

Video Streaming over Variable Bit Rate Channels: From End User's Perspective

Hao Luan, Lin X. Cai, Xuemin (Sherman) Shen

Department of Electrical and Computer Engineering

University of Waterloo, Waterloo, ON N2L 3G1, Canada

{hluan, lcai, xshen}@bbr.uwaterloo.ca

Abstract

In this paper, we integrate the video transmission over the variable bit rate (VBR) channels. An analytical framework is developed to study the impacts of varying network delays on the user perceived video quality. Our analysis stands from the end user's perspective. In specific, we model the playback buffer at the receiver by a $G/G/1$ queue with arbitrary packet arrivals and playback, and examine the transient evolution of the queue using the diffusion approximation. We obtain the closed-form expressions of the video quality in terms of start-up delay and fluency of playback represented by the statistics of networks, *i.e.*, the average network throughput and delay jitters. Based on the closed-form expressions, we propose an adaptive playout buffer management scheme to optimally control the threshold of video playback towards the maximal user utility. The proposed framework is validated by extensive simulations.

I. INTRODUCTION

The revolutionary advances in broadband wireless access technologies provide great potential for broadband multimedia applications, such as video conferencing, live multimedia streaming, and IPTV broadcasting. However, high quality video streaming over packet switched networks is still fraught with fundamental challenges [1]. This attributes to the real-time nature of video traffic and the inherent dynamics of the packet networks. Specifically, video applications have strict deadlines of presentation. In this case, the delay variations in the network may create significant jitters to packet arrivals at the end user, which lead to the failure of packets to meet the deadline and consequently the jerkiness or even frozen of playback. On the other hand, the past decade has been seen a widespread adoption of various dynamic

networks, notably peer-to-peer networks, mobile ad hoc networks, *etc.* The dynamic topology change, coupled with time-varying wireless channel and multihop relay, could result in severe variations in the end-to-end throughput and delay [2]. As a result, how to overcome the extraordinary network dynamics and provide end users with the steady high-quality video experience are crucial for the next generation networks.

A typical way to combat the network dynamics in packet video streaming is by employing a playout buffer (or dejitter buffer) at the receiver. To eliminate the effects of variable arrival delays (or delay jitters), the playout buffer postpones the start of video playback by a short period, namely start-up delay, and buffers the downloaded video packets in the cache until a certain threshold of playback is reached. In this case, as long as the playout buffer is kept nonempty during the video presentation, the playback can always sustain. However, if more packets are buffered to ensure smooth playout, users inevitably experience longer start-up delay which may be intolerable due to the long wait. Hence, the selection of playback threshold needs to strike a balance between the start-up delay and the fluency of playback, in the presence of intensive network jitters.

Even with broad research attention [6]-[12], how to intelligently control playout buffer and determine optimal playout threshold is still an open issue because of the user's blindness on the impacts of the complicated network dynamics to the perceived video quality. To address this issue, in this paper, we develop an analytical framework to study the relationship between network dynamics and the video quality perceived by the end user. We apply the diffusion approximation approach to examine the transient evolution of the playout buffer at the receiver, and obtain the QoS performance metrics of video quality in terms of start-up delay and fluency of playback in closed-form expressions based on the general statistics of network conditions, *i.e.*, mean network throughput and the delay jitters. We show the tradeoff between the start-up delay and the playback smoothness, in terms of the frequency that playback freezes during the video presentation, characterized by the playback threshold. We then propose an adaptive playout buffer management scheme to optimally control the playback threshold towards the maximal user utility. Our main contributions are in three-fold.

- ▷ *General Model:* We consider a general network setting and model the playout buffer at the receiver as a $G/G/1$ queue characterized by the first two moments of statistics, *i.e.*, the mean and variance of traffic arrival rate and the video playback rate. In this way, the proposed analytical model is general and suitable for a diverse range of video coding schemes and networking scenarios, such

as peer-to-peer network and wireless mesh networks.

- ▷ *Compact Solution*: We study the transient evolution of the playout buffer using the diffusion approximation and derive closed-form expressions to show the impacts of network statistics on the perceptual video quality.
- ▷ *Optimal Control*: Given the network statistics as an input to our analytical model, the end user can determine the tradeoff between the start-up delay and fluency of playback, and employ various playback strategies to fit different requirements of video applications. Based on this idea, we design an adaptive playout buffer management scheme to optimally select the playback threshold. The proposed scheme is employed at the user end via local estimation only without any assistance from the networks, which is hence particularly suitable for large scale deployments.

The remainder of this paper is organized as follows: Section II discusses the related work in the literature. Section III presents the system model and defines the main performance metrics. The analytical framework is developed in Section IV, and the adaptive video playback scheme is described in Section V. Extensive simulations are conducted to validate the proposed framework in Section VI, followed by the concluding remarks in Section VII.

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, with nice reviews provided in [1], [3], including the rate-distortion optimized packets scheduling and routing [4], power control and adaptive coding at the transmitter [5], playback rate and strategy adaption at the receiver [6]-[12]. Our work belongs to the scope of end-system centric solutions [3] 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 user end based on the occupancy of playout buffer or user's bandwidth [6]. Liu *et al.* [7] propose an end-to-end playback rate adaptation scheme based on the layer coding technique. In their work, each end-system actively measures its local available bandwidth and informs 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. As a result, users can

enjoy smooth playback. Galluccio *et al.* [8] describe an adaptive MPEG video streaming framework in which the wireless channel is modeled as a Rayleigh fading channel represented by a 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 [5] where channel coding is adapted in different channel conditions which are evaluated using FSMC. However, the source adaptation schemes suffers from the scalability issue, as the server needs to response to each individual user and perform transcoding, *etc.*, to resolve different quality requirements. When the network scales to a large size, the server can be easily overloaded. Moreover, most of the previous work only consider a single-hop wireless channel which can be well modeled by FSMA. However, if multi-hop wireless transmissions are considered, the analysis becomes 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 works propose to adapt the video playback at the end users. Kalman *et al.* [9] introduces 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 enhanced to avoid the overflow of playout buffer. This leads to the effects of fast forward to the users. Laoutaris *et al.* [10] adopt the same mechanism but use the Markov decision process (MDP) to optimally determine the video playback rate at different channel conditions. Such scheme does not suffer from the scalability issue, but the perceptual visual quality by the end user may fluctuate due to time-varying rhythm of playback.

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.* [11] establish a Markovian model to study the tradeoff between playback continuity and start-up delay. In this work, the wireless channel is modeled as a FSMC and the interplay between the channel statistics and playout buffer is provided under different buffer strategies. However, the work only considers the single-hop scenario and can not be applied for multihop transmissions. Dua *et al.* [12] propose to adapt the playback threshold through a MDP. The channel is also modeled as a FSMC, where each successful transmission will incur certain profit. The playback buffer is managed to determine an optimal playback threshold to maximize the overall profit.

Unlike those previous efforts, in this paper, we consider the a general channel status by modeling the

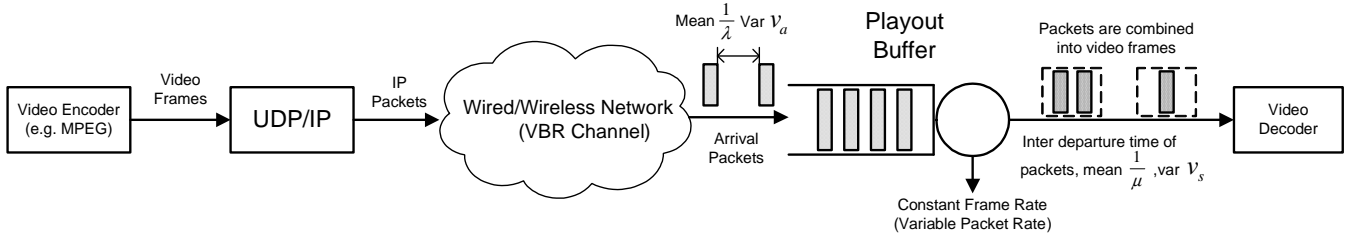


Fig. 1. Process of video streaming

playout buffer at the user end as a $G/G/1$ queue. Thus, 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 channel can be highly dynamic with intensive variance, we explicitly take delay jitter and variation of playback rate into consideration and show their impacts on the perceived video quality.

