A Fluid Model Analysis of Streaming Media in the Presence of time-varying Bandwidth

J.W. Bosman*†, R.D. van der Mei*† and R. Nunez-Queija‡*

*CWI, Probability and Stochastic Networks, Amsterdam
†VU University, Faculty of Sciences, Amsterdam
‡University of Amsterdam, Korteweg - de Vries Insitute

***Abstract*—This paper is motivated by the increasing popularity of streaming media applications that are offered via the Internet. Packet streams generated by media application servers are distorted due to variability in the available bandwidth; at the receiving end, a play-out buffer compensates for the distortion. We study the dynamics of this system in a queuing-theoretical setting, using fluid analysis. To this end, we consider a model in which a constant bit-rate (CBR) media application is streamed over an unreliable network. Our model consists of a tandem of two fluid queues. The first queue is a Markov Modulated fluid queue that models the network congestion, and the second queue represents the play-out buffer. For this model the distribution of the total amount of fluid in the congestion and play-out buffers corresponds to the distribution of the maximum attained level of the first buffer. We show that the distribution of the total amount of fluid converges to a Gumbel extreme value distribution. From this result, we derive a simple closed-form expression for the initial playout-buffer level that provides a probabilistic guarantee for undisturbed playback.**

I. Introduction

Over the past few years, the tremendous popularity of smart mobile end devices and services (like YouTube) has boosted the demand for streaming media applications offered via the Internet. One of the key requirements for the success of providers of such services is the ability to deliver services at competitive price-quality ratios. However, the Internet provides no more than best-effort service quality. Therefore, the packet streams generated by streaming media applications are distorted by fluctuations in the available bandwidth on the Internet, which may be significant over the duration of a typical streaming application (whose duration may range from a few minutes to tens of minutes). To cope with these distortions, play-out buffers temporarily store packets so as to reproduce the signal with a fixed delay offset (see Figure 1). For smooth reproduction of the packet stream the playout buffer should not empty, as the stream will stall whenever packets do not arrive in time. For that reason, it is beneficial to start the play-out of a streaming media application *only when the playout buffer content exceeds some safety threshold value*. In this context, our main goal is to determine a proper choice for the initial playout-buffer level, providing a given probabilistic guarantee on undisturbed playback. Our objective in this paper is to contribute to the understanding of the performance implications of the playout-buffer settings for streaming applications over unreliable networks such as the Internet, by relating the proper buffer level to network variability parameters. Congestion is modeled by a fluid queue with fixed input rate and output rate determined by a stochastic process that is modeled as a Continuous Time Markov Chain (CTMC). This CTMC represents the IP network dynamics that causes congestion and fluctuations in available bandwidth. If this model is applied to a real network, the CTMC parameters must be estimated in order to capture the network behavior. Our approach relies on a queuing-theoretical fluid model analysis.

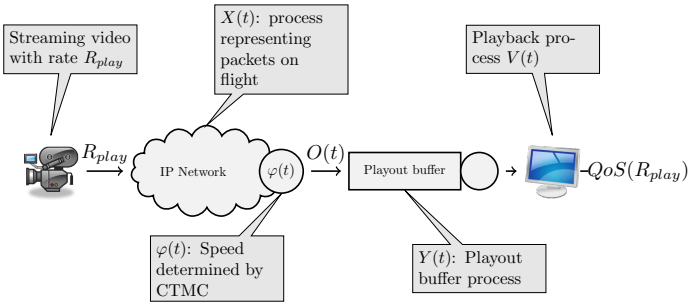

Fig. 1. Tandem of fluid buffers representing video streaming at rate R_{play} through an IP-network with variable output rate $O(t)$.

Fluid queues have proven to be a powerful modeling paradigm in a wide range of applications and have therefore received much attention in literature. On the one hand fluid models often capture the key characteristics that determine the performance of communication networks with complex packet-level dynamics, while on the other hand they remain mathematically tractable (hiding largely irrelevant details). Many analytic results have been obtained, and we refer to Scheinhardt [1] and Kulkarni [2] for excellent overviews of results on fluid queues that are directly relevant to our analysis. Asmussen and Bladt [3] propose a sample-path approach to study mean busy periods in Markov Modulated fluid queues, and derive a simple way of calculating mean busy periods in terms of steady-state quantities. In [4], Asmussen shows that the probability of buffer overflow within a busy cycle is approximately exponential, gives an explicit expression for the Laplace Transform of the busy period, and moreover, derives several inequalities and approximations for the transient behavior. Boxma and Dumas [5] study the busy

period of a fluid queue fed by N ON/OFF sources with exponential OFF periods and heavy tailed activity durations (more specifically, with regularly varying activity duration distributions). Scheinhardt and Zwart [6] study a two-node tandem with gradual input, and computer the steady-state joint buffer-content distribution using martingale methods. Kulkarni and Tzenova [7] study fluid queuing systems with different fluid-arrival rates governed by a CTMC and constant service rate. For this model, they derive a system of first-order non-homogeneous linear differential equations for the mean passage time. Sericola and Remiche [8] propose a method to analyze the maximum level and the hitting probabilities in a Markov driven fluid queue for various initial condition scenarios, allowing for both finite and infinite buffers. Their analysis leads to matrix differential Ricatti equations for which there is a unique solution.

Our approach leads to a dimensioning rule for the play-out buffer, based on an extreme value distribution approximation. The analysis was strongly motivated by the classical papers of Berman [9] and Iglehart [10]. Berman [9] studies the limiting distribution of the maximum in sequences of random variables satisfying certain dependence conditions. Iglehart [10] derived asymptotic distributions for the extreme value of the buffer content and the number of customers in the GI/G/1 queue. We refer to Asmussen [11] for an excellent survey on extreme-value theory for queues. Our paper provides an alternative approach for the analysis in Asmussen [4] specified to our model. Through our approach, we obtain more explicit results for the targeted dimensioning rules.

