Optimal Quality of Service Control in Communication Systems

Bosman, J.W.

2014

document version
Publisher's PDF, also known as Version of record

A Fluid Model Analysis of Streaming Media in the Presence of Time-Varying Bandwidth

2

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 2.1).

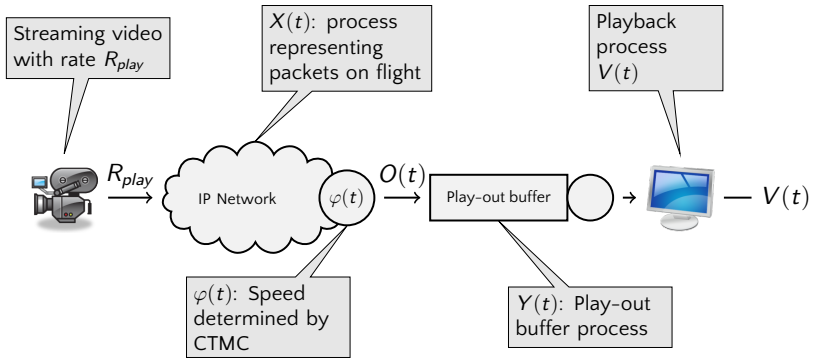

Figure 2.1: *Streaming video through an IP-network.*

Figure 2.1 clarifies the connection of the model described in Figure 2.2 with the application to streaming over an unreliable medium. 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, providing a given probabilistic guarantee on undisturbed playback. Our objective in this chapter 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

[2] This chapter is based on [22].

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.

The precise object of study in this chapter is a fluid model for constant bit-rate streaming media applications in the presence of bandwidth that varies over time (see Remark 2.6.1 for use of the model for variable bit rate applications). Motivated from Figure 2.1, 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.

2.1 Background

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 Bléfari-Melazzi et al. [15], and Kontovassilis et al. [71] 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 in Wu et al. [108]. 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 [108]. 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 this chapter. A third such example is in the context of closed-loop control for wireless streaming: Dua and Ambos [41], 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. The setting studied in Kim et al. [70] is closest in nature to that in this chapter. 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 dy-

namic deterministic optimization, e.g. Tabrizi et al. [84], and Zhang et al. [115]. 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 [37]. 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 [36] 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. Particularly 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 Figure 2.2: 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 [98], and Kulkarni [74] for excellent overviews of results on fluid queues that are directly relevant to our analysis. Asmussen and Bladt [6] 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 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 [23] 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 [99] study a two-node tandem with gradual input, and compute the steady-state joint buffer-content distribution using martingale methods. Kulkarni and Tzenova [75] 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 [101] 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 [4] investigates a more general setting than the one considered in this chapter, which focuses on the streaming video setting. In our work we use an alternative matrix-theoretic analysis technique and obtain more explicit dimensioning rules than can be derived directly from specializing [4] to our model.

We derive a dimensioning rule for the play-out buffer, based on an extreme value distribution approximation. Our analysis is strongly motivated by the classical papers of Berman [13] and Iglehart [65]. Berman [13] studies the limiting distribution of the maximum in sequences of random variables satisfying certain dependence conditions. Iglehart [65] 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 [5] for an excellent survey on extreme-value theory for queues. This chapter 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.

Our analysis proceeds as follows: We use results from [101] 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 [75] to obtain the mean expected cycle time. Next we apply an approach similar to [65] in order to show that the maximum buffer level converges to a Gumbel extreme value distribution. From this result the correct initial play-out buffer level can be estimated. As mentioned previously, our work 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 chapter we independently establish this result in a more intuitive manner. 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.

2.2 Model

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 $\{s_i,\ i = 1, 2, \dots, n\}$. 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 $\mathcal{S} = \{1, \dots, n\}$. 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 $n_\downarrow + n_0 + n_\uparrow = n$:

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

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:

$$T = \begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0} & T_{\downarrow\uparrow} \\ T_{0\downarrow} & T_{00} & T_{0\uparrow} \\ T_{\uparrow\downarrow} & T_{\uparrow 0} & T_{\uparrow\uparrow} \end{pmatrix}. \tag{2.1}$$

The combination of network congestion and play-out buffering is represented by a tandem of two fluid queues. See Figure 2.2 for an illustration of our model.

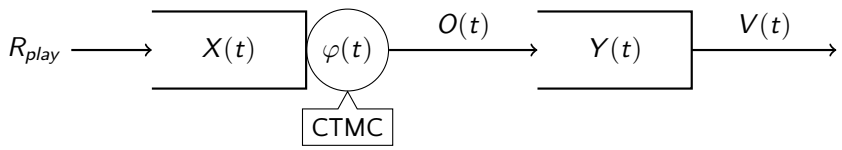

Figure 2.2: *Tandem of fluid buffers representing streaming video through an IP-network.*

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 we define rates of change (of the first buffer contents) by $r_i := R_{play} - s_i$ ($i = 1, \ldots, n$), when $\varphi(t) = i$. Conversely 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 $V(t) = R_{play}$. For the second fluid buffer the rate of change is directly proportional whenever $Y(t) > 0$, so that $V(t)$ can sustain play-out at rate 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.