III. SYSTEM MODEL

A. System Model

A typical structure of the video transmission system is shown in Fig. 1. The video sequence generated by the video source is compressed and encoded into video frames with various frame sizes (e.g., by Motion Picture Experts Group (MPEG) encoder), at the constant frame rate. The video frames are then packetized using a User Datagram Protocol (UDP)/IP protocol suite. Since each video frame may consists of one or multiple IP packets, the packet rate of video flows is variable. The video packets are transmitted over a VBR wired/wireless network which introduces variable delays to the arrived video packets. At the end user, the downloaded packets are first stored at the playout buffer, then recovered into video frames and injected into the video player at the same constant frame rate (i.e., variable packet rate) which the encoder generates.

The envisioned evolution of the playout buffer is shown in Fig. 2. In general, the video playback process can be divided into two iterative phases, namely the charging phase and playback phase. The charging phase starts once the buffer becomes empty. In this case, the buffer is charged with continuously downloaded packets and the playback is kept frozen until b packets are filled. Henceforth, we refer to b as the threshold of playback. Let a R.V.(random variable) \mathcal{D} denote the duration of charging phase, i.e., start-up delay. The playback phase starts after the 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 a R.V. \mathcal{T} be the duration of the video playback phase. The charging and playback phases iterate until the whole video is downloaded.

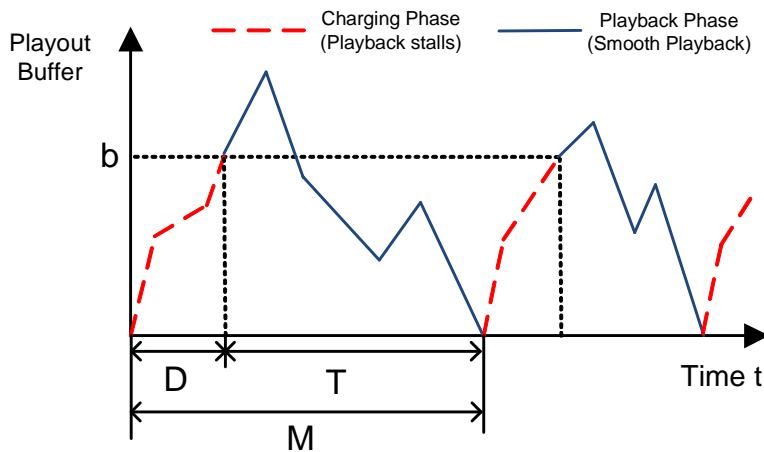


Fig. 2. Evolution of the playout buffer during media playout

B. User-oriented Metrics of Video Quality

We define the following performance metrics to evaluate video quality from the end user's perspective.

▷ Start-up Delay \mathcal{D} and Duration of Video Playback \mathcal{T}

Intuitively, both R.V.s \mathcal{D} and \mathcal{T} monotonically increase with the playback threshold b because a larger b will not only prolong the duration of playback \mathcal{T} but also enlarge the annoying start-up delay \mathcal{D} . To provide smooth playback with tolerable start-up delay, we need to strike a balance between \mathcal{D} and \mathcal{T} by properly setting the threshold b .

▷ Stopping Probability \mathcal{P}

The stopping probability \mathcal{P} is defined as the probability that the playback stops in the middle of the video presentation given that b packets are buffered initially, which is given by

$$\mathcal{P} = \Pr(t < S | B(t) = 0, B(0) = b), \quad (1)$$

where S denotes the video length and $B(t)$ denotes the number of packets stored in the playout buffer at time t .

▷ Number of Playback Frozen \mathcal{F}

The number of playback frozen \mathcal{F} is defined as the total number of playback interruptions encountered by the end user during the interval $[0, t]$. Intuitively, \mathcal{F} is non-decreasing with time t and non-increasing with threshold b .

Henceforth, we refer to the four performance metrics, \mathcal{D} , \mathcal{T} , \mathcal{P} and \mathcal{F} as the user-oriented QoS metrics for video flows. In the following section, we develop an analytical framework to reveal the impacts of

network dynamics and initial buffers storage b on these QoS metrics. We study from the user's perspective by analyzing the evolution of playout buffer.

IV. ANALYTICAL FRAMEWORK

A. Buffer Model

Without loss of generality, we assume that the inter-arrival time of video packets follows a given but arbitrary distribution with mean $\frac{1}{\lambda}$ and variance v_a , as shown in Fig. 1. The packets are combined into variable size frames and served for playback. As video frames are played at the constant rate, the service rate in terms of video packets is variable. We assume that the inter-departure time of video packets which is determined by the instantaneous video playback rate also follow a general distribution with mean $\frac{1}{\mu}$ and variance v_s . The playout buffer is thus modeled as a $G/G/1$ queue which is applicable for various network scenarios and video codecs. We assume the buffer capacity is sufficiently large to accommodate the whole video file. This is practically true for current PCs and most end-system device.

Given the statistics of packet arrival, λ , v_a , and packet playback or departure, μ and v_s , we analyze the transient evolution of the buffer size $B(t)$ in the charging and playback phases, and obtain the user-oriented performance metrics in terms of the initial buffer size b and the stopping probability of playback. However, it is well known that the exact analysis of the $G/G/1$ queue is very difficult to derive, not mention the transient evolution of the queue at any instant. Thus, we resort to the diffusion approximation for compact solutions [13].

B. Diffusion Approximation

The diffusion approximation method [14], [15] consists in replacing the discrete process, *e.g.*, the variation of queue length, with an appropriate diffusion process with continuous-path. The diffusion process is typically modeled by a Brownian motion process which has independent normally distributed increments with the same mean and variance as the original discrete one. In this case, the probability density of the queue length can then be characterized by partial differential equations and solved explicitly with given boundary and initial conditions. Assuming that arrivals and departures are mutually independent, we approximate the discrete buffer size $B(t)$ by a continuous process $X(t)$ and model it as the Brownian motion,

$$dX(t) = X(t + dt) - X(t) = \beta dt + G\sqrt{\alpha dt}, \quad (2)$$

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

$$\begin{cases} \beta = E\left(\lim_{\Delta t \rightarrow 0} \frac{X(t)}{\Delta t}\right) = \lambda - \mu \\ \alpha = Var\left(\lim_{\Delta t \rightarrow 0} \frac{X(t)}{\Delta t}\right) = \lambda^3 v_a + \mu^3 v_s. \end{cases} \quad (3)$$

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

$$p(x|t, x_0) = \Pr(x \leq X(t) < x + dx | X(0) = x_0), \quad (4)$$

where x_0 is the initial queue length.

According to the diffusion approximation, $p(x|t, x_0)$ satisfies the (forward) diffusion equation, and we have

$$\frac{\partial p(x|t, x_0)}{\partial t} = \frac{\alpha}{2} \frac{\partial^2 p(x|t, x_0)}{\partial x^2} - \beta \frac{\partial p(x|t, x_0)}{\partial x}, \quad (5)$$

with the initial condition

$$p(x|0, x_0) = \delta(x - x_0). \quad (6)$$

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 . This approach substantially facilitates the computation of the start-up delay and playback duration. In the following subsections, we invoke the diffusion approximation to analyze the charging phase and playback phase respectively to study the user-oriented performance metrics.

C. Start-up Delay \mathcal{D}

We first analyze the charging phase to evaluate the start-up delay. In the charging phase, the buffer is initially empty, *i.e.*, $x_0 = 0$, and the playback is frozen, *i.e.*, $\mu = v_s = 0$. The buffer is charged with continuous arrival packets until b packets are stored. After that, the charging phase terminates and playback phase starts. The start-up delay is then given by

$$\mathcal{D} = \min\{t | X(0) = 0, X(t) = b, t > 0\}. \quad (7)$$

The evolution of the buffer in this phase can be modeled as a diffusion process with drift $\beta_D = \lambda$ and diffusion coefficient $\alpha_D = \lambda^3 v_a$, which are obtained from (3).

