A Statistical-based Rate Adaptation Approach for Short Video Service

Chao Zhou, Shucheng Zhong, Yufeng Geng, Bing Yu
Beijing Kuaishou Technology Co., Ltd
{zhouchao, zhongshucheng, gengyufeng, yubing}@kuaishou.com

***Abstract*—Dynamic adaptive streaming has been recently widely adopted for providing uninterrupted video streaming services to users with dynamic network conditions and heterogeneous devices in Live and VoD (Video on Demand). However, to the best of our knowledge, no rate adaptation work has been done for the new arisen short video service, where a user generally watches many independent short videos with different contents, quality, bitrate, and length (generally about several seconds). In this work, we are the first to study the rate adaptation problem for this scenario and a Statistical-based Rate Adaptation Approach (SR2A) is proposed. In SR2A, each short video is transcoded into several versions with different bitrate. Then, when a user watches the short videos, the network conditions and player status are collected, and together with the to be requested video's information, the best video version (bitrate or quality) will be selected and requested. Thus, the user will experience the short videos with the most suitable quality depending on the current network conditions. We have collected the network trace and user behavior data from Kuaishou[1], the largest short video community in China. By the collected data set, the users' watching behavior is analyzed, and a statistical model is designed for bandwidth prediction. Then, combined with the video information derived from the manifest, the maximal video bitrate is selected under the condition that the probability of play interruption is smaller than a predefined threshold during the whole playback process. The trace based experiments show that SR2A can greatly improve the user experience in quality and fluency of watching short videos.**

I. Introduction

In recently, the explosive growth of short video service has demonstrated that short video is one of the most popular medium for people to record and share their experience, such as Kuaishou (or Kwai) [1], Instagram [2], and so on. In this kind of short video service, the users shoot short videos, do some process (such as adding special effects, magic emotion), and upload them to the media servers. At the server side, when the videos are authenticated and validated, they can be transcoded into several quality (bitrates) to satisfy the heterogeneous networks conditions and diverse end devices. Due to the time-varying network conditions and limited bandwidth, the users may experience interruption when watching a certain video, and this heavily deteriorates the quality of experience (QoE). To solve this problem, dynamic rate adaptation has proven to be one of the most effective techniques since it can automatically throttle the visual quality to match the available bandwidth so that a user receives the video at the maximum possible quality. To improve the streaming QoE, dynamic rate adaptation (such as MPEG-DASH, HLS) has been widely adopted by the video service providers, such as Youtube [3], Netflix [4], and so on. And also many high-efficiency rate adaptation algorithms have been designed [5], [6], [7], [8], [9]. However, all of the existing applications and research work are focusing on the scenario of Live and VoD (Video on Demand). To the best of our knowledge, no rate adaptation work has been done for the new arisen short video service.

In short video service, whenever a user begin to watch a new short video, the best video bitrate can be selected by the rate adaptation algorithm to maximal the visual quality under the current network conditions. However, design a high-efficiency rate adaptation algorithm is very challenging in the scenario of short video service, and it is different from that in Live or VoD. First, there is no buffered video data can be used to compensate the bandwidth variation as done in Live and VoD [9], [10], [5], [6], [11]. Then, the different videos are viewed at discrete and random instants of time (depending on the user's behavior), leading to the serious ON-OFF phenomenon [12], where "ON" denotes the network is downloading, and "OFF" denote the network is idle. During the "OFF" period, there is no information about the network conditions can be collected, make the bandwidth estimation challenging. At last, to improve the encoding efficiency, generally, Variable Bit Rate (VBR) is adopted that the actual instantaneous bitrate may vary heavily and it is depending on the video content. This also makes rate adaptation challenging since the video quality can only be switched at the beginning of downloading.