The precise object of study in this paper is a fluid model for constant bit-rate streaming media applications in the presence of bandwidth that varies over time. More specifically, we consider a tandem model consisting of two fluid queues. The first queue is a Markov Modulated fluid queue that models the congestion in the network caused by bandwidth fluctuations. The second buffer represents the play-out buffer. For this model, we first show that the distribution of the total amount of fluid in the congestion and play-out buffer corresponds to the distribution of the maximum level of the first buffer. We then prove that the distribution of the total amount of fluid converges to a Gumbel extreme value distribution. Based on this result, we derive an explicit expression for the initial level of the playout-buffer at which the playout can best be started to guarantee undisturbed playback with sufficiently high probability.

Our analysis proceeds as follows: We use results from [8] for the analysis of the maximum in a busy period. Furthermore, we show that the busy period maximum has an exponential tail and the maximum grows logarithmically. We apply a result on mean busy periods from [7] to obtain the mean expected cycle time. Next we apply an approach similar to [10] in order to show that the maximum buffer level converges to a Gumbel extreme value distribution. From this result the correct initial playout buffer level can be estimated. As mentioned previously, our paper shows strong similarities with [4]. Like us, Asmussen shows that the maximum fluid level grows logarithmically over time and under proper scaling converges to random variable with a Gumbel extreme value distribution. In this paper we independently establish this result in a more intuitive manner. Furthermore we provide an explicit recipe to calculate the asymptotic behavior of the maximum level in the Markov Modulated fluid queue. This can directly be applied to dimension the initial playout buffer size.

The organization of this paper is as follows. In Section II we describe the fluid more in detail, introduce the proper notation and formulate the problem. In Section III we discuss the details of a fluid-queue analysis, and in Section IV we translate the results in Section III to derive a simple rule for the proper value of the initial buffer content at which the playout should be started. In Section V, we provide a numerical validation of the model by means of simulations. Section VI contains a discussion of the results and looks out to future work.

II. Model and problem statement

In our model we mimic a video stream that has fixed data rate R_{play}. Video is streamed through an IP network with fluctuating speed. From the IP network packets arrive to the play-out buffer with a rate that can take values from a finite set $\{s_i,\, i = 1, 2, \ldots, n\}$. The actual output rate of the network is determined by a stochastic process $\varphi(t)$ that is modeled by an n-state CTMC. The CTMC has generator matrix T and state-space $\mathcal{S} = \{1, \ldots, n\}$. States are arranged in increasing order such that $s_1 > \cdots > s_n$. State-space $\mathcal{S}$ can be separated into two subsets:

$$\mathcal{S}_- = \{i : s_i > R_{play}, i = 1, \ldots, n_-\},$$
$$\mathcal{S}_+ = \{i : s_i < R_{play}, i = n_- + 1, \ldots, n_- + n_+\}.$$

States in $\mathcal{S}_-$ imply a *decreasing* number of packets in flight, while for states in $\mathcal{S}_+$ this number increases. We write $n_- := |\mathcal{S}_-|$ and $n_+ := |\mathcal{S}_+|$. We assume that $\varphi(t)$ can modeled such that there exists a stationary distribution $\pi = (\pi_1, \ldots, \pi_n)$. Furthermore we partition the generator matrix T of $\varphi(t)$ as an $(n_- + n_+) \times (n_- + n_+)$ matrix according to:

$$T = \begin{pmatrix} T_{--} & T_{-+} \\ T_{+-} & T_{++} \end{pmatrix}.$$

The combination of network congestion and play-out buffering is represented by a tandem of two fluid queues (see Figure 2). The first fluid buffer models the network congestion (packets on flight), and has corresponding fluid level $X(t)$. The second fluid buffer models the play-out buffering process at the client with corresponding fluid level $Y(t)$. Process $V(t)$ represents the video output rate of the play-out buffer. For the first fluid buffer we define rates of change by $r_i := R_{play} - s_i$ $(i = 1, \ldots, n)$, when $\varphi(t) = i$. Also for $\varphi(t) = i$, the rate of change in the second fluid buffer is exactly $-r_i$ whenever $Y(t) > 0$. Indeed, if $Y(t) > 0$, the play-out buffer can sustain output rate $V(t) = R_{play}$. On the contrary, when $Y(t) = 0$ and $s_i < R_{play}$ the play-out buffer stays empty and $V(t) = s_i$. In this case the video stream is disturbed. We define the following

$(n_-+n_+)\times(n_-+n_+)$ rate-of-change-matrix that is partitioned like generator matrix T:

$$R := \begin{pmatrix} R_- & 0 \\ 0 & R_+ \end{pmatrix}.$$

The entries are defined by:

$$\begin{aligned} R_- &:= \mathrm{diag}(r_i), && i \in S_-, \\ R_+ &:= \mathrm{diag}(r_i), && i \in S_+. \end{aligned}$$

So that the model represents a stable system, we assume that the average throughput S_{res} satisfies:

$$S_{res} = \sum_{i=1}^{n} s_i \pi_i > R_{play},$$

which is equivalent to requiring a negative drift $d < 0$ defined by:

$$d := \sum_{i=1}^{n} r_i \pi_i = R_{play} - S_{res}.$$

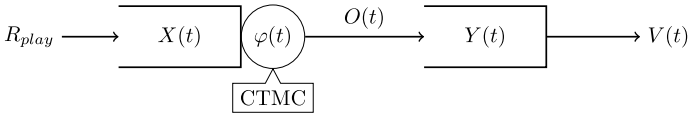

Fig. 2. Tandem of fluid buffers representing streaming video through an IP-network.

