Noname manuscript No.
(will be inserted by the editor)

A spectral theory approach for extreme value analysis in a tandem of fluid queues

J.W. Bosman · R. Núñez-Queija

January 27, 2014

Abstract We consider a model for to evaluate performance of streaming media over an unreliable network. Our model consists of a tandem of two fluid queues. The first fluid 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 buffer corresponds to the distribution of the maximum attained level of the first buffer. We show that, under proper scaling and when we let time go to infinity, 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 play-out buffer level that provides a probabilistic guarantee for undisturbed play-out.

1 Motivation and literature

Over the past few years, the tremendous popularity of smart phone 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

J.W. Bosman · R. Núñez-Queija
Centrum Wiskunde & Informatica (CWI),
P.O. Box 94079,
1090 GB Amsterdam
Tel.: +31 20 592 9333
E-mail: {J.W.Bosman, sindo}@cwi.nl

J.W. Bosman
Faculty of Sciences,
VU University Amsterdam

R. Núñez-Queija
Korteweg-de Vries Institute for Mathematics,
University of Amsterdam,
E-mail: nunezqueija@uva.nl

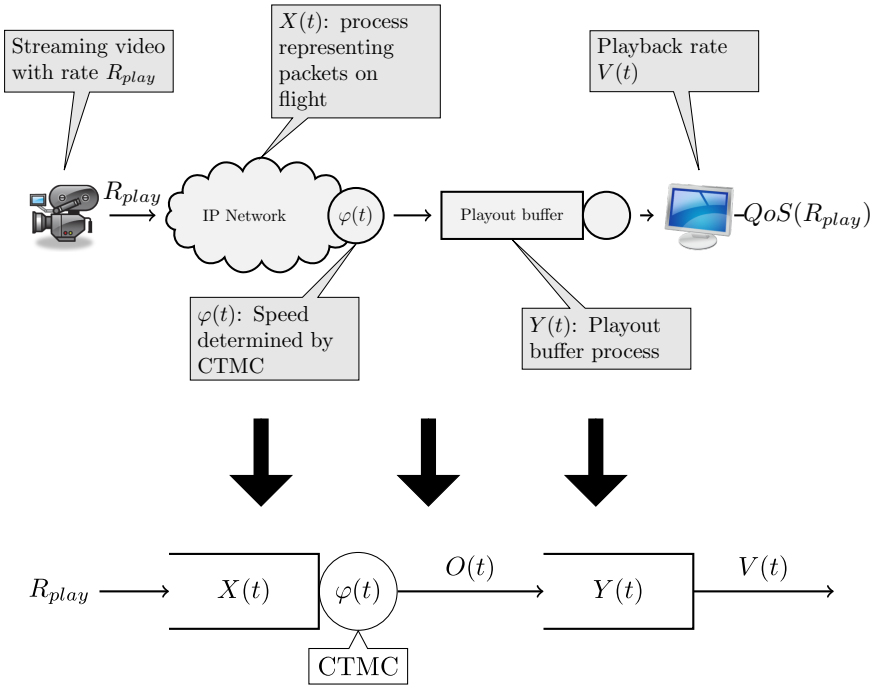


Fig. 1: Streaming video through an IP-network represented by a tandem of fluid queues.

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. Fig. 1 clarifies the connection of the model with the application to video streaming. The upper part of the figure depicts the network setting, whereas the lower part displays the abstract queueing model. The content of the first queue represents the part of the video that flows through the network and the second queue models the play-out buffer. For smooth reproduction of the packet stream the play-out 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 play-out buffer content exceeds some safety threshold value*. In this context, our main goal is to determine a proper choice for the initial play-out buffer level b_{min}, providing a given probabilistic guarantee on undisturbed play-out. Our objective in this paper is to contribute

to the understanding of the performance implications of the play-out buffer settings for streaming applications over unreliable networks such as the Internet, by relating the proper buffer level to network variability parameters. Our approach relies on a queueing-theoretical fluid model analysis. This work extends [6] by allowing multiple network throughput rates.

Buffer dimensioning for streaming video over variable rate networks has already received considerable attention in literature over the past two decades. Most work focused on balancing play-out buffer overflow and underrun probabilities, and develop dimensioning rules for the play-out buffer at the receiver end using analytic models. A particular large collection of work emerged in the 1990s. For example [15,5] determine the probability of overflow at the play-out buffer. This metric is particularly relevant for interactive video with stringent delay requirements, but less so for non-interactive streaming. More recently, play-out buffer engineering regained interest in the context of Voice over IP (VoIP), with popular examples such as Skype and Google Talk [23]. Again, VoIP play-out buffer dimensioning must balance between conversational interactivity and speech quality. Proper dimensioning of the play-out buffer is known to have a decisive impact on conversation quality [23]. The *real time interactive* character of VoIP, however, poses again stringent restrictions on the buffer size, making the trade-off very different from non-interactive (video) streaming, which is the objective of our paper. A third such example is in the context of closed-loop control for wireless streaming: [11], for example, investigate dynamic rules for play-out buffer management to avoid both the overflow of the play-out buffer and stalling of the streaming application. Of these papers, the setting studied in [14] is closest in nature to that in our paper. Their approach, however, builds on a "square root" formula to approximate the throughput of TCP and the stalling probability is obtained through a fixed-point solution. Somewhat tangent to the above mentioned literature, there are works that concentrate on dynamic deterministic optimization, e.g. [18,24]. In our model, network unreliability is captured by a stochastic (Markovian) process and buffer dimensioning is tailored to the variability of the network.

Despite the large volume of literature devoted to play-out buffer dimensioning, the problem is still highly timely because of tremendous popularity of video streaming services such as YouTube. This popularity is catalyzed by two main developments. One is the continuing rise of streaming media usage on mobile devices, who suffer from highly unpredictable channel conditions, making an accurate buffer dimensioning rule crucial for viability of such services. Cisco's global mobile data traffic forecasts predict that mobile video will make up for 66% of all mobile data traffic in 2017, amounting to an approximate monthly 7 Exabytes of mobile video worldwide, from less than 1 Exabyte in 2013 [9]. Second, the market for video traffic over the Internet shows tremendous growth as well, in terms of numbers of users as well as in traffic volume. Cisco [8] predicts that in 2017, every second, nearly a million minutes of video content will cross the global IP network, making up for 69% of all consumer Internet traffic (from 57% in 2012). This number further increases to nearly 90% if video exchanged through peer-to-peer file sharing is included. Particu-

larly relevant is *Internet video to TV*, which doubled in 2012 and continues to grow at a rapid pace, increasing fivefold by 2017.

Both in the context of wireless streaming and video on demand, a natural performance metric is the probability of uninterrupted video play out. The non-interactive nature of these services and the fact that memory is not the limiting factor on modern devices (naturally, mobile devices have much less memory, but videos played on mobile devices are streamed at a much lower bit rate also), make the memory usage a secondary consideration. The foremost important tradeoff is then between the initial play-out delay and the probability of stalling. We therefore set off to determine the *smallest* initial play-out delay (i.e. initial size of the play-out buffer) that gives a probabilistic guarantee on uninterrupted play out.

The above mentioned papers all focus on an engineering perspective. From a theoretical angle there is a considerable volume of related research too. Our modeling approach was already depicted in Fig.1: We will use a tandem of fluid queues (one with variable rate) to capture the most essential ingredients that determine the stalling probability (a detailed model description follows later). Fluid queues have proven to be a powerful modeling paradigm in a wide range of applications and have received much attention in literature. On one hand fluid model often capture the key characteristics that determine the performance of e.g. communication networks with complex packet-level dynamics (hiding largely irrelevant details), while on the other hand they remain mathematically tractable. Many analytic results have been obtained, and we refer to Scheinhardt [20] and Kulkarni [16] 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 [1], Asmussen shows that the probability of buffer overflow within a busy cycle has an exponential tail, gives an explicit expression for the Laplace Transform of the busy period and, moreover, derives several inequalities and approximations for the transient behaviour. Boxma and Dumas [7] 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 [21] study a two-node tandem with gradual input, and compute the steady-state joint buffer-content distribution using martingale methods. Kulkarni and Tzenova [17] study a 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 [22] propose a method to analyse 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. Asmussen [1] investigates a more general setting than the one considered in this paper, which focuses on

the streaming video setting. In our work we use an alternative matrix-theoretic analysis technique that is better suited for standard computational methods.

Our analysis was strongly motivated by the classical papers of Berman [4] and Iglehart [13]. Berman [4] studies the limiting distribution of the maximum in sequences of random variables satisfying certain dependence conditions. Iglehart [13] 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 [2] for an excellent survey on extreme-value theory for queues.

The precise object of study in this paper is a fluid model for streaming media applications in the presence of bandwidth that varies over time. More precisely, 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 show 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 play-out buffer at which the play-out can best be started so as to guarantee undisturbed play-out with sufficient certainty.

In our model the input rate into the first queue and the play-out rate are constant. This directly connects with constant bit rate streaming services, but can equally well describe variable bit rate services – which are more realistic from an application perspective – if the unit of transmission is interpreted as *time* rather than *number of bits* or *packets*. We will come back to this in our final discussion in Section 5, but for the purpose of the theoretical development and clarity of exposition we will focus on constant bit rate streaming throughout the analysis.

Our analysis proceeds as follows: We use results from [22] 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 from [17] on mean busy periods to obtain the mean expected cycle time. Next we apply an approach similar to [13]. This leads to our main result, as stated in Theorem 1 (Section 3.2):

Let $M^*(t) := \sup_{0\leq s\leq t} X(s)$ *be the supremum of first fluid queue level process* $X(t)$. *The limiting distribution of* $\kappa M^*(t) - \log(\frac{bt}{c})$ *when* $t \to \infty$ *converges to the standard Gumbel extreme value distribution where* b, c *and* κ *are constants that will be determined in our analysis in Section 3.*

Using this result the correct initial play-out buffer level can be estimated. As mentioned previously, our paper shows strong similarities with [1]. 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 direct manner using spectral analysis. Furthermore we provide an explicit recipe to calculate the asymptotic behaviour of the maximum level in

the Markov Modulated fluid queue. This can directly be applied to dimension the initial play-out buffer size.

The organization of the remainder of this paper is as follows. In Section 2 we describe the mathematical model followed by a fluid-queue analysis in Section 3. Moreover, in Section 3 we translate the results to derive a rule for the proper value of the initial buffer content at which the play-out should be started. In Section 4, we provide a numerical validation of the proposed dimensioning rule by means of simulations. Section 5 contains a discussion of the results, particularly in the context of *variable bit rate* streaming and looks out to future work.

