Impact of Network Dynamics on User's Video Quality: Analytical Framework and QoS Provision

Tom H. Luan, Lin X. Cai, and Xuemin (Sherman) Shen, *Fellow, IEEE*

***Abstract*—We develop an analytical framework to investigate the impacts of network dynamics on the user perceived video quality. Our investigation stands from the end user's perspective by analyzing the receiver playout buffer. In specific, we model the playback buffer at the receiver by a $G/G/1/\infty$ and $G/G/1/N$ queue, respectively, with arbitrary patterns of packet arrival and playback. We then examine the transient queue length of the buffer using the diffusion approximation. We obtain the closed-form expressions of the video quality in terms of the start-up delay, fluency of video playback and packet loss, and represent them by the network statistics, i.e., the average network throughput and delay jitter. Based on the analytical framework, we propose adaptive playout buffer management schemes to optimally manage the threshold of video playback towards the maximal user utility, according to different quality-of-service requirements of end users. The proposed framework is validated by extensive simulations.**

***Index Terms*—Diffusion approximation, playout buffer, quality-driven.**

I. Introduction

MULTIMEDIA streaming services such as live video broadcasting, video on demand, video conferencing, etc., are expecting to be extensively deployed in the future with the evolution of broadband communication networks. However, as real-time multimedia services have stringent quality-of-service (QoS) requirements to maintain user's satisfaction, high quality video streaming over variable bit rate (VBR) channels still represents several fundamental challenges in engineering [1]. In specific, the packet communication networks are by nature dynamic with network channels stochastically shared by various media flows of different traffic characteristics. This leads to the dynamic changes of end-to-end throughput and delays. As video packets have strict deadlines of presentation, the varying network delays may result in the missing of playback deadline and consequently the jerkiness or even frozen of video playback. With the widespread adoption of wireless access technologies (e.g., IEEE 802.16 WiMAX and IEEE 802.11 WLAN) as the last-mile multimedia service distribution, the network dynamics are further intensified due to the inherent time-varying wireless channels. Therefore, to understand how the network dynamics affect the user's perceived video quality and effectively accommodate the dynamic network delays in the video playback are crucial.

There are a variety of techniques appeared in the literature to address the problem of packet video delivery over time-varying VBR channels [1], [2], including the rate-distortion optimized packet scheduling [3] and routing [4]–[6], power control and adaptive coding at the transmitter [7]–[9], playback rate and strategy adaption at the receiver [10], etc. However, most of these approaches focus on the network-side issues which consider the network optimization in terms of throughput, delay and delay jitter; nevertheless they fail to fully investigate the resultant video quality from the end user's perspective, which motivates our work.

In this study, we analyze and optimize the user perceived video quality in terms of the start-up delay, fluency of video playback and packet loss rate. By intelligently using the user-side resources which are dependent on network dynamics, we design adaptive quality-driven video streaming schemes towards the maximal user utility. To this end, we first develop an analytical framework to reveal the impacts of network dynamics on the video quality perceived by the end user. In specific, to overcome the network dynamics, the playout buffer is usually deployed as shown in Fig. 1, which stores the received video packets and delays the initial media playout for a short period until a certain playback threshold is reached. This short period constitutes the start-up delay. During the media playout, the packets are discharged from the playout buffer and injected to the media player for playback. As long as the playout buffer is non-empty, the continuous media playout is always guaranteed. Since the occupancy of the playout buffer is closely related to the user's viewing experience in terms of start-up delay and smoothness of playback, we study the video quality by analyzing the evolution of the queue length under the network dynamics. Secondly, we propose adaptive schemes to optimally determine the playback threshold driven by the users' required video quality. We strike a trade-off between the start-up delay, smoothness of playback and video packet loss, by adjusting the playback threshold. Specifically, to ensure smooth media playout, enough packets should be cached initially in the playout buffer to absorb the variations of packet arrivals. This, however, may incur an intolerable long waiting time to end users. Meanwhile, to prevent packet loss due to overflow of the finite buffer, the queue length should be maintained at a relatively low level with less packets buffered

Manuscript received October 01, 2008; revised April 01, 2009. First published November 13, 2009; current version published December 16, 2009. This work was supported jointly by the Natural Sciences and Engineering Council (NSERC) of Canada under Strategic Grant and Research In Motion (RIM). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Aggelos K. Katsaggelos.

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: hluan@bbcr.uwaterloo.ca; lcai@bbcr.uwaterloo.ca; xshen@bbcr.uwaterloo.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2009.2036294

1520-9210/$26.00 © 2009 IEEE

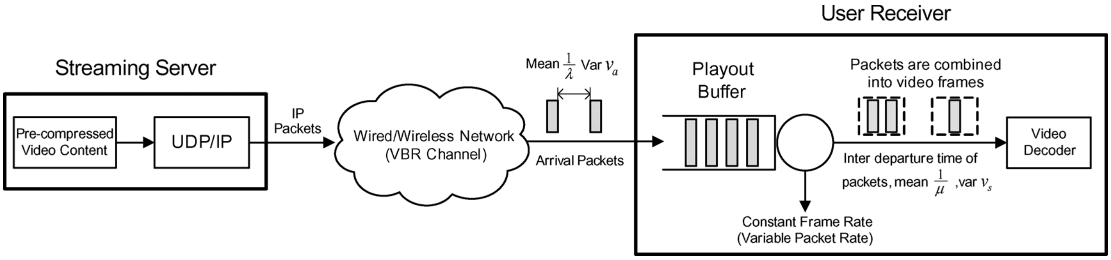

User Receiver
Streaming Server
Pre-compressed Video Content
UDP/IP
IP Packets
Wired/Wireless Network (VBR Channel)
Mean $\frac{1}{\lambda}$ Var v_a
Arrival Packets
Playout Buffer
Packets are combined into video frames
Video Decoder
Inter departure time of packets, mean $\frac{1}{\mu}$,var v_s
Constant Frame Rate (Variable Packet Rate)

Fig. 1. Process of video streaming.

initially. However, low queue occupancy may cause frequent playback interruptions. Thus, the playback threshold should be adaptively and optimally determined under the specific requirements of video quality metrics. Based on the analytical framework, we formulate the playback threshold selection as a stochastic optimization problem driven by the specific video quality requirements, and provide the optimal solutions to guarantee the stochastic video performance to end users.

Our main contributions are in three-fold.

- *General Modeling*: We consider a general network setting characterized by the first two moments of statistics, i.e., the mean and variance of traffic arrival rate and the video playback rate. We study both infinite buffer and finite buffer cases by modeling the playout buffer as $G/G/1/\infty$ and $G/G/1/N$ queues, respectively. In this way, the proposed analytical model is general and suitable for a diverse range of video codecs, video streaming applications, and network scenarios.
- *Compact Solution*: We apply the diffusion approximation to derive the closed-form expressions of the video quality metrics in terms of the start-up delay, the number of playback frozens, and the packet loss rate, and represent the video quality metrics by the network statistics, i.e., the average throughput and delay jitters. With the obtained results, we can evaluate the impacts of network statistics on the user's video quality. In a reverse manner, given the user's specific requirements on video quality, we can also conveniently determine the demanded network throughput to support the required video quality. In this way, the achieved compact solutions pave the way for quality-driven network resource allocation.
- *Distributed Optimal Control*: With the network statistics as an input to our analytical model, we design adaptive playout buffer management schemes to optimally select the playback threshold to cater to users with different quality requirements. The proposed schemes are employed in a distributed manner via local estimations of users only without any assistance from the networks, which is hence particularly suitable for large scale network deployments.

The remainder of this paper is organized as follows: Section II reviews the related works. The analytical framework is presented in details in Section III by considering both the infinite buffer and finite buffer cases. Section IV describes adaptive playout buffer management schemes, and Section V validates the achieved analytical results and optimal control schemes by extensive simulations. Section VI closes the paper with the concluding remarks.

II. Related Work

Streaming media over unreliable and time-varying VBR channels has attracted an extensive research attention in the last decade. Various network-adaptive schemes have been proposed [1]. Our work belongs to the scope of end-system centric solutions [2] which adapt the video from the receiver's perspective.

The end-system centric solutions refer to the adaptive video streaming mechanisms which adaptively modify the visual quality via the playback rate control or playout buffer management at the end-systems based on the occupancy of playout buffer or user's available bandwidth [10]. Liu *et al.* [11] propose an end-to-end playback rate adaptation scheme based on the layer coding technique. Each receiver actively measures its local available bandwidth and pass that to the server. Based on the echo information, the server then determines the appropriate number of layered streams conveyed to users and hence adapts the video compression rate according to the available bandwidth. By doing so, the visual quality degrades with enhanced compression ratio when the end-to-end bandwidth is insufficient; nevertheless, users can enjoy smooth playback. Galluccio *et al.* [12] describe an adaptive MPEG video streaming framework in which the wireless channel is modeled as a Rayleigh fading channel represented by an FSMC (finite state Markov chain). By analyzing the channel status via the Markovian model, the available bandwidth can be computed and the appropriate video playback rate is determined accordingly. Similar approach is adopted in [7] where channel coding is adapted in different channel conditions which are evaluated using FSMC. However, the source adaption schemes suffer from the scalability issue, as the server needs to respond to each individual user to resolve different quality requirements. When the network scales to a large size, the server can be easily overloaded. Moreover, most of the previous works on wireless multimedia transmission only consider a single-hop wireless channel which can be well modeled by FSMC. However, if multi-hop wireless transmissions and heterogeneous networks are considered, the analysis become invalid as accurate channel model in this case is generally not available.

To distribute the computation burden to the end users and hence enhance the network scalability, some approaches have been proposed to adapt the video playback at the end users. Kalman *et al.* [13] introduce the adaptive media playback (AMP) scheme at the end user, which can adaptively tune the video playout rate according to the playout buffer occupancy to ensure the smooth video playback. In specific, when the occupancy of playout buffer is above some threshold, the video playback rate will be increased to avoid the overflow of playout buffer. This leads to the effects of fast forward to the users. Laoutaris *et al.* [14] adopt the same mechanism but use the Markov decision process (MDP) to optimally determine the video playback rate at different channel conditions.

Another prevailing way of adaptive video streaming is by playout buffer management. In this scenario, the key issue is how to optimally determine the playback threshold to maximize the duration of continuous playback while minimizing the start-up delay. Liang *et al.* [15] establish a Markovian model to study the tradeoff between playback continuity and start-up delay. The wireless channel is modeled as an FSMC and the interplay between the channel statistics and playout buffer is provided under different buffer strategies. However, only the single-hop scenario is considered which cannot be applied to multi-hop transmissions. Dua *et al.* [16] propose to adapt the playback threshold through a MDP. The channel is also modeled as an FSMC, where each successful transmission incurs certain profit. The playout buffer is managed to determine an optimal playback threshold to maximize the overall profit. In [17], the upper bound and lower bound of the jitter-free probability have been derived. However, this work only provides the probability of smooth playback without interruptions and does not capture the interruption frequency, which is of more interest. Instead of providing performance bounds, we provide the closed form expressions of the video quality.