Due to congestion the play-out buffer level $Y(t)$ fluctuates. When the play-out buffer is empty video playback will be disturbed as only a rate of $V(t) < R_{play}$ is supported. We consider a video stream of length T_{play} and assume that the network has an average throughput $S_{res} > R_{play}$. Due to fluctuations in traffic volumes the bit-rate R_{play} may not be guaranteed at all times during T_{play}. At periods with high traffic, congestion builds up resulting in temporary throughput $s_i < R_{play}$. Therefore the video needs to be buffered at client side. When the play-out buffer is empty, video playback will be disturbed as the play-out rate of R_{play} can not be sustained. The result is that the video is quickly alternating between buffering and playback. This is commonly experienced as being very disturbing. We want to guarantee a certain Quality of Service (QoS) on the video playback. The QoS objective is to find an initial buffer level b_{init} such that the probability of disturbed playback during T_{play} is smaller than a given value p_{empty}:

$$\mathbb{P}\{\min_{0\le s\le t} V(s) < R_{play} \mid Y(0) = b_{init}\} < p_{empty}.$$

Of course the probability that play-out will be disturbed equals zero $p_{empty} = 0$ if a stream is fully buffered. However the larger the play-out pre-buffering, the larger loading time are. We want the play-out buffer to strike the right balance between both objectives, so that we aim for the minimal buffering threshold that guarantees undisturbed play-back with probability at least $1 - p_{empty}$. To do so, we develop a guideline that maps video parameters T_{play}, R_{play}, network characteristics (captured in generator matrix T and s_i), and QoS objective p_{empty} to an initial buffer level b_{init}.

III. Analysis

We are interested in a mapping from network, video characteristics and distortion probability p_{empty} to a minimal buffer level b_{init}:

$$\mathbb{P}\{\min_{0\le s\le t} V(s) < R_{play} \mid X(0) = 0, \; Y(0) = b_{init}\} < p_{empty}. \quad (1)$$

To this end we analyze the interaction between the network congestion buffer level $X(t)$ and the play-out buffer level $Y(t)$. In our analysis four different scenarios can be identified. These are depicted in Figure 3. Each scenario is represented by a time interval t_i:

1) During interval t_1 the network achieves a transfer rate lower than the video bit-rate $r_i < 0$ ($s_i < R_{play}$), while the play-out buffer level is positive $Y(t) > 0$. In this case the level of X increases while the level of Y decreases.
2) Within interval t_2 the network transfer rate is lower than video bit-rate $r_i < 0$, while the play-out buffer level is zero $Y(t) = 0$. Now the video playback will be disturbed and the play-out buffer remains empty $Y(t) = 0$ while the back-log in the network $X(t)$ continues to grow.
3) Next, in interval t_3 we have a network transfer rate higher than the video bit-rate $r_i > 0$ ($s_i > R_{play}$), while $X(t) > 0$. The level $X(t)$ decreases while $Y(t)$ increases.
4) Finally, during interval t_4 there is a network transfer rate higher than the video bit-rate $r_i > 0$, without any back-log in the network, $X(t) = 0$. Although a higher transfer rate $r_i > 0$ is supported, an effective network rate of R_{play} will be achieved as the fluid entering X directly flows to the play-out buffer Y.

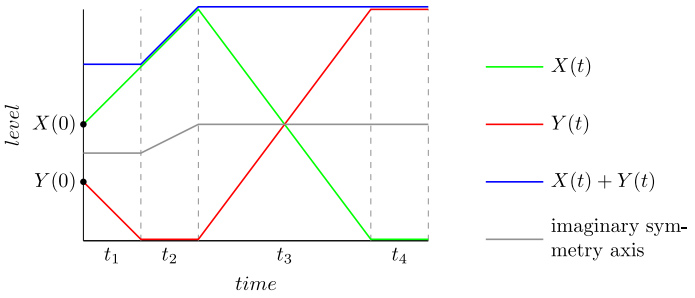

Fig. 3. Different behaviors of the stochastic processes.

Observe that in Figure 3 for intervals t_1, t_3 and t_4 $X(t)+Y(t)$ will remain constant. Therefore, in these cases an imaginary symmetry axis can be drawn between $X(t)$ and $Y(t)$. Moreover, within these intervals $V(t) = R_{play}$ and the CTMC determines how the constant level $X(t) + Y(t)$ is distributed over the first and second fluid buffer. In scenario

2 (corresponding to t_2 in Figure 3) the second buffer will remain empty ($Y(t) = 0$) while the first buffer continues to grow. In that case $X(t)$ attains a new maximum, and obviously $X(t) = X(t) + Y(t)$ since $Y(t) = 0$. Because $X(t)+Y(t)$ grows whenever $X(t)$ attains a new maximum, we can conclude that the total fluid buffer contents $X(t)+Y(t)$ is a not stationary process. However the growth of the maximum becomes an increasingly rare event each time a new maximum level is reached.

Definition We define the maximum level process as

$$M^*(t) := \sup_{0\le s\le t} X(s). \tag{2}$$

Lemma 3.1: Let $\big(X(t), Y(t)\big)$ be the stochastic process describing fluid levels in the tandem system. Then

$$X(t) + Y(t) = \sup_{0\le s\le t} X(s) = M^*(t). \tag{3}$$

Proof: Unless $Y(t) = 0$ and $\varphi(t) \in \mathcal{S}_+$ the total amount of fluid in $X(t)+Y(t)$ remains constant. Only de partition of fluid between $X(t)$ and $Y(t)$ changes as the rates of change for the two buffers have the same magnitude but opposite signs. Whenever $Y(t) = 0$ and $\varphi(t) \in \mathcal{S}_+$ the amount of fluid in $X(t)$ will grow while the the second buffer remains $Y(t) = 0$. In fact this is the moment where a new maximum level $M^*(t)$ for $X(t)$ will be reached. Therefore we can conclude that the total amount $X(t)+Y(t)$ must be equal to the maximum level $M^*(t)$. ∎

Using Lemma 3.1 we can rewrite equation (1) to:

$$\mathbb{P}\{M^*(T_{play}) > b_{init}\} < p_{empty}. \tag{4}$$

Lemma 3.1 focuses our problem on identifying the maximum level of packets on flight. Therefore we may neglect $Y(t)$ and consider the process $X(t)$ only. This process is driven by a CTMC and the process has negative drift. This results in a behavior where semi regenerative busy cycles are formed consisting of a busy period where $X(t) > 0$ that is followed by an idle period with $X(t) = 0$. In the special case $n_- = n_+ = 1$ when a busy period is initiated, the CTMC must be in the single state in $\mathcal{S}_+$, and when the busy period is terminated, the CTMC must be in the single state in $\mathcal{S}_-$. Therefore, in that special case the busy cycles constitute a regenerative process. This property is not restricted to this special case only, and we will assume it to hold when applying extreme value analysis in the next subsection. The analysis can be extended without this assumption, but is not contained in this paper.