2 Model

We consider a video stream with fixed data rate R_{play} (see Section 5 for use of the model for variable bit rate applications). Video is transported (through an IP network) with fluctuating speed. From the IP network packets arrive to the play-out buffer with rate s_i. The behaviour of rate s_i is determined by a stochastic process $\varphi(t)$ that is modelled by a n-state CMTC (Continuous Time Markov Process). 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 three subsets $\mathcal{S}_\downarrow$, $\mathcal{S}_0$ and $\mathcal{S}_\uparrow$, where $n_- := |\mathcal{S}_-|$, $n_0 := |\mathcal{S}_0|$, and $n_+ := |\mathcal{S}_+|$ and $n_\downarrow + n_0 + n_\uparrow = n$:

$$\begin{aligned}
\mathcal{S}_\downarrow &= \{i : s_i > R_{play}\} = \{1, \ldots, n_\downarrow\},\\
\mathcal{S}_0 &= \{i : s_i = R_{play}\} = \{n_\downarrow + 1, \ldots, n_\downarrow + n_0\},\\
\mathcal{S}_\uparrow &= \{i : s_i < R_{play}\} = \{n_\downarrow + n_0 + 1, \ldots, n_\downarrow + n_0 + n_\uparrow\}.
\end{aligned}$$

In short, $\mathcal{S}_\downarrow$ represents the states with decreasing number of packets in flight, $\mathcal{S}_0$ represents the states with stable number of packets in flight, and $\mathcal{S}_\uparrow$ states with increasing number of packets in flight. We assume that $\varphi(t)$ can modeled such that there exists a stationary distribution π. We partition the generator matrix T as a $(n_\downarrow + n_0 + n_\uparrow) \times (n_\downarrow + n_0 + n_\uparrow)$ matrix according to:

$$T = \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} & T_{\downarrow\uparrow} \\ T_{0\downarrow} & T_{00} & T_{0\uparrow} \\ T_{\uparrow\downarrow} & T_{\uparrow 0} & T_{\uparrow\uparrow} \end{pmatrix}. \tag{1}$$

As illustrated by Fig. 1, the combination of network congestion and play-out buffering is represented by a tandem of two fluid queues. 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 play-out rate that is achieved from the play-out buffer. For the first fluid buffer when $\varphi(t) = i$ we define rates of change by $r_i := R_{play} - s_i,\ i = 1, \ldots, n$. Whenever i is not explicitly specified, we use the value $i = \varphi(t)$. For the second fluid buffer the rate of change is directly proportional to the rate of

change of the first fluid buffer whenever $Y(t) > 0$, so that $V(t)$ can sustain play-out at rate R_{play}. However when $Y(t) = 0$ and $s_i < R_{play}$ the rate of change for the play-out buffer equals 0 and so $V(t) = s_i$. In this case video has a disturbed play-out. In practice the video may be stalled instead of continuously buffering and playing back. In that case the disturbed playback period in our model may be seem as a measure for the severity of distortion. We define the rate of change matrix that is partitioned like generator matrix T, i.e. a $(n_\downarrow + n_0 + n_\uparrow) \times (n_\downarrow + n_0 + n_\uparrow)$ matrix:

$$R := \begin{pmatrix} R_\downarrow & 0 & 0 \\ 0 & R_0 & 0 \\ 0 & 0 & R_\uparrow \end{pmatrix}. \tag{2}$$

The entries are defined by:

$$\begin{aligned} R_\downarrow &:= \mathrm{diag}(r_i), && i \in S_\downarrow, \\ R_0 &:= \mathrm{diag}(r_i) = 0 \text{ and} && i \in S_0, \\ R_\uparrow &:= \mathrm{diag}(r_i), && i \in S_\uparrow. \end{aligned}$$

In order for the first buffer to be stable the average potential throughput S_{res} must satisfy:

$$S_{res} := \sum_{i=1}^{n} s_i \pi_i > R_{play}. \tag{3}$$

The drift of the process is expressed in terms of rates of change r_i and is defined as:

$$d := \sum_{i=1}^{n} r_i \pi_i = R_{play} - S_{res}. \tag{4}$$

Stability condition (3) is equivalent to having a negative drift $d < 0$.

Due to congestion the play-out buffer level $Y(t)$ fluctuates. When the play-out buffer is empty video play-out will be disturbed as only a rate of $V(t) < R_{play}$ is supported. We consider a video stream of length $t = T_{play}$. Although we assume $S_{res} > R_{play}$ due to fluctuations in traffic the bitrate R_{play} cannot be guaranteed at all times t $(0 \leq t \leq T_{play})$ during T_{play}. At periods with high traffic, congestion in the network builds up resulting in a temporary throughput $O(t) = s_i < R_{play}$. Therefore the video needs to be buffered at play out. When the play-out buffer is empty video play out will be disturbed as a play-out rate of R_{play} can not be sustained. The result is that the video is alternating between buffering and play-out. This is commonly experienced as being very disturbing. In our analysis we assume in that case there will be a (disturbed) play out at a rate $s_i < R_{play}$ such that $Y(t)$ remains equal to 0. We want to guarantee a certain Quality of Service (QoS) on the video play-out. The QoS objective is to find an initial buffer level b_{min} such that the probability of disturbed play-out during T_{play} is smaller than p_{empty}:

$$\mathbb{P}\{\exists s \in [0,t] : V(s) < R_{play} \mid X(0) = 0, Y(0) = b_{min}\} < p_{empty}. \tag{5}$$

Of course the probability that play-out will be disturbed equals zero if a stream is fully buffered. However the larger the play-out buffer the longer the loading time. Therefore, we want the play-out buffer to have a minimal size. In order to minimize the initial buffer level b_{min} while meeting the QoS requirements, we develop a procedure that maps video parameters T_{play}, R_{play}, network characteristics and QoS objective p_{empty} onto a initial buffer level b_{min}.

3 Analysis

We are interested in a mapping from network, video characteristics and distortion probability p_{empty} to a minimal buffer level b_{min} such that the condition in Equation (5) is satisfied. 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 Fig. 2. 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 $s_i < R_{play}$ ($r_i < 0$), 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 $s_i < R_{play}$ ($r_i < 0$), while the play-out buffer level is zero $Y(t) = 0$. Now the video playback will be disturbed and the play-out buffer level will remain zero $Y(t) = 0$ while the network content $X(t)$ continues to grow.
3. Next, in interval t_3 we have a network transfer rate higher than the video bit-rate $s_i > R_{play}$ ($r_i > 0$), while the network content is positive $X(t) > 0$. The level of X decreases while the level of Y increases.
4. Finally, during interval t_4 there is a network transfer rate higher than the video bit-rate $s_i > R_{play}$ ($r_i > 0$), without any backlog in the network, $X(t) = 0$. Although higher transfer rate $r_i > 0$ is supported, an effective rate of R_{play} will be achieved as the fluid entering X directly flows to the play-out buffer Y.

Observe in Fig. 2 that within intervals t_1, t_3 and t_4, $X(t) + Y(t)$ remains constant. Therefore, in these cases an artificial 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 Fig. 2) the second buffer remains 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 $Y(t) = 0$. Each time $X(t)$ attains a new maximum, $X(t)+Y(t)$ grows. We can conclude that the total fluid buffer contents $X(t)+Y(t)$ is not a stationary process. However the growth of the maximum becomes an increasingly rare event each time a new maximum level is reached.

Lemma 1 *Let* $\big(X(t), Y(t)\big)$ *be the stochastic process describing fluid levels in the tandem system. Then, if* $Y(0) = 0$,

$$X(t) + Y(t) = \sup_{0 \le s \le t} X(s) = M^*(t). \tag{6}$$

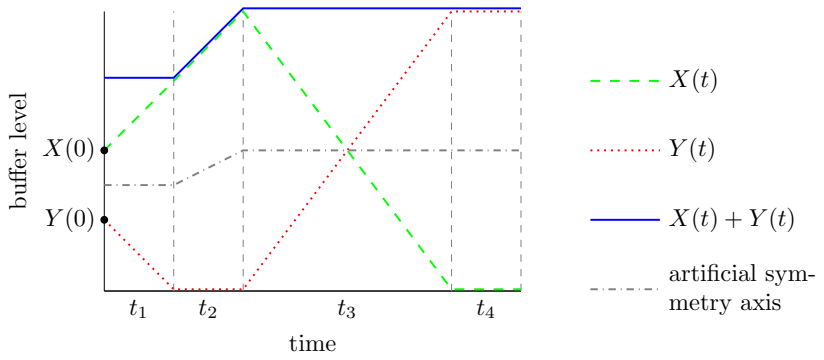


Fig. 2: Different phases of the stochastic processes $X(t)$ and $Y(t)$.

Proof Obviously, the initial conditions ensure that $M^*(0) = X(0) + Y(0)$. We will show that the maximum and the sum remain equal throughout time, because the maximum can only increase when $Y(t) = 0$. From the construction it is clear that, unless $Y(t) = 0$ and $\varphi(t) \in \mathcal{S}_+$, the total amount of fluid in $X(t) + Y(t)$ remains equal. Only the partition of fluid over $X(t)$ and $Y(t)$ changes as the rates of change for both buffers only differ in sign. On the contrary, when $Y(t) = 0$ and $\varphi(t) \in \mathcal{S}_+$ the amount of fluid in $X(t)$ will grow while the the second buffer remains $Y(t) = 0$ (because the inflow into the second buffer is below R_{play}). Beyond this point, both the maximum level $M^*(t)$ for $X(t)$ and $X(t)$ itself increase, as long as $Y(t)$ remains empty. We can conclude that the total amount $X(t) + Y(t)$ must always be equal to the maximum level $M^*(t)$. □

In equation (5) we use an initial buffer level of $Y(0) = b_{min}$, while in Lemma 1 we assume $Y(0) = 0$. However, setting $Y(0) = b_{min}$ and $X(0) = 0$ corresponds to the case where $X(0)$ has a virtual (initial) supremum equal to b_{min}. Thus we are interested in the probability that new supremum $M^*(t) > b_{min}$ is attained in time interval $[0, t]$ given that the initial supremum level is set to $M^*(0) = b_{min}$. Using the connection of the initial buffer level b_{min} to the supremum level $M^*(t)$ and Lemma 1 we can rewrite equation (5) to:

$$\mathbb{P}\{M^*(t) > b_{min}\} < p_{empty}. \tag{7}$$

This corresponds to the probability that $M^*(t)$ exceeds b_{min} when no initial-buffering is applied. We assume here and throughout the remainder of the paper the initial condition to be $X(0) = Y(0) = 0$.

Lemma 1 targets our problem on identifying the maximum level of packets on flight. Therefore we consider the process $X(t)$. The process is driven by a CTMC and the process has negative drift. This results in a behaviour where semi regenerative busy cycles are formed each consisting of a busy period with $X(t) > 0$ that is followed by an idle period.

We first will analyse in Section 3.2 the asymptotic behavior of the busy cycles leading to an extreme value distribution in Lemma 8. In Section 3.1 we analyse the mean busy cycle length $\mathbb{E}[C]$ in order to map the extreme value analysis of Section 3.1 from busy cycles to time.