Different from the previous efforts, we consider a general network scenario by modeling the playout buffer at the user end as a $G/G/1/\infty$ and $G/G/1/N$ queue, with both finite and infinite buffer cases. The analytical model could be applied to not only the single-hop wireless networks but also the multi-hop wired/wireless networks. Furthermore, since the network can be highly dynamic with intensive variance, we explicitly take delay jitter into consideration and show its impacts on the perceived video quality.

III. Analytical Framework

In this section, we first present the architecture of video streaming system. After that, we describe a general model for the playout buffer and develop an analytical framework to study the video quality perceived by the user, considering both infinite and finite buffer cases.

A. Model of Playout Buffer

Fig. 1 shows a typical architecture of media streaming [18]. In Fig. 1, the raw video contents are pre-compressed and saved in the storage devices. Upon the user's request, the media server retrieves the pre-stored content and then segmented it into packets. The video packets are transmitted over a lossy variable bit rate (VBR) heterogeneous network using the User Datagram Protocol (UDP)/IP protocol suite. With network dynamics, packets arrive at the user with variable delays. Without loss of generality, we assume that the inter-arrival time of video packets follows a given but arbitrary distribution with mean 1λ and variance v_a. At the user end, the downloaded packets are first stored at the playout buffer, then combined into video frames and injected into the video player at the same cadence of frame rate that the video encoder generates. As the video frames are played at the constant rate, the service rate in terms of video packets is hence variable. We consider that the inter-departure time of video packets, determined by the instantaneous video playback rate, also follows a general distribution with the constant mean $1/\mu$ and variance v_s.

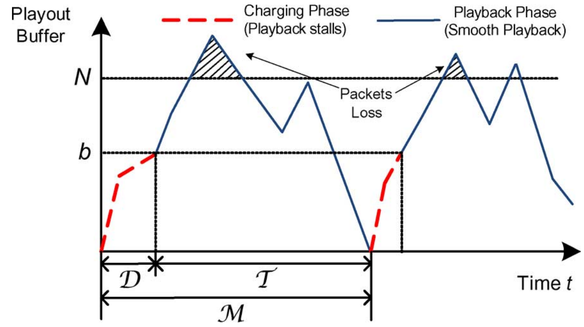


Fig. 2. Evolution of the playout buffer during media playout with buffer size N.

The playout buffer thus can be modeled as a $G/G/1/\infty$ queue when the buffer size is infinite or a $G/G/1/N$ queue when the buffer is finite. For the remaining part of this section, we analyze the evolution of the playout buffer in the infinite buffer and finite buffer cases, respectively, given network and playback statistics, i.e., λ, μ, v_a and v_s.

B. Infinite Buffer Case

We first consider the infinite playout buffer case, i.e., the buffer is infinitely large or large enough to accommodate the whole video file. This is typically true when the end host systems are personal computers with a large hard disk.

In general, the video playback process can be divided into two iterative phases, namely the charging phase and playback phase, as shown in Fig. 2. The charging phase starts once the playout buffer becomes empty. In this case, the buffer is charged with continuously downloaded packets and the media playback is kept frozen until b packets are filled. Henceforth, we refer to b as the threshold of playback. Let R.V. (random variable) $\mathcal{D}$ denote the duration of the charging phase. The playback phase starts once the playback threshold b is reached and packets are discharged from the buffer for playback. Due to dynamic packet arrivals and departures of the buffer, the playback phase may stall when the playout buffer becomes empty again. Let R.V. $\mathcal{T}$ denote the duration of the video playback phase. The charging and playback phases iterate until the whole video is downloaded.

In this work, we evaluate the user's video quality in following two aspects: the *start-up delay* and *smoothness of video playback*. The former refers to the time period that users have to wait before video playback starts, which is the duration of the charging phase $\mathcal{D}$ in Fig. 2. The latter is evaluated by the likelihood or frequency of playback frozens during the media playout. The trade-off between the two aspects of video quality is adapted by the playback threshold b. A larger threshold b results in a longer start-up delay, but makes the playback less likely freeze during the media playback. In what follows, we develop a mathematical framework to investigate this trade-off and evaluate the impacts of the threshold b and network statistics on the two quality metrics.

1) Diffusion Approximation: To evaluate the length of start-up delay $\mathcal{D}$ and the frequency of playback frozens, we model the playout buffer as a $G/G/1/\infty$ queue and resort to the diffusion approximation [19], [20] for compact solutions [21].

Denote the buffer size at time instant t by $B(t)$. The diffusion approximation method consists in replacing the discrete buffer size $B(t)$ by a continuous process $X(t)$ and model it as the Brownian motion

$$dX(t) = X(t+dt) - X(t) = \beta dt + G\sqrt{\alpha dt} \qquad (1)$$

where $G \sim N(0,1)$ is a normally distributed random variable with zero mean and unit variance. β and α are called drift and diffusion coefficients, respectively, defined by

$$\begin{cases} \beta = E\left(\lim\limits_{\Delta t \to 0} \frac{X(t)}{\Delta t}\right) = \lambda - \mu \\ \alpha = Var\left(\lim\limits_{\Delta t \to 0} \frac{X(t)}{\Delta t}\right) = \lambda^3 v_a + \mu^3 v_s. \end{cases} \qquad (2)$$

Let $p(x,t|x_0)$ denote the conditional probability density function (p.d.f.) of the buffer size $X(t)$ at time t

$$p(x,t|x_0) = \Pr\left(x \le X(t) < x + dx | X(0) = x_0\right) \qquad (3)$$

where x_0 is the initial queue length. With the diffusion approximation, $p(x,t|x_0)$ can be characterized by the (forward) diffusion equation

$$\frac{\partial p(x,t|x_0)}{\partial t} = \frac{\alpha}{2}\frac{\partial^2 p(x,t|x_0)}{\partial x^2} - \beta\frac{\partial p(x,t|x_0)}{\partial x} \qquad (4)$$

with the initial condition

$$p(x,0|x_0) = \delta(x - x_0). \qquad (5)$$

By applying the diffusion approximation, we can exploit the transient solution of the queue length by obtaining its p.d.f. at any time instant t.

2) Start-Up Delay $\mathcal{D}$: We first evaluate the start-up delay by analyzing the charging phase. In the charging phase shown in Fig. 2, the buffer is initially empty, i.e., $x_0 = 0$, and the playback is frozen, i.e., $\mu = v_s = 0$. This phase terminates when b packets are stored. The duration of charging phase or the start-up delay is thus given by

$$\mathcal{D} = \min\left\{t | X(0) = 0, X(t) = b, t > 0\right\}. \qquad (6)$$

Note that $\mathcal{D}$ is a random variable. In what follows, we evaluate it by showing its density function and statistics.

To this end, we model the charging phase as a diffusion process with drift $\beta_D = \lambda$ and diffusion coefficient $\alpha_D = \lambda^3 v_a$ based on (2). Define $P_D(x,t|0)$ as the conditional CDF of the buffer size $X(t)$ in the charging phase. During this phase, the initial buffer is empty and the queue length $X(t)$ is less than b. Thus, we have

$$P_D(x,t|0) = \Pr\left\{X(t) \le x | X(0) = 0, X(\tau) < b \text{ for } 0 < \tau < t\right\}. \qquad (7)$$

The CDF of the start-up delay is given by

$$G_D(t) = \Pr\{\mathcal{D} \le t\} = 1 - P_D(b,t|0) = 1 - \int_0^b p_D(y,t|0)dy \qquad (8)$$

as $P_D(b,t|0)$ represents the probability that $X(t)$ is still below b at time t, i.e., $\mathcal{D}$ is greater than t. $p_D(x,t|0) = \Pr\{x \le X(t) < x + dx | X(0) = 0, X(\tau) < b \text{ for } 0 < \tau < t\}$ is the p.d.f. of $X(t)$ in the charging phase.

The p.d.f. of $\mathcal{D}$ is hence obtained as

$$g_D(t) = \frac{dG_D(t)}{dt} = -\frac{d}{dt}P_D(b,t|0) = -\frac{d}{dt}\int_0^b p_D(y,t|0)dy. \qquad (9)$$

As $p_D(x,t|0)$ can be described by the diffusion approximation with the queue length never exceeding b [20], it follows the diffusion (4), as

$$\frac{\partial p_D(x,t|0)}{\partial t} = \frac{\alpha_D}{2}\frac{\partial^2 p_D(x,t|0)}{\partial x^2} - \beta_D\frac{\partial p_D(x,t|0)}{\partial x}, \quad x < b \qquad (10)$$

coupled with the initial condition

$$p_D(x,0|0) = \delta(x) \qquad (11)$$

and the boundary condition

$$p_D(b,t|0) = 0. \qquad (12)$$

Equation (12) is obtained by the event that the diffusion process terminates when $X(t) = b$. This is imposed by the absorbing barrier in the diffusion process [22].

Solving (10) with (11) and (12) yields[1]

$$p_D(x,t|0) = \frac{1}{\sqrt{2\pi\alpha_D t}}\left[\exp\left\{-\frac{(x-\beta_D t)^2}{2\alpha_D t}\right\} - \exp\left\{\frac{2\beta_D b}{\alpha_D} - \frac{(x-2b-\beta_D t)^2}{2\alpha_D t}\right\}\right] \qquad (13)$$

and

$$P_D(x,t|0) = \Phi\left(\frac{x-\beta_D t}{\sqrt{\alpha_D t}}\right) - \exp\left\{\frac{2\beta_D b}{\alpha_D}\right\}\Phi\left(\frac{x-2b-\beta_D t}{\sqrt{\alpha_D t}}\right) \qquad (14)$$

where $\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x} e^{-(y^2/2)}dy$.

Substituting (14) into (8) and (9), we can obtain the CDF of $\mathcal{D}$

$$G_D(t) = 1 - \Phi\left(\frac{b-\beta_D t}{\sqrt{\alpha_D t}}\right) + \exp\left\{\frac{2\beta_D b}{\alpha_D}\right\}\Phi\left(-\frac{b+\beta_D t}{\sqrt{\alpha_D t}}\right) \qquad (15)$$

[1] The solution is obtained by the method of images as shown in [22], [23].

and its p.d.f.

$$g_D(t) = \frac{b}{\sqrt{2\pi\alpha_D t^3}} \exp\left\{-\frac{(b-\beta_D t)^2}{2\alpha_D t}\right\}. \qquad (16)$$

The moment generating function (m.g.f.), represented by the Laplace transform, of $g_D(t)$ is [24]

$$g_D^*(s) = E(e^{-st}) = \exp\left[\frac{b}{\alpha_D}\left\{\beta_D - \sqrt{\beta_D^2 + 2s\alpha_D}\right\}\right]. \qquad (17)$$

Based on the m.g.f. $g_D^*(s)$, the mean and variance of the start-up delay with the playback threshold b can be derived accordingly

$$E(\mathcal{D}) = -\frac{d}{ds} g_D^*(s)\bigg|_{s=0} = \frac{b}{\lambda} \qquad (18)$$

$$Var(\mathcal{D}) = \frac{d^2}{ds^2} g_D^*(s)\bigg|_{s=0} - E^2(\mathcal{D}) = bv_a. \qquad (19)$$

Equations (18) and (19) indicate that the expected value and variance of start-up delay increase linearly with the playback threshold b.