A. Maximum of busy period in a Markov Modulated fluid queue

In Sericola and Remiche [8] the distribution of the maximum level reached in a busy period is derived using matrix exponential analysis. The resulting equations are rewritten such that they can be transformed into matrix differential Riccati equations. Recall that we have state space S, generator matrix T with corresponding rate matrix R. To determine the distribution of the maximum level in a busy period, fluid rates are uniformized resulting in a modified $(n_-+n_+)\times(n_-+n_+)$ matrix:

$$Q = R^{-1}T = \begin{pmatrix} Q_{--} & Q_{-+} \\ Q_{+-} & Q_{++} \end{pmatrix}$$

where the entries are defined by:

$$\begin{aligned} Q_{--} &= R_-^{-1}T_{--}, \\ Q_{-+} &= R_-^{-1}T_{-+}, \\ Q_{+-} &= R_+^{-1}T_{+-}, \\ Q_{++} &= R_+^{-1}T_{++}. \end{aligned}$$

We use M_+ to denote the maximum level of the process $X(t)$ in the first busy period after time $t = 0$ and will determine its distribution through

$$\begin{aligned} \Psi_{i,j}(x) := \mathbb{P}\big\{\varphi\big(\tau_0\big) = j, M_+ \le x \mid \varphi(0) = i, \\ X(0) = 0\big\}, \end{aligned} \tag{5}$$

$$\tau_0 := \inf\{t > 0 : X(t) = 0\},$$

This is the joint distribution for the maximum M_+ and the final state of the environment $\varphi(\tau_0)$ at the end of a busy period, given that the process starts empty at time $t = 0$ and $\varphi(0) = i$. This joint distribution is determined by solving a matrix differential Riccati equation. For the solution, the modified matrix Q is transformed into matrix exponential form:

$$e^{Qx} = \exp\left[\begin{pmatrix} Q_{--} & Q_{-+} \\ Q_{+-} & Q_{++} \end{pmatrix}\right] = \begin{pmatrix} A(x) & B(x) \\ C(x) & D(x) \end{pmatrix}. \tag{6}$$

The expression for $\Psi(x)$ is given by:

$$\Psi(x) = C(x)A(x)^{-1}. \tag{7}$$

In general, the maxima in consecutive busy periods are not independent, because the starting states of the environment may induce correlation. For the two-state model with $n_- = n_+ = 1$, busy cycles constitute regenerative sequences, implying that maxima in consecutive busy periods *are independent*. The non-regenerative nature of the general case implies several technical complications that, while we can handle them largely analogously using semi-regenerative processes, the technical details are not part of the scope of this paper. Instead, we specialize only for the two-state model and refer to future work for details on extensions to the semi-regenerative case.

For a system with a two-state environment process with transmission rates $s_1 > R_{play}$ and $s_2 < R_{play}$, generator matrix:

$$T = \begin{bmatrix} -\alpha_1 & \alpha_1 \\ \alpha_2 & -\alpha_2 \end{bmatrix},$$

rate matrix:

$$R = \begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix},$$

and generator matrix with uniformized fluid rates:

$$Q = \begin{bmatrix} \frac{-\alpha_1}{r_1} & \frac{\alpha_1}{r_1} \\ \frac{\alpha_2}{r_2} & -\frac{\alpha_2}{r_2} \end{bmatrix}.$$

For a system with two rates r_1, r_2 the solution is given by:

$$\Psi(x) = 1 - \frac{r_2\alpha_1 + r_1\alpha_2}{r_2\alpha_1 + r_1\alpha_2 e^{x(\frac{\alpha_1}{r_1}+\frac{\alpha_2}{r_2})}}.$$

The maximum of a busy cycle is given by:

$$\mathbb{P}\{M_+ \leq x\} = \Psi(x), \tag{8}$$

where M_+ represents the stochastic variable corresponding to the maximum in a busy cycle. The distribution of the maximum of a busy period for the two-state model has an exponential decaying tail, and when $x \to \infty$:

$$\begin{aligned} 1 - \Psi(x) &= \frac{r_2\alpha_1 + r_1\alpha_2}{r_2\alpha_1 + r_1\alpha_2 e^{x(\frac{\alpha_1}{r_1}+\frac{\alpha_2}{r_2})}} \\ &\sim \Big(\frac{r_1\alpha_2 + r_2\alpha_1}{r_1\alpha_2}\Big) e^{-x(\frac{\alpha_1}{r_1}+\frac{\alpha_2}{r_2})}. \end{aligned} \tag{9}$$

Similar to Iglehart [10, Lemma 1] we obtain an expression for

$$\mathbb{P}\{M_+ > x\} \sim be^{-\kappa x},\ x \to \infty. \tag{10}$$

In our case $b = \Big(\frac{r_1\alpha_2 + r_2\alpha_1}{r_1\alpha_2}\Big)$ and $\kappa = (\frac{\alpha_1}{r_1} + \frac{\alpha_2}{r_2})$.