3.1 Maximum over busy cycles

In Sericola and Remiche [22] the distribution of the maximum level reached in a busy period is derived using matrix exponential forms. The resulting equations are rewritten such that they can be transformed into matrix differential Riccati equations. For calculation of the distribution of the maximum level in a busy period, only the rates that change the buffer level ($r_i, i \notin S_0$) contribute to the solution. Moreover time is not considered in the distribution of the maximum level in a busy period. Therefore the rates can be uniformised resulting in $(n_\downarrow + n_\uparrow) \times (n_\downarrow + n_\uparrow)$ matrix Q that implicitly takes in account the states that are in $\mathcal{S}_0$:

$$Q = \begin{pmatrix} Q_{\downarrow\downarrow} & Q_{\downarrow\uparrow} \\ Q_{\uparrow\downarrow} & Q_{\uparrow\uparrow} \end{pmatrix},$$

and where the entries are defined by:

$$\begin{aligned} Q_{\downarrow\downarrow} &= R_\downarrow^{-1}(T_{\downarrow\downarrow} - T_{\downarrow 0}T_{00}^{-1}T_{0\downarrow}), \\ Q_{\downarrow\uparrow} &= R_\downarrow^{-1}(T_{\downarrow\uparrow} - T_{\downarrow 0}T_{00}^{-1}T_{0\uparrow}), \\ Q_{\uparrow\downarrow} &= R_\uparrow^{-1}(T_{\uparrow\downarrow} - T_{\uparrow 0}T_{00}^{-1}T_{0\downarrow}), \\ Q_{\uparrow\uparrow} &= R_\uparrow^{-1}(T_{\uparrow\uparrow} - T_{\uparrow 0}T_{00}^{-1}T_{0\uparrow}). \end{aligned}$$

Definition 1 With $\Psi_{i,j}(x)$ we define the joint distribution for M_+, the maximum level in a busy period, given that a busy period starts in state $\varphi(0) = i, (i \in \mathcal{S}_\uparrow)$ at level $X(0)=0$ and finishes in state $\varphi(\tau_0) = j, (j \in \mathcal{S}_\downarrow)$:

$$\begin{aligned} \Psi_{i,j}(x) &:= \mathbb{P}\{\varphi(\tau_0) = j, M_+ \le x \mid \varphi(0) = i, X(0) = 0\}, \quad i \in \mathcal{S}_\uparrow, j \in \mathcal{S}_\downarrow \quad (8) \\ \tau_0 &:= \inf\{t > 0 : X(t) = 0\}, \\ M_+ &:= M^*(\tau_0). \end{aligned}$$

The joint distribution of the maximum in a busy period $\Psi_{i,j}(x)$ is calculated by solving a matrix differential Riccati equation [22]. Function $\Psi_{i,j}(x)$ can be expressed in terms of the matrix exponential form of matrix Q:

$$e^{Qx} = \exp\left[\begin{pmatrix} Q_{\downarrow\downarrow} & Q_{\downarrow\uparrow} \\ Q_{\uparrow\downarrow} & Q_{\uparrow\uparrow} \end{pmatrix}\right] = \begin{pmatrix} A(x) & B(x) \\ C(x) & D(x) \end{pmatrix}. \quad (9)$$

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

$$\Psi(x) = C(x)A(x)^{-1}. \quad (10)$$

In general we are interested in the distribution of the busy cycle $\beta(x)$ which we describe in Definition 3 below. First we introduce some further notation.

Definition 2 Matrix U is the transition matrix from an empty system to the start of a new busy cycle and is defined by:

$$U_{i,j} := \mathbb{P}\{\varphi(\tau_{S_\uparrow}) = j \mid \varphi(0) = i,\ X(0) = 0\}, \qquad i \in \mathcal{S}_\downarrow, j \in \mathcal{S}_\uparrow,$$
$$\tau_{S_\uparrow} := \inf\{t > 0 : \varphi(t) \in \mathcal{S}_\uparrow\}.$$

Definition 3 We define $\beta_{i,j}(x)$ as the joint distribution for the maximum level M_+ in a busy cycle, given that the busy cycle starts in state $\varphi(0) = i, (i \in \mathcal{S}_\uparrow)$ at level $X(0)$=0 and finishes in state $\varphi(\tau_{0\uparrow}) = j, (j \in \mathcal{S}_\uparrow)$:

$$\beta_{i,j}(x) := \mathbb{P}\big\{\varphi\big(\tau_{0\uparrow}\big) = j, M_+ \le x \mid \varphi(0) = i, X(0) = 0\big\}, \quad i \in \mathcal{S}_\uparrow, j \in \mathcal{S}_\uparrow, \tag{11}$$

$$\tau_{0\uparrow} := \inf\{t > \tau_0 : \varphi(t) \in \mathcal{S}_\uparrow\},$$
$$\tau_0 := \inf\{t > 0 : X(t) = 0\}.$$

Observation 1 The function $\beta(x)$ can be written as $\beta(x) = \Psi(x)U$ where $\Psi(x)$ is the joint stationary distribution of the maximum level in a busy period from Definition 1. Matrix U is the transition matrix from start of an idle period to start of a busy period from Definition 2 and is given by (see for example [19, Example 1.4.4]):

$$U = -\begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0}, \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \begin{pmatrix} T_{\downarrow\uparrow} \\ T_{0\uparrow} \end{pmatrix}. \tag{12}$$

We show that the expression for the distribution of the maximum in a busy cycle has an exponential tail. Moreover we can derive an explicit expression for the asymptotic tail. In the expression for $\Psi(x)$ from (10) function $C(x)$ is a $n_\uparrow \times n_\downarrow$ matrix and $A(x)$ is a $n_\downarrow \times n_\downarrow$ matrix. For the case $n_\downarrow > 1$ we have to take the inverse of a matrix that contains exponential terms with exponents corresponding to the eigenvalues of Q. However using Sylvester's formula [10, Page 87] the matrix exponential e^{Qt} can be decomposed as:

$$e^{Qx} = e^{\lambda_1 x}\widetilde{Q}_1 + \ldots + e^{\lambda_{n_\downarrow + n_\uparrow} x}\widetilde{Q}_{n_\downarrow + n_\uparrow}, \tag{13}$$

where the eigenvalues $\lambda_1, \ldots, \lambda_{n_\downarrow + n_\uparrow}$ of Q are the solution of

$$\det[Q - \lambda I] = 0, \tag{14}$$

and the matrices $\widetilde{Q}_i, \quad i = 1, \ldots, n_\downarrow + n_\uparrow$ are the Frobenius covariants. Let ϕ_i^l and ϕ_i^r be the normalised left and right eigenvector corresponding to eigenvalue λ_i:

$$\phi_i^l Q = \lambda_i \phi_i^l \text{ and} \tag{15}$$
$$Q\phi_i^r = \lambda_i \phi_i^r \tag{16}$$

respectively. The corresponding Frobenius covariants are given by $\widetilde{Q}_i = \phi_i^r \phi_i^l$. These describe how the exponentials $e^{\lambda_i x}$ with corresponding eigenvalues λ_i contribute to the matrix exponential e^{Qx}. If we consider the partitioning of e^{Qx} in Equation (9) then $C(x)$ and $A(x)$ can be represented as:

$$A(x) = e^{\lambda_1 x}\widetilde{A}_1 + \ldots + e^{\lambda_{n_\downarrow+n_\uparrow} x}\widetilde{A}_{n_\downarrow+n_\uparrow} \quad \text{and} \tag{17}$$

$$C(x) = e^{\lambda_1 x}\widetilde{C}_1 + \ldots + e^{\lambda_{n_\downarrow+n_\uparrow} x}\widetilde{C}_{n_\downarrow+n_\uparrow}. \tag{18}$$

Now we decompose (10) into:

$$\Psi(x) = \frac{C(x)\,\mathrm{adj}\left[A(x)\right]}{\det\left[A(x)\right]}. \tag{19}$$

As $A(x)$ is $n_\downarrow \times n_\downarrow$ both determinant of $A(x)$ and the product $C(x)\,\mathrm{adj}\left[A(x)\right]$ will contain terms that are products of $n_\downarrow$ exponentials. The resulting exponential terms have exponents that are sums of $n_\downarrow$ eigenvalues.

Definition 4 Let c be a vector with n elements. In summations we denote with

$$\sum_{k\in c} := \sum_{\substack{k=c_i,\\ i=1,\ldots,n}}$$

that we iterate k over the elements from vector c.

Lemma 2 *Let A be an $n \times n$ matrix and $m \geq 1$:*

$$A = \sum_{k=1}^{m} b_k A_k,$$

with:

$$A_k = \boldsymbol{r}_k^T \boldsymbol{c}_k = \begin{bmatrix} r_{k,1}c_{k,1} & \cdots & r_{k,1}c_{k,n} \\ \vdots & \ddots & \vdots \\ r_{k,n}c_{k,1} & \cdots & r_{k,n}c_{k,n} \end{bmatrix}.$$

Then the following holds:

$$\mathrm{adj}\,[A] = \sum_{c\in\mathcal{C}} \left(\prod_{k\in c} b_k\right) \mathrm{adj}\left[\sum_{k\in c} A_k\right],$$

where $\mathcal{C}$ is the set with all combinations of length $n-1$ from the set $\{1, 2, \ldots, m\}$.

Proof The proof can be found in Appendix A where the preliminaries can be found in Appendix A.1 and the actual proof can be found in Appendix A.2.

Lemma 3 *Let*

$$A(x) = e^{\lambda_1 x}\widetilde{A}_1 + \ldots + e^{\lambda_{n_\downarrow+n_\uparrow} x}\widetilde{A}_{n_\downarrow+n_\uparrow} \quad \textit{and}$$
$$C(x) = e^{\lambda_1 x}\widetilde{C}_1 + \ldots + e^{\lambda_{n_\downarrow+n_\uparrow} x}\widetilde{C}_{n_\downarrow+n_\uparrow},$$

with:

$$\widetilde{Q}_k = \phi_k^r \phi_k^l = \begin{bmatrix} \phi_{k,1}^r \phi_{k,1}^l & \cdots & \phi_{k,1}^r \phi_{k,n_\downarrow+n_\uparrow}^l \\ \vdots & \ddots & \vdots \\ \phi_{k,n_\downarrow+n_\uparrow}^r \phi_{k,1}^l & \cdots & \phi_{k,n_\downarrow+n_\uparrow}^r \phi_{k,n_\downarrow+n_\uparrow}^l \end{bmatrix} = \begin{bmatrix} \widetilde{A}_k & \widetilde{B}_k \\ \widetilde{C}_k & \widetilde{D}_k \end{bmatrix},$$

where $\widetilde{A}_k$ is $n_\downarrow \times n_\downarrow$, $\widetilde{B}_k$ is $n_\downarrow \times n_\uparrow$, $\widetilde{C}_k$ is $n_\uparrow \times n_\downarrow$ and $\widetilde{D}_k$ is $n_\uparrow \times n_\uparrow$. Furthermore, let $\mathcal{C}$ be the set of combinations of length n from the set $\{1, \ldots, n\}$. Then the following holds:

$$C(x)\,\mathrm{adj}\,[A(x)] = \sum_{c\in\mathcal{C}} \left(\prod_{k\in c} e^{\lambda_k x}\right) \sum_{k\in c} \widetilde{C}_k\,\mathrm{adj}\left[\sum_{k\in c} \widetilde{A}_k\right]. \tag{20}$$

Proof By applying Lemma 2 we obtain:

$$C(x)\,\mathrm{adj}\,[A(x)] = \sum_{j=1}^{n_\downarrow+n_\uparrow} e^{\lambda_j x}\widetilde{C}_j \sum_{c\in\overline{\mathcal{C}}} \left(\prod_{k\in c} e^{\lambda_k x}\right) \mathrm{adj}\left[\sum_{k\in c} \widetilde{A}_k\right],$$

with $\overline{\mathcal{C}}$ the set of combinations of $n_\downarrow - 1$ elements from the set $\{1, \ldots, n_\downarrow + n_\uparrow\}$. From (9) and Observation 1 we find that both $\widetilde{A}_k$ and $\widetilde{C}_k$ share the same row vector. As all sums of $n_\downarrow - 1$ matrices $\widetilde{A}_k$ have rank $n_\downarrow - 1$ the following is true:

$$\sum_{k\in c} \widetilde{C}_k\,\mathrm{adj}\left[\sum_{k\in c} \widetilde{A}_k\right] = 0, \quad \forall c \in \overline{\mathcal{C}}. \tag{21}$$

