORMANCE EVALUATION

In this section, we evaluate the proposed fairness-aware smooth rate adaptation approach over our network testbed, as illustrated in Fig.1. Two DASH clients are connected to the same HTTP streaming server and the bottleneck link stays from the HTTP streaming server to the router. Dummynet[18] is used to control the available bandwidth of the bottleneck link. In our experiments, same with Netflix, the server provides five different versions of video bit-rates $R = \{r_1 = 300, r_2 = 700, r_3 = 1500, r_4 = 2500, r_5 = 3500\}$(kbps). Each video version is segmented into the same equal-length video chunks, which contains 2 seconds worth of video. Moreover, the buffer size is set to 30s and the thresholds are set as $q_{\min} = 15$s, $q_{\max} = 25$s. For performance comparison, besides our proposed method, the PANDA (Probe AND Adapt) proposed in [12] is also implemented.

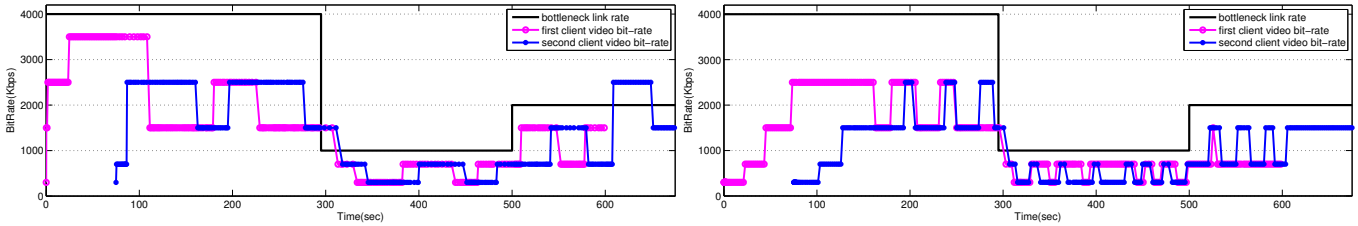
Firstly, as shown in Fig.2, we evaluate the performance of the two schemes where two clients are competing for the bandwidth during instant 200s to 400s (the second client joint the system at 200s and leave it at around 400s), and the available bandwidth is set to 4Mbps. Fig.2(a)-(b) show that our method is able to converge the probed bandwidth to the estimated bandwidth very quickly through the carefully designed logarithmic law based increase probing scheme. Note that it takes a long time for PANDA's probed bandwidth to converge due to the linear increasing scheme. The figures also show that when two clients compete with each other, PANDA fails to track the fair-share bandwidth well. As a result, the two clients fail to request the fair video bit-rates as shown in Fig.2(d). On the other hand, our method is able to well track the fair-share bandwidth and the probed bandwidth fluctuates around 2Mbps (which is equal to the fair-share bandwidth). Based on the probed bandwidth, the clients request the video bit-rates in a much fairer way, as shown in Fig.2(c). Besides, Fig.2(c-d) also show that our approach provides much smoother video bit-rates than PANDA when two clients com-



(a) Our (b) PANDA

(c) Our (d) PANDA

Fig. 2. Two identical players compete for available bandwidth. One player occupies a bottleneck with available bandwidth of 4Mbps, and another player joins at around 200s and leaves at around 400s.(a)-(b):estimated bandwidth and probed bandwidth; (c)-(d): video bit-rate and buffer occupancy evolution



(a) Our (b) PANDA

Fig. 3. Two identical players compete for available bandwidth. The players start at t=0s, and t=75s, respectively. The available bandwidth is initially set to 4Mbps, and drops to 1Mbps at around t=300s and increases to 2Mbps at around t=500s. (a)-(b): bottleneck link rate and video bit-rates

pete together. This is mainly because that our approach smooths out all the short-term variations of the probed bandwidth by employing a dual threshold which is similar to a low-pass filter. Whereas, the PANDA adapts the video bit-rate according to the probed bandwidth so as to maintain a stable buffer occupancy, as shown in Fig.2(d).

Finally, another set of experiments is conducted to evaluate the two schemes under the scenario that the available bandwidth varies between 1-4Mbps, as shown in Fig.3. Fig.3(a)-(b) show that the video rate in our proposed method is much smoother than in PANDA, this is mainly because that besides the probed bandwidth, a dual-threshold based rate selection scheme is adopted. Moreover, compared with PANDA, the requested video rate is also much fairer due to the high effective bandwidth probing scheme. Overall, the results have shown that multiple clients share the bandwidth fairly in our proposed rate adaption scheme while smooth video rate is also achieved with high bandwidth utilization.