Let $M_+(k)$ be the maximum of the kth busy cycle. Using similar arguments as in Iglehart [10, Lemma 2] we obtain

$$\lim_{n\to\infty} \mathbb{P}\{\kappa \max_{1\leq k\leq n} M_+(k) - \log(bn) \leq x\} = \Lambda(x), \tag{11}$$

where

$$\Lambda(x) = \exp[-e^{-x}]. \tag{12}$$

Here, we use the following extreme value theorem argument:

$$\begin{aligned} &\mathbb{P}\{\max_{1\leq k\leq n} M_+(k) \leq \frac{x + \log(bn)}{\kappa}\} \\ &= \mathbb{P}^n\{M_+(1) \leq \frac{x + \log(bn)}{\kappa}\} \\ &= \Big[1 - b\exp[-(x + \log(x + bn))] \\ &\qquad + o(\exp[-(x + \log(n))])\Big]^n. \end{aligned}$$

B. Maximum over time

Rather than the asymptotics for the busy cycles, we are interested in the evolution of the maximum over time. For this we use a result in Kulkarni and Tzenova [7]. In this paper an expression is derived for the joint mean first passage time in a Markov Modulated fluid queue:

$$\begin{aligned} &\mathbb{E}[\tau_{\mathcal{S}_-} \mid X(0) = x, \varphi(0) = i], \qquad i \in \mathcal{S}, \\ &\tau_{\mathcal{S}_-} := \inf\{t > 0 : X(t) = 0, \varphi(t) \in \mathcal{S}_-\}. \end{aligned} \tag{13}$$

The joint mean first passage time will be represented by the function $f_i(x)$:

$$f_i(x) := \mathbb{E}[\tau_{\mathcal{S}_-} \mid X(0) = x, \varphi(0) = i], \qquad i \in \mathcal{S}. \tag{14}$$

An expression for the joint mean first passage time can be obtained by solving the system of differential equations

$$R\frac{\mathrm{d}f(x)}{\mathrm{d}x} + Tf(x) + \mathbf{e} = 0, \tag{15}$$

with boundary condition

$$f_i(x) = 0, \qquad \forall i \in \mathcal{S}_-, \tag{16}$$

where $R = \mathrm{diag}(r_1, \ldots, r_n)$ is the diagonal matrix of rates of change, T is the generating matrix and where $\mathbf{e}$ is a column vector of ones. Here eigenvalues λ_j are the solution to

$$\det[R - \lambda T] = 0, \tag{17}$$

and the corresponding right eigenvectors ϕ_j satisfy:

$$\lambda_i R\phi_j = T\phi_j. \tag{18}$$

According to Kulkarni [2, Theorem 11.5] the eigenvalues can be ordered as follows:

$$\begin{aligned} Re(\lambda_1) \leq Re(\lambda_2) \leq \ldots \leq Re(\lambda_{n_+}) < 0 \\ < Re(\lambda_{n_++2}) \leq Re(\lambda_{n_++n_-}). \end{aligned} \tag{19}$$

There are n solutions to Equation (17) of which there are $n_- - 1$ eigenvalues with positive real part, one eigenvalue has real part equal to 0 and there are n_+ eigenvalues with negative real part. In Kulkarni and Tzenova [7, Theorem 4.2] the solution for (15) is given by:

$$f(x) = \sum_{j=n_++1}^{n_++n_-} a_j\phi_j e^{-\lambda_j x} - \frac{\mathbf{e}x}{d} + g. \tag{20}$$

In this expression g is a solution to

$$Tg = -(cR + I)\mathbf{e}. \tag{21}$$

Note that $\mathrm{rank}(T) = n - 1$ therefore we fix one free variable $g_n = 0$ in order to get a unique solution to (21). The coefficients a_j are obtained as a solution to:

$$\sum_{j=n_++1}^{n_++n_-} a_j\phi_{ij} + g_i = 0, \qquad \forall i \in \mathcal{S}_-, \quad r_i < 0, \tag{22}$$

where ϕ_{ij} is the ith entry of eigenvector ϕ_j. We are interested in the solution for the two-state model where $n_- = n_+ = 1$. Plugging in T and R into the results of Kulkarni [2, Example 1] gives:

$$\begin{aligned} &d = \frac{\alpha_2 r_1 + \alpha_1 r_2}{\alpha_1 + \alpha_2}, \\ &\lambda_1 = 0, \quad \lambda_2 = \frac{\alpha_2 r_1 + \alpha_1 r_2}{r_1 r_2}, \\ &\phi_1 = \big[1, 1\big]^t, \quad \phi_2 = \Big[-\frac{\alpha_1 r_2}{\alpha_2 r_1}, 1\Big]^t, \\ &g_1 = \frac{r_2 - r_1}{\alpha_1 r_2 + \alpha_2 r_1}, \quad g_2 = 0, \end{aligned}$$

which gives

$$\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} 0 \\ -\dfrac{r_2 - r_1}{\delta} \end{bmatrix} + \begin{bmatrix} -\dfrac{\alpha_1 + \alpha_2}{\delta} \\ -\dfrac{\alpha_1 + \alpha_2}{\delta} \end{bmatrix} x, \tag{23}$$

with

$$\delta := r_2\alpha_1 + r_1\alpha_2.$$

In the two-state model the only way that a busy period can be initiated is whenever $X(t) = 0$ and the state with $r_2 > 0$ is reached. The expected length of this busy period is equal to the first mean passage time:

$$\begin{aligned} f_2(0) &= \mathbb{E}[\tau_B \mid X(0) = 0, \varphi(0) = 2] \\ &= -\frac{r_2 - r_1}{r_2\alpha_1 + r_1\alpha_2}, \\ \tau_B &:= \inf\{t > 0 : X(t) = 0, \varphi(t) = 1\}. \end{aligned} \tag{24}$$