Using (21) we can rewrite:

$$\begin{aligned} C(x)\,\mathrm{adj}\,[A(x)] &= \sum_{j=1}^{n_\downarrow+n_\uparrow} e^{\lambda_j x}\widetilde{C}_j \sum_{c\in\overline{\mathcal{C}}} \left(\prod_{k\in c} e^{\lambda_k x}\right) \mathrm{adj}\left[\sum_{k\in c} \widetilde{A}_k\right] \\ &= \sum_{c\in\overline{\mathcal{C}}} \sum_{j\notin c} e^{\lambda_j x}\widetilde{C}_j \left(\prod_{k\in c} e^{\lambda_k x}\right) \mathrm{adj}\left[\sum_{k\in c} \widetilde{A}_k\right] \\ &= \sum_{c\in\mathcal{C}} \sum_{j\in c} e^{\lambda_j x}\widetilde{C}_j \left(\prod_{k\in c\setminus j} e^{\lambda_k x}\right) \mathrm{adj}\left[\sum_{k\in c\setminus j} \widetilde{A}_k\right] \\ &= \sum_{c\in\mathcal{C}} \left(\prod_{k\in c} e^{\lambda_k x}\right) \sum_{j\in c} \widetilde{C}_j\,\mathrm{adj}\left[\sum_{k\in c\setminus j} \widetilde{A}_k\right]. \end{aligned}$$

By using again (21) we obtain:

$$\sum_{c\in\mathcal{C}}\left(\prod_{k\in c}e^{\lambda_k x}\right)\sum_{j\in c}\widetilde{C}_j\,\mathrm{adj}\left[\sum_{k\in c\setminus j}\widetilde{A}_k\right]$$
$$=\sum_{c\in\mathcal{C}}\left(\prod_{k\in c}e^{\lambda_k x}\right)\sum_{k\in c}\widetilde{C}_k\,\mathrm{adj}\left[\sum_{k\in c}\widetilde{A}_k\right].$$

□

Observation 2 There are $m=\binom{n_\downarrow+n_\uparrow}{n_\downarrow}$ unique combinations of $n_\downarrow$ eigenvalues from $n_\downarrow+n_\uparrow$ eigenvalues.
Let $c\in\mathcal{C}$ be the set of combinations of $n_\downarrow$ indices from the set $\{1,2,\ldots,n_\downarrow+n_\uparrow\}$. We define the sums of eigenvalues $\lambda_{k_1},\ldots,\lambda_{k_{n_\downarrow}}$ corresponding to combination c_k with index k by:

$$\widehat{\lambda}_k:=\sum_{j\in c_k}\lambda_j,\quad k=1,\ldots,m,\quad c_k\in\mathcal{C},$$

where the set $\mathcal{C}$ is ordered in decreasing order according to the real parts of $\hat{\lambda}_k$ such that:

$$\mathrm{Re}\left(\widehat{\lambda}_1\right)\geq\mathrm{Re}\left(\widehat{\lambda}_2\right)\geq\ldots\geq\mathrm{Re}\left(\widehat{\lambda}_m\right).$$

Lemma 4 *Equation (19) can be rewritten as:*

$$\Psi(x)=\frac{\widehat{C}_1e^{\widehat{\lambda}_1}+\widehat{C}_2e^{\widehat{\lambda}_2}+\ldots+\widehat{C}_me^{\widehat{\lambda}_m}}{\widehat{A}_1e^{\widehat{\lambda}_1}+\widehat{A}_2e^{\widehat{\lambda}_2}+\ldots+\widehat{A}_me^{\widehat{\lambda}_m}},\tag{22}$$

with values $\hat{\lambda}_k$ as defined in Observation 2, and

$$\widehat{C}_k:=\sum_{j\in c_k}\widetilde{C}_j\,\mathrm{adj}\Big[\sum_{j\in c_k}\widetilde{A}_j\Big],\quad c_k\in\mathcal{C},$$

and

$$\widehat{A}_k:=\det\Big[\sum_{j\in c_k}\widetilde{A}_j\Big],\quad c_k\in\mathcal{C},$$

where the elements c_k from set $\mathcal{C}$ are ordered according to Observation 2 such that:

$$\mathrm{Re}\left(\widehat{\lambda}_1\right)\geq\mathrm{Re}\left(\widehat{\lambda}_2\right)\geq\ldots\geq\mathrm{Re}\left(\widehat{\lambda}_m\right).$$

Proof Due to the determinant and adjoint matrix in Equation (19), there will be exponential terms in both numerator and denominator that result from products of $n_\downarrow$ exponentials $e^{\lambda_i x}$ with eigenvalues $\lambda_i,\ i\in\mathcal{S}$. First consider the terms in the denominator. Remember that the Frobenius covariants $\widetilde{Q}_i$ (and also $\widetilde{A}_i$, $\widetilde{C}_i$) have rank 1. Therefore only linear combinations of $n_\downarrow$ distinct

Frobenius covariants, defined by $c_k \in \mathcal{C}$, will result in positive determinants. Combination $c_k \in \mathcal{C}$ is element of the set containing al combinations of length $n_\downarrow$ from the set $\{1, \ldots, n_\downarrow + n_\uparrow\}$ as defined in Observation 2. Considering the numerator, the adjoint matrix of a linear combination of Frobenius covariants $\widetilde{A}_i$ will only have positive entries when it is a linear combination of $n_\downarrow - 1$ distinct Frobenius covariants as the adjoint matrix contains minors of degree $n_\downarrow - 1$. By applying Lemma 3 we observe that only remaining exponential terms in the numerator are those that correspond to sums over combinations $c_k \in \mathcal{C}$ of $n_\downarrow$ eigenvalues. □

As $\widehat{\lambda}_k$ is ordered in decreasing order the leading exponential term is $e^{\widehat{\lambda}_1}$. Considering (19) the limiting distribution Ψ^∞ becomes:

$$\Psi^\infty := \lim_{x\to\infty} \Psi(x) = \frac{\widehat{C}_1}{\widehat{A}_1}. \tag{23}$$

Using this we can derive the tail behaviour of $\Psi(x)$:

Lemma 5 *$\Psi(x)$ has an exponential tail that behaves as*

$$\Psi^\infty - \Psi(x) \to Ge^{-\kappa x}, \qquad x \to \infty, \tag{24}$$

where

$$\kappa = \lambda_{n_\uparrow}, \quad G = \frac{\widehat{C}_1\widehat{A}_2 - \widehat{A}_1\widehat{C}_2}{\widehat{A}_1^2}$$

and κ is the maximal (least) negative eigenvalue of Q.

Proof Subtracting Ψ^∞ from the expression of $\Psi(x)$ in Lemma 4 gives:

$$\begin{aligned}\Psi^\infty - \Psi(x) &= \frac{\widehat{C}_1}{\widehat{A}_1} - \frac{\widehat{C}_1 e^{\widehat{\lambda}_1} + \widehat{C}_2 e^{\widehat{\lambda}_2} + \ldots + \widehat{C}_m e^{\widehat{\lambda}_m}}{\widehat{A}_1 e^{\widehat{\lambda}_1} + \widehat{A}_2 e^{\widehat{\lambda}_2} + \ldots + \widehat{A}_m e^{\widehat{\lambda}_m}}, \\ &= \frac{[\widehat{C}_1\widehat{A}_2 - \widehat{A}_1\widehat{C}_2]e^{\widehat{\lambda}_2} + \ldots + [\widehat{C}_1\widehat{A}_m - \widehat{A}_1\widehat{C}_m]e^{\widehat{\lambda}_m}}{\widehat{A}_1[\widehat{A}_1 e^{\widehat{\lambda}_1} + \widehat{A}_2 e^{\widehat{\lambda}_2} + \ldots + \widehat{A}_m e^{\widehat{\lambda}_m}]}.\end{aligned}$$

When $x \to \infty$ the two leading exponential terms $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ remain:

$$\Psi^\infty - \Psi(x) \to \frac{\widehat{C}_1\widehat{A}_2 - \widehat{A}_1\widehat{C}_2}{\widehat{A}_1^2} e^{\widehat{\lambda}_2 - \widehat{\lambda}_1}, \qquad x \to \infty. \tag{25}$$

According to Kulkarni [16, Theorem 11.5] the eigenvalues of Q, resulting from $\det[R - \lambda T] = 0$ can be ordered as follows:

$$Re(\lambda_1) \le Re(\lambda_2) \le \ldots \le Re(\lambda_{n_\uparrow}) < 0 < Re(\lambda_{n_\uparrow+2}) \le \cdots \le Re(\lambda_{n_\uparrow+n_\downarrow}).$$

there are $n_\uparrow$ eigenvalues with negative real part, one eigenvalue is equal to zero and there are $n_\downarrow - 1$ eigenvalues with positive real part. In Definition 2 we defined $\widehat{\lambda}_k$ as the sum of $n_\downarrow$ unique eigenvalues. Consider $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$:

$$\widehat{\lambda}_1 = 0 \quad + \lambda_{n_\uparrow+2} + \ldots + \lambda_{n_\uparrow+n_\downarrow},$$
$$\widehat{\lambda}_2 = \lambda_{n_\uparrow} + \lambda_{n_\uparrow+2} + \ldots + \lambda_{n_\uparrow+n_\downarrow}.$$

Observe that $\widehat{\lambda}_1$ consists of $n_\downarrow - 1$ eigenvalues with positive real part and one eigenvalue equal to zero. The next $\widehat{\lambda}_2$ is obtained by replacing the eigenvalues equal to zero with the eigenvalue with least negative real part $\lambda_{n_\uparrow}$. Therefore

$$\widehat{\lambda}_2 - \widehat{\lambda}_1 = \max_{i\in\{i:\lambda_i<0\}} \lambda_i = \lambda_{n_\uparrow}.$$

Plugging this in (25) gives:

$$\Psi^\infty - \Psi(x) \to Ge^{\kappa x}, \qquad x \to \infty,$$

with

$$G := \frac{\widehat{C}_1\widehat{A}_2 - \widehat{A}_1\widehat{C}_2}{\widehat{A}_1^2} \text{ and}$$
$$\kappa := \max_{i\in\{i:\lambda_i<0\}} \lambda_i = \lambda_{n_\uparrow}.$$

□

From Lemma 5 we established that $\Psi(x)$ has an exponential tail $Ge^{-\kappa x}$. Here G is a matrix while we are interested in the general case averaging over all transitions. Therefore we define the following transition matrices:

Definition 5

$$P_{BI} := \Psi^\infty = \frac{\widehat{C}_1}{\widehat{A}_1}, \tag{26}$$

$$P_{IB} := U = -\begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \begin{pmatrix} T_{\downarrow\uparrow} \\ T_{0\uparrow} \end{pmatrix}, \tag{27}$$

$$P_{BB} := P_{BI}P_{IB} = -\begin{pmatrix} \Psi^\infty & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \begin{pmatrix} T_{\downarrow\uparrow} \\ T_{0\uparrow} \end{pmatrix}, \tag{28}$$

where P_{BI} is the transition matrix from a busy to an idle period, P_{IB} is the transition matrix from an idle to a busy period and P_{BB} is the transition matrix between states that initiate busy cycles. In P_{BI}, Ψ^∞ is transition matrix from a state that initiates a busy period to the state that terminates the busy period. Recall that U is the transition matrix from Definition 2 for transitions from idle period states to busy period initiating states.

We use transition matrix P_{BB} for calculating the stationary distribution π_B over states ($i \in \mathcal{S}_\uparrow$) that initiate a busy period. The stationary distribution π_B is the solution of:

$$\begin{aligned} &\pi_B P_{BB} = \pi_B, \\ &\sum \pi_B = 1 \end{aligned} \tag{29}$$

Corollary 1 *The overall expected tail of the distribution on the maximum is given by:*

$$\mathbb{P}\{M_+ > z\} \to be^{-\kappa z}, \qquad x \to \infty, \tag{30}$$

where $b = \pi_B \frac{\widehat{C}_1 \widehat{A}_2 - \widehat{A}_1 \widehat{C}_2}{\widehat{A}_1^2} \mathbf{e}$ *and* $\kappa = \lambda_{n_\uparrow}$.