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 has the same block partitioning as generator matrix T, i.e. R is a $(n_\downarrow + n_0 + n_\uparrow) \times (n_\downarrow + n_0 + n_\uparrow)$ matrix:

$$R := \begin{pmatrix} R_\downarrow & 0 & 0 \\ 0 & R_0 & 0 \\ 0 & 0 & R_\uparrow \end{pmatrix}. \tag{2.2}$$

The entries are defined by:

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

Here diag (r_i) $i \in S$ is the diagonal matrix with on the diagonal all elements of set S. In order for the first buffer to be stable the average potential throughput S_{res} must satisfy:

$$S_{res} := \sum_{i=1}^{n} s_i \pi_i > R_{play}. \tag{2.3}$$

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

$$d := \sum_{i=1}^{n} r_i \pi_i = R_{play} - S_{res}. \tag{2.4}$$

Stability condition (2.3) 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 bit rate R_{play} cannot be guaranteed at all times 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 client side. 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 quickly alternating between buffering and play-out. This is commonly experienced as being very disturbing. 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_{init} such that the probability of disturbed play-out during T_{play} is smaller than p_{empty}:

$$\mathbb{P}\{\exists s \in [0, T_{play}] : V(s) < R_{play} \mid X(0) = 0, Y(0) = b_{init}\} < p_{empty}. \tag{2.5}$$

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. Second a large buffering delay causes a too large lag before the event is displayed on screen. Therefore, we want the play-out buffer to have a minimal size.

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 $1 - p_{empty}$. In order to minimize the initial buffer level b_{init} 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_{init}.

2.3 Analysis

We are interested in a mapping from network, video characteristics and distortion probability p_{empty} to a minimal buffer level b_{init} such that Equation (2.5) 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 Figure 2.3. 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 $s_i < R_{play}$ ($r_i < 0$), 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.

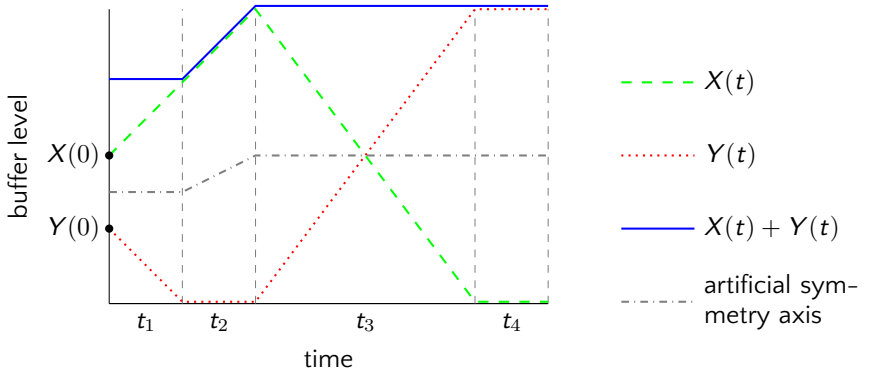

Figure 2.3: *Different phases of the stochastic processes $X(t)$ and $Y(t)$.*

Observe in Figure 2.3 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 Figure 2.3) 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.

2.3.1. Definition. We define the maximum level process as

$$M^*(t) := \sup_{0 \le s \le t} X(s). \tag{2.6}$$

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

$$X(t) + Y(t) = \sup_{0 \le s \le t} X(s) = M^*(t). \tag{2.7}$$

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 (2.5) we use an initial buffer level of $Y(0) = b_{init}$, while in Lemma 2.3.2 we assume $Y(0) = 0$. However, setting $Y(0) = b_{init}$ and $X(0) = 0$ corresponds to the case where $X(0)$ has a virtual (initial) supremum equal to b_{init}. Thus we are interested in the probability that new supremum $M^*(t) > b_{init}$ is attained in time interval $[0, t]$ given that the initial supremum level is set to $M^*(0) = b_{init}$. Using the connection of the initial buffer level b_{init} to the supremum level $M^*(t)$ and Lemma 2.3.2 we can rewrite Equation (2.5) to:

$$\mathbb{P}\{M^*(t) > b_{init}\} < p_{empty}. \tag{2.8}$$

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

Lemma 2.3.2 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.