There is only one state that can end a busy period and that is $\varphi(t) = 1$ when $X(t) = 0$. This initiates an idle period that continues until the state $\varphi(t) = 2$ with rate r_2 is reached. The duration until the initiation of a consecutive busy period is exponentially distributed with mean $1/\alpha_1$. By combining the expected busy cycle with the expected idle period we obtain an expression for the total expected busy cycle:

$$\begin{aligned} \mathbb{E}[C] =& \mathbb{E}[\tau_B \mid X(0) = 0, \varphi(0) = 2] \\ &+ \mathbb{E}[\tau_I \mid X(0) = 0, \varphi(0) = 1] \\ =& -\frac{r_2 - r_1}{r_2\alpha_1 + r_1\alpha_2} + \frac{1}{\alpha_1} \\ =& \Big(\frac{r_1}{\alpha_1}\Big) \cdot \frac{\alpha_1 + \alpha_2}{r_2\alpha_1 + r_1\alpha_2}, \end{aligned} \tag{25}$$

with:

$$\begin{aligned} \tau_B &:= \inf\{t > 0 : X(t) = 0, \varphi(t) = 1\}, \\ \tau_I &:= \inf\{t > 0 : \varphi(t) = 2\}. \end{aligned}$$

In Equation (11) we stated that the asymptotic distribution of the maximum of a sequence of busy cycles converges to an extreme value distribution. We now derive the asymptotic distribution over time. Define $\{c(t) : t \geq 0\}$ as the counting process of busy cycles. Then $M^*(t)$ satisfies:

$$\begin{aligned} \max_{0 \leq k \leq c(t)} \{M_+(k) \leq x\} \leq M^*(t) \\ \leq \max_{0 \leq k \leq c(t)+1} \{M_+(k) \leq x\}. \end{aligned} \tag{26}$$

According to the weak law of large numbers we have:

$$\frac{c(t)}{t} \to \frac{1}{\mathbb{E}[C]}, \; t \to \infty. \tag{27}$$

Using Berman [9, Theorem 3.2] and equation (11) the limiting distribution becomes:

$$\lim_{t\to\infty} \mathbb{P}\{\kappa M^*(t) - \log(bt) \leq x\} = \Lambda^{\frac{1}{\mathbb{E}[C]}}(x). \tag{28}$$

In equation (28) the term $\frac{1}{\mathbb{E}[C]}$ from (27) represents the expected number of busy cycles per time unit (this corresponds to the c in Berman [9, Theorem 3.2]). The expression for the asymptotic distribution for the maximum of the two-state fluid queue

$$\mathbb{P}\{M^*(t) > b_{init}\} < p_{empty} \tag{29}$$

can now be expressed as:

$$\mathbb{P}\{\kappa M^*(t) - \log(bt) > x\} \approx 1 - \Lambda^{\frac{1}{\mathbb{E}[C]}}(x), \tag{30}$$

$$\mathbb{P}\{M^*(t) > b_{init}\} \approx 1 - \Lambda^{\frac{1}{\mathbb{E}[C]}}(\kappa b_{init} - \log(bt)), \tag{31}$$

whenever we have a sufficiently large b_{init} such that at least $b_{init} > \frac{\log(bt)}{\kappa}$.

Using the fact that when $t \to \infty$ the distribution of the maximum $M^*(t)$ converges to a Gumbel distribution, we can also establish the following asymptotic expectation of the maximum level:

$$\mathbb{E}[M^*(t)] \to \frac{\log\left(\frac{bt}{\mathbb{E}[C]}\right) + \gamma}{\kappa}, \qquad t \to \infty, \tag{32}$$

where $\gamma \approx 0.577215665$ is the Euler-Mascheroni constant. The behavior with respect to the real process is illustrated in Figure 6. Observe that $\mathbb{E}[M^*(t)]$ grows logarithmically over time with logarithmic slope $\frac{1}{\kappa}$.

IV. Dimensioning the initial buffer size

In section III we showed that the probability of an empty playout buffer corresponds to the maximum level reached by the first fluid buffer representing the number of packets in flight. Given the parameters that capture the network behavior (s and T) for a video stream with bit-rate R_{play} and duration T_{play} the initial buffer level b_{init} can be determined. Given the video playback QoS parameter p_{empty}, that represents the maximum probability a video is disturbed during T_{play}, the initial buffer size b_{init} should be chosen such that:

$$b_{init} > \frac{-\log[-\frac{\mathbb{E}[C]}{bT_{play}}\log(1 - p_{empty})]}{\kappa}. \tag{33}$$

This holds when we have T_{play} sufficiently large such that

$$\frac{-\log[-\frac{\mathbb{E}[C]}{bT_{play}}\log(1 - p_{empty})]}{\kappa} > 0.$$

This is a reasonable assumption, since we are considering video streams that have typically long durations (minutes and longer) compared to the time scale of fluctuations in the network transmission speed (typically in the order of seconds).

V. Numerical evaluation

In the previous sections we derived a mapping from the QoS parameter p_{empty} and streaming video duration T_{play} to minimal initial buffer level b_{init}. We will now run simulations in order to evaluate the accuracy of our mapping. Our parameter setting is as follows:

$$\begin{aligned} T &= \begin{bmatrix} -\alpha_1 & \alpha_1 \\ \alpha_2 & -\alpha_2 \end{bmatrix}, \\ \alpha_1 &= 0.1, & \alpha_2 &= 0.2, \\ s_1 &= 8Mbps, & s_2 &= 2Mbps, \\ R_{play} &= 4Mbps, \\ r_1 &= -4, & r_2 &= 2 \\ R &= \mathrm{diag}(\begin{bmatrix} r_1 & r_2 \end{bmatrix}). \end{aligned}$$

The simulation consists of 10.000.000 sample paths. Figure 4 represents the relative difference between target tail probability p_{empty} and the actual fraction of sample paths that exceed the

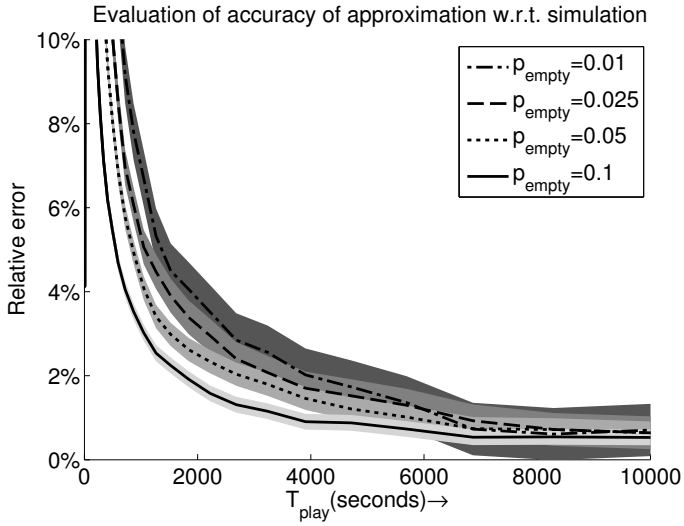

Fig. 4. Relative difference of buffer under-run probability to simulation. The gray bands around the lines are the 95% confidence intervals of the simulation.