Proof The stationary distribution of states that initiate a busy period is given by π_B. The marginal distribution of the maximum in a busy period is given by $\Psi(x)$ and is conditioned on the states $i \in \mathcal{S}_\uparrow$ that initiate a busy period. The overall distribution of the maximum is given by:

$$\pi_B \Psi(x) \mathbf{e}.$$

We have to add the rows and weight the sums according to the stationary distribution π_B. The same holds for the exponential tail parameter G from Lemma 5:

$$b := \pi_B G \mathbf{e}. \tag{31}$$

We define the maximum of an arbitrary busy cycle by:

$$\mathbb{P}\{M_+ \leq x\}$$

where M_+ represents the stochastic variable corresponding to the maximum of the busy cycle. Similar to Iglehart [13, Lemma 1] we obtain an expression for

$$\mathbb{P}\{M_+ > z\} \to be^{-\kappa z}, \qquad x \to \infty.$$

In our case $b = \pi_B \frac{\widehat{C}_1 \widehat{A}_2 - \widehat{A}_1 \widehat{C}_2}{\widehat{A}_1^2} \mathbf{e}$ and $\kappa = \lambda_{n_\uparrow}$. □

Lemma 6 *Let* $M_+(k)$ *be the maximum of the* k*th busy cycle. Then the following holds:*

$$\lim_{n \to \infty} \mathbb{P}\{\kappa \max_{1 \leq k \leq n} M_+(k) - \log(bn) \leq x\} = \Lambda(x), \tag{32}$$

where

$$\Lambda(x) = \exp[-e^{-x}]. \tag{33}$$

Proof In Corollary 1 we showed that the maximum of a busy cycle has an exponential tail according to:

$$\mathbb{P}\{M_+ > z\} \to be^{-\kappa z}, \qquad x \to \infty.$$

Using the same arguments as in Iglehart [13, Lemma 2] we can derive that

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

The following extreme value theorem argument can be used:

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

□

3.2 Maximum with respect to 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 [17]. In this paper an expression is derived for the joint mean first passage time in a Markov Modulated fluid queue:

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

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

$$f_i(x) := \mathbb{E}[\tau_{\mathcal{S}_\downarrow} \mid X(0) = x, \varphi(0) = i], \qquad i \in \mathcal{S}. \tag{34}$$

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

$$R\frac{\mathrm{d}f(x)}{\mathrm{d}x} + Tf(x) + \mathbf{e} = 0. \tag{35}$$

with boundary condition:

$$f_i(x) = 0, \qquad \forall i \in \mathcal{S}_\downarrow. \tag{36}$$

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

$$\det[R - \lambda T] = 0 \tag{37}$$

and corresponding right eigenvectors ϕ_j^r for which holds:

$$\lambda_i R\phi_j^r = T\phi_j^r. \tag{38}$$

Note that the eigenvalues are equal to the eigenvalues obtained in (14). Recall that the eigenvalues of Q, ordered in increasing order, have the following property:

$$\mathrm{Re}\left(\lambda_1\right) \leq \mathrm{Re}\left(\lambda_2\right) \leq \cdots \leq \mathrm{Re}\left(\lambda_{n_\uparrow}\right) < 0 < \mathrm{Re}\left(\lambda_{n_\uparrow+2}\right) \leq \cdots \leq \mathrm{Re}\left(\lambda_{n_\uparrow+n_\downarrow}\right).$$

In Kulkarni and Tzenova [17, Theorem 4.2] the solution for (35) is given by:

$$f(x) = \sum_{j=1+n_\uparrow}^{n_\downarrow+n_\uparrow} a_j \phi_j^r e^{-\lambda_j x} - \frac{\mathbf{e}x}{d} + g. \tag{39}$$

In this expression g a solution of

$$Tg = -(cR + I)\mathbf{e}. \tag{40}$$

Note that $\mathrm{rank}(T) = n - 1$ therefore we have one free variable in g and fix $g_n = 0$ in order to get a solution to (40). Coefficients a_j are obtained from the solution to:

$$\sum_{j=1+n_\uparrow}^{n_\downarrow+n_\uparrow} a_j \phi_{ij}^r + g_i = 0, \qquad \forall i \in \mathcal{S}_\downarrow, \quad r_i < 0, \tag{41}$$

where ϕ_{ij}^r is the ith entry of eigenvector ϕ_j^r.

Resulting from the equation (38) we obtain eigenvectors that are partitioned into:

$$\phi^r = \begin{pmatrix} \phi_\downarrow^r \\ \phi_0^r \\ \phi_\uparrow^r \end{pmatrix}.$$

For the sake of readability we omit the index j in his expression. There are n_0 states with $r_i = 0$ therefore we write:

$$\lambda \begin{pmatrix} R_\downarrow & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & R_\uparrow \end{pmatrix} \begin{pmatrix} \phi_\downarrow^r \\ \phi_0^r \\ \phi_\uparrow^r \end{pmatrix} = \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} & T_{\downarrow\uparrow} \\ T_{0\downarrow} & T_{00} & T_{0\uparrow} \\ T_{\uparrow\downarrow} & T_{\uparrow 0} & T_{\uparrow\uparrow} \end{pmatrix} \begin{pmatrix} \phi_\downarrow^r \\ \phi_0^r \\ \phi_\uparrow^r \end{pmatrix}$$

and obtain:

$$\phi_0^r = -T_{00}^{-1} T_{0\downarrow} \phi_\downarrow^r - T_{00}^{-1} T_{0\uparrow} \phi_\uparrow^r. \tag{42}$$

Plugging in (42) gives:

$$\lambda \begin{pmatrix} R_\downarrow & 0 \\ 0 & R_\uparrow \end{pmatrix} \begin{pmatrix} \phi_\downarrow^r \\ \phi_\uparrow^r \end{pmatrix} = \begin{pmatrix} T_{\downarrow\downarrow} - T_{\downarrow 0} T_{00}^{-1} T_{0\downarrow} & T_{\downarrow\uparrow} - T_{\downarrow 0} T_{00}^{-1} T_{0\uparrow} \\ T_{\uparrow\downarrow} - T_{\uparrow 0} T_{00}^{-1} T_{0\downarrow} & T_{\uparrow\uparrow} - T_{\uparrow 0} T_{00}^{-1} T_{0\uparrow} \end{pmatrix} \begin{pmatrix} \phi_\downarrow^r \\ \phi_\uparrow^r \end{pmatrix}.$$

The resulting eigenvectors will become:

$$\phi^r = \begin{pmatrix} \phi^r_\downarrow \\ -T_{00}^{-1}\left[T_{0\downarrow}\phi^r_\downarrow + T_{0\uparrow}\phi^r_\uparrow\right] \\ \phi^r_\uparrow \end{pmatrix}. \tag{43}$$

Observe that this is equivalent using the eigenvalues and vectors from matrix Q (see Equations 14-16) and plugging this into (43).

In order to have a valid solution only positive eigenvalues can contribute to (39). Let Φ be the matrix consisting of all right-eigenvectors ordered according to all corresponding eigenvalues with non negative real parts $\text{Re}\left(\lambda_{n_\uparrow+1}\right) = 0 \leq \cdots \leq \text{Re}\left(\lambda_{n_\downarrow+n_\uparrow}\right)$. We now partition matrix Φ into

$$\Phi = \begin{pmatrix} \Phi_\downarrow \\ \Phi_0 \\ \Phi_\uparrow, \end{pmatrix}, \tag{44}$$

where $\Phi_\downarrow$ is $n_\downarrow \times n_\downarrow$, Φ_0 is $n_0 \times n_\downarrow$ and $\Phi_\uparrow$ is $n_\uparrow \times n_\downarrow$.

Definition 6 We define the conditional expected duration of a busy period and idle period by:

$$\mathbb{E}[C_B] := \Big(\mathbb{E}[\tau_{\mathcal{S}_\downarrow} \mid X(0) = 0, \varphi(0) = i],\ i \in \mathcal{S}_\uparrow\Big), \tag{45}$$

$$\tau_{\mathcal{S}_\downarrow} := \inf\{t > 0 : X(t) = 0, \varphi(t) \in \mathcal{S}_\downarrow\},$$

$$\mathbb{E}[C_I] := \Big(\mathbb{E}[\tau_{\mathcal{S}_\uparrow} \mid X(0) = 0, \varphi(0) = i],\ i \in \mathcal{S}_\downarrow\Big), \tag{46}$$

$$\tau_{\mathcal{S}_\uparrow} := \inf\{t > 0 : \varphi(t) \in \mathcal{S}_\uparrow\}.$$

Lemma 7 *The mean duration of a busy period starting in state $i \in \mathcal{S}_\uparrow$ is given by:*

$$\mathbb{E}[C_B] = \Phi_\uparrow \Phi_\downarrow^{-1} g_\downarrow + g_\uparrow \tag{47}$$

where Φ is the block partitioned matrix with right eigen vectors from (44) corresponding to non negative eigenvalues, g is the solution to:

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

with vector g partitioned in

$$g = \begin{pmatrix} g_\downarrow \\ g_0 \\ g_\uparrow \end{pmatrix}.$$

Proof The solution for (35) is given by:

$$f(x) = \sum_{j=1+n_\uparrow}^{n_\downarrow+n_\uparrow} a_j \Phi_j e^{-\lambda_j x} - \frac{\mathbf{e}x}{d} + g. \tag{48}$$

Coefficients a_j are obtained from the solution to:

$$\sum_{j=1+n_\uparrow}^{n_\downarrow+n_\uparrow} a_j \Phi_{ij} + g_i = 0, \qquad \forall i \in \mathcal{S}_\downarrow, \quad r_i < 0, \tag{49}$$

where Φ_{ij} is the ith entry of jth eigenvector Φ_j in eigenvector matrix Φ. We are interested in the mean first passage time for a busy period started at $x = 0$. Therefore we take $f(0)$:

$$f(0) = \sum_{j=1+n_\uparrow}^{n_\downarrow+n_\uparrow} a_j \Phi_j + g_\uparrow.$$

Switching to matrix notation gives:

$$f(0) = \Phi_\uparrow a + g, \tag{50}$$

where

$$\Phi_\downarrow a + g_\downarrow = 0.$$

Matrix $\Phi_\downarrow$ is invertable, therefore we can write:

$$a = -\Phi_\downarrow^{-1} g_\downarrow. \tag{51}$$

Plugging (51) in (50) gives:

$$f(0) = \Phi_\uparrow \Phi_\downarrow^{-1} g_\downarrow + g_\uparrow. \tag{52}$$

□

Definition 7 We define the expected busy cycle time conditioned on starting in a state $i \in \mathcal{S}_\uparrow$ by:

$$\begin{aligned} \mathbb{E}[C_{BB}] &:= \mathbb{E}[\tau_B \mid \varphi(0) = i, X(0) = 0], \qquad i \in \mathcal{S}_\uparrow, \\ \tau_B &:= \inf\{t > \tau_{\mathcal{S}_\downarrow} : \varphi(t) \in \mathcal{S}_\uparrow\}, \\ \tau_{\mathcal{S}_\downarrow} &:= \inf\{t > 0 : X(t) = 0, \varphi(t) \in \mathcal{S}_\downarrow\}. \end{aligned}$$

Lemma 8 *The overall mean expected busy cycle length is given by:*

$$\mathbb{E}[C] = \pi_B \Big[\mathbb{E}[C_B] + P_{BI}\mathbb{E}[C_I]\Big],$$

where

$$\mathbb{E}[C_I] = -\begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \mathbf{e}, \tag{53}$$

resulting in

$$\mathbb{E}[C] = \pi_B \left[\Phi_{\uparrow\downarrow}\Phi_{\downarrow\downarrow}^{-1} g_\downarrow + g_\uparrow - \begin{pmatrix} \Psi^\infty & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \mathbf{e}\right],$$

with $\Psi^\infty = \frac{\widehat{C}_1}{\widehat{A}_1}$ *as defined in* (23).