2.3.1 Maximum over busy cycles

In Sericola and Remiche [101] 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. Recall that the state space $\mathcal{S}$ is be partitioned into:

$$\begin{aligned}
&\mathcal{S}_{\downarrow} := \{1, \dots, n_{\downarrow}\},\\
&\mathcal{S}_0 := \{n_{\downarrow}+1, \dots, n_{\downarrow}+n_0\} \quad \text{and}\\
&\mathcal{S}_{\uparrow} := \{n_{\downarrow}+n_0+1, \dots, n_{\downarrow}+n_0+n_{\uparrow}\},
\end{aligned}$$

with corresponding rate matrices

$$\begin{aligned}
R_{\downarrow} &:= \text{diag}\,(r_i), && i \in \mathcal{S}_{\downarrow},\\
R_0 &:= \text{diag}\,(r_i) = 0 \quad \text{and}, && i \in \mathcal{S}_0,\\
R_{\uparrow} &:= \text{diag}\,(r_i), && i \in \mathcal{S}_{\uparrow},
\end{aligned}$$

that contain rates that are negative, zero or positive, respectively. 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 uniformized resulting in modified matrix Q:

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

where the entries are defined by:

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

2.3.3. Definition. With $\Psi_{i,j}(x)$ we define the joint distribution for the maximum level in a busy period M_+, 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 $\varphi(\tau_0) = j, (j \in \mathcal{S}_{\downarrow})$:

$$\begin{aligned}
\Psi_{i,j}(x) &:= \mathbb{P}\big\{\varphi(\tau_0) = j, M_+ \le x \mid \varphi(0) = i, X(0) = 0\big\},\ i \in \mathcal{S}_{\uparrow}, j \in \mathcal{S}_{\downarrow}, \qquad (2.9)\\
\tau_0 &:= \inf\{t > 0 : X(t) = 0\},\\
M_+ &:= M^*(\tau_0).
\end{aligned}$$

The joint distribution of the maximum in a busy period $\Psi_{i,j}(x)$ is calculated by solving a matrix differential equation [101]. Function $\Psi_{i,j}(x)$ can be expressed in terms of the matrix exponential form of matrix Q:

$$e^{Qx} = \exp\left[\begin{pmatrix} Q_{\downarrow\downarrow} & Q_{\downarrow\uparrow} \\ Q_{\uparrow\downarrow} & Q_{\uparrow\uparrow} \end{pmatrix}\right] = \begin{pmatrix} A(x) & B(x) \\ C(x) & D(x) \end{pmatrix}. \tag{2.10}$$

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

$$\Psi(x) = C(x)A(x)^{-1}. \tag{2.11}$$

In general we are interested in the distribution of the busy cycle $\beta(x)$ which we describe in Definition 2.3.5 below. First we introduce some further notation.

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

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

2.3.5. Definition. We define $\beta_{i,j}(x)$ as the joint distribution for M_+, the maximum level in a busy cycle, 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 $\varphi(\tau_{0\uparrow}) = j$ ($j \in \mathcal{S}_\uparrow$):

$$\begin{aligned} \beta_{i,j}(x) &:= \mathbb{P}\big\{\varphi\big(\tau_{0\uparrow}\big) = j, M_+ \le x \mid \varphi(0) = i, X(0) = 0\big\}, i \in \mathcal{S}_\uparrow, j \in \mathcal{S}_\uparrow, \\ \tau_{0\uparrow} &:= \inf\{t > \tau_0 : \varphi(t) \in \mathcal{S}_\uparrow\}, \\ \tau_0 &:= \inf\{t > 0 : X(t) = 0\}. \end{aligned} \tag{2.12}$$

2.3.6. Observation. 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 2.3.3. Matrix U is the transition matrix from start of an idle period to start of a busy period from Definition 2.3.4 and is given by (see for example [86, Example 1.4.4]):

$$U = -\begin{pmatrix} I & 0 \end{pmatrix}\begin{pmatrix} T_{\downarrow\downarrow} & T_{\downarrow 0}, \\ T_{0\downarrow} & T_{00} \end{pmatrix}^{-1}\begin{pmatrix} T_{\downarrow\uparrow} \\ T_{0\uparrow} \end{pmatrix}. \tag{2.13}$$

In this chapter we start with the analysis for the case where $n_\downarrow = n_\uparrow = 1$ and $n_0 = 0$. In Chapter 3 we extend the analysis for the case where $n_\downarrow \ge 1$, $n_\uparrow \ge 1$ and $n_0 \ge 0$, In general, the maxima in consecutive busy periods are not independent,

because the starting states of the environment may induce correlation. However, for the two-state model with $n_\downarrow = n_\uparrow = 1$ we have $U = [1]$. Therefore 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 chapter. Instead, we specialize only for the two-state model and refer to future work for details on extensions to the semi-regenerative case.