3) Playback Duration $\mathcal{T}$: The video playback starts immediately after the charging phase. With a longer playback duration $\mathcal{T}$, less playback frozens will be encountered, and hence the length of $\mathcal{T}$ is critical to the smoothness of media playback. Without loss of generality, we focus on one playback phase and model it as a diffusion process starting at time $t = 0$. As the playback phase terminates when the buffer becomes empty again, the playback duration is thus given by

$$\mathcal{T} = \min\{t | X(0) = b, X(t) = 0, t > 0\}. \qquad (20)$$

Denote $g_T(t)$ and $G_T(t)$ as the p.d.f. and CDF of $\mathcal{T}$, respectively. Same as the start-up delay $\mathcal{D}$, we evaluate $\mathcal{T}$ by showing its density function.

Given that the buffer size is b at the beginning of the charging phase, the probability that the buffer size $X(t)$ is larger than x at time t is given by

$$\begin{aligned} P_T(x,t|b) &= \Pr\{X(t) > x | X(0) = b \\ &\qquad X(\tau) > 0 \text{ for } 0 < \tau < t\} \\ &= \int_x^\infty p_T(y|t,b)dy \end{aligned} \qquad (21)$$

where $p_T(y,t|b) = \Pr\{y \le X(t) < y + dy | X(0) = b, X(\tau) > 0 \text{ for } 0 < \tau < t\}$ is the p.d.f. of $X(t)$ at time t in the playback phase, given the initial buffer size b.

Similar to the computation of start-up delay, we have

$$g_T(t) = -\frac{d}{dt}\int_0^\infty p_T(x,t|b)dx \qquad (22)$$

where $p_T(x,t)$ follows the diffusion equation

$$\frac{1}{2}\alpha_T \frac{\partial^2 p_T(x,t|b)}{\partial x^2} - \beta_T \frac{\partial p_T(x,t|b)}{\partial x} = \frac{\partial p_T(x,t|b)}{\partial t} \qquad (23)$$

subject to the initial and boundary conditions

$$p_T(x,0|b) = \delta(x-b), \quad t = 0 \qquad (24)$$

$$p_T(0,t|b) = 0, \quad t > 0. \qquad (25)$$

Equation (25) is dictated by the events that the playback phase terminates when the buffer becomes empty. β_T and α_T can be derived from (2).

Solving the diffusion (23)–(25), we have

$$\begin{aligned} p_T(x|t,b) = &\frac{\exp\left\{\frac{\beta_T}{\alpha_T}(x-b) - \frac{\beta_T^2}{2\alpha_T}t\right\}}{\sqrt{2\pi\alpha_T t}} \\ &\times\left[\exp\left\{-\frac{(x-b)^2}{2\alpha_T t}\right\} - \exp\left\{-\frac{(x+b)^2}{2\alpha_T t}\right\}\right]. \end{aligned} \qquad (26)$$

Substitute (26) into (22), we have

$$g_T(t) = \frac{b}{\sqrt{2\pi\alpha_T t^3}} \exp\left\{-\frac{(\beta_T t + b)^2}{2\alpha_T t}\right\} \qquad (27)$$

and its m.g.f.

$$g_T^*(s) = \exp\left\{-\frac{b}{\alpha_T}\left(\beta_T + \sqrt{\beta_T^2 + 2\alpha_T s}\right)\right\}. \qquad (28)$$

4) Smoothness of Playback: With the p.d.f. of $\mathcal{T}$ in hand, we are now ready to evaluate the smoothness of playback in terms of two metrics, namely the stopping probability $\mathcal{P}$ and frequency of playback frozens $\mathcal{F}$.

a) Stopping probability $\mathcal{P}$: The stopping probability $\mathcal{P}$ represents the probability that the playback freezes in the middle of media playout, mathematically

$$\mathcal{P} = \Pr(t < S | B(0) = b, B(t) = 0) \qquad (29)$$

where S denotes the length of the video file.

Substituting (27) into (29), we have

$$\mathcal{P} = \int_0^S g_T(t)dt. \qquad (30)$$

To obtain the closed-form expression on $\mathcal{P}$, we approximate S to be infinity as

$$\begin{aligned} \mathcal{P} &\approx \lim_{S\to\infty}\int_0^S g_T(t)dt = \lim_{s\to 0} g_T^*(s) \\ &= \begin{cases} 1, & \text{if } \beta_T \le 0 \\ \exp\left\{-\frac{2b}{\alpha_T}\beta_T\right\}, & \text{if } \beta_T > 0. \end{cases} \end{aligned} \qquad (31)$$

Note that the obtained stopping probability is conservative as in reality S is limited. However, this approximation does not generate much difference as S is considerably large compared to the video frame intervals.

Substitute (2) into (31), we have

$$\mathcal{P}=\begin{cases}1, & \text{if } \lambda \leq \mu \\ \exp\left\{-\frac{2b}{\lambda^3 v_a+\mu^3 v_s}(\lambda-\mu)\right\}, & \text{if } \lambda > \mu.\end{cases} \tag{32}$$

Equation (32) indicates that the video playback stops with probability 1 when the mean download rate λ is less than or equal to the video playback rate μ. In the mean time, even if the mean traffic arrival rate or video download rate λ exceeds the average video playback rate μ, it is still possible that video playback stops due to the variance of packet arrivals and playback. In the real-world deployments, $\lambda-\mu$ is normally controlled small to admit more users in the system. In this case, the stopping probability $\mathcal{P}$ is heavily dependent on the threshold b and the statistics of the network.

b) Number of playback frozens $\mathcal{F}$: In (32), we have shown that when the mean traffic arrival rate λ is smaller than the average video playback rate μ, the video playout will stop with probability one. To shed light on how serious the interruptions of playback are in this case, we derive the overall number of playback frozens, denoted by $\mathcal{F}$, encountered during the media playback. Let R.V. $\mathcal{M}$ denote the duration between two consecutive playback frozen events, and we have $\mathcal{M}=\mathcal{D}+\mathcal{T}$, as shown in Fig. 2. It is obvious that $\mathcal{F}$ is negatively proportional to $\mathcal{M}$. In what follows, we show its density function and statistics.

Denote the p.d.f. of $\mathcal{M}$ as $g_M(t)$. Hence

$$g_M(t)=g_D(t)\otimes g_T(t) \tag{33}$$

where $\otimes$ denotes convolution. The m.g.f. of $g_M(t)$ is thus given by

$$g_M^*(s)=g_D^*(s)\cdot g_T^*(s). \tag{34}$$

Substitute (17) and (28) into (34), we can obtain the mean and variance of $\mathcal{M}$ as

$$E(\mathcal{M})=-\frac{d}{ds}g_M^*(s)\bigg|_{s=0}=\frac{-b\mu}{\lambda(\lambda-\mu)}, \quad \lambda<\mu \tag{35}$$

and

$$\begin{aligned}Var(\mathcal{M})=&\frac{d^2}{ds^2}g_M^*(s)\bigg|_{s=0}-E^2(\mathcal{M})\\ =&-b\frac{\mu^3(v_s+v_a)+3v_a\lambda\mu(\lambda-\mu)}{(\lambda-\mu)^3}, \quad \lambda<\mu.\end{aligned} \tag{36}$$

Next, we use the diffusion approximation to obtain the p.d.f. of $\mathcal{F}$. Specifically, we assume that there is a virtual event buffer B_F which counts the events of playback frozen. Whenever an event of playback frozen happens, we increase the queue length of B_F by one. Thus, the buffer size of B_F at time t, denoted by $X_F(t)$, represents the number of playback frozens up to time t. The interarrival time between two consecutive increments of $X_F(t)$ is $\mathcal{M}$, where $X_F(t)$ is a non-decreasing function of time t. Denote by $P_F(x,t|0)$ the conditional CDF of $X_F(t)$ at time t, given the initial buffer size 0

$$P_F(x,t|0)=\Pr\{X_F(t)\leq x|X_F(0)=0\}. \tag{37}$$

Similarly, $X_F(t)$ can be approximated as a continuous function by applying diffusion equation, and its CDF is governed by

$$\frac{\partial P_F(x,t|0)}{\partial t}=\frac{\alpha_F}{2}\frac{\partial^2 P_F(x,t|0)}{\partial x^2}-\beta_F\frac{\partial P_F(x,t|0)}{\partial x} \tag{38}$$

coupled with the boundary condition

$$\begin{cases}\lim\limits_{x\to\infty} P_F(x,t|0)=1, & t\geq 0\\ \lim\limits_{x\to 0} P_F(x,t|0)=0, & t\geq 0\end{cases} \tag{39}$$

where $\beta_F=1/E(M)$ and $\alpha_F=Var(M)/E^3(M)$ can be derived from (2), (35), and (36).

Solving (38) and (39), we have

$$P_F(x,t|0)=\Phi\left(\frac{x-\beta_F t}{\sqrt{\alpha_F t}}\right)-\exp\left\{\frac{2\beta_F x}{\alpha_F}\right\}\Phi\left(-\frac{x+\beta_F t}{\sqrt{\alpha_F t}}\right). \tag{40}$$

The mean and variance of the number of playback frozens at time t, when $\lambda<\mu$, can be approximated as

$$E(\mathcal{F})\approx\beta_F t=-\frac{\lambda(\lambda-\mu)}{\mu b}t \tag{41}$$

$$Var(\mathcal{F})\approx\alpha_F t=\frac{\mu^2\lambda^3(v_s+v_a)+3v_a\lambda^4(\lambda-\mu)}{b^2\mu^2}t \tag{42}$$

as $\exp\{2\beta_F x/\alpha_F\}\Phi(-((x+\beta_F t)/\sqrt{\alpha_F t}))$ decreases dramatically when t is large.

C. Finite Buffer Case

In this subsection, we consider the case that the playout buffer is limited compared with the volume of video file. This is typical when the end users use personal devices with limited buffer size and hard disk such as handsets.