Proof In Lemma 7 we obtained an expression for the mean busy period. For the idle period using standard first passage time calculations for a CTMC we obtain:

$$\mathbb{E}[C_I] = -\begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1}.$$

Using $\mathbb{E}[C_B]$ and $\mathbb{E}[C_B]$ the expected cycle time can be obtained:

$$\mathbb{E}[C_{BB}] = \mathbb{E}[C_B] + P_{BI}\mathbb{E}[C_I].$$

When a busy period is initiated for a given initiating state $i \in \mathcal{S}_\uparrow$ the expected passage time is given by $\mathbb{E}[C_B]$. Remember that in Definition 5, Equation (26) we defined $P_{BI} = \Psi^\infty$. From this the expected idle time after a busy period that has been initiated by state $i \in \mathcal{S}_\uparrow$ is obtained:

$$\begin{aligned} &\mathbb{E}[\tau_{\mathcal{S}_\uparrow} - \tau_0 \mid X(0) = 0, \varphi(0) = i] = P_{BI}\mathbb{E}[C_I], \qquad i \in \mathcal{S}_\uparrow, \qquad (54) \\ &\tau_{\mathcal{S}_\uparrow} := \inf\{t > \tau_0 : \varphi(t) \in \mathcal{S}_\uparrow\}, \\ &\tau_0 := \inf\{t > 0 : X(t) = 0\}. \end{aligned}$$

This corresponds to taking the expectation over $\mathbb{E}[C_I]$ with respect to the transition matrix P_{BI}. Combining (53) and (54) gives the expected cycle time given the busy cycle started in state $i \in \mathcal{S}_\uparrow$:

$$\mathbb{E}[C_{BB}] = \mathbb{E}[C_B] + P_{BI}\mathbb{E}[C_I], \qquad i \in \mathcal{S}_\uparrow.$$

In (29) we defined the distribution π_B of states that initiate a busy period. The mean cycle time becomes:

$$\begin{aligned} \mathbb{E}[C] &= \pi_B \Big[\mathbb{E}[C_B] + P_{BI}\mathbb{E}[C_I]\Big] \\ &= \pi_B \left[\Phi_{\uparrow\downarrow}\Phi_{\downarrow\downarrow}^{-1} g_\downarrow + g_\uparrow - \begin{pmatrix} \Psi^\infty & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \mathbf{e}\right]. \end{aligned}$$

□

Theorem 1 *Let* $M^*(t) := \sup_{0 \le s \le t} \{M(s)\}$. *The limiting distribution of* $M^*(t)$ *is given by:*

$$\lim_{t\to\infty} \mathbb{P}\{\kappa M^*(t) - \log(\frac{b}{\mathbb{E}[C]}t) \le x\} = \Lambda(x). \qquad (55)$$

where:

$$b = \pi_B \frac{\widehat{C}_1\widehat{A}_2 - \widehat{A}_1\widehat{C}_2}{\widehat{A}_1^2}\mathbf{e},$$

$$\mathbb{E}[C] = \pi_B \left[\Phi_{\uparrow\downarrow}\Phi_{\downarrow\downarrow}^{-1} g_\downarrow + g_\uparrow - \begin{pmatrix} \Psi^\infty & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \mathbf{e}\right]$$

and

$$\kappa = \lambda_{n_\uparrow}.$$

Proof The proof is similar to that of Iglehart [13, Theorem 3]. In Lemma 6 we showed that:

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

Define $\{c(t) : t \ge 0\}$ as the renewal process associated with the length of busy cycles. Then $M^*(t)$ satisfies:

$$\lim_{t\to\infty} \mathbb{P}\{\max_{0\le k\le c(t)} M_+(k) \le x\} \le M^*(t) \le \lim_{t\to\infty} \mathbb{P}\{\max_{0\le k\le c(t)+1} M_+(k) \le x\}. \tag{56}$$

From Lemma 8 we know that:

$$\mathbb{E}[C] = \pi_B \left[\Phi_{\uparrow\downarrow}\Phi_{\downarrow\downarrow}^{-1} g_\downarrow + g_\uparrow - \begin{pmatrix} \Psi^\infty & 0 \end{pmatrix} \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1} \mathbf{e} \right].$$

Considering $c(t)$ the following weak law of large numbers can be derived:

$$\frac{c(t)}{t} \to \frac{1}{\mathbb{E}[C]},\ t \to \infty. \tag{57}$$

Using Berman [4, Theorem 3.2] and Lemma 6 the limiting distribution becomes:

$$\lim_{t\to\infty} \mathbb{P}\{\kappa M^*(t) - \log(\frac{b}{\mathbb{E}[C]}t) \le x\} = \Lambda(x). \tag{58}$$

The term $\frac{1}{\mathbb{E}[C]}$ from (57) represents the expected number of busy cycles per time unit and corresponds to the c in Berman [4, Theorem 3.2]. □

From Theorem 1 the expression for the asymptotic distribution for the maximum of fluid queue

$$\mathbb{P}\{M^*(t) > b_{min}\} < p_{empty} \tag{59}$$

can now be used to approximate the tail probabilities:

$$\mathbb{P}\{\kappa M^*(t) - \log(\frac{b}{\mathbb{E}[C]}t) > x\} \approx 1 - \Lambda(x), \tag{60}$$

$$\mathbb{P}\{M^*(t) > b_{min}\} \approx 1 - \Lambda(\kappa b_{min} - \log(\frac{b}{\mathbb{E}[C]}t)), \tag{61}$$

whenever we have a sufficiently large b_{min} such that at least $b_{min} > \frac{\log(\frac{b}{\mathbb{E}[C]}t)}{\kappa}$.

Define p_{empty} as the maximal allowed probability that a buffer, with initial contents b_{min}, will become empty during play-out of a video stream of length $t = T_{play}$. Given p_{empty}, that represents the maximum probability a video is disturbed during T_{play}, the initial buffer size b_{min} should be chosen such that:

$$b_{min} > \frac{-\log[-\frac{\mathbb{E}[C]}{bT_{play}}\log(1-p_{empty})]}{\kappa}. \tag{62}$$

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

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

Furthermore $M^*(T_{play})$ represents the limiting distribution on the maximum congestion over time. Then if we consider $M^*(T_{play})$ there should hold that:

$$\mathbb{P}\{M^*(T_{play}) > b_{min}\} < p_{empty}. \tag{63}$$

Using the fact that when $t \to \infty$ the maximum M^* converges to a Gumbel distribution the following asymptotic expectation of the maximum level can be derived:

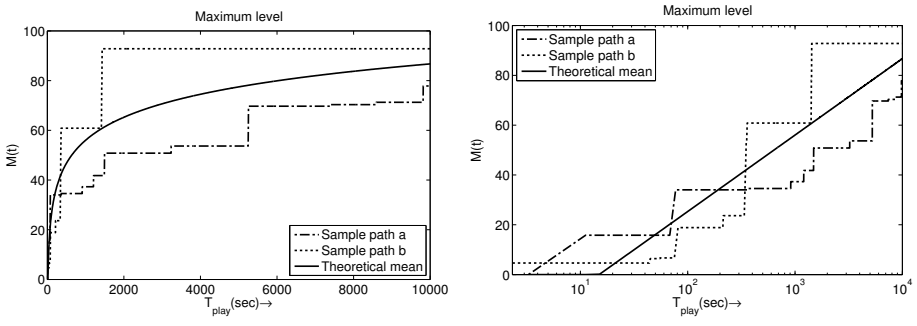
$$\mathbb{E}[M^*(t)] \to \frac{\log\left(\frac{bt}{\mathbb{E}[C]}\right)+\gamma}{\kappa}, \qquad t \to \infty, \tag{64}$$

where $\gamma \approx 0.577215665$ is the Euler-Mascheroni constant. Observe that $\mathbb{E}[M^*(t)]$ grows logarithmically over time with logarithmic slope $\frac{1}{\kappa}$.

4 Numerical examples

In Section 3 we derived that the combined buffer contents, that is congested and in the play-out buffer $X(t) + Y(t) = M^*(t)$, equals the maximum of the congestion process $X(t)$. Moreover the distribution $M^*(t)$ can be approximated by an extreme value distribution for sufficiently large t. From this we derived a mapping from the maximum buffer under-run probability p_{empty} and streaming video duration T_{play} to minimal initial buffer level b_{min}. We will now run simulations in order to evaluate the accuracy of our mapping. Our parameter setting is as follows:

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



(a) Linear time scale. (b) Logarithmic time scale.

Fig. 3: Sample paths of $M^*(t)$ compared to asymptotic mean as expressed in (64). Sample paths a and b correspond to realisations of $X(t)+Y(t)$ from the fluid model simulation.

The simulation consists of 1.000.000 sample paths. Examples of realizations of sample paths are represented in Fig. 3. In these figures we observe that the sample paths follow the asymptotic mean quite well. Fig. 3b is the logarithmic time scale variant of Fig. 3a. On the logarithmic time scale in Fig. 3b the logarithmic growth behavior of the sample paths with respect to time t, determined by parameter κ, can be observed.

In Fig. 4 simulations ran for different values of T_{play} while keeping R_{play} at a fixed level. On the vertical axis the required buffer (in seconds) is matched against the corresponding tail probability p_{empty} on the horizontal axis. The lines indicate for a given T_{play} and tail probability what maximal bit rate is supported such that the QoS requirement is met, as stated in Equation (5). In addition we determine the required buffer using our asymptotic result. The marked and unmarked lines thow the comparison of determining the required buffer using simulation to determining the required buffer using our asymptotic result. Here we observe that for reasonably long T_{play} (minutes) the asymtotic result gives a good handle for determining the required buffer time.

In Figs. 5a-5i the buffer time is set to a fixed level, while the maximum supported video bitrate is determined. In this setting the network parameters remain fixed while the play-out rate R_{play} is varied from 2.1 Mbps to 5.9 Mbps. This range is determined by the fact that for the given parameters a minimal bitrate of 2 Mbps is archieved and the average bitrate is equal to 6Mbps. The maximum supported level determined by simulation is compared to the theoretical maximum supported bitrate. Using (61) for given parameters (including R_{play}) the empty buffer probability p_{empty} can be approximated. Note that κ, b and $\mathbb{E}[C]$ all depend on R_{play}. Finding a supported R_{play} using (61) is done by applying a search method.

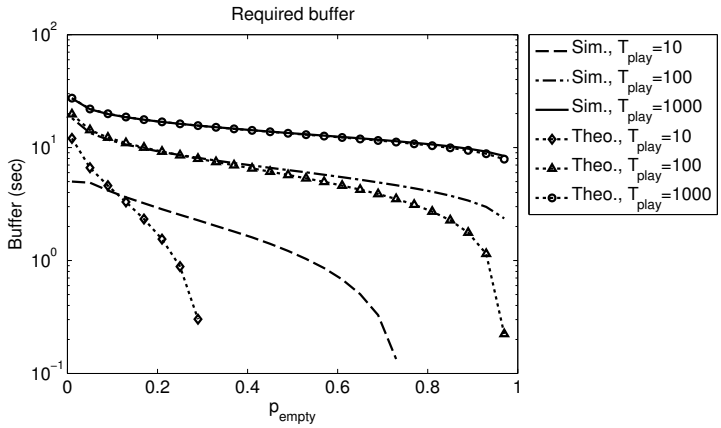


Fig. 4: Required buffer (from simulations) for given T_{play} and p_{empty} compared to theoretical required level given by (62).