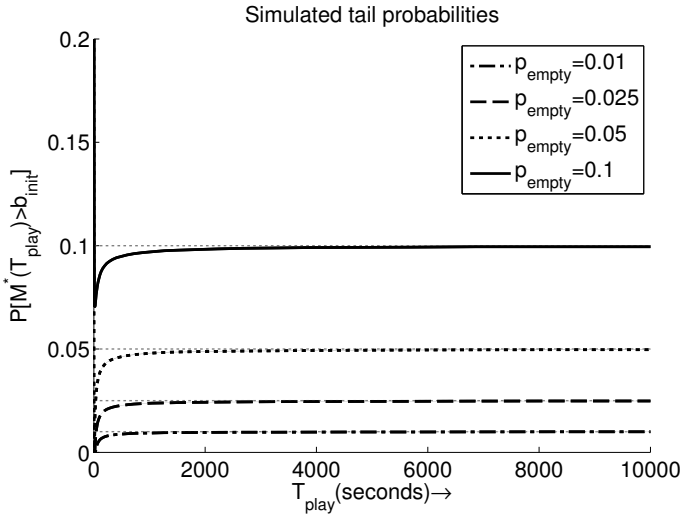

Fig. 5. Tail probabilities using empirical distribution based on simulation, evaluated on theoretical asymptotic percentiles.

buffer level approximation. We define the relative difference of approximation (app) and simulation (sim) by:

$$\text{diff}_{relative}(app, sim) = \left| \frac{app - sim}{sim} \right|. \tag{34}$$

From Figure 4 it can be observed that the error of the tail probability $\mathbb{P}\{M(T_{play}) > b_{init}\}$ quickly approaches the region below 5%. Figure 5 represents the actual fraction of sample paths that exceeds the theoretical asymptotic percentiles. The theoretical percentiles are based on equation (33). The straight thin dashed lines represent the desired tail probability.

The tail probabilities in Figure 5 indicate that the buffer level, derived from asymptotics, gives a conservative estimate, i.e., an overestimation of the tail probability. So using the asymptotics, depending on the duration of the video stream, the estimated buffer level is slightly higher than strictly needed.

In Section III-B we derived the asymptotic mean in Equation (32). We compare the asymptotic mean to the simulation results in Figure 6. This figure has a logarithmic time scale because we expect the mean maximum level to asymptotically converge to logarithmic growth with respect to time. From Figure 6 we observe that this is indeed the case.

In Figure 7 percentiles from simulation are compared to the theoretical asymptotic percentiles. Black lines represent simulation percentiles while gray lines represent the theoretical percentiles as expressed in Equation (33). On a linear time scale, simulation and asymptotic percentiles coincide quite closely.

Figure 8 presents the percentiles on logarithmic time scale. On small time scale we observe a "notch" in the simulation percentiles. This is caused by the fact that the figure is presented in logarithmic time scale. From the buffer process we can derive a coarse upper bound. A percentile at time t can not exceed $t \max(R - s_i)$ as $\max(R - s_i)$ is the maximal

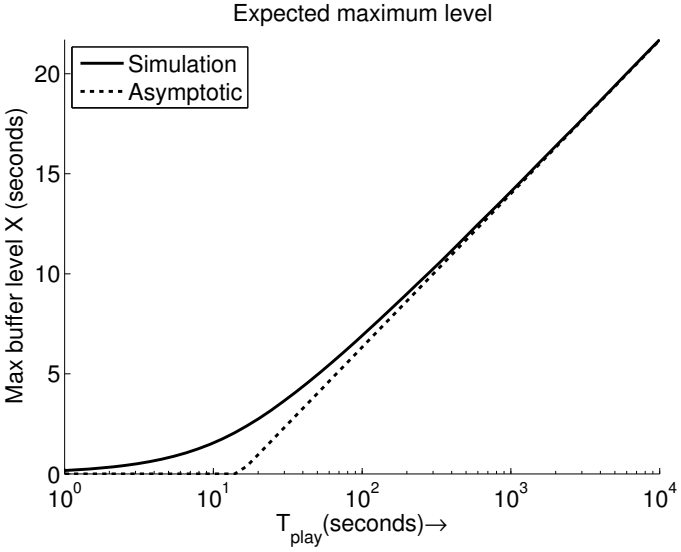

Fig. 6. Simulated and theoretical (asymptotic, see Equation (32)) expectation of the maximum level $M^*(t)$ on logarithmic time-scale.

possible growth rate of $X(t)+Y(t)$. The "notch" corresponds to the upper bound (which is curved due to logarithmic time scale).

VI. Discussion

We studied a model for a constant bit-rate video stream over an IP network with a play-out buffer at the client side. The network is modeled as a Markov Modulated fluid queue in which a CTMC determines the actual transmission rate through the network. For the play-out buffer an initial buffer level b_{init} was determined such that the probability that the video will stall during play-out will not exceed an agreed service level probability p_{empty}. We have shown that the probability of this event corresponds to the event of the maximum congestion level $M(t)$ exceeding the initial buffer level b_{init}.