The start-up delay $\mathcal{D}$ obtained in the previous subsection is also valid in the finite buffer case as the start and termination conditions of the charging phase in both cases are the same. As shown in Fig. 2, in the playback phase, the queue length of the playout buffer is upper bounded by the buffer size, denoted by N $(N>b)$. When the playout buffer is full, the arrival video packets will be dropped, which not only degrades the user's video quality but also results in the bandwidth waste. Therefore, a key performance metric in this case is the packet loss probability due to buffer overflow. In this paper, the packet loss probability and the buffer overflow probability are interchangeably used.

Let $\mathcal{L}$ denote the packet loss probability of the playout buffer

$$\mathcal{L}=\lim_{t\to\infty}\Pr(B(t)=N). \tag{43}$$

The smoothness of playback is evaluated by the charging probability, denoted by $\mathcal{C}$, which is defined as the probability that the playback is frozen and the playout buffer is in the charging phase at any time instant.

We invoke the diffusion approximation to analyze playback phase in the finite buffer case and evaluate $\mathcal{L}$ and $\mathcal{C}$ in terms of network statistics and threshold of playback, as

$$\frac{\partial p(x,t|b)}{\partial t} = \frac{1}{2}\alpha_T \frac{\partial^2 p(x,t|b)}{\partial x^2} - \beta_T \frac{\partial p(x,t|b)}{\partial x} + \frac{\lambda}{b}\mathcal{C}\delta(x-b) + \mu\mathcal{L}\delta(x-N+1) \quad (44)$$

$$\lim_{x\to 0}\left[\frac{\alpha_T}{2}\frac{\partial p(x,t|b)}{\partial x} - \beta_T p(x,t|b)\right] = \frac{\lambda}{b}\mathcal{C} \quad (45)$$

$$\lim_{x\to N}\left[\frac{\alpha_T}{2}\frac{\partial p(x,t|b)}{\partial x} - \beta_T p(x,t|b)\right] = -\mu\mathcal{L} \quad (46)$$

subject to the initial and boundary conditions

$$\lim_{x\to 0^+} p(x,t|b) = 0 \quad t > 0$$
$$\lim_{x\to N^-} p(x,t|b) = 0 \quad t > 0$$

where $\delta(x)$ is the Dirac delta function; $p(x,t|b) = \Pr\{x \le X(t) < x + dx | X(0) = b\}$ is the conditional p.d.f. of the queue length $X(t)$.

The probability density in (44) is composed of two parts, the p.d.f. of the queue length $p(x,t|b)$ when $x \in (0, N)$ and the p.m.f. $\mathcal{L}$ and $\mathcal{C}$ on the two boundaries when buffer is full and in the charging phase, respectively. $(\lambda/b)\mathcal{C}\delta(x-b)$ in (44) evaluates the probability that the queue changes from the charging phase to the playout phase, where λ/b is the mean rate of the change computed as $1/E(\mathcal{D})$. $\mu\mathcal{L}\delta(x-N+1)$ evaluates the probability that the queue length jumps from N to $N-1$ with packets being served at the mean rate μ. More details about the rational behind the equations can be found in [25].

It is important to note that the diffusion process of playback phase in (44) is different from that in (23). Equation (23) describes a single playback phase that starts when b packets are stored in the charging phase and terminates once the playout buffer becomes empty. However, (44) represents the whole media playback process from the first playback to the instant when video is wholly downloaded. The reason we consider the whole session of video playback as one diffusion process in this case is that it facilitates to evaluate the long-term buffer overflow and underflow probability in the steady state. While in the infinite buffer case, we are more interested in the transient behavior of the queue to evaluate the duration of each start-up delay and playback phase.

Solving (44) at the steady state when $\lim_{t\to\infty}(\partial p(x,t|b)/\partial t) = 0$, we have

$$p(x,\infty|b) = \begin{cases} \frac{\lambda\mathcal{L}}{b\beta_T}(e^{rx}-1), & 0 < x \le b \\ \frac{\lambda\mathcal{L}}{b\beta_T}(1-e^{-rb})e^{rx}, & b < x \le N-1 \\ \frac{\mu\mathcal{C}}{\beta_T}\left(1-e^{r(x-N)}\right), & N-1 < x < N \end{cases} \quad (47)$$

where $r = 2\beta_T/\alpha_T$, and the packet loss probability $\mathcal{L}$ and the charging probability $\mathcal{C}$ are given by

$$\mathcal{L} = \left(\frac{-(1-e^{-r})\mu^2 b}{\lambda\beta_T(1-e^{-rb})e^{r(N-1)}} + \frac{\lambda}{\beta_T}\right)^{-1} \quad (48)$$

$$\mathcal{C} = \left(-\frac{\mu}{\beta_T} + \frac{\lambda^2}{\beta_T b\mu}\frac{e^{r(N-1)}(1-e^{-rb})}{1-e^{-r}}\right)^{-1}. \quad (49)$$

The infinite playout buffer could be regarded as a special case of the finite playout buffer when $N \to \infty$. In specific, when $\beta_T < 0$, i.e., $\lambda < \mu$, with (48) and (49), we have

$$\lim_{N\to\infty}\mathcal{L} = 0, \quad \lambda < \mu \quad (50)$$

as no packets will be lost when $N \to \infty$, and

$$\lim_{N\to\infty}\mathcal{C} = -\frac{\beta_T}{\mu}, \quad \lambda < \mu. \quad (51)$$

Equation (51) matches the results of infinite buffer case as

$$\lim_{N\to\infty}\mathcal{C} = \frac{E(\mathcal{F}) \times E(\mathcal{D})}{t}, \quad \lambda < \mu \quad (52)$$

where $E(\mathcal{D})$ and $E(\mathcal{F})$ are given in (18) and (41), respectively.

Therefore, we show the video quality in terms of start-up delay and smoothness of video playback with given statistics of the download and playback rates of video packets. The mean video playback rate μ is usually fixed and given for non-scalable video. When scalable video coding, e.g., layer-encoded video streaming [26], is used, the playback rate can be adjustable when different video layers are downloaded. The statistics of playback rate in this case are no longer predefined and therefore need to be measured at the receiver.

IV. Quality Driven Playout Buffer Management

In this section, by exploiting the obtained video quality metrics from the analytical framework, we determine the optimal playback threshold to achieve the maximal user utility, based on different video requirements of end users. Towards this goal, we formulate the playback threshold selection as a stochastic optimization problem.

A. Infinite Buffer Case

Let $\widehat{\mathcal{D}}$ and $\widehat{\mathcal{F}}$ denote the maximum tolerable start-up delay and number of playback frozens input by the users, respectively. Our objective is to manage the threshold of playback b to maximize the user perceived video quality within the tolerable range specified by $\widehat{\mathcal{D}}$ and $\widehat{\mathcal{F}}$, mathematically

$P1$: if $\lambda > \mu$

$$\begin{aligned} \min_b \quad & \mathcal{P} + \varpi_1\left(E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})\right) \\ s.t. \quad & \Pr\left\{\mathcal{D} > \widehat{\mathcal{D}}\right\} \le \zeta \\ & b > 0. \end{aligned} \quad (53)$$

$P2$: if $\lambda \le \mu$

$$\begin{aligned} \min_b \quad & E(\mathcal{F}) + \vartheta_F Var(\mathcal{F}) + \varpi_2\left(E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})\right) \\ s.t. \quad & \Pr\{\mathcal{D} > \widehat{\mathcal{D}}\} \le \zeta \\ & \Pr\{\mathcal{F} > \widehat{\mathcal{F}}\} \le \eta \\ & b > 0 \end{aligned} \quad (54)$$

where $\omega_1, \omega_2 > 0$ are the weighting factors and $\vartheta_D, \vartheta_F \ge 0$ are called risk aversion factors which are adjustable with respect to different user requirements. ζ, η are predefined scalers such that $0 < \zeta, \eta \ll 1$.

Scheme $P1$ is implemented when the mean packet arrival rate λ is larger than the mean video playback rate μ. In this case, with probability $1-\mathcal{P}$ the video playback can be finished without any interruptions. The objective is hence to avoid playback frozens while minimizing the start-up delay. ϖ_1 in the utility function is a knob to balance the requirements between smooth playback and start-up delay. A larger ϖ_1 represents that users are more sensitive to the start-up delay, e.g., when watching a live soccer match. ϑ_D is called risk aversion factor which models the user's attitude to the variance of start-up delay.[2] When ϑ_D is large, the users are conservative and require more strict start-up delay with low variance. The constraint is represented by a stochastic bound that the resulting start-up delay must be within the tolerable region $\widehat{\mathcal{D}}$ imposed by the user with a high probability. The stochastic QoS is considered because providing absolute QoS guarantee may not be feasible and is typically difficult and costly for implementation in the time-varying environment [27].

The scheme $P2$ is employed when the mean packet arrival rate is insufficient to meet the playback. In this case, interruptions of playback are inevitable as shown by (32). The objective is to minimize the number of playback frozens and the incurred start-up delay. The utility functions and constraints are defined in the same fashion of $P1$.

Both $P1$ and $P2$ are probability-constrained stochastic optimization (also referred to as chance constrained programming) [28]. By substituting (15), (18), (19), and (32) into $P1$; (15), (18), (19), (40), (41) and (42) into $P2$, we have

$P1'$: if $\lambda > \mu$

$$\begin{aligned}
\min_{b} \quad & \exp\left\{-\frac{2b}{\lambda^3 v_a + \mu^3 v_s}(\lambda-\mu)\right\} + \varpi_1 b\left(\frac{1}{\lambda} + \vartheta_D v_a\right)\\
s.t. \quad & \Phi\left(\frac{b-\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a \widehat{\mathcal{D}}}}\right) - \exp\left\{\frac{2b}{\lambda^2 v_a}\right\}\Phi\left(-\frac{b+\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a\widehat{\mathcal{D}}}}\right) \le \zeta\\
& b \ge 0
\end{aligned} \tag{55}$$

$P2'$: if $\lambda \le \mu$

$$\begin{aligned}
\min_{b} \quad & \frac{A}{b} + \frac{\vartheta_F}{b^2}B + \varpi_2 b\left(\frac{1}{\lambda} + \vartheta_D v_a\right)\\
s.t. \quad & \Phi\left(\frac{b-\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a \widehat{\mathcal{D}}}}\right) - \exp\left\{\frac{2b}{\lambda^2 v_a}\right\}\Phi\left(-\frac{b+\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a\widehat{\mathcal{D}}}}\right) \le \zeta\\
1 - {} & \Phi\left(\frac{\widehat{\mathcal{F}}-\beta_F S}{\sqrt{\alpha_F S}}\right) + \exp\left\{\frac{2\beta_F\widehat{\mathcal{F}}}{\alpha_F}\right\}\Phi\left(-\frac{\widehat{\mathcal{F}}+\beta_F S}{\sqrt{\alpha_F S}}\right) \le \eta\\
& b \ge 0
\end{aligned} \tag{56}$$

where $A = -(\lambda(\lambda-\mu)/\mu)S$, $B = ((\mu^2\lambda^3(v_s + v_a) + 3v_a\lambda^4(\lambda-\mu))/\mu^2)S$ are positive scalers. Here, the statistics of network and video playback rate, i.e., λ, v_a, μ and v_s, and the video length S are known and used as inputs to the control scheme. This is reasonable as those network statistics can be measured in real time at the user end.