Let $g_D(t)$ denote the p.d.f. of start-up delay \mathcal{D} and $G_D(t)$ the corresponding CDF.

Define $P_D(x|t, 0)$ as the conditional CDF of $X(t)$ in the charging phase. In this case, the queue length $X(t)$ has never reached b before and the initial buffer is empty, mathematically,

$$P_D(x|t, 0) = \int_0^x p_D(y|t) = \Pr\{X(t) \leq x | X(0) = 0, X(\tau) < b \text{ for } 0 < \tau < t\}, \quad (8)$$

where $p_D(x|t) = \Pr\{x \leq X(t) < x + dx | X(0) = 0, X(\tau) < b \text{ for } 0 < \tau < t\}$ is the p.d.f. of $P_D(x, t)$.

Thus, we have

$$G_D(t) == \Pr\{\mathcal{D} \leq t\} = 1 - \Pr\{\mathcal{D} > t\} = 1 - P_D(b|t, 0) = 1 - \int_0^b p_D(y|t, 0) dy. \quad (9)$$

as $P_D(b|t, 0)$ computes the probability that $X(t)$ is still below b at time t , i.e., \mathcal{D} is greater than t .

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

$$g_D(t) = \frac{dG_D(t)}{dt} = -\frac{d}{dt} P_D(b|t, 0) = -\frac{d}{dt} \int_0^b p_D(y|t, 0) dy. \quad (10)$$

As $p_D(x|t, 0)$ can be described by the diffusion approximation with the queue length never exceeding b , as shown in Appendix A, it follows the diffusion equation (5), as

$$\frac{\partial p_D(x|t, 0)}{\partial t} = \frac{\alpha_D}{2} \frac{\partial^2 p_D(x|t, 0)}{\partial x^2} - \beta_D \frac{\partial p_D(x|t, 0)}{\partial x}, \quad x < b, \quad (11)$$

coupled with the initial condition

$$p_D(x|0, 0) = \delta(x), \quad (12)$$

and the boundary condition

$$p_D(b|t, 0) = 0. \quad (13)$$

(13) 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 [16].

Solving (11), (12) and (13) yields¹

$$p_D(x|t, 0) = \frac{1}{\sqrt{2\pi\alpha_D t}} \left[\exp\left\{-\frac{(x - \beta_D t)^2}{2\alpha_D t}\right\} - \exp\left\{\frac{2\beta_D b}{\alpha_D} - \frac{(x - 2b - \beta_D t)^2}{2\alpha_D t}\right\} \right] \quad (14)$$

¹The solution is obtained by the method of images as shown in [16], [17].

and

$$P_D(x|t, 0) = \Phi\left(\frac{x - \beta_D t}{\sqrt{\alpha_D t}}\right) - \exp\left\{\frac{2\beta_D b}{\alpha_D}\right\} \Phi\left(\frac{x - 2b - \beta_D t}{\sqrt{\alpha_D t}}\right), \quad (15)$$

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

Substituting (15) into (9) and (14), we can obtain the CDF of \mathcal{D} ,

$$G_D(t) = 1 - \Phi\left(\frac{b - \beta_D t}{\sqrt{\alpha_D t}}\right) + \exp\left\{\frac{2\beta_D b}{\alpha_D}\right\} \Phi\left(-\frac{b + \beta_D t}{\sqrt{\alpha_D t}}\right). \quad (16)$$

and its p.d.f.

$$g_D(t) = -\frac{\partial}{\partial t} P_D(b, t) = \frac{b}{\sqrt{2\pi\alpha_D t^3}} \exp\left\{-\frac{(b - \beta_D t)^2}{2\alpha_D t}\right\}. \quad (17)$$

The moment generating function (m.g.f.), *i.e.*, the Laplace transform, of $g_D(t)$ is,

$$g_D^*(s) = E(e^{-st}) = \exp\left[\frac{b}{\alpha_D} \left\{\beta_D - \sqrt{\beta_D^2 + 2s\alpha_D}\right\}\right]. \quad (18)$$

The details are shown in Appendix B.

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

$$E(\mathcal{D}) = -\frac{d}{ds} g_D^*(s) \Big|_{s=0} = \frac{b}{\lambda}, \quad (19)$$

$$Var(\mathcal{D}) = \frac{d^2}{ds^2} g_D^*(s) \Big|_{s=0} - E^2(\mathcal{D}) = bv_a, \quad (20)$$

which indicate that the expected value and variance of start-up delay increase linearly with the playback threshold b .

D. Playback Duration \mathcal{T} and Stopping Probability \mathcal{P}

The video playback starts immediately after the charging phase. Without loss of generality, we focus on one playback phase and model it as a diffusion process starting at time $t = 0$. Playback phase terminates when the buffer becomes empty. In this case, the playback duration can be represented as

$$\mathcal{T} = \min\{t | X(0) = b, X(t) = 0, t > 0\}. \quad (21)$$

Denote $g_T(t)$ and $G_T(t)$ as the p.d.f. and CDF of \mathcal{T} , respectively. Define the conditional probability

that the buffer size is larger than x at time t ,

$$P_T(x|t, b) = \int_x^\infty p_T(y|t, b) dy = \Pr\{X(t) > x | X(0) = b, X(\tau) > 0 \text{ for } 0 < \tau < t\}, \quad (22)$$

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

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

$$g_T(t) = -\frac{d}{dt} \int_{0^+}^\infty p_T(y|t, b) dy. \quad (23)$$

where $p_T(x, t)$ is governed by following diffusion equation,

$$\frac{1}{2}\alpha_T \frac{\partial^2 p_T(x|t, b)}{\partial x^2} - \beta_T \frac{\partial p_T(x|t, b)}{\partial x} = \frac{\partial p_T(x|t, b)}{\partial t}, \quad (24)$$

subject to the initial and boundary conditions

$$p_T(x|0, b) = \delta(x - b), \quad t = 0, \quad (25)$$

$$p_T(0|t, b) = 0, \quad t > 0. \quad (26)$$

(26) is dictated by the events that the playback phase terminates when the buffer becomes empty. β_T and α_T can be derived from (3).

Solving the diffusion equations (24), (25) and (26), we have

$$p_T(x|t, b) = \frac{\exp\left\{\frac{\beta_T}{\alpha_T}(x - b) - \frac{\beta_T^2}{2\alpha_T}t\right\}}{\sqrt{2\pi\alpha_T t}} \left[\exp\left\{-\frac{(x - b)^2}{2\alpha_T t}\right\} - \exp\left\{-\frac{(x + b)^2}{2\alpha_T t}\right\} \right]. \quad (27)$$

Substitute (27) into (23), we have

$$g_T(t) = \frac{b}{\sqrt{2\pi\alpha_T t^3}} \exp\left\{-\frac{(\beta_T t + b)^2}{2\alpha_T t}\right\}, \quad (28)$$

and its m.g.f.

$$g_T^*(s) = \exp\left\{-\frac{b}{\alpha_T}(\beta_T + \sqrt{\beta_T^2 + 2\alpha_T s})\right\}. \quad (29)$$

The stopping probability \mathcal{P} is then obtained as

$$\begin{aligned}
\mathcal{P} &= \lim_{S \rightarrow \infty} \Pr(t < S | B(t) = 0, B(0) = b) \\
&= \lim_{t \rightarrow \infty} \int_0^t g_T(\tau) d\tau = \lim_{s \rightarrow 0} g_T^*(s) \\
&= \begin{cases} 1, & \text{if } \beta_T \leq 0 \\ \exp\left\{-\frac{2b}{\alpha_T} \beta_T\right\}, & \text{if } \beta_T > 0 \end{cases}.
\end{aligned} \tag{30}$$

Note that the obtained stopping probability is conservative as the whole video length S is considered to be infinity. In reality, S is limited. However, this approximation does not generate much difference. This is because the video is more likely to stop in the early period of the video session.