5 Discussion and future work

We studied a model where video is streamed over an IP network and buffered before it is rendered at a client. The network is modelled as a Markov Modulated fluid queue where a CTMC determines the actual transmission rate through the network. For the playback buffer an initial buffer level b_{min} has to be chosen such that the probability that the video will stop before the end has been reached will not exceed the agreed *service level*.

In our exposition, we assumed that the video application was streamed at constant bit rate. For practical application, however, it is more realistic to assume that the video produces variable bit rate flows. Our model still applies to this case, if we take the transport unit to be *time* rather than *bits* or *packets*. The streaming and play-out rate are then $R_{play} = 1$ (one unit of time is played each unit of time). To incorporate the variable bit rate into our model, we modify the network throughput process $\varphi(t)$ as follows. We construct it from two independent components $\varphi(t) = (\varphi^1(t), \varphi^2(t))$. The first component is a CTMC and again determines the network capacity at time t in *bits per time unit*, say speed s_i^1 if $\varphi^1(t) = i$. The second component $\varphi_2(t)$ is also a CTMC, independent of $\varphi_1(t)$, and determines the *length of time encoded per bit* for the video segments transported through the network at time t, say s_j^2 if $\varphi^2(t) = j$. Setting the network speeds as $s_{i,j} = s_i^1 s_j^2$ whenever $\varphi(t) = (i, j)$, our original model can be directly used. Of course, exploiting the structure of the process $\varphi(t)$ (its generator, for example, can be written as the Kronecker product of the generators of φ^1 and φ^2) was not part of the scope of our analysis here. Incorporating this structure may further enhance efficient computations.

We have shown that the probability of this event corresponds to the event where maximum congestion level $M(t) := \sup_{0 \le s \le t} X(s)$ exceeds the initial buffer level b_{min}. Moreover we derived that the asymptotic distribution of the

maximum level $M(t), t \to \infty$ has a Gumbel distribution. For smaller t the expression of the asymptotic distribution can be used to approximate the tail probability $\mathbb{P}\{M(T) > b_{min}\}$. From this expression a formula is derived that maps p_0, t_{play} and the network and video parameters to a minimal buffer level b_{min}. Simulation results indicate that the buffer level that is obtained from the asymptotics is an overestimation of the real necessary buffer level. The longer the video stream the more accurately the asymptotic distribution of the maximum corresponds to the real distribution of the maximum.

The convergence to the extreme value distribution depends on the rate in which transitions (of the CTMC that models throughput) occur. In the examples we observe that for small timescale the model is less accurate. An improvement would be adding an approximation for the behavior on shorter time scale. 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 timescale linear behavior model and the long time scale extreme value model to become more accurate.

For practical purposes it may be difficult to estimate the transition probabilities of the modulating process $\varphi(t)$. In principle, this can be done using the classical maximum likelihood estimators as described for example in [19, Section 1.10]. For the choice of the state space it is natural to let the state of the modulating process coincide with the measured network rate; the granularity then determines the dimension of the transition matrix. In practice, one may however not want to go into estimation of the network characteristics, but rather try to adapt the coefficients κ and $\mathbb{E}[C]/b$ in the dimensioning rule formulated in relation 62. Through live measurements, one may decide on adapting the estimates for these coefficients so as to improve quality when the stall probability is too large, or reduce the initial delay, when the buffer is never close to empty.

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.

A Proof of Lemma 2

A.1 Preliminaries

Definition 8 Let A be a $n \times m$ matrix:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{pmatrix}.$$

Then any order–p minor of A will be denoted as:

$$A\begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ k_1 & k_2 & \cdots & k_p \end{pmatrix} := \det\left[\begin{pmatrix} a_{i_1,k_1} & a_{i_1,k_2} & \cdots & a_{i_1,k_p} \\ a_{i_2,k_1} & a_{i_2,k_2} & \cdots & a_{i_2,k_p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_p,k_p} & a_{i_p,k_2} & \cdots & a_{i_p,k_p} \end{pmatrix}\right],$$

provided that

$$1 \leq i_1 < i_2 < \cdots < i_p \leq m,$$
$$1 \leq k_1 < k_2 < \cdots < k_p \leq n,$$
$$p \leq m, n.$$

The Binet-Cauchy formula on minors [12](Page 12):
Let A be an $m \times n$ matrix, B be a $n \times q$ matrix and C be an $m \times q$ matrix and $C = AB$. Then any minor of C of order p is the sum of the products of all possible minors of A with order p and corresponding minors of the same order of B:

$$C\begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix} = \sum_{1 \leq k_1 < k_2 < \cdots < k_m \leq n} A\begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ k_1 & k_2 & \cdots & k_p \end{pmatrix} B\begin{pmatrix} k_1 & k_2 & \cdots & k_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix}.$$

Lemma 9 *Let A be a $n \times n$ matrix:*

$$A = \sum_{k=1}^{m} A_k$$

with:

$$A_k = \begin{bmatrix} a_{k,1,1} & \cdots & a_{k,1,n} \\ \vdots & \ddots & \vdots \\ a_{k,n,1} & \cdots & a_{k,n,n} \end{bmatrix}.$$

Define $\mathbf{A}$ as a $n \times mn$ matrix with:

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_k \end{bmatrix},$$

and $\mathbf{I}$ is a $mn \times n$ matrix (consisting of m $n \times n$ identity matrices I_n) defined by:

$$\mathbf{I} = \begin{bmatrix} I_n & I_n & \cdots & I_n \end{bmatrix}^T.$$

Let $\mathcal{V}$ be the set of subsets with exactly $n-1$ elements from the set $\{1, 2, \ldots, mn\}$ which is defined by:

$$\mathcal{V} = \{(k_1, k_2, \ldots, k_{n-1}) : 1 \leq k_1 < k_2 < \cdots < k_{n-1} \leq mn\}.$$

Then the following holds:

$$\mathrm{adj}(A) = \sum_{v \in \mathcal{V}} \mathrm{adj}\left(F_C(\mathbf{A}, v) F_R(\mathbf{I}, v)\right),$$

with operators:

$$F_C(\mathbf{A}, v) = \begin{bmatrix} \mathbf{a}_{1,v_1} & \cdots & \mathbf{a}_{1,v_{n-1}} \\ \vdots & \ddots & \vdots \\ \mathbf{a}_{n,v_1} & \cdots & \mathbf{a}_{n,v_{n-1}} \end{bmatrix},$$

$$F_R(\mathbf{I}, v) = \begin{bmatrix} \mathbf{i}_{v_1,1} & \cdots & \mathbf{i}_{v_1,n} \\ \vdots & \ddots & \vdots \\ \mathbf{i}_{v_{n-1},1} & \cdots & \mathbf{i}_{v_{n-1},n} \end{bmatrix},$$

Operator $F_C(\mathbf{A}, v)$ selects the columns from $\mathbf{A}$ according to vector v, while operator $F_R(\mathbf{I}, v)$ selects rows from $\mathbf{I}$ according to vector v.

Proof We write $\sum_{k=1}^{m} A_k = \mathbf{AI} = \begin{bmatrix} A_1 & A_2 & \cdots & A_k \end{bmatrix} \begin{bmatrix} I_n & I_n & \cdots & I_n \end{bmatrix}^T$. Using the Binet-Cauchy formula on minors, this can be rewritten to:

$$\begin{aligned} \operatorname{adj}[A] = \operatorname{adj}[\mathbf{AI}] &= \begin{pmatrix} \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{1,1}(v) & \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{1,2}(v) & \cdots & \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{1,n}(v) \\ \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{2,1}(v) & \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{2,2}(v) & \cdots & \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{2,n}(v) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{n,1}(v) & \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{n,2}(v) & \cdots & \sum_{v\in\mathcal{V}} \bar{\mathbf{a}}_{n,n}(v) \end{pmatrix} \\ &= \sum_{v\in\mathcal{V}} \begin{pmatrix} \bar{\mathbf{a}}_{1,1}(v) & \bar{\mathbf{a}}_{1,2}(v) & \cdots & \bar{\mathbf{a}}_{1,n}(v) \\ \bar{\mathbf{a}}_{2,1}(v) & \bar{\mathbf{a}}_{2,2}(v) & \cdots & \bar{\mathbf{a}}_{2,n}(v) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{a}}_{n,1}(v) & \bar{\mathbf{a}}_{n,2}(v) & \cdots & \bar{\mathbf{a}}_{n,n}(v) \end{pmatrix} \\ &= \sum_{v\in\mathcal{V}} \operatorname{adj}\Big(F_C(\mathbf{A}, v) F_R(\mathbf{I}, v)\Big), \end{aligned}$$

with:

$$\bar{\mathbf{a}}_{i,j}(v) = \mathbf{A} \begin{pmatrix} 1 & \cdots & i-1 & i+1 & \cdots & n \\ v_1 & \cdots & v_{i-1} & v_i & \cdots & v_{n-1} \end{pmatrix} \mathbf{I} \begin{pmatrix} v_1 & \cdots & v_{j-1} & v_j & \cdots & v_{n-1} \\ 1 & \cdots & j-1 & j+1 & \cdots & n. \end{pmatrix}.$$

□

A.2 Proof of Lemma 2

Proof Let $\mathcal{V}$ be the set of subsets with exactly $n-1$ elements from the set $\{1, 2, \ldots, mn\}$ which is defined by:

$$\mathcal{V} = \{(k_1, k_2, \ldots, k_{n-1}) : 1 \le k_1 < k_2 < \cdots < k_{n-1} \le mn\}.$$

We define $\mathcal{P}$ as the set containing all k-permutations of $n-1$ elements from the set $\{1, \ldots, n\}$. Furthermore we define $\mathcal{C}$ as the set with all combinations of $n-1$ elements from the set $\{1, \ldots, m\}$. For each combination $c \in \mathcal{C}$ we define:

$$\mathbf{A}_c = \begin{bmatrix} A_{c_1} & A_{c_2} & \cdots & A_{c_{n-1}} \end{bmatrix},$$

$$\mathbf{I}_c = \begin{bmatrix} I_n & I_n & \cdots & I_n \end{bmatrix}^T,$$

and thus:

$$\mathbf{A}_c \mathbf{I}_c = \sum_{k\in c} A_k.$$

Next we apply Lemma 9:

$$\operatorname{adj}[A] = \sum_{v\in\mathcal{V}} \left(\prod_{k\in v} b_{\lceil k/n \rceil} \right) \operatorname{adj}\Big(F_C(\mathbf{A}, v) F_R(\mathbf{I}, v)\Big),$$

with

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_k \end{bmatrix},$$

and

$$\mathbf{I} = \begin{bmatrix} I_n & I_n & \cdots & I_n \end{bmatrix}^T.$$

Because all matrices A_k have rank 1 the only adjugates that remain are those where there are $n-1$ columns, at $n-1$ different positions, from $n-1$ different A_k matrices. All other combinations of columns result in a matrix with rank $< n-1$ for which the minors of order $n-1$ are zero. Thus the only elements from $\mathcal{V}$ that contribute are those that correspond to any k-permutation of $n-1$ columns from the set $\{1, \ldots n\}$ where each column is selected from a distinct matrix A_k, $k = 1, \ldots, m$. Note that each selected column remains exactly on its originating column position in the A_k matrix. As the only combinations consisting of $n-1$ columns at unique positions from $n-1$ unique matrices contribute to non-zero minors it holds that:

$$\begin{aligned} &\sum_{v \in \mathcal{V}} \left(\prod_{k \in v} b_{\lceil k/n \rceil} \right) \text{adj} \left(F_C(\mathbf{A}, v) F_R(\mathbf{I}, v) \right) \\ =&\sum_{c \in \mathcal{C}} \sum_{p \in \mathcal{P}} \left(\prod_{k \in c} b_k \right) \text{adj} \left(F_C(\mathbf{A}_c, v_p) F_R(\mathbf{I}_p, v_p) \right), \end{aligned}$$

where v_p is the vector that selects the p_ith column from matrix A_{c_i}:

$$(v_p)_i := n(i-1) + p_i, \qquad i \in \{1, \ldots, n-1\},\ p \in \mathcal{P}.$$