Both $P1'$ and $P2'$ are nonlinear programming problems which may be prohibitively expensive for practical real-time streaming systems. To reduce the computation complexity, we apply the one-sided Chebyshev inequality, which states that for any R.V. χ and any positive real number x

$$\Pr\{\chi - E(\chi) \ge x\} \le \frac{Var(\chi)}{Var(\chi) + x^2}, \text{ for } \chi > E(\chi). \tag{57}$$

Using the Chebyshev inequality, together with (18), (19), (41), and (42), the constraints of $P2$ become

$$b \le \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) - \sqrt{\frac{4\widehat{\mathcal{D}}\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2} \tag{58}$$

$$b \ge \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}} \tag{59}$$

where A and B are the same as those in (56). The details are shown in Appendix A.

By replacing the constraints of $P1$ and $P2$ with (58) and (59), both $P1$ and $P2$ become convex optimization problems which could be solved efficiently. Note that computation complexity is reduced at the expanse of user's utility, because comparing with $P1'$ and $P2'$, the new constraints obtained with the Chebyshev inequality is more conservative, resulting in a smaller feasible region. However, a conservative but fast algorithm is desirable for practical use.

To ensure that the resultant video performance is within the tolerable region, the threshold of playback b must be within the range specified by (58) and (59). To make this condition satisfied, we could apply call admission control at the user end. In this way, the request of playback is reject directly at agent of the end host without sending it to the media server if there is no positive b to meet both (58) and (59). Thus, the network resources can be efficiently utilized to provision video quality for all admitted videos.

B. Finite Buffer Case

We optimize the video playback in the finite buffer case. Our objective is to control the threshold of playback b to minimize the interruptions of video playback due to buffer empty and the packet loss caused by the buffer overflow. The minimization problem, in this case, can be represented as

$$\begin{aligned}
\min_{b} \quad & \rho_1\mathcal{L} + \rho_2\frac{\mathcal{C}\times S}{E(\mathcal{D})} + \varpi_1\left(E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})\right)\\
s.t. \quad & \Pr\{\mathcal{D} > \widehat{\mathcal{D}}\} \le \zeta\\
& b > 0
\end{aligned} \tag{60}$$

where ρ_1 and ρ_2 are the weighting factors of packet loss and charging probabilities, respectively. ϖ_1, ϑ_D, and ζ are defined in the same manner as those in the infinite buffer case.

In (60), $(\mathcal{C}\times S)/E(\mathcal{D})$ represents the mean number of playback frozens where $\mathcal{C}\times S$ computes the overall time spent in the charging phase. The objective is to balance the trade-off between the video quality metrics, i.e., the smoothness of playback, packet loss and the encountered start-up delay.

In a summary, this section provides examples to apply the achieved analytical results to the optimal receiver buffer design.

[2] This utility function is defined in the fashion of Markowitz mean-variance model which is widely used in portfolio optimization.

TABLE I
STATISTICS OF VIDEO FRAMES

Video Clip	Frame	Frame Size		Bit Rate		Inter-departure of pkts	
Name	Number	Mean (bytes)	Variance	Mean (bit/sec)	Peak	$\frac{1}{\mu}$(msec)	Variance
Aladdin	89998	7.7e+02	5.8e+05	1.5e+05	1.3e+06	33.6	102
Susi & Strolch	89998	5.8e+02	3.9e+05	1.2e+05	1.3e+06	36.2	70.4

This leverages the property that the analytical results bridge the network throughput with the user perceived video quality as

$$(\mathcal{D},\mathcal{F}) = f(\lambda, v_a, \mu, v_s, b) \tag{61}$$

where the mapping function $f(\cdot)$ could be represented by the constraints of $P1$ and $P2$ [or (58) and (59)]. In a reverse manner, we can also obtain the desired network resource with given user requirements as

$$(\lambda, v_a) = f^{-1}(\mathcal{D},\mathcal{F},\mu, v_s, b). \tag{62}$$

This is useful as the guideline of the network resource allocation to achieve specific video quality requirements. For non-scalable video coding, the video playback rate μ is usually fixed and only the playback threshold b is adjusted to adapt to the required video quality. In this case, the presented optimization framework can be applied directly. When layered-encoded video coding is used, the video playback rate could also be adjustable [26] and the problem can be extended to a joint optimization framework of playback threshold and playback rate (or video layers) selections. We will pursue the joint optimization problem in our future work.

V. Simulation Results

In this section, we verify the achieved analytical results using extensive simulations, based on a trace-driven discrete event simulator coded in C++.

A. Simulation Setup

We use two real VBR video clips, "Aladdin" and "Susi & Strolch", from [29] encoded by MPEG-4 with diverse frame statistics. Each video clip lasts $S = 1$ hour and the sequences are encoded at a constant frame rate of 25 frames per second in the Quarter Common Intermediate Format (QCIF) resolution (176 × 144). The statistics of video frames are summarized in Table I.

The simulated network is shown in Fig. 1. In each simulation run, the simulator loads video frames from the video trace file and segment the variable size video frames into IP packets with the maximum size of 1400 Bytes. The available bandwidth of the network varies over time during which the overall variable bit rates of the channels are 10 Kbps, 500 Kbps, 2 Mbps, and 4 Mbps with probability 0.02, 0.48, 0.30, and 0.20, respectively. The average throughput is thus 1.64 Mbps; the mean and standard deviation of network delay are $1/\lambda = 35.4$ ms and $\sqrt{v_a} = 155.2$ ms, respectively. The video file is played at a constant rate 25 frames per second by default and variable packet rates as shown in Table I. For each scenario, we conduct 30 simulation runs and plot the mean results with the 95% confidence intervals.

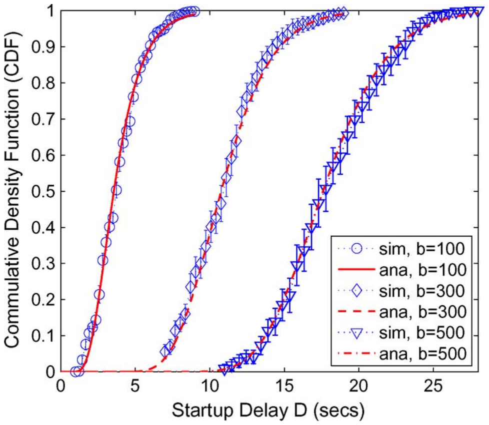


Fig. 3. CDF of the start-up delay D for $b = 100$, 300, and 500 packets, respectively.

B. Infinite Buffer Case

We first examine the case of infinite playout buffer.

1) Start-Up Delay $\mathcal{D}$ and Playback Duration $\mathcal{T}$: In the first simulation, we verify the analysis of the start-up delay $\mathcal{D}$ and playback duration $\mathcal{T}$. We use the trace "Aladdin" in which $\lambda < \mu$ according to Table I. In this case, the playback frozens are inevitable as indicated by (32).

The CDF of the start-up delays and playback durations with different buffer thresholds b are shown in Figs. 3 and 4, respectively. In Fig. 3, the mean start-up delay increases with b and the corresponding CDF moves to the left. In addition, the variance of start-up delay increases accordingly as the CDF curve expands in width. Similarly, it can be seen in Fig. 4 that both mean and variance of the playback duration increase with the threshold b. The simulation results well validate our analysis.

2) Stopping Probability $\mathcal{P}$: We verify the stopping probability $\mathcal{P}$ in the second simulation using the clip "Susi & Strolch" where $\lambda > \mu$. In this case, the video packets are downloaded at a faster rate than that of the playback, and the stopping probability $\mathcal{P}$ is less than one as shown in (32). We conduct 30 experiments, and for each experiment we increase the buffer threshold b by 70 packets starting from one packet. Within each experiment, we conduct 500 simulation runs with each run terminated either when the playback frozen occurs or after the whole video is played without any interruptions. The simulation stopped by playback frozens are called frozen events. The probability of stopping is then computed as the total number of frozen events divided by 500. It is observed in Fig. 5 that the probability of stopping decreases exponentially with the increase of buffer threshold b. The analytical results are slightly larger than the

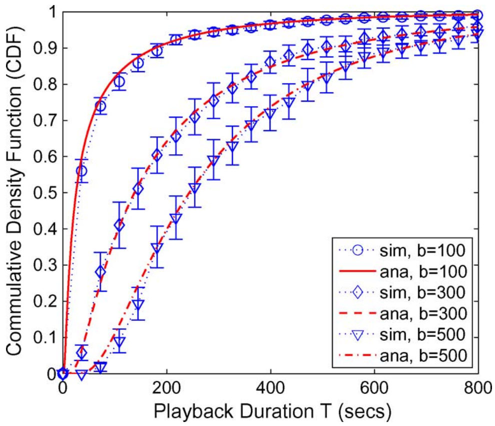


Fig. 4. CDF of the playback duration T for $b = 100$, 300, and 500 packets, respectively, at time $t = S$.

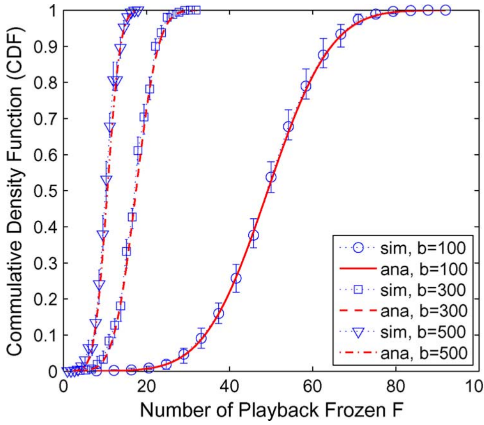


Fig. 6. Number of playback frozens $\mathcal{F}$ for $S = 1\ \mathrm{hour}$, and $b = 100$, 300, and 500 packets, respectively.

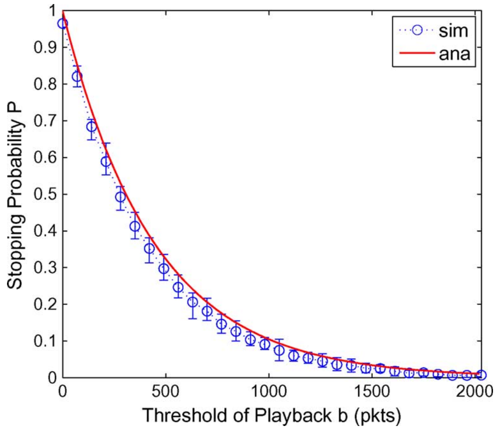


Fig. 5. Simulated stopping probability $\mathcal{P}$.

simulation results because the video length S is assumed infinity for analysis while S is 1 h in the simulations.