Substitute (3) into (30), we have

$$\mathcal{P} = \begin{cases} 1, & \text{if } \lambda \leq \mu \\ \exp\left\{-\frac{2b}{\lambda^3 v_a + \mu^3 v_s} (\lambda - \mu)\right\}, & \text{if } \lambda > \mu \end{cases} \tag{31}$$

As shown in (31), the video playback stops with probability 1 when the mean downloading rate is less than or equal to the video playback rate. On the other hand, even if $\lambda > \mu$, *i.e.*, the mean traffic arrival rate or video downloading rate exceeds the average video playback rate, (31) shows that it is still possible that video playback stops due to the variance of packet arrivals and playback. In real-world deployments, $\lambda - \mu$ should be controlled to be a small value to economically use the bandwidth and accept more users for video streaming. In this case, the stopping probability \mathcal{P} is significantly affected by the threshold b and the variations of the network.

E. Number of Playback Frozen \mathcal{F}

We have shown that when the mean traffic arrival rate is smaller than the average video playback rate, the video playout will stop with probability one. People may be curious about how seriously the interruptions of playback are. To address this question, we derive the number of playback frozen \mathcal{F} encountered in this case. Let a R.V. M denote the duration between two consecutive playback frozen events, and we have $M = \mathcal{D} + \mathcal{T}$, as shown in Fig. 2. The durations between video frozen events are i.i.d. random variables as the traffic arrivals and departures are assumed independent.

Again, we invoke the diffusion approximation to obtain the p.d.f. of the number of playback frozen \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 frozen up to time t . The interarrival time between two consecutive increments of $X_F(t)$ is M , where $X_F(t)$ is a non-decreasing function of time t .

Denote the p.d.f. of M as $g_M(t)$. Hence,

$$g_M(t) = g_D(t) \otimes g_T(t), \quad (32)$$

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

$$g_M^*(s) = g_D^*(s) \cdot g_T^*(s). \quad (33)$$

Substitute (18) and (29) into (33), we can obtain the mean and variance of M as

$$E(M) = -\left. \frac{d}{ds} g_M^*(s) \right|_{s=0} = \frac{-b\mu}{\lambda(\lambda - \mu)}, \quad \lambda < \mu, \quad (34)$$

and

$$Var(M) = \left. \frac{d^2}{ds^2} g_M^*(s) \right|_{s=0} - E^2(M) = -b \frac{\mu^3(v_s + v_a) + 3v_a\lambda\mu(\lambda - \mu)}{(\lambda - \mu)^3}, \quad \lambda < \mu. \quad (35)$$

Denote $P_F(x|t, 0)$ the conditional CDF of $X_F(t)$ at time t , given the initial buffer size to be 0,

$$P_F(x|t, 0) = \Pr\{X_F(t) \leq x | X_F(0) = 0\}. \quad (36)$$

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

$$\frac{\partial P_F(x|t, 0)}{\partial t} = \frac{\alpha_F}{2} \frac{\partial^2 P_F(x|t, 0)}{\partial x^2} - \beta_F \frac{\partial P_F(x|t, 0)}{\partial x}, \quad (37)$$

coupled with the boundary condition

$$\begin{cases} \lim_{x \rightarrow \infty} P_F(x|t, 0) = 1, & t \geq 0, \\ \lim_{x \rightarrow 0} P_F(x|t, 0) = 0, & t \geq 0. \end{cases} \quad (38)$$

where $\beta_F = \frac{1}{E(M)}$ and $\alpha_F = \frac{Var(M)}{E^3(M)}$ can be derived from (3), (34) and (35).

Solving (37) and (38) we have

$$P_F(x|t, 0) = \Phi\left(\frac{x - \beta_F t}{\sqrt{\alpha_F t}}\right) - \exp\left\{\frac{2\beta_F x}{\alpha_F}\right\} \Phi\left(-\frac{x + \beta_F t}{\sqrt{\alpha_F t}}\right). \quad (39)$$

The mean and variance of the number of playback frozen at time t can be approximated as

$$E(\mathcal{F}) = \int_0^\infty x dP_F(x, t) \approx \beta_F t = -\frac{\lambda(\lambda - \mu)}{\mu b} t, \quad \lambda < \mu \quad (40)$$

$$Var(\mathcal{F}) = \int_0^\infty x^2 dP_F(x, t) \approx \alpha_F t = \frac{\mu^2 \lambda^3 (v_s + v_a) + 3v_a \lambda^4 (\lambda - \mu)}{b^2 \mu^2} t, \quad \lambda < \mu \quad (41)$$

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

V. QoS CONTROL

In this section, we show how to exploit the analytically obtained user-oriented QoS metrics, *i.e.*, \mathcal{D} , \mathcal{T} , \mathcal{P} and \mathcal{F} , to optimally control the video playout the maximal user utility.

A. Optimal Playout Buffer Control

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

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

$$\begin{aligned} \min_b \quad & \mathcal{P} + \varpi_1 (E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})) \\ \text{s.t.} \quad & \Pr\{\mathcal{D} \geq \hat{\mathcal{D}}\} \leq \zeta, \\ & b > 0. \end{aligned} \quad (42)$$

$P2$: if $\lambda \leq \mu$,

$$\begin{aligned} \min_b \quad & E(\mathcal{F}) + \vartheta_F Var(\mathcal{F}) + \varpi_2 (E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})) \\ \text{s.t.} \quad & \Pr\{\mathcal{D} > \hat{\mathcal{D}}\} \leq \zeta, \\ & \Pr\{\mathcal{F} > \hat{\mathcal{F}}\} \leq \eta, \\ & b > 0. \end{aligned} \quad (43)$$

where $\omega_1, \omega_2 > 0$ are the weighting factors and $\vartheta_D, \vartheta_F \geq 0$ are called risk aversion factors which are adjustable with respect to different user requirements. ζ, η are predefined scalars 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 interruption. The objective is hence to avoid the playback frozen 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. It is large if users are sensitive to the start-up delay, *e.g.*, when watching football match. ϑ_D is called risk aversion

factor which models the user's attitude to the variance of start-up delay². When ϑ_D is large, the users are conservative and require strict start-up delay. The constraint is represented by a stochastic bound that the resulting start-up delay must be within the tolerable region \widehat{D} imposed by the user with high probability. The reason we apply the stochastic QoS is because providing absolute QoS guarantee may not be feasible and is typically difficult and costly to implement in the time-varying environment [18].

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 (31) and the objective is then to minimize the number of playback frozen 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) [19]. By substituting (16), (19), (20) and (31) into $P1$; (16), (19), (20), (39), (40) and (41) into $P2$, we have

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

$$\begin{aligned} \min_b \quad & \exp\left\{-\frac{2b}{\lambda^3 v_a + \mu^3 v_s}(\lambda - \mu)\right\} + \varpi_1 b \left(\frac{1}{\lambda} + \vartheta_D v_a\right) \\ \text{s.t.} \quad & \Phi\left(\frac{b - \lambda \widehat{D}}{\sqrt{\lambda^3 v_a \widehat{D}}}\right) - \exp\left\{\frac{2b}{\lambda^2 v_a}\right\} \Phi\left(-\frac{b + \lambda \widehat{D}}{\sqrt{\lambda^3 v_a \widehat{D}}}\right) \leq \zeta, \\ & b \geq 0. \end{aligned} \quad (44)$$