In this work, we have studied the rate adaptation problem in short video service and a Statistical-based Rate Adaptation Approach (SR2A) is proposed. In SR2A, each short video is transcoded into several versions with different bitrate after it has been uploaded to the server. The details of the versions, such as the Uniform Resource Locator (URL) of each version, the duration, the instantaneous bitrate, and so on, are all described by a manifest file. For the users, they request and parse the manifest at first, and then they can selectively watch the video with a certain bitrate which is dynamic determined by SR2A depending on the user's network conditions and behavior. SR2A is designed depending on the data set collected from Kuaishou [1], including the network condition data and user behavior data. By the data, we find that when the

[1]https://www.kuaishou.com

978-1-5386-4458-4/18/$31.00 ©2018 IEEE

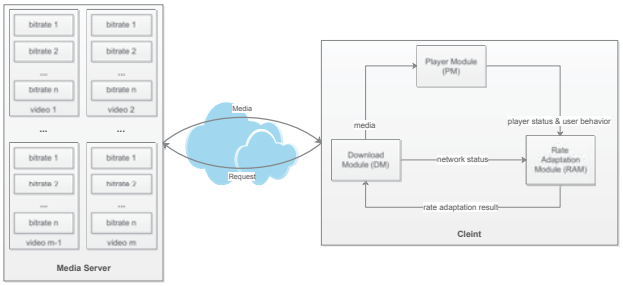

Fig. 1. Sketch of system structure for short video service

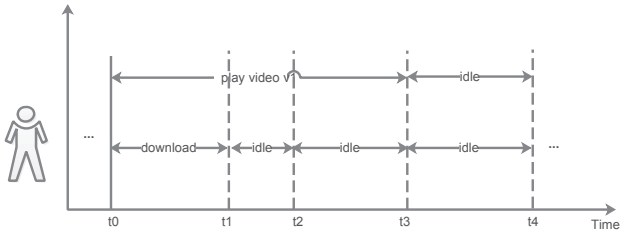

Fig. 2. Typical user watching behavior

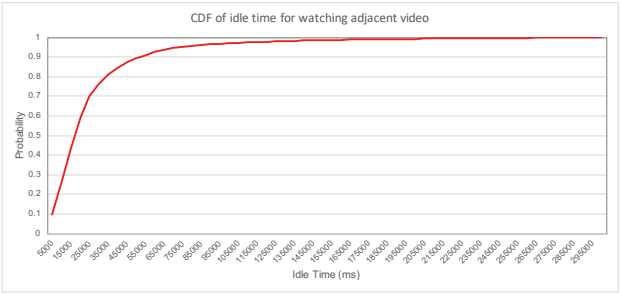

Fig. 3. Idle time distribution

users begin to watch the short videos, generally, they watch several videos continuously, and the interval for watching the adjacent videos is about several seconds. After that, there may have a relative long gap, typical several hours, that no video is requested. Thus, we have designed a statistical model to predict the distribution of the network bandwidth. Then, combined with the users behavior, i.e., the distribution of the interval for watching the adjacent videos, we aim to maximize the video bitrate while guaranteeing the risk of interruption is below a given threshold. We have further formulated it into an optimization problem. At last, we have evaluated the performance of SR2A over the data set, where the network is simulated by the collected bandwidth trace, and the user watching behavior is imitated from the collected intervals for watching adjacent videos. The results have shown that SR2A can achieve a good trade-off between the video quality and risk of interruption.

II. System Structure

In this section, we give a sketch about the system structure for short video service. Fig. 1 shows the infrastructure of a typical short video service system with dynamic rate adaptation. At the server side, video content is transcoded into multiple discrete bitrates, and the different encodings are stored as independent files for each bitrate. The client includes three modules: download module (DM), player module (PM), and rate adaptation module (RAM). The DM downloads the suitable video version from the server based on the rate adaptation decision from RAM, and sends the media data to the PM, in which the media data will be decoded and rendered. The RAM continuously collects the network conditions (such as throughput) from DM, and player status (such as if playback interruption happens) from PM. Based on the collected data, the RAM dynamically make the best rate adaptation decision for the DM whenever a new short video is requested.

It is worth to note that in short video service, the downloading process is very different from that in traditional Live or Video on Demand (VoD). For Live or VoD, a session is related long, and the client will continuously download the media data unless it reaches the end or the user close the session. Therefore, during the downloading process, the client can continuously collect the status of network and player. Besides, it can also prefetch and buffer some media data to compensate the network variations. For example, when watching a film by a DASH client, the client can download and buffer several fragments when the throughput is high. This is very useful to compensate the network variations, and it is also the precondition when design some high-efficiency rate adaptation schemes [5], [7], [13], [14].