3) Number of Playback Frozens $\mathcal{F}$: We study the number of playback frozens using the clip "Aladdin" with $\lambda < \mu$. In this simulation, we conduct 500 runs and measure the number of playback frozens. Fig. 6 plots the CDF of the number of playback frozens when b is 100, 300, and 500 packets, respectively, at time $t = S$. The analysis obtained from (40) well match the simulation result. Meanwhile, we can see that when b increases, the CDF curve shifts to the left which implies that on average fewer events of playback frozens are encountered. However, the step size of each shift is different; the mean number of playback frozens decreases dramatically when b is initially small. The width of the CDF curves also becomes smaller with a larger b, which implies that the variance of the number of playback frozens decreases when b increases.

4) Optimal Selection of Playback Threshold b: Based on the above analytical results, we apply Matlab `fmincon` to solve (53) and (54) subject to the constraints (58) and (59) for optimal playback threshold management. The default setting of parameters are in Table II.

We first show the impacts of weighting factors on the optimal selection of playback threshold. Fig. 7 plots the optimal threshold b^* of $P1$ using trace "Aladdin". In this scenario, we can see that the optimal threshold b^* decreases monotonically with the increasing weighting factor ϖ_1. This is because when ϖ_1 in (53) increases, the utility of start-up delay, evaluated as $E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})$, becomes more important in the objective and overwhelms the stopping probability. Therefore, the optimal threshold b^* is reduced accordingly to shrink the start-up delay at the cost of a higher playback frozen probability. The resultant stopping probability and utility of start-up delay at different optimal thresholds b^* are shown in Fig. 8, where the utility of start-up delay is computed as $b^*((1/\lambda)+\vartheta_D v_a)$ which is the portion of utility characterized by the start-up delay in (55). Fig. 9 plots the optimal thresholds b^* of $P2$ with the increasing weighting factor ϖ_2. We can see that the corresponding b^* decreases, as the utility of start-up delay becomes more critical when ϖ_2 increases. When ϖ_2 is very small, b^* is upper bounded by 2500 packets to guarantee that the resulting start-up delay is within the tolerable value $\widehat{\mathcal{D}}$. When ϖ_2 is very large, b^* is lower bounded by 800 packets to assure that the tolerable value $\widehat{\mathcal{F}}$ is not violated. The resultant utilities of playback frozens and start-up delay at the different optimal thresholds of playback b^* are shown in Fig. 10 where the utility of playback frozens is characterized by playback frozen in (56) computed as $(A/b^*) + \vartheta_F(B/(b^*)^2)$. The utility of start-up delay is computed in the same manner as that in Fig. 8.

The impacts of risk aversion factors are shown in Figs. 11–13. Figs. 11 and 12 show the impacts of ϑ_D in schemes $P1$ and $P2$, respectively, with all the other parameters the same as in Table II. With an increasing ϑ_D, the optimal playback threshold b decreases in both schemes, $P1$ and $P2$. This is because users require a strictly small variance of start-up delay. Fig. 13 shows the impact of ϑ_F in $P2$. With ϑ_F increasing, the optimal playback threshold increases monotonically which hence reduce the variance of interruption frequency of the playback.

C. Finite Buffer Case

1) Packet Loss Probability $\mathcal{L}$ and Charging Probability $\mathcal{C}$: We verify the analytical results of the packet loss probability $\mathcal{L}$ and charging probability $\mathcal{C}$ when the buffer size is finite. In each simulation, $\mathcal{L}$ is computed as the dropped packets due to buffer overflow divided by the total number of transmitted packets. $\mathcal{C}$ is evaluated as the overall time spent in the charging phase divided by the whole video session length, i.e., 1 h. We set the buffer size

TABLE II
PARAMETERS IN OPTIMAL PLAYOUT BUFFER MANAGEMENT (INFINITE BUFFER)

Scheme	Video Trace	$\widehat{\mathcal{D}}$	$\widehat{\mathcal{F}}$	ζ	η	ϖ_1	ϖ_2	ϑ_D	ϑ_F
$P1(\lambda > \mu)$	Susi & Strolch	120 sec	N/A	5%	N/A	0.01	N/A	1	N/A
$P2(\lambda \leq \mu)$	Aladdin	120 sec	20	5%	5%	N/A	0.1	1	1

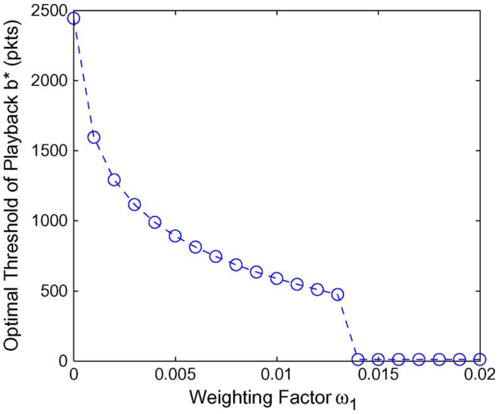


Fig. 7. Optimal playback threshold b^* with the increasing weighting factor ϖ_1.

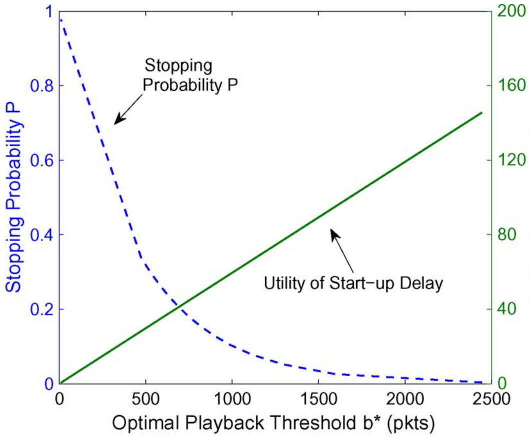


Fig. 8. Trade-off between the stopping probability and start-up delay at different optimal playback thresholds b^*.

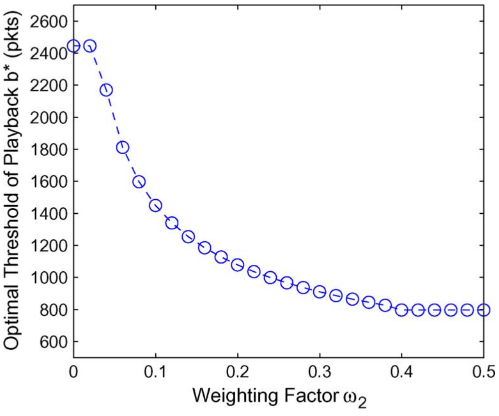


Fig. 9. Optimal playback threshold b^* with the increasing weighting factor ϖ_2.

N to be 500 packets by default and use the trace “Aladdin” in the experiment.

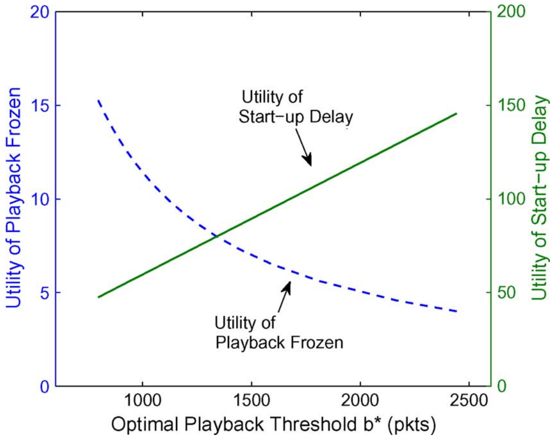


Fig. 10. Trade-off between the number of playback frozens and start-up delay at different optimal playback thresholds b^*.

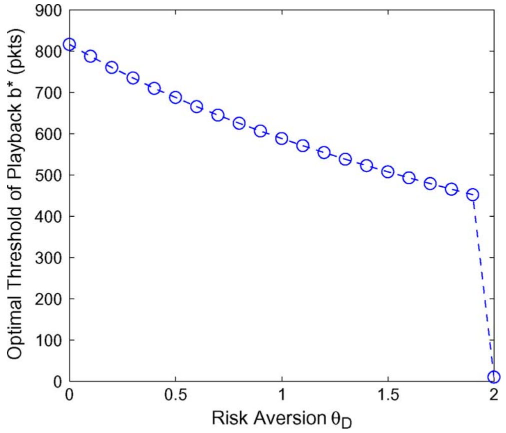


Fig. 11. Impact of θ_D on the optimal selection of playback threshold b in $P1$.

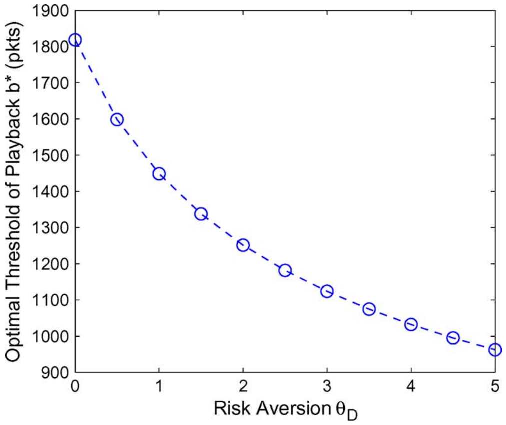


Fig. 12. Impact of θ_D on the optimal selection of playback threshold b in $P2$.

Fig. 14 plots the packet loss probability $\mathcal{L}$ with increasing threshold b. With a larger b, more packets are buffered in the

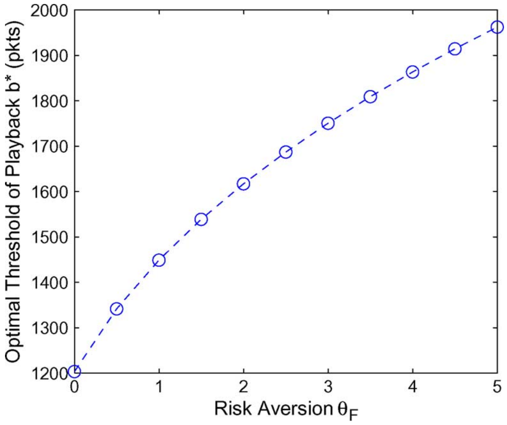


Fig. 13. Impact of θ_F on the optimal selection of playback threshold b in $P2$.

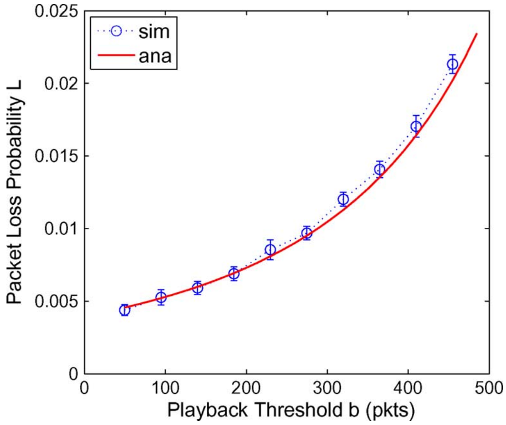


Fig. 14. Packet loss probability with the increasing playback threshold b.