$P2'$: if $\lambda \leq \mu$,

$$\begin{aligned} \min_b \quad & \frac{A}{b} + \frac{\vartheta_F}{b^2} B + \varpi_2 b \left(\frac{1}{\lambda} + \vartheta_D v_a\right) \\ \text{s.t.} \quad & \Phi\left(\frac{b - \lambda \widehat{D}}{\sqrt{\lambda^3 v_a \widehat{D}}}\right) - \exp\left\{\frac{2b}{\lambda^2 v_a}\right\} \Phi\left(-\frac{b + \lambda \widehat{D}}{\sqrt{\lambda^3 v_a \widehat{D}}}\right) \leq \zeta, \\ & 1 - \Phi\left(\frac{\widehat{\mathcal{F}} - \beta_F S}{\sqrt{\alpha_F S}}\right) + \exp\left\{\frac{2\beta_F \widehat{\mathcal{F}}}{\alpha_F}\right\} \Phi\left(-\frac{\widehat{\mathcal{F}} + \beta_F S}{\sqrt{\alpha_F S}}\right) \leq \eta, \\ & b \geq 0. \end{aligned} \quad (45)$$

where $A = -\frac{\lambda(\lambda - \mu)}{\mu} S$, $B = \frac{\mu^2 \lambda^3 (v_s + v_a) + 3v_a \lambda^4 (\lambda - \mu)}{\mu^2} S$ are positive scalars. Here, we consider 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 input to the control scheme. This is reasonable as those 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 use the one-sided Chebyshev inequality, which states that for any random variable χ and any positive real number x ,

$$\Pr\{\chi - E(\chi) \geq x\} \leq \frac{\text{Var}(\chi)}{\text{Var}(\chi) + x^2}, \text{ for } \chi > E(\chi) \quad (46)$$

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

Using the Chebyshev inequality, together with (19), (20), (40) and (41), the constraints of $P2$ become

$$b \leq \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) - \sqrt{\frac{4\widehat{\mathcal{D}}\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2}, \quad (47)$$

$$b \geq \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}}, \quad (48)$$

where A and B are the same as in (45). The details are shown in Appendix

By replacing the constraints of $P1$ and $P2$ with (47) and (48), both $P1$ and $P2$ become convex optimization which could be solved efficiently. Note that the reduced complexity is at the expense 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.

In addition, to ensure that the resultant video performance is within the tolerable region, the threshold of playback b must be within the range specified by (47) and (C-7). To make this condition satisfied, we apply call admission control at the user end. In this case, the users will reject the request of playback directly without sending it to the media server if there is no positive b to meet both (47) and (C-7). This can save the waste of precious bandwidth in serving intolerable video playback.

B. Adaptive Media Playback

The above scheme assumes that the video contents are pre-stored with fixed video playback rate. In some cases, however, the server is able to modify the video playback adaptively based on the user's requirements. In this context, we can jointly consider the video playback adaption and the playout buffer management to achieve best visual quality using the following optimization framework,

$$\begin{aligned} \min_{b,\mu} \quad & \mathcal{P} + \varpi_1 (E(\mathcal{D}) + \vartheta_D \text{Var}(\mathcal{D})) - \varpi_3 \log \mu \\ \text{s.t.} \quad & \Pr \{ \mathcal{D} \geq \widehat{\mathcal{D}} \} \leq \zeta \\ & b \geq 0, 0 < \mu < \lambda \end{aligned} \quad (49)$$

where $\varpi_3 > 0$ is the weighting factor of the visual quality.

In this problem, the objective is to tradeoff between the fluency of video playback, start-up delay and the visual quality. In specific, as μ increases, the video is compressed at a lower ratio and the visual quality can be improved. However, this may harm the fluency of video playback and start-up delay as it demands more bandwidth resource for the video transmission. Hence, the optimization problem select an

optimal video playback rate μ which is affordable by the download bandwidth of users with $\mu < \lambda$, and meanwhile users achieve the best visual quality. At the same time, the playback threshold b is adjusted to tradeoff between start-up delay and stopping probability of playback.

Replacing the constraint (49), the above optimization problem is a convex problem which can be solved efficiently at the receiver. However, to implement this scheme in practice, the users need to inform the server the optimal playback rate, which introduces extra overhead traffic and workload to the server.

C. Network Planning

A salient feature of our analytical results is that they can bridge the network metrics, throughput and delay variations, with the user-oriented video quality metrics. Therefore, it can serve as a guideline for not only the receiver design, as shown above, but also the network protocol design, as shown below.

The network is generally shared by multiple users carrying multimedia traffics with diverse QoS requirements on the network resource. Even for the same video application, as users have different demands on the video quality, such as start-up delay and playback fluency, the required network resource is also different. In general, the network resource allocation can be represented by the following programming problem,

$$\begin{aligned}
 & \max_{\lambda_i, v_i} \quad \sum_i U_i(D, T, P, F) \\
 & s.t. \quad (D, T, P, F) \sim \mathcal{F}(\lambda_i, v_i), \\
 & \quad \quad (\lambda_i, v_i) \in \mathcal{N}, \\
 & \quad \quad \lambda_i, v_i \geq 0.
 \end{aligned} \tag{50}$$

The decision variables are λ_i and v_i which denote the end-to-end flow rate and delay variance of user i , respectively. The problem is to construct the network to assign various users with their desired throughput and delay variance. Towards this goal, the objective is defined to maximize the global network welfare where $U_i(D, T, P, F)$ denotes the utility of node i as a function of the user-oriented QoS metrics. Examples of U_i are the objectives of $P1$ and $P2$ while different users may have various weighting and risk aversion factors. The constraints $(D, T, P, F) \sim \mathcal{F}(\lambda_i, v_i)$ represents the relationship between resource allocation and the achieved video performance characterized by our analytical results. Examples are the constraints of problem $P1$ and $P2$. The constraint $(\lambda_i, v_i) \in \mathcal{N}$ models the network resource allocation scheme where \mathcal{N} denotes the feasible set of λ_i, v_i constrained by the network capacities and other ingredients, such as flow conservation. In a summary, problem (50) thus jointly considers the network construction

with the application-layer performance. In (50), the required video performance is first translated to the demanded throughput and variance guarantee, and then the network is construct to meet this demand and maximize the overall user utility.

VI. SIMULATION AND NUMERICAL RESULTS

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

A. Simulation Setup

We use four real VBR video clips encoded by MPEG-4 with diverse frame statistics from [20]. Each video clip lasts one 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.

TABLE I
STATISTICS OF VIDEO FRAMES

Video Clip Name	Frame Number	Frame Size (Bytes)		Bit Rate (bit/sec)	
		Mean	Variance	Mean	Peak
Jurassic Park I	89998	1.3e+03	1.3e+06	2.7e+05	1.7e+06
The Firm		4.2e+02	2.7e+06	8.4e+04	9.5e+05
Star Wars IV		3.9e+02	2.1e+05	7.8e+04	9.4e+05
Mr. Bean		9.2e+02	8.0e+05	1.8e+05	1.5e+06

The network scenario is shown in Fig. 1. In each simulation run, the simulator loads video frames from the video trace file. Video frames with various size are then segmented into IP packets with the maximum size of 1,400 Bytes. The available bandwidth of the network varies over time, *i.e.*, 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. Thus, the average throughput is 1.64 Mbps; the mean and standard deviation of network delay are $\frac{1}{\lambda} = 35.4$ ms and $v_a = 155.2$ ms, respectively. When the network condition changes, the mean and the variance of the packet delay varies accordingly. The proposed analytical framework uses the first and second moments of the network statistics and is applicable to general network settings. The video file is played at the constant rate 25 frames/sec by default. As each frame is comprised of variable number of packets, the video file is played at variable rate in terms of packets. For each experiment, we conduct 30 simulation runs and plot the mean results with the 95% confidence intervals.

B. Experimental Results

1) *Start-up Delay D and Playback Duration T* : In the first experiment, we verify the analysis of start-up delay D and playback duration T . We sequentially load the four video trace files which make the video session last 4 hours. This enable us to collect plenty samples. The mean inter departure time of downloaded packets of the trace files is measured as 31.2 msec, and the standard deviation is 11.5 msec.