5. CONCLUSION

In this paper, we proposed a fairness-aware smooth rate adaption approach for DASH under the scenario that multiple clients are competing for the network resources. To avoid the unfair bandwidth estimated by the client induced by the off-intervals, we have designed a probe-based bandwidth estimation method to obtain the fair-share bandwidth by probing mechanism. With the carefully designed logarithmic law based increase probing scheme and the conservative back-off based decrease probing scheme, our method is able to quickly converge the probed bandwidth to the fair-share bandwidth whilst avoiding over-probing. Then, combining the probed bandwidth and the buffered video duration, we proposed a dual-threshold based video rate switching scheme, by which smooth video rate is obtained with high bandwidth utilization and playback freeze is also avoided. The extensive experiments on our network testbed demonstrate that the proposed approach outperforms the existing schemes significantly.

6. REFERENCES

[1] T. Stockhammer, "Dynamic Adaptive Streaming over HTTP: Standards and Design Principles," in *Proc. ACM MMSys11*, 2011.

[2] M. Watson, "HTTP Adaptive Streaming in Practice," *Netflix, Tech. Rep.*, 2011.

[3] C. Zhou, X. Zhang, L. Huo, and Z. Guo, "A Control-theoretic Approach to Rate Adaptation for Dynamic HTTP Streaming," in *Proc. VCIP*, 2012.

[4] A. Zambelli, "IIS Smooth Streaming Technical Overview," *Microsoft Corp.*, 2009.

[5] James F.Kurose and Keith W.Ross, *Computer Networking: A Top-Down Approach*, Addison-Wesley, 6th edition, 2012.

[6] T.Y. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari, "Confused, Timid, and Unstable: Picking A Video Streaming Rate is Hard," in *Proc. ACM IMC*, 2012.

[7] Saamer Akhshabi, Lakshmi Anantakrishnan, Constantine Dovrolis, and Ali C. Begen, "What happens when HTTP adaptive streaming players compete for bandwidth," in *Proc. ACM Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV12)*, 2012.

[8] B. Zhou, J. Wang, Z. Zou, and J. Wen, "Bandwidth Estimation and Rate Adaptation in HTTP Streaming," in *Proc. IEEE ICNC*, 2012.

[9] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate Adaptation for Adaptive HTTP Streaming," in *Proc. ACM MMSys11*, 2011.

[10] L. De Cicco, S. Mascolo, and C.T. Abdallah, "An Experimental Evaluation of Akamai Adaptive Video Streaming over HSDPA Networks," in *Proc. IEEE Int. Syp. Computer-Aided Control System Design*, 2011.

[11] L. De Cicco, S. Mascolo, and P. Vittorio, "Feedback Control for Adaptive live Video Streaming," in *Proc. ACM MMSys11*, 2011.

[12] Zhi Li, Xiaoqing Zhu, Joshua Gahm, Rong Pan, Hao Hu, Ali C. Begen, and David Oran, "Probe and Adapt: Rate Adaptation for HTTP Video Streaming At Scale," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, vol. 32, pp. 719–733, APRIL 2014.

[13] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An Experimental Evaluation of Rate-Adaptation Algorithms in Adaptive Streaming over HTTP," in *Proc. ACM MMSys11*, 2011.

[14] T. Cloonan and J. Allen, "Competitive analysis of adaptive video streaming implementations," in *Proc. SCTE Cable-Tec Expo Technical Workshop*, 2011.

[15] Saamer Akhshabi, Lakshmi Anantakrishnan, Constantine Dovrolis, and Ali C. Begen, "Server-based traffic shaping for stabilizing oscillating adaptive streaming players," in *Proc. ACM Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV13)*, 2013.

[16] N. Cranley, P. Perry, and L. Murphy, "User perception of adaption video quality," *Int. J. Human-Comput. Stud.*, vol. 64, no. 8, pp. 637647, 2006.

[17] Chao Zhou, Chia-Wen Lin, Xinggong Zhang, and Zongming Guo, "A Control-Theoretic Approach to Rate Adaption for DASH over Multiple Content Distribution Servers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, pp. 681–694, APRIL 2014.

[18] L. Rizzo, "Dummynet: A Simple Approach to the Evaluation of Network Protocols," *SIGCOMM CCR*, vol. 27, no. 1, pp. 31–41, 1997.