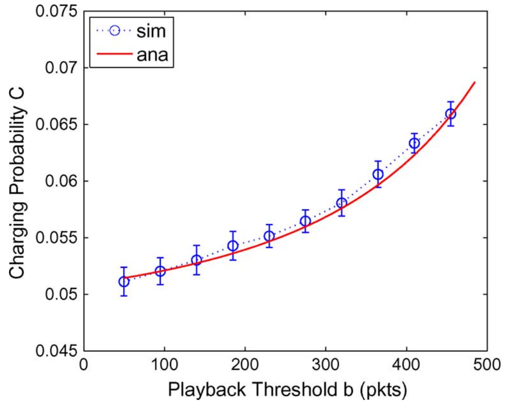


Fig. 15. Charging probability with the increasing playback threshold b.

charging phase and therefore the buffer becomes more easily to get filled. As a result, $\mathcal{L}$ increases monotonically with b. Fig. 15 plots the charging probability $\mathcal{C}$ under various thresholds b. It can be seen that $\mathcal{C}$ also increases monotonically with b. This is because the charging probability $\mathcal{C}$ represents the probability that at any point the buffer is in the charging phase; with a larger b, each charging phase is elongated, making $\mathcal{C}$ increase accordingly. However, the mean number of playback frozens, computed as $(\mathcal{C} \times S)/E(\mathcal{D})$, reduces when b increases, as shown in Fig. 16.

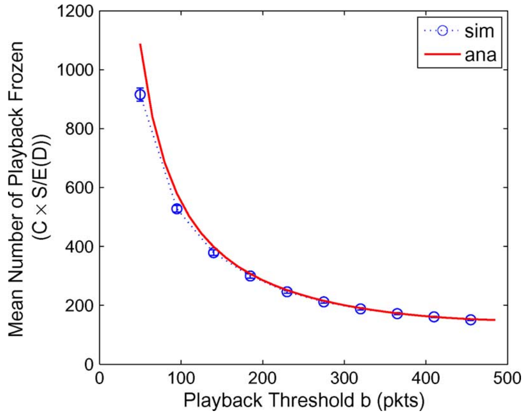


Fig. 16. Mean number of playback frozens with the increasing playback threshold b.

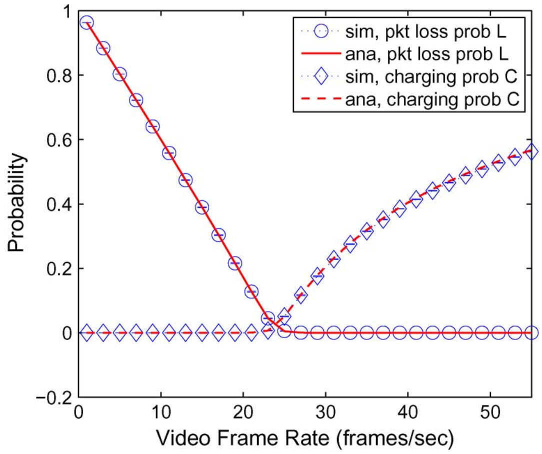


Fig. 17. Packet loss and charging probabilities with the increasing frame rate.

Fig. 17 plots the packet loss probability $\mathcal{L}$ and charging probability $\mathcal{C}$ under different video frame rates (and playback rate). It can be seen that, with an increasing playback rate μ which implies more video frames are played in a unit time, $\mathcal{L}$ decreases and $\mathcal{C}$ increases. This is because with a faster playback, the buffer is more likely to become empty and less likely to overflow. The simulation verifies our analysis with various values of μ. Fig. 18 shows the impacts of the playout buffer size N on the overflow and charging probabilities when playback threshold b is 50 packets. It can be seen that as the buffer size increases, both $\mathcal{L}$ and $\mathcal{C}$ decrease monotonically. This is because with enhanced buffer capacity, less packets will be dropped due to buffer overflow and more packets are served for playback. This reduces the frequency that buffer becomes empty. However, unlike $\mathcal{L}$ which becomes 0 when N is large, $\mathcal{C}$ approaches to a nonzero value as $\lim_{N\to\infty} \mathcal{C} = -((\lambda - \mu)/\mu) = 0.0536$, as derived in (50) and (51).

2) Optimal Selection of Playback Threshold b: After verifying the correctness of our analysis, we show how to invoke the analytical results to optimally determine the optimal playback threshold b^* using numerical examples. We solve (60) using the `fmincon` function of Matlab. The video trace used is "Susi & Strolch" and the default setting of the parameters are: $\widehat{\mathcal{D}} = 15\,\mathrm{s}$, $\rho_1 = 50$, $\rho_2 = 0.05$, $\varpi_1 = 0.1$, $\vartheta_D = 1$ and $\zeta = 0.05$.

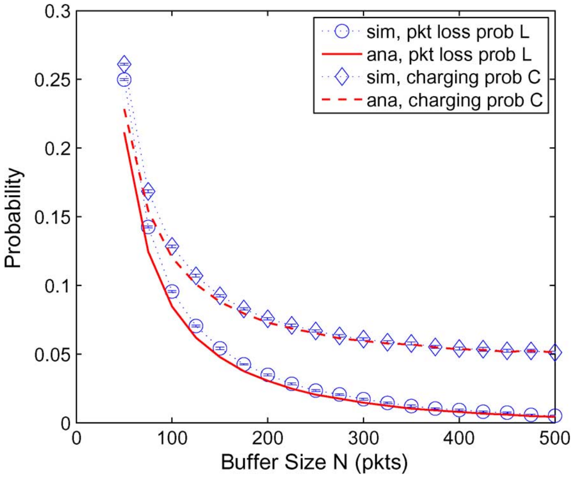


Fig. 18. Packet loss and charging probabilities with the increasing buffer size.

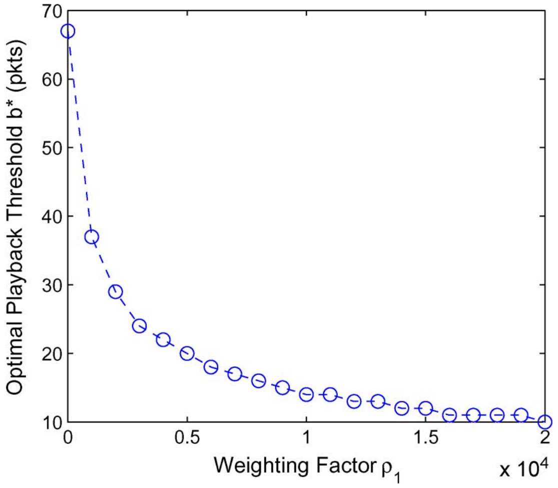


Fig. 19. Optimal playback threshold b^* with the increasing weighting factor ρ_1.

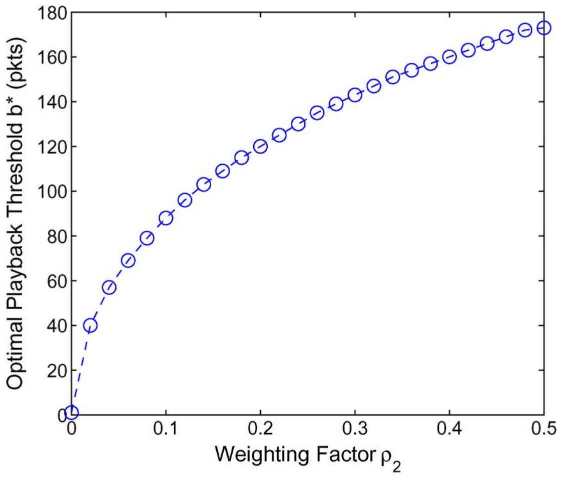


Fig. 20. Optimal playback threshold b^* with the increasing weighting factor ρ_2.

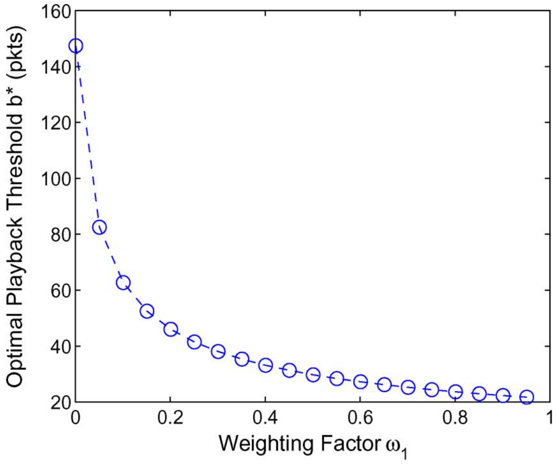


Fig. 21. Optimal playback threshold b^* with the increasing weighting factor ϖ_1.

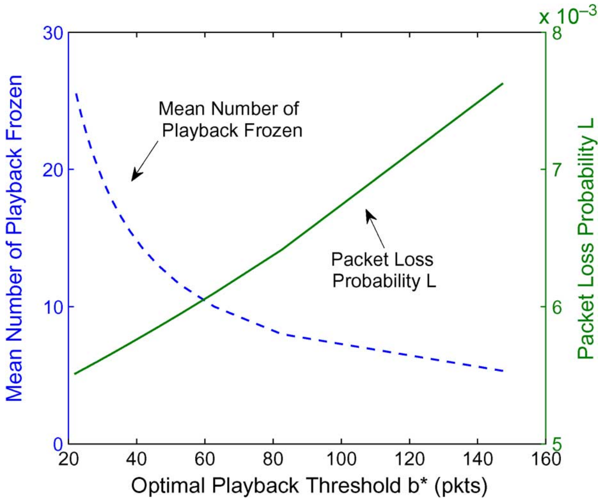


Fig. 22. Trade-off between the mean number of playback frozens and the packet loss probability at different optimal playback thresholds b^*.

Fig. 19 shows the optimal playback threshold b^* with different values of ρ_1 and default setting of other parameters. It can be seen that the increasing ρ_1 leads to the decrease of b^* because a smaller playback threshold is preferred in order to avoid buffer overflow. Fig. 20 shows the impact of ρ_2. When ρ_2 increases, the objective function is sensitive to the playback frozen and hence a larger b is desirable. Fig. 21 plots the optimal selection of playback threshold b with the increasing ϖ_1. Similar to the case of infinite buffer, with ϖ_1 increasing, the optimal playback threshold b^* decreases monotonically to keep reducing the start-up delay. Fig. 22 shows the tradeoff between the mean number of playback frozens computed by $(\mathcal{C} \times S)/E(\mathcal{D})$ and the packet loss probability $\mathcal{L}$.