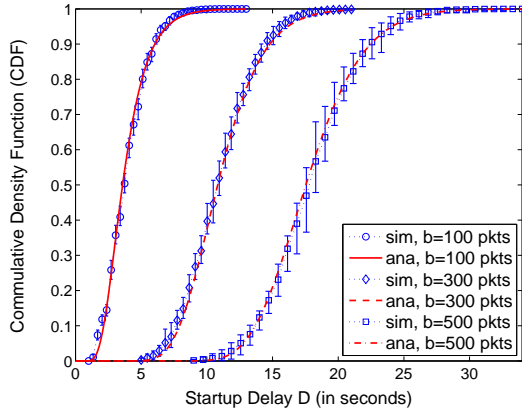


Fig. 3. CDF of the start-up delay D when b is 100, 300 and 500 packets, respectively

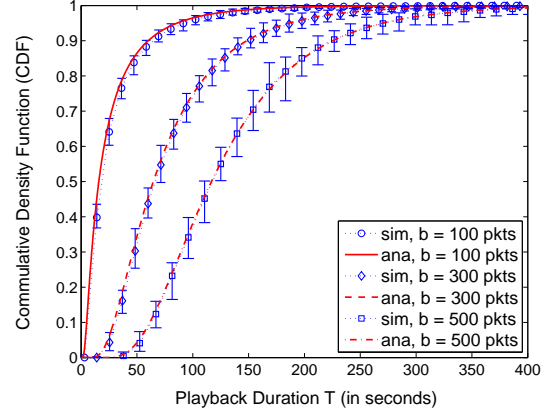


Fig. 4. CDF of the playback duration T when b is 100, 300, 500 packets, respectively

The CDF of the start-up delays and playback durations with different buffer thresholds b are shown in Fig. 3 and Fig. 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}* : In the second experiment, we select the clip "Jurassic Park I", and adjust the video playback rate to 20 frame/sec. This makes the mean packet arrival rate greater than the mean playback rate such that the stopping probability \mathcal{P} is less than one as shown in (31). The resulting mean and variance of the inter departure time of video packets are measured to be 39.7 msec and 15.2 msec, respectively. In this case, the video packets are downloaded at a faster rate than that of the playback. We conduct 30 experiments, and for each experiment we increase the buffer threshold b by 18 packets starting from 1 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 interruption. The simulation runs due to playback frozen are called frozen events. The probability of stopping is then computed as the total number of frozen events divided by 500. The result is plotted in Fig. 5. It is observed that the probability of stopping decreases exponentially with the increase of buffer threshold b .

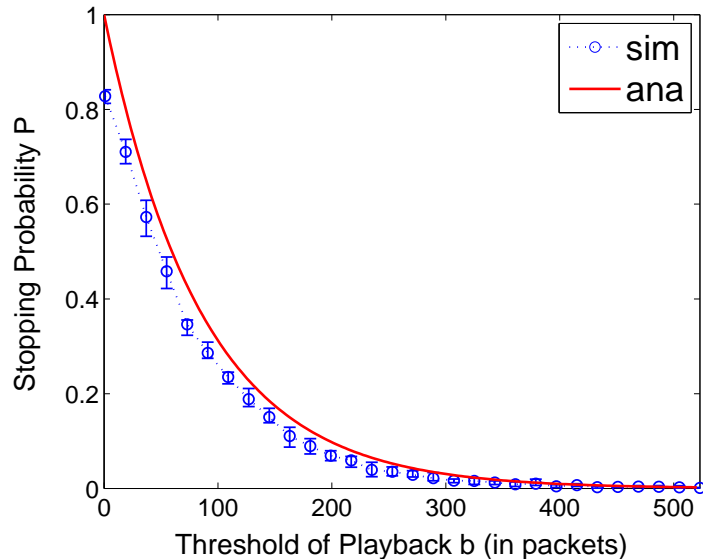


Fig. 5. The simulated stopping probability P

The analytical results are slightly larger than the simulation results because the video length S is assumed infinity for analysis while S is 1 hour in the experiments.

3) *Number of Playback Frozen \mathcal{F}* : We study the number of playback frozen using the clip "Jurassic Park I" with the default 25 frames/sec. The mean and standard deviation of the inter departure time of played packets are measured to be 26.4 msec and 12.2 msec, respectively. Fig. 6 plots the CDF of the number of playback frozen when b is 300 packets and the video length S is 1 hour. The analysis obtained from (39) well matches the simulation result. Fig 7 shows the comparison of the CDF curves of the number of playback frozen with different thresholds b . When b increases, the CDF curve shifts to the left which means that on average fewer events of playback frozen are encountered. However, the step size of the shifts is different and the mean number of playback frozen decreases dramatically when b is relatively small. Moreover, the width of the CDF curves becomes smaller for a larger b , which means that the variance of the number of playback frozen decreases when b increases.

4) *Optimal Playout Buffer Management*: Finally, we show the optimal selection of playback threshold b by solving $P1$ and $P2$ subject to the constraints (47) and (48). As $P1$ and $P2$ are applied at different channel conditions, we play the select the video clip "Jurassic Park I" at different frame rates as summarized in Table II. Fig. 8 shows the optimal threshold b of $P1$ with the increasing weighting factor ϖ_1 . We can see that the optimal b decreases monotonically with ϖ_1 increasing. This is because when ϖ_1 in (42) increases, the metric of start-up delay becomes more and more important and overwhelms the stopping probability. In this case, the optimal b is reduced accordingly to shrink the start-up delay at

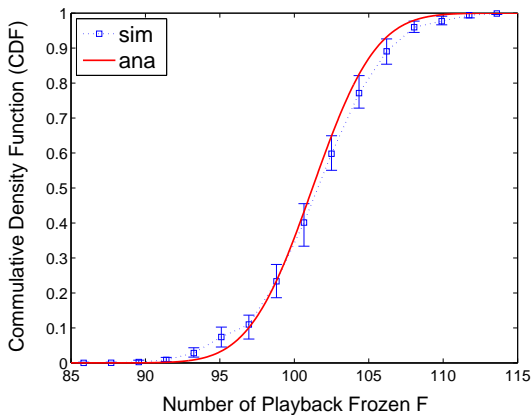


Fig. 6. Number of playback frozen F when $S = 1$ hr, b is 300 packets

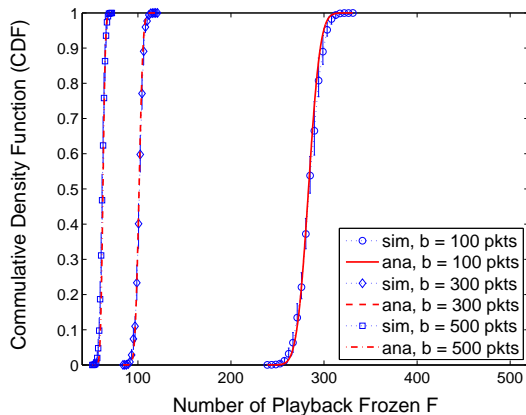


Fig. 7. Number of playback frozen F when $S = 1$ hr and b is 100, 300, 500 packets, respectively

TABLE II
SETTINGS OF PARAMETERS IN EXPERIMENTS OF OPTICAL PLAYOUT BUFFER MANAGEMENT

Scheme	Frame Rate	$1/\mu$	v_s	\hat{D}	\hat{F}	ζ	η	ϑ_D	ϑ_F
$P1(\lambda > \mu)$	20 frames/sec	39.7 msec	15.2 msec	30 sec	N/A	5%	N/A	1	N/A
$P2(\lambda \leq \mu)$	25 frames/sec	26.4 msec	12.2 msec	30 sec	100	5%	5%	1	1

the cost of a high playback frozen probability. The resultant stopping probability and utility of start-up delay at different optimal thresholds b are shown in Fig. 9, where the utility of start-up delay is computed as $b\left(\frac{1}{\lambda} + \vartheta_D v_a\right)$ which is the portion of utility characterized by start-up delay in (44). Fig. 10 plots the optimal thresholds b of $P2$ with the increasing weighting factor ϖ_2 . We can see that the corresponding optimal b decreases, as the metric of start-up delay becomes more and more critical when ϖ_2 increases. When ϖ_2 is very small, b is upper bounded by 450 packets. This is to guarantee that the resulting start-up delay is within the tolerable value \hat{D} . When ϖ_2 is very large, b is lower bounded by 300 packets which guarantees that the tolerable value \hat{F} is not violated. The resultant utilities of playback frozen and start-up delay at the different optimal thresholds of playback b are shown in Fig. 11 where the utility of playback frozen is the utility characterized by playback frozen in (45) computed as $\frac{A}{b} + \frac{\vartheta_F}{b^2} B$. The utility of start-up delay is computed in the same manner as in Fig. 9.