While in short video service, since each video sequence is very short, no fragmentation is implemented, otherwise it will loss the encoding efficiency by adding at least one I-frame for each fragment. Though we can treat each short video as a fragment (as that in DASH). However, the essence is very different that each short video is independent in short video service, therefore, when a new short video is requested, the buffer at the client is empty. This is similar to the scenario that for each video, the user only requests and watches the initial (first) fragment, and changes to the others. No buffer makes the rate adaptation challenging that, generally, the lowest bitrate is requested for the initial fragments to decrease the start-up delay and quickly fill the buffer, and then it may changed to request the next fragments with high bitrate. But this cannot be done in short video service, since it will result that the video with the lowest bitrate is requested all the time.

Besides, when a short video is completely downloaded, the network is idle unless a new video is requested (user clicks a new video). The idle time is affected by many factors, such as the video bitrate, network bandwidth, user behavior, and so on. A typical user watching behavior is shown in Fig. 2. At t_0, the user click the short video, and it begins downloading and rendering. Since the start-up delay in short video service is very short (generally about 200ms), we roughly ignore this delay and treat that the downloading and rendering process begins at the same time. At t_1, the video is completely downloaded, but the video is rendered completely at t_2 ($t_2 \geq t_1$) for the first time. Then the user

may playback this video in loop until t_3 ($t_3 \geq t_2$) that the video is obtained from the cache other than downing again. After that, the user may go to the homepage and select the next video at t_4 ($t_4 \geq t_3$). Then we need to make the rate adaptation decision at t_4 for the next request, while we know nothing about the network from t_1 to t_4 as the network is idle during this period. Missing the knowledge about the network condition also makes the rate adaptation challenging. Fig. 3 has shown the cumulative distribution function (CDF) of the idle time distribution of about one hundred thousand users' ten million watching behaviors (requests) that the idle time mainly range from 5000ms to 35000ms.

III. Statistical-based Rate Adaptation Approach

In this section, we describe our proposed SR2A for short video service. In SR2A, the client makes a rate adaptation decision to select the maximal available video bitrate under the condition that the probability of interruption is smaller than a predefined threshold during the whole playback process. We continuously collect the average throughput for each τ seconds, and stored them in a sliding window with length of T seconds. The sampled average throughput is quantized into several discrete values denoted as $(b_1, b_2, ..., b_n)$ where for a given throughput b, it is quantized to b_i as:

$$\begin{aligned} &\max b_i \\ &\text{s.t. } b \leq b_i \end{aligned} \tag{1}$$

Then, the histogram is used to denote the distribution of the throughput as Fig. 4. The probability of the throughput b_i is p_i with $\sum p_i = 1$.

For watching each short video, the buffered video time process, represented as $q(t)$, can be modeled as a queue with a constant service rate of unity, i.e., in each second, a piece of video with playback length of one second is dequeued from the buffer and then played. The enqueue process is driven by the video download rate and the downloaded video version. Specifically, we assume a short video is encoded into L different average bitrate $V_1 < V_2 < ... < V_L$. Generally, the video is encoded/transcoded with Variable Bit Rate (VBR) to improve the coding efficiency compared with Constant Bit Rate (CBR). Therefore, the actual bitrate is time-varying. For the j_{th} bitrate of the i_{th} short video, the actual bitrate of t_{th} second is denoted as v_{ij}^t with $t = 1, 2, ..., d_i$ where d_i is the duration of i_{th} short video. Therefore, the buffered video time evolution becomes

$$q(t) = \Delta + t - \sum \frac{v_{ij}^t}{b(t)} \tag{2}$$

where $b(t)$ is the throughput with the statistic histogram distribution, and Δ is the buffered video date in time (typical set to the duration of about $3 \sim 5$ frames) when begin to playback.