VI. Conclusion

We have developed a mathematical framework to study the impacts of network dynamics on the perceived video quality of end users. We have evaluated the user's perceptual quality, in terms of the start-up delay, playback smoothness, and the packet loss probability, and represented them by the network statistics and the threshold of playback in closed-form expressions. We have shown how to invoke the analytical framework to guide the playout buffer design. The analytical results have been verified by extensive simulations.

We envision the future research in two dimensions. First, we will study efficient measurement schemes of the statistics of download and playback rates at the receiver and extend the analysis by taking the scalable video coding and time-dependent packet arrivals into account. Second, we will implement the proposed buffer management schemes in real-world applications and test them under different network scenarios, such as the buffer management for peer-to-peer on demand video streaming and handset buffer management for 3G cellular video streaming.

APPENDIX

A. Derivation of (58) and (59)

We show how to apply the Chebyshev inequality (57) to derive (58) and (59), respectively.

Based on (57), we have

$$\Pr\{\mathcal{D} > \widehat{\mathcal{D}}\} \le \frac{Var(\mathcal{D})}{Var(\mathcal{D}) + \left(\widehat{\mathcal{D}} - E(\mathcal{D})\right)^2} = \frac{bv_a}{bv_a + \left(\widehat{\mathcal{D}} - \frac{\mathsf{b}}{\lambda}\right)^2}. \tag{A1}$$

To satisfy the constraint $\Pr\{\mathcal{D} > \widehat{\mathcal{D}}\} \le \zeta$, we make

$$\frac{bv_a}{bv_a + \left(\widehat{\mathcal{D}} - \frac{\mathsf{b}}{\lambda}\right)^2} \le \zeta \tag{A2}$$

which implies

$$b \le \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) - \sqrt{\frac{2D\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2} \quad \text{or}$$
$$b \ge \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) + \sqrt{\frac{2D\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2}. \tag{A3}$$

As $\widehat{\mathcal{D}} \ge E(\mathcal{D}) = b/\lambda$, we have $b \le \widehat{\mathcal{D}}\lambda$. Together with (A3), we have

$$b \le \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) - \sqrt{\frac{2\widehat{\mathcal{D}}\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2}. \tag{A4}$$

Apply the Chebyshev inequality to bound $\Pr\{\mathcal{F} > \widehat{\mathcal{F}}\}$, we have

$$\Pr\{\mathcal{F} > \widehat{\mathcal{F}}\} \le \frac{Var(\mathsf{F})}{Var(\mathcal{F}) + \left(\widehat{\mathcal{F}} - E(\mathcal{F})\right)^2} = \frac{\frac{B}{b^2}}{\frac{B}{b^2} + \left(\widehat{\mathcal{F}} - \frac{\mathsf{A}}{\lambda}\right)^2}. \tag{A5}$$

To satisfy the constraint $\Pr\{\mathcal{F} > \widehat{\mathcal{F}}\} \le \eta$, we make

$$\frac{\frac{B}{b^2}}{\frac{B}{b^2} + \left(\widehat{\mathcal{F}} - \frac{\mathsf{A}}{\lambda}\right)^2} \le \eta \tag{A6}$$

which implies

$$b \le \frac{A}{\widehat{\mathcal{F}}} - \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}} \quad \text{or} \quad b \ge \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}}. \tag{A7}$$

In addition, b be set such that $\widehat{\mathcal{F}} \ge E(\mathcal{F}) = A/b$, i.e., $b \ge A/\widehat{\mathcal{F}}$. Substitute it into (A7), we have

$$b \ge \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}}. \tag{A8}$$

REFERENCES

[1] B. Girod, J. Chakareski, M. Kalman, Y. J. Liang, E. Setton, and R. Zhang, "Advances in network-adaptive video streaming," *Wireless Commun. Mobile Comput.*, vol. 2, no. 6, pp. 549–552, 2002.

[2] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "End-to-end QoS for video delivery over wireless Internet," *Proc. IEEE*, vol. 93, no. 1, pp. 123–134, Jan. 2005.

[3] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 390–404, Apr. 2006.

[4] J. Chakareski and P. Frossard, "Rate-distortion optimized distributed packet scheduling of multiple video streams over shared communication resources," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 207–218, Apr. 2006.

[5] S. Mao, Y. T. Hou, X. Cheng, H. D. Sherali, S. F. Midkiff, and Y.-Q. Zhang, "On routing for multiple description video over wireless Ad Hoc networks," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 1063–1074, Oct. 2006.

[6] X. Tong, Y. Andreopoulos, and M. van der Schaar, "Distortion-driven video streaming over multihop wireless networks with path diversity," *IEEE Trans. Mobile Comput.*, vol. 6, no. 12, pp. 1343–1356, Dec. 2007.

[7] J. Xu, X. Shen, J. W. Mark, and J. Cai, "Adaptive transmission of multi-layered video over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 6, p. 2305, Jun. 2007.

[8] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[9] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and transmission power management for energyefficient wireless video communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 411–424, Jun. 2002.

[10] N. Laoutaris and I. Stavrakakis, "Intrastream synchronization for continuous media streams: A survey of playout schedulers," *IEEE Netw.*, vol. 16, no. 3, pp. 30–40, May–Jun. 2002.

[11] J. Liu, B. Li, and Y.-Q. Zhang, "An end-to-end adaptation protocol for layered video multicast using optimal rate allocation," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 87–102, Feb. 2004.

[12] L. Galluccio, G. Morabito, and G. Schembra, "Transmission of adaptive MPEG video over time-varying wireless channels: Modeling and performance evaluation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2777–2788, Nov. 2005.

[13] M. Kalman, E. Steinbach, and B. Girod, "Adaptive media playout for low-delay video streaming over error-prone channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 841–851, Jun. 2004.

[14] N. Laoutaris, B. V. Houdt, and I. Stavrakakis, "Optimization of a packet video receiver under different levels of delay jitter: An analytical approach," *Perform. Eval.*, vol. 55, no. 3–4, pp. 251–275, 2004.

[15] G. Liang and B. Liang, "Balancing interruption frequency and buffering penalties in VBR video streaming," in *Proc. IEEE Infocom*, 2007.

[16] A. Dua and N. Bambos, "Buffer management for wireless media streaming," in *Proc. IEEE GLOBECOM*, 2007.

[17] G. Liang and B. Liang, "Effect of delay and buffering on jitter-free streaming over random VBR channels," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1128–1141, Oct. 2008.

[18] D. Wu, Y. T. Hou, W. Zhu, Y. Q. Zhang, and J. M. Peha, "Streaming video over the internet: Approaches and directions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 282–300, Mar. 2001.

[19] L. Kleinrock, *Queueing Systems, Volume II: Computer Applications*. NewYork: Wiley, 1976.

[20] G. Louchard and G. Latouche*, Probability Theory and Computer Science*. San Diego, CA: Academic, 1983.

[21] A. Duda, "Transient diffusion approximation for some queueing systems," in *Proc. ACM Sigmetrics*, 1983.

[22] D. R. Cox and H. D. Miller*, The Theory of Stochastic Processes*. London, U.K.: Chapman & Hall/CRC, 1977.

[23] F. P. T. Czachorski, "Diffusion approximation as a modelling tool in congestion control and performance evaluation," in *Proc. HET-NET*, 2004.

[24] , M. Abramowitz and I. A. Stegun, Eds.*, Handbook of Mathematical Functions*. New York: Dover, 1965.

[25] E. Gelenbe, "On approximate computer system models," *J. ACM*, vol. 22, no. 2, pp. 261–269, 1975.

[26] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Scalable video coding and transport over broadband wireless networks," *Proc. IEEE*, vol. 89, no. 1, pp. 6–20, Jan. 2001.

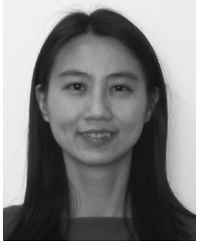
[27] W. Kumwilaisak, Y. T. Hou, Q. Zhang, W. Zhu, C.-C. J. Kuo, and Y.-Q. Zhang, "A cross-layer quality-of-service mapping architecture for video delivery in wireless networks," *IEEE J. Select. Areas Commun.*, vol. 21, no. 10, pp. 1685–1698, Dec. 2003.

[28] J. R. Birge and F. Louveaux*, Introduction to Stochastic Programming*. New York: Springer, 1997.

[29] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H. 263 video traces for network performance evaluation," *IEEE Netw.*, vol. 15, no. 6, pp. 40–54, Nov.–Dec. 2001.

Tom H. Luan received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 2004 and the M.Phil. degree in electronic engineering from the Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2007. He is now pursuing the Ph.D. degree at the University of Waterloo, Waterloo, ON, Canada.

His current research interests focus on wired and wireless multimedia streaming, QoS routing in multihop wireless networks, peer-to-peer streaming, and vehicular network design.



Lin X. Cai received the B.E. degree in computer science from Nanjing University of Science and Technology, Nanjing, China, in 1996 and the M.A.Sc. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2005. She is currently pursuing the Ph.D. degree in the same field at the University of Waterloo.

Her research interests include network performance analysis and protocol design for multimedia applications over broadband wireless networks.

Xuemin (Sherman) Shen (M'97–SM'02–F'09) received the B.Sc. degree from Dalian Maritime University, Dalian, China, in 1982, and the M.Sc. and Ph.D. degrees from Rutgers University, New Brunswick, NJ, in 1987 and 1990, respectively, all in electrical engineering.

He is a Professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. Hiss research focuses on mobility and resource management in interconnected wireless/wired networks, UWB wireless communications networks, wireless network security, wireless body area networks, and vehicular ad hoc and sensor networks. He is a coauthor of three books and has published more than 400 papers and book chapters in wireless communications and networks, control, and filtering.

Dr. Shen served as the Tutorial Chair for IEEE ICC'08, the Technical Program Committee Chair for IEEE Globecom'07, the General Co-Chair for Chinacom'07 and QShine'06, and the Founding Chair for IEEE Communications Society Technical Committee on P2P Communications and Networking. He also serves as a Founding Area Editor for the IEEE Transactions on Wireless Communications; Editor-in-Chief for *Peer-to-Peer Networking and Application*; Associate Editor for the IEEE Transactions on Vehicular Technology; KICS/IEEE Journal of Communications and Networks, Computer Networks; ACM/Wireless Networks; Wireless Communications and Mobile Computing (Wiley), etc. He has also served as Guest Editor for IEEE Journal on Selected Areas in Communication, IEEE Wireless Communications, IEEE Communications Magazine, *ACM Mobile Networks and Applications*, etc. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004 and 2008 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. He is a registered Professional Engineer of Ontario, Canada and a Distinguished Lecturer of the IEEE Communications Society.