VII. CONCLUSION

We have developed a mathematical framework to study the impacts of the time-varying network channels on the perceived video quality of end users. We characterized the user's perceptual quality with four metrics, in terms of start-up delay and playback smoothness, and represented them by the network statistics and the threshold of playback in closed-form expressions. After that, we showed how to invoke the

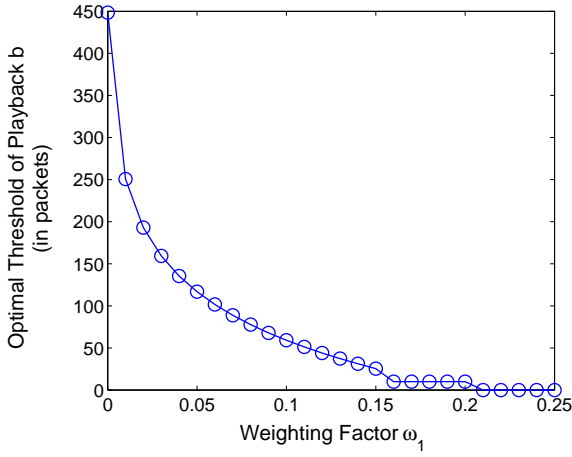


Fig. 8. The optimal playback threshold b with the increasing weighting factor ϖ_1

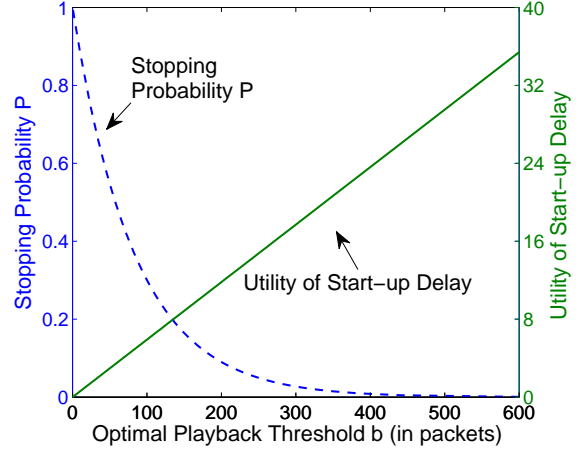


Fig. 9. Tradeoff between the stopping probability and start-up delay

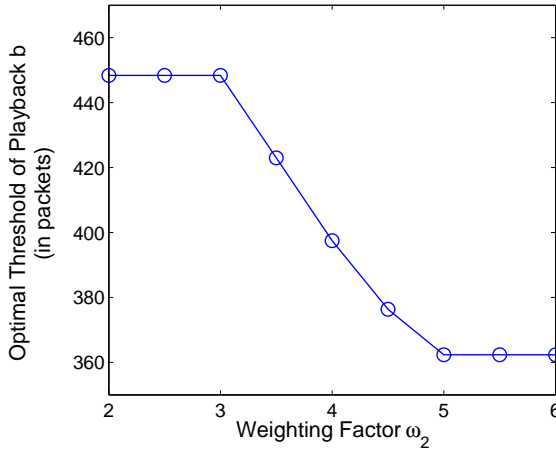


Fig. 10. The optimal playback threshold b with the increasing weighting factor ϖ_2

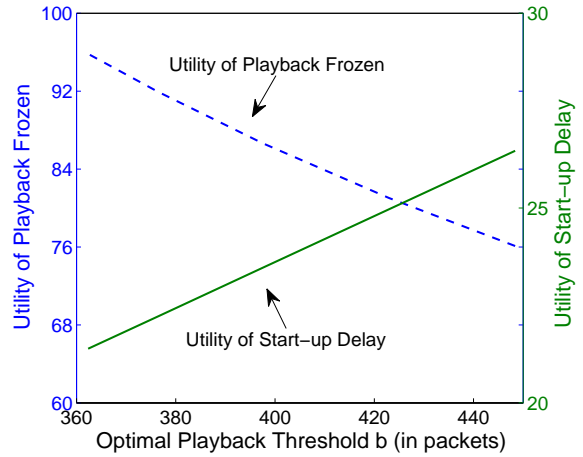


Fig. 11. Tradeoff between the number of playback frozen and start-up delay

analytical results to guide the playout buffer design and the network resource allocation. The analytical results were finally verified by extensive simulations.

We envision the future research in two dimensions. First, we intend to further extend the analysis by taking the time-dependent packet arrivals into account. Second, we will implement the proposed control scheme in real-world applications and test it under different network scenarios, *e.g.*, peer-to-peer streaming or cellular 3G networks.

APPENDIX

A. Diffusion Approximation of $p_D(x, t)$

Let $0 < \tau_1 < \tau_2 < \dots$ be the intervals between the arrival packets in the renew phase, which are i.i.d. random variables with mean and variance $\frac{1}{\lambda}$ and v_a , respectively. Let $S(n)$ represent the cumulative number of n arrivals, *i.e.*,

$$S(n) = \sum_{i=1}^n \tau_i, \quad (\text{A-1})$$

Let $N(t)$ be the number of arrivals until time t , or equivalently, the queue length at time t . Let us consider the short period $[0, \Delta]$ with start of video downloading synchronized to time 0. Apparently,

$$\Pr\{S(n) < \Delta\} = \Pr\{N(\Delta) \geq n\}. \quad (\text{A-2})$$

As $S(n)$ is the sum of i.i.d. random variables $\{\tau_i; i = 1, \dots, n\}$, when Δ is sufficiently large, with central limit theory,

$$\Pr\{S(n) < \Delta\} \approx \Phi\left(\frac{\Delta - n\frac{1}{\lambda}}{\sqrt{nv_a}}\right), \quad (\text{A-3})$$

where $\Phi(x)$ is the unit normal distribution with $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\{-\frac{y^2}{2}\} dy$.

With (A-2), we have

$$\Pr\{N(\Delta) < n\} = 1 - \Pr\{S(n) < \Delta\} \approx 1 - \Phi\left(\frac{\Delta - n\frac{1}{\lambda}}{\sqrt{nv_a}}\right) = \Phi\left(\frac{n - \lambda\Delta}{\sqrt{n\lambda^2 v_a}}\right). \quad (\text{A-4})$$

For sufficiently long period Δ , the distribution of $N(\Delta)$ is concentrated around $n \approx \lambda\Delta$. Substitute it into (A-4), yielding

$$\Pr\{N(t) < n\} \approx \Phi\left(\frac{n - \lambda t}{\sqrt{\lambda^3 t v_a}}\right). \quad (\text{A-5})$$

Without the boundary conditions $N(t) > 0$ and $N(t) < b$ imposed, the p.d.f. of $N(t)$ is

$$p(x, t) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left\{-\frac{(x - \lambda t)^2}{2\alpha t}\right\}, \quad (\text{A-6})$$

where $\alpha = \lambda^3 v_a$.

Apparently, $p(x, t)$ follows the diffusion equation as in (5). With the boundary condition $N(t) > 0$ and the absorbing state b imposed, the p.d.f. of queue length in the renew phase can hence be characterized by the diffusion equations shown in (11), (12) and (13).