Since play interruption greatly deteriorates user experience, we denote ε the maximal acceptable interruption probability when watching a short video, therefore, the video bitrate is selected with $P(q(t) < 0) \leq \varepsilon$ where $P(q(t) < 0)$ denotes the probability that interruption happens at t. Combined with (2), we have

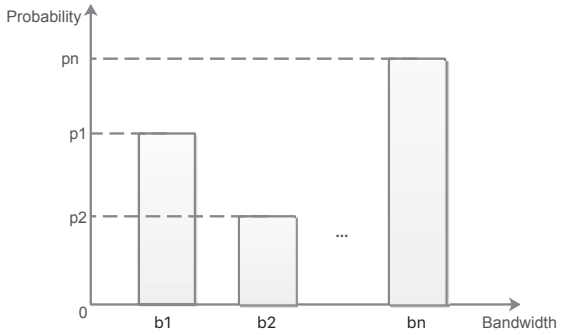

Fig. 4. Histogram distribution of the throughput

$$P(\sum \frac{v_{ij}^t}{b(t)} - t > \Delta) < \varepsilon,\ \forall 0 \leq t \leq d_i \tag{3}$$

At last, the SR2A can be formulated as

$$\begin{aligned} &arg\max i \\ &\text{s.t. } (3) \\ &1 \leq i \leq L \end{aligned} \tag{4}$$

where v_{ij}^t and V_i are obtain from the manifest, L is the total number of available video versions, and $b(t)$ is throughput with the statistic histogram distribution in window T.

IV. Experiment

In this section, we evaluate SR2A based on the date collected from Kuaishou [1]. In Kuaishou, for each day, more than ten million short videos are uploaded, and the short video are downloaded more than ten billion times by over one hundred million users. We have collected one hundred thousand users' ten million watching behaviors, including the throughput, the idle time, the interruption evens, and so on. Based on the data set, the parameters are set as $\tau = 0.5$, $T = 80$, $\varepsilon = 0.05$. And each video is transcoded into three bitrate $\{1500kbps, 2000kbps, 2700kbps\}$, it is worth to noting that the bitrate is $vbv\_maxbitrate$ in the coding configuration with vbv is adopted for video encoding, and the actual bitrate depends on the video content and may vary heavily as explained in Section II. We have randomly selected three thousands short videos to simulate 100 users' watching behavior (30 video for each user) under the throughput trace with about one hundred thousand seconds. For each user, the idle time between any adjacent videos is also obtained from the data set as Fig. 3 shows.

We compared SR2A with fixed rate approach (FRA, which is used by nearly all short video applications) and random rate selection approach (RRSA). For all approaches, we have compared the average video bitrate (AVB), the interrupted video count (IVC, i.e., the total number of videos which have experienced interruption during the whole watching process), the interruption times count (ITC, the total interruption times

TABLE I
PERFORMANCE SUMMARY

		AVB (kbps)	IVC	ITC	BT
FRA	2700kbps	1495.55	12.86	62.29	61.11
	1500kbps	827.61	2.96	9.2	7.88
RRSA		1046.06	7.65	32.73	31.26
SR2A		1172.45	3.54	10.27	8.72

TABLE II
DETAIL PERFORMANCE FOR ONE USER

		AVB (kbps)	IVC	ITC	BT
FRA	2700kbps	1370.56	9	29	25.27
	1500kbps	759.09	1	1	0.88
RRSA		1020.25	4	9	7.24
SR2A		1192.12	1	1	0.71

during the whole watching process), and the buffering time (BT, the time consumed for buffering).