We now define $\mathcal{V}_c$ as be the set of subsets with exactly $n-1$ elements from the set $\{1, 2, \ldots, n(n-1)\}$ which is defined by:

$$\mathcal{V}_c = \{(k_1, k_2, \ldots, k_{n-1}) : 1 \leq k_1 < k_2 < \cdots < k_{n-1} \leq n(n-1)\}.$$

For each combination $c \in \mathcal{C}$ we can do the opposite: add again the terms (corresponding to zero valued minors) from the set $\mathcal{V}_c$ corresponding to columns of $\mathbf{A}_c = \begin{bmatrix} A_{c_1} & \cdots & A_{c_{n-1}} \end{bmatrix}$:

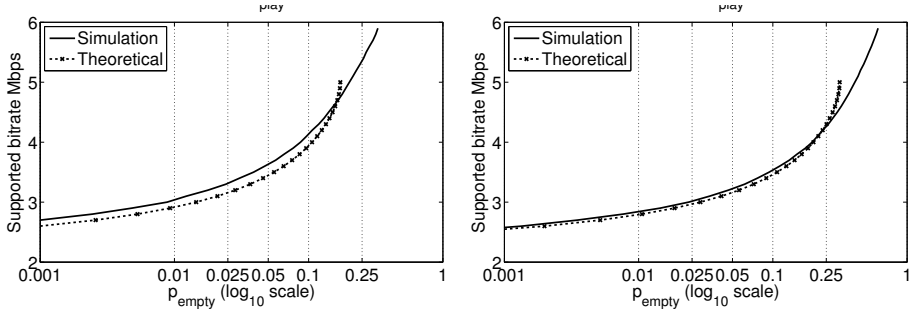
$$\begin{aligned} &\sum_{c \in \mathcal{C}} \sum_{p \in \mathcal{P}} \left(\prod_{k \in c} b_k \right) \text{adj} \left(F_C(\mathbf{A}_c, v_p) F_R(\mathbf{I}_p, v_p) \right), \\ =&\sum_{c \in \mathcal{C}} \left(\prod_{k \in c} b_k \right) \sum_{v \in \mathcal{V}_c} \text{adj} \left(F_C(\mathbf{A}_c, v) F_R(\mathbf{I}_c, v) \right), \\ =&\sum_{c \in \mathcal{C}} \left(\prod_{k \in c} b_k \right) \text{adj} \left[\sum_{k \in c} A_k \right]. \end{aligned}$$

□

References

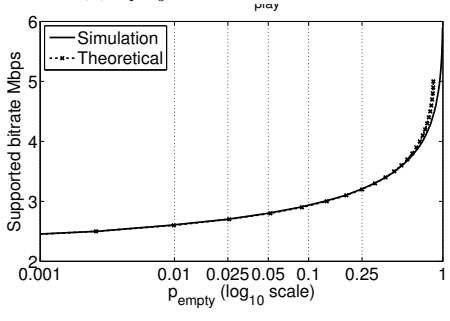
1. Asmussen, S.: Busy period analysis, rare events and transient behavior in fluid flow models. Journal of Applied Mathematics and Stochastic Analysis **7**(3), 269–299 (1994)
2. Asmussen, S.: Extreme value theory for queues via cycle maxima. Extremes **1**(2), 137–168 (1998)
3. Asmussen, S., Bladt, M.: A sample path approach to mean busy periods for markov-modulated queues and fluids. Advances in applied probability pp. 1117–1121 (1994)

4. Berman, S.: Limiting distribution of the maximum term in sequences of dependent random variables. The Annals of mathematical statistics **33**(3), 894–908 (1962)
5. Blέfari-Melazzi, N., Eramo, V., Listanti, M.: Dimensioning of play-out buffers for real-time services in a b-isdn. Computer Communications **21**(11), 980 – 995 (1998). DOI http://dx.doi.org/10.1016/S0140-3664(98)00171-6. URL http://www.sciencedirect.com/science/article/pii/S0140366498001716
6. Bosman, J., van der Mei, R., Nunez-Queija, R.: A fluid model analysis of streaming media in the presence of time-varying bandwidth. In: 24th International Teletraffic Congress (ITC 2012). Krakow, Poland (2012)
7. Boxma, O., Dumas, V.: The busy period in the fluid queue **26**(1) (1998)
8. Cisco: Cisco Visual Networking Index: Forecast and Methodology, 20122017. Cisco white paper, Cisco (2013)
9. Cisco: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 20122017. Cisco white paper, Cisco (2013)
10. Claerbout, J.: Fundamentals of geophysical data processing (1985)
11. Dua, A., Bambos, N.: Buffer management for wireless media streaming. In: Global Telecommunications Conference, 2007. GLOBECOM '07. IEEE, pp. 5226–5230 (2007). DOI 10.1109/GLOCOM.2007.991
12. Gantmacher, F.: Matrix Theory vol. 1. AMS Chelsea Publishing (2000)
13. Iglehart, D.: Extreme values in the gi/g/1 queue. The Annals of Mathematical Statistics pp. 627–635 (1972)
14. Kim, T., Avadhanam, N., Subramanian, S.: Dimensioning receiver buffer requirement for unidirectional vbr video streaming over tcp. In: Image Processing, 2006 IEEE International Conference on, pp. 3061–3064 (2006). DOI 10.1109/ICIP.2006.313086
15. Kontovassilis, K., Tsiligaridis, J., Stassinopoulos, G.: Buffer dimensioning for delay- and loss-sensitive traffic. Computer Communications **18**(5), 315 – 328 (1995). DOI http://dx.doi.org/10.1016/0140-3664(95)96833-C. URL http://www.sciencedirect.com/science/article/pii/014036649596833C
16. Kulkarni, V.: Fluid models for single buffer systems. Frontiers in Queueing: Models and Applications in Science and Engineering pp. 321–338 (1997)
17. Kulkarni, V., Tzenova, E.: Mean first passage times in fluid queues. Operations Research Letters **30**(5), 308–318 (2002)
18. Molazem Tabrizi, F., Peters, J., Hefeeda, M.: Dynamic control of receiver buffers in mobile video streaming systems. Mobile Computing, IEEE Transactions on **12**(5), 995–1008 (2013). DOI 10.1109/TMC.2012.56
19. Norris, J.R.: Markov chains. 2008. Cambridge university press (1998)
20. Scheinhardt, W.: Markov-modulated and feedback fluid queues. Ph.D. thesis, Faculty of Mathematical Sciences, University of Twente, Enschede, The Netherlands, 1998, http://www. ub. utwente. nl/webdocs/tw/1/t0000008. pdf (1998)
21. Scheinhardt, W., Zwart, B.: A tandem fluid queue with gradual input. Probability in the Engineering and Informational Sciences **16**(1), 29–45 (2002)
22. Sericola, B., Remiche, M.: Maximum level and hitting probabilities in stochastic fluid flows using matrix differential riccati equations. Methodology and Computing in Applied Probability **13**(2), 307–328 (2011)
23. Wu, C., Chen, K., Huang, C., Lei, C.: An empirical evaluation of VoIP playout buffer dimensioning in Skype. In: Proceedings of ACM NOSSDAV 2009 (2009)
24. Zhang, L., Fu, H.: Dynamic bandwidth allocation and buffer dimensioning for supporting video-on-demand services in virtual private networks. Computer Communications **23**(1415), 1410 – 1424 (2000). DOI http://dx.doi.org/10.1016/S0140-3664(00)00186-9. URL http://www.sciencedirect.com/science/article/pii/S0140366400001869

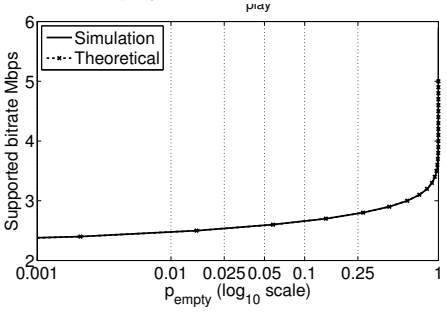


(a) $T_{play} = \mathbf{60}s$, buffer= $\mathbf{10}s$.

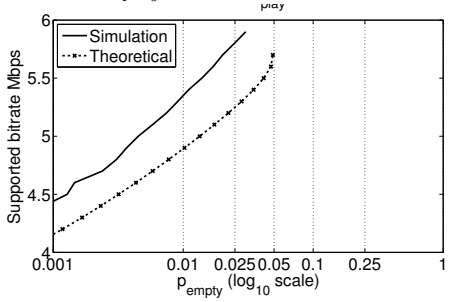
(b) $T_{play} = \mathbf{120}s$, buffer= $\mathbf{10}s$.



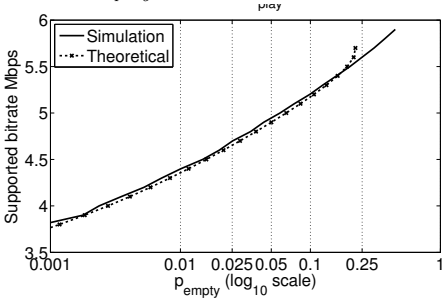
(c) $T_{play} = \mathbf{600}s$, buffer= $\mathbf{10}s$.



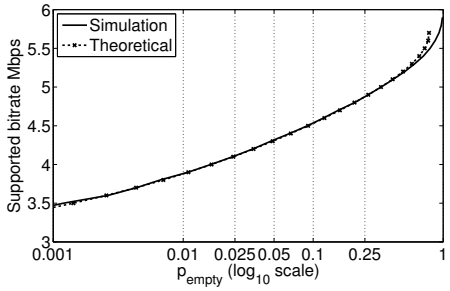
(d) $T_{play} = \mathbf{3600}s$, buffer= $\mathbf{10}s$.



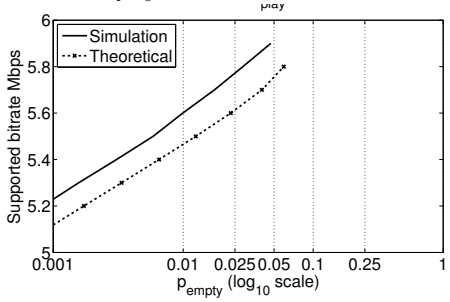
(e) $T_{play} = \mathbf{120}s$, buffer= $\mathbf{30}s$.



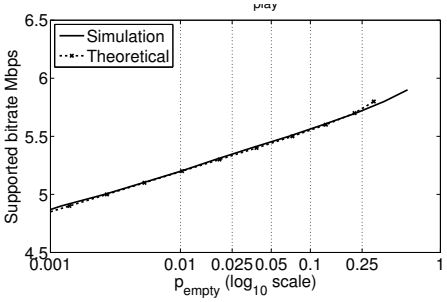
(f) $T_{play} = \mathbf{600}s$, buffer= $30s$.



(g) $T_{play} = \mathbf{3600}s$, buffer= $\mathbf{30}s$.



(h) $T_{play} = \mathbf{600}s$, buffer= $\mathbf{60}s$.



(i) $T_{play} = \mathbf{3600}s$, buffer= $\mathbf{60}s$.

Fig. 5: Supported bitrate for given T_{play} and initial buffer level (in seconds) with respect to p_{empty}.