B. Derivation of $g_D^*(s)$

The derivation of $g_D^*(s)$ is based on a standard result [21] that

$$\mathcal{L}\left(\frac{k}{2\sqrt{\pi t^3}} \exp\left\{-\frac{k^2}{4t}\right\}\right) = \exp\{-k\sqrt{s}\}, \quad (\text{B-1})$$

with $\mathcal{L}(f(t))$ denoting the Laplace transform of function $f(t)$. Denote by

$$f(k, t) = \frac{k}{2\sqrt{\pi t^3}} \exp\left\{-\frac{k^2}{4t}\right\}. \quad (\text{B-2})$$

and its Laplace transform

$$f^*(k, s) = \exp\{-k\sqrt{s}\}. \quad (\text{B-3})$$

As

$$g_D(t) = \frac{b}{\sqrt{2\pi\alpha_D t^3}} \exp\left\{-\frac{(b - \beta_D t)^2}{2\alpha_D t}\right\} = f\left(\frac{b}{\sqrt{\alpha_D/2}}, t\right) \cdot e^{-\frac{\beta_D^2}{2\alpha_D} t} \cdot e^{\frac{\beta_D b}{\alpha_D}}, \quad (\text{B-4})$$

by the property of Laplace transform, we hence have

$$g_D^*(s) = f^*\left(\frac{b}{\sqrt{\alpha_D/2}}, s + \frac{\beta_D^2}{2\alpha_D}\right) \cdot e^{\frac{\beta_D b}{\alpha_D}} = \exp\left[\frac{b}{\alpha_D} \left\{\beta_D - \sqrt{\beta_D^2 + 2s\alpha_D}\right\}\right]. \quad (\text{B-5})$$

C. Derivation of (47) and (48)

We show how to apply the Chebyshev inequality (46) to derive (47) and (48), respectively.

Based on (46), we have

$$\Pr\{\mathcal{D} > \widehat{\mathcal{D}}\} \leq \frac{\text{Var}(\mathcal{D})}{\text{Var}(\mathcal{D}) + \left(\widehat{\mathcal{D}} - E(\mathcal{D})\right)^2} = \frac{bv_a}{bv_a + \left(\widehat{\mathcal{D}} - \frac{b}{\lambda}\right)^2}. \quad (\text{C-1})$$

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

$$\frac{bv_a}{bv_a + \left(\widehat{\mathcal{D}} - \frac{b}{\lambda}\right)^2} \leq \zeta, \quad (\text{C-2})$$

which implies

$$\begin{aligned} b &\leq \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) - \sqrt{\frac{2D\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2} \quad \text{or} \\ b &\geq \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) + \sqrt{\frac{2D\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2}. \end{aligned} \quad (\text{C-3})$$

As $\widehat{\mathcal{D}} \geq E(\mathcal{D}) = \frac{b}{\lambda}$, we have $b \leq \widehat{\mathcal{D}}\lambda$. Together with (C-3), we have

$$b \leq \widehat{\mathcal{D}}\lambda + \frac{v_a(1-\zeta) - \sqrt{\frac{2\widehat{\mathcal{D}}\zeta}{\lambda}v_a(1-\zeta) + v_a^2(1-\zeta)^2}}{2\zeta/\lambda^2}. \quad (\text{C-4})$$

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

$$\Pr\{\mathcal{F} > \widehat{\mathcal{F}}\} \leq \frac{\text{Var}(\mathcal{F})}{\text{Var}(\mathcal{F}) + \left(\widehat{\mathcal{F}} - E(\mathcal{F})\right)^2} = \frac{\frac{B}{b^2}}{\frac{B}{b^2} + \left(\widehat{\mathcal{F}} - \frac{A}{\lambda}\right)^2}. \quad (\text{C-5})$$

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

$$\frac{\frac{B}{b^2}}{\frac{B}{b^2} + \left(\widehat{\mathcal{F}} - \frac{A}{\lambda}\right)^2} \leq \eta, \quad (\text{C-6})$$

which implies

$$b \leq \frac{A}{\widehat{\mathcal{F}}} - \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}} \quad \text{or} \quad b \geq \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}}. \quad (\text{C-7})$$

In addition, b be set such that $\widehat{\mathcal{F}} \geq E(\mathcal{F}) = \frac{A}{b}$, i.e., $b \geq \frac{A}{\widehat{\mathcal{F}}}$. Substitute it into (C-7), we have

$$b \geq \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta(1-\eta)}}{\eta\widehat{\mathcal{F}}} \quad (\text{C-8})$$

REFERENCES

- [1] B. Girod, J. Chakareski, M. Kalman, Y. J. Liang, E. Setton, and R. Zhang. Advances in network-adaptive video streaming. *Wireless Communications and Mobile Computing*, 2(6):549–552, 2002.
- [2] Y. Sun, I. Sheriff, E. M. Belding-Royer, and K. C. Almeroth. An experimental study of multimedia traffic performance in mesh networks. In *Proceedings of ACM WiTMeMo*, 2005.
- [3] Q. Zhang, W. Zhu, and Y.-Q. Zhang. End-to-end QoS for video delivery over wireless Internet. *Proceedings of the IEEE*, 93(1):123–134, 2005.
- [4] P. A. Chou and Z. Miao. Rate-distortion optimized streaming of packetized media. *IEEE Trans. on Multimedia*, 8(2):390–404, 2006.
- [5] J. Xu, X. Shen, J. W. Mark, and J. Cai. Adaptive transmission of multi-layered video over wireless fading channels. *IEEE Trans on Wireless Communications*, 6(6):2305, 2007.
- [6] N. Laoutaris and I. Stavrakakis. Intra-stream synchronization for continuous media streams: a survey of playout schedulers. *Network, IEEE*, 16(3):30–40, 2002.
- [7] J. Liu, B. Li, and Y.-Q. Zhang. An end-to-end adaptation protocol for layered video multicast using optimal rate allocation. *IEEE Trans. on Multimedia*, 6(1):87–102, Feb. 2004.
- [8] L. Galluccio, G. Morabito, and G. Schembra. Transmission of adaptive MPEG video over time-varying wireless channels: modeling and performance evaluation. *IEEE Trans. on Wireless Communications*, 4(6):2777–2788, 2005.
- [9] M. Kalman, E. Steinbach, and B. Girod. Adaptive media playout for low-delay video streaming over error-prone channels. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(6):841–851, 2004.

- [10] N. Laoutaris, B. V. Houdt, and I. Stavrakakis. Optimization of a packet video receiver under different levels of delay jitter: an analytical approach. *Performance Evaluation*, 55(3-4):251–275, 2004.
- [11] G. Liang and B. Liang. Balancing interruption frequency and buffering penalties in VBR video streaming. In *Proc. of IEEE Infocom*, 2007.
- [12] A. Dua and N. Bambos. Buffer management for wireless media streaming. In *Proc. of IEEE GLOBECOM*, 2007.
- [13] A. Duda. Transient diffusion approximation for some queueing systems. In *Proc. of ACM Sigmetrics*, 1983.
- [14] L. Kleinrock. *Queueing systems, volume II: computer applications*. John Wiley & Sons, 1976.
- [15] G. Louchard and G. Latouche. *Probability theory and computer science*. Academic Press Professional, Inc. San Diego, CA, USA, 1983.
- [16] D. R. Cox and H. D. Miller. *The theory of stochastic processes*. Chapman & Hall/CRC, 1977.
- [17] F. Pekergin T. Czachorski. Diffusion approximation as a modelling tool in congestion control and performance evaluation. In *Proc. of HET-NET*, 2004.
- [18] W. Kumwilaisak, Y. T. Hou, Q. Zhang, W. Zhu, C.-C. Jay Kuo, and Y.-Q. Zhang. A cross-Layer quality-of-service mapping architecture for video delivery in wireless networks. *Selected Areas in Communications, IEEE Journal on*, 21(10):1685–1698, 2003.
- [19] J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer, 1997.
- [20] F. H. P. Fitzek and M. Reisslein. MPEG-4 and H. 263 video traces for network performance evaluation. *Network, IEEE*, 15(6):40–54, 2001.
- [21] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions*. Dover, New York, 1965.