The summary results are shown in Table I which has shown the average performance for all one hundred users. We can find that in FRA, when the selected bitrate ($vbv\_maxbitrate$) is set to the maximal available bitrate, i.e., 2700kbps, the maximal average bitrate is obtained with the highest number of interrupted videos and interruption times, also the highest BT. In this case, the users will experience serious interruption which deteriorates the QoE. On the other hand, when the bitrate is fixed to 1500kbps (the minimal available bitrate), the best performance in IVC, ITC, and BT is obtained, but the video quality (AVB) is the worst. Under this scheme, though the risk of interruption is the lowest, the user cannot get the best possible quality. It is easy to find that any fixed bitrate is hard to adapt to the time-varying and diversified network conditions. The mainly reason is that there is a inherent tradeoff between the video quality and playback interruption, and even the random scheme can obtain a better tradeoff than FRA as Table I shows. This is also the motivation for us to design SR2A, a rate adaptation approach to dynamically select the suitable video birtate to obtain the best tradeoff between the video quality and playback interruption. The results in Table I shows that with nearly no performance loss in IVC, ITC and BT, SR2A obtains much larger average bitrate than FRA of $vbv\_maxbitrate = 1500kbps$. While on the other hand, compared FRA of $vbv\_maxbitrate = 2500kbps$, SR2A has much lower IVC, ITC and BT with acceptable average bitrate loss. When compared with FRA, SR2A performance better in all the metrics, this also demonstrates the high efficiency of out proposed scheme.

Moreover, we have shown the results of some particular one user's detail performance in Table II, it shows similar characters with that in Table I. But it is worth to noting that when comparing SR2A with FRA (with $vbv\_maxbitrate = 1500$), we can find that the video quality (bitrate) is greatly improved with no interruption performance loss. This also demonstrates the importance of rate adaptation on short video service to improve users' QoE, and this work is the beginning to make up the gap in this domain.

V. CONCLUSION

In this work, we have studied the rate adaptation problem for short video service and a Statistical-based Rate Adaptation Approach is designed to improve the streaming QoE. By the data set collected from Kuaishou, we have analyzed the user watching behavior and designed a statistical model for bandwidth prediction. Then, combined with the video information derived from the manifest, the maximal video bitrate is selected under the condition that the probability of play interruption is smaller than a predefined threshold during the whole playback process. The trace based results have demonstrated the high-efficiency of the proposed rate adaptation approach. This work is the beginning to make up the gap in rate adaptation for short video service. For future work, the smoothness among the videos, together with more effective bandwidth estimation model can be studied to improve the QoE.

REFERENCES

[1] [Online]. Available: https://www.kuaishou.com/
[2] [Online]. Available: https://www.instagram.com/
[3] [Online]. Available: https://www.youtube.com/
[4] [Online]. Available: https://www.netflix.com/
[5] C. Zhou, C.-W. Lin, X. Zhang, and Z. Guo, "A Control-Theoretic Approach to Rate Adaption for DASH Over Multiple Content Distribution Servers," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 24, no. 4, pp. 681–694, Apr. 2014.
[6] M. Xing, S. Xiang, and L. Cai, "A Real-Time Adaptive Algorithm for Video Streaming over Multiple Wireless Access Networks," *IEEE J. Select. Areas Commun.*, vol. 32, no. 4, pp. 795–805, Apr. 2014.
[7] C. Zhou, C. W. Lin, and Z. Guo, "mdash: A markov decision-based rate adaptation approach for dynamic http streaming," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 738–751, 2016.
[8] L. De Cicco, S. Mascolo, and P. V., "Feedback Control for Adaptive Live Video Streaming," *in Proc. ACM Multimedia Syst.*, pp. 145–156, Feb. 2011.
[9] 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*, vol. 32, no. 4, pp. 719–733, 2014.
[10] A. Begen, T. Akgul, and M. Baugher, "Watching video over the web: Part 1: Streaming Protocols," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 54–63, Mar. 2011.
[11] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate adaptation for adaptive HTTP streaming," *in Proc. ACM Multimedia Syst.*, pp. 169–174, Feb. 2011.
[12] S. Akhshabi, L. Anantakrishnan, A. C. Begen, and C. Dovrolis, "What happens when http adaptive streaming players compete for bandwidth?" in *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video*. ACM, 2012, pp. 9–14.
[13] T.-Y. Huang, R. Johari, and N. McKeown, "Downton abbey without the hiccups: Buffer-based rate adaptation for http video streaming," *in Proc. ACM SIGCOMM workshop on Future human-centric multimedia networking*, pp. 9–11, Aug. 2013.
[14] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive," *IEEE/ACM Transactions on Networking (TON)*, vol. 22, no. 1, pp. 326–340, 2014.