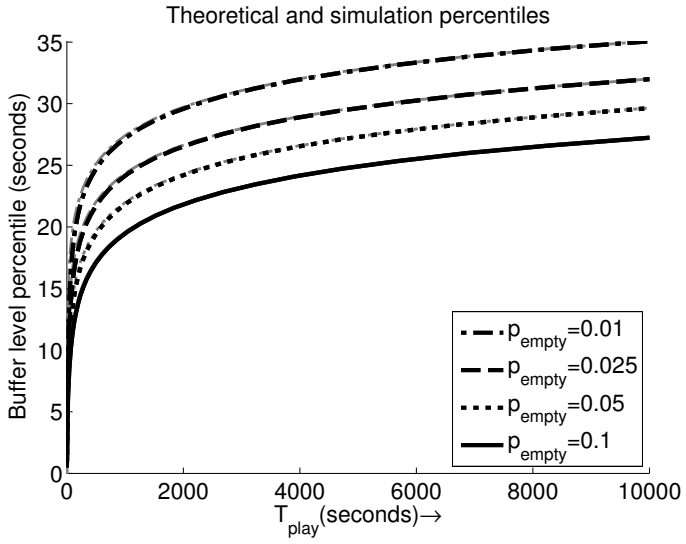

Fig. 7. Black lines represent simulation percentiles, gray lines represent theoretical asymptotic percentiles.

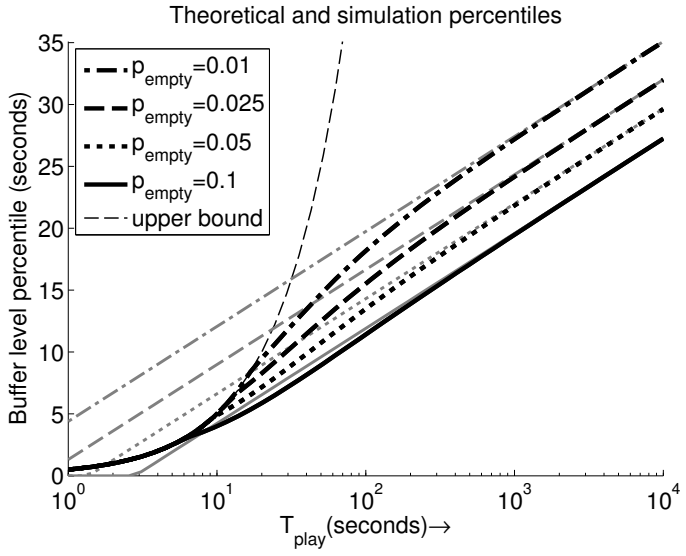

Fig. 8. Percentiles on logarithmic time scale. Black lines represent percentiles from simulation, gray lines represent theoretical asymptotic percentiles.

As a by-product, we found that the asymptotic distribution of the maximum level $M(t), t \to \infty$ has a Gumbel distribution, which is in agreement with earlier results in [4]. For smaller t the expression of the asymptotic distribution can be used to approximate the tail probability $\mathbb{P}\{M(T) > b_{init}\}$. From this expression we derived a formula that maps p_{empty}, T_{play} and the network and video parameters to a minimal buffer level b_{init}. Simulation results indicate that the buffer level that is obtained from the asymptotic analysis is a conservative estimate, i.e., it overestimates the true minimal required buffer level. The longer the video stream the more accurate the asymptotic prediction is. In adaptive media streaming, streaming servers tend to adapt R_{play} to the fluctauting available bandwidth. Our analysis facilitates proper parameter selection with respect to the altered network parameters.

In continuation of this work we have extended our analysis to networks with more than two throughput "modes" and general modulating Markov chains. Our results indicate that the convergence to the extreme value distribution depends on the rate of transitions in the modulating CTMC. In the examples we observe that for small time-scales the model is less accurate. An improvement could be achieved by adding an approximation for the behavior on shorter time-scales. We know that when $t \approx 0$ the distribution quantiles grow linearly with respect to transmission rate and initial distribution. We expect a mix of the small time-scale linear behavior model and the long time-scale extreme value model to be accurate across all time scales.

ACKNOWLEDGMENT

This work has been carried out in the context of the IOP GenCom project Service Optimization and Quality (SeQual), which is supported by the Dutch Ministry of Economic Affairs, Agriculture and Innovation via its agency Agentschap NL.

REFERENCES

[1] W. Scheinhardt, *Markov-modulated and feedback fluid queues*. Universiteit Twente, 1998.

[2] V. Kulkarni, "Fluid models for single buffer systems," *Frontiers in Queueing: Models and Applications in Science and Engineering*, pp. 321–338, 1997.

[3] S. Asmussen and M. Bladt, "A sample path approach to mean busy periods for markov-modulated queues and fluids," *Advances in Applied Probability*, pp. 1117–1121, 1994.

[4] S. Asmussen, "Busy period analysis, rare events and transient behavior in fluid flow models," *Journal of Applied Mathematics and Stochastic Analysis*, vol. 7, no. 3, pp. 269–299, 1994.

[5] O. Boxma and V. Dumas, "The busy period in the fluid queue," vol. 26, no. 1, 1998.

[6] W. Scheinhardt and B. Zwart, "A tandem fluid queue with gradual input," *Probability in the Engineering and Informational Sciences*, vol. 16, no. 1, pp. 29–45, 2002.

[7] V. Kulkarni and E. Tzenova, "Mean first passage times in fluid queues," *Operations Research Letters*, vol. 30, no. 5, pp. 308–318, 2002.

[8] B. Sericola and M. Remiche, "Maximum level and hitting probabilities in stochastic fluid flows using matrix differential riccati equations," *Methodology and Computing in Applied Probability*, vol. 13, no. 2, pp. 307–328, 2011.

[9] S. Berman, "Limiting distribution of the maximum term in sequences of dependent random variables," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 894–908, 1962.

[10] D. Iglehart, "Extreme values in the gi/g/1 queue," *The Annals of Mathematical Statistics*, pp. 627–635, 1972.

[11] S. Asmussen, "Extreme value theory for queues via cycle maxima," *Extremes*, vol. 1, no. 2, pp. 137–168, 1998.