For the two-state model with transmission rates $s_1 > R_{play}$ and $s_2 < R_{play}$, we use generator matrix:

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

and rate matrix:

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

to obtain the generator matrix with uniformized fluid rates:

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

The solution the the differential equation 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:

$$\mathbb{P}\{M_+ \leq x\} = \Psi(x), \tag{2.14}$$

where M_+ represents the r 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 $x \to \infty$:

$$1 - \Psi(x) = \frac{r_2\alpha_1 + r_1\alpha_2}{r_2\alpha_1 + r_1\alpha_2 e^{x(\frac{\alpha_1}{r_1} + \frac{\alpha_2}{r_2})}} \sim \left(\frac{r_1\alpha_2 + r_2\alpha_1}{r_1\alpha_2}\right) e^{-x(\frac{\alpha_1}{r_1} + \frac{\alpha_2}{r_2})}. \tag{2.15}$$

Similar to Iglehart [65, Lemma 1] we obtain an expression for

$$\mathbb{P}\{M_+ > x\} \sim be^{-\kappa x},\ x \to \infty. \tag{2.16}$$

In our case, $b = \left(\frac{r_1\alpha_2 + r_2\alpha_1}{r_1\alpha_2}\right)$ 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 [65, Lemma 2] we obtain:

$$\lim_{n\to\infty} \mathbb{P}\{\kappa \max_{1\le k\le n} M_+(k) - \log(bn) \le x\} = \Lambda(x), \tag{2.17}$$

where

$$\Lambda(x) = \exp[-e^{-x}]. \tag{2.18}$$

Here, we use the following extreme value theorem argument:

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

2.3.2 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 [75]. In this chapter an expression is derived for the joint mean first passage time in a Markov Modulated fluid queue:

$$\begin{aligned}
&\mathbb{E}[\tau_{\mathcal{S}_\downarrow} \mid X(0) = x, \varphi(0) = i], && i \in \mathcal{S}, \\
&\tau_{\mathcal{S}_\downarrow} := \inf\{t > 0 : X(t) = 0, \varphi(t) \in \mathcal{S}_\downarrow\}.
\end{aligned} \tag{2.19}$$

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

$$f_i(x) := \mathbb{E}[\tau_{\mathcal{S}_\downarrow} \mid X(0) = x, \varphi(0) = i], \qquad i \in \mathcal{S}. \tag{2.20}$$

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

$$R\frac{\mathrm{d}f(x)}{\mathrm{d}x} + Tf(x) + \overline{e} = 0, \tag{2.21}$$

with boundary condition

$$f_i(x) = 0, \qquad \forall i \in \mathcal{S}_{\downarrow}, \tag{2.22}$$

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

$$\det[R - \lambda T] = 0, \tag{2.23}$$

and the corresponding right eigenvectors ϕ_j^r satisfy:

$$\lambda_i R\phi_j^r = T\phi_j^r. \tag{2.24}$$

The eigenvalues of Q, ordered such that the real parts are in increasing order:

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

There are n solutions to Equation (2.23) of which there are $n_\downarrow - 1$ eigenvalues with positive real part, one eigenvalue has real part equal to 0 and there are $n_\uparrow$ eigenvalues with negative real part. In Kulkarni and Tzenova [75, Theorem 4.2] the solution for (2.21) is given by:

$$f(x) = \sum_{j=n_\uparrow+1}^{n_\uparrow+n_\downarrow} a_j\phi_j^r e^{-\lambda_j x} - \frac{\overline{e}x}{d} + g. \tag{2.25}$$

In this expression g is a solution to

$$Tg = -(cR + I)\overline{e}. \tag{2.26}$$

We are interested in the solution for the two-state model where $n_\downarrow = n_\uparrow = 1$. Plugging in T and R into the results of Kulkarni [74, Example 1] gives:

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

which gives

$$
\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} 0 \\ -\dfrac{r_2 - r_1}{\delta} \end{bmatrix} + \begin{bmatrix} -\dfrac{\alpha_1 + \alpha_2}{\delta} \\ -\dfrac{\alpha_1 + \alpha_2}{\delta} \end{bmatrix} x, \tag{2.27}
$$

with

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

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

$$
\begin{aligned}
\mathbb{E}[C_B] &:= \mathbb{E}[\tau_{\mathcal{S}_\downarrow} \mid X(0) = 0, \varphi(0) = i], && i \in \mathcal{S}_\uparrow, && (2.28) \\
\tau_{\mathcal{S}_\downarrow} &:= \inf\{t > 0 : X(t) = 0, \varphi(t) \in \mathcal{S}_\downarrow\}, \\
\mathbb{E}[C_I] &:= \mathbb{E}[\tau_{\mathcal{S}_\uparrow} \mid X(0) = 0, \varphi(0) = i], && i \in \mathcal{S}_\downarrow, && (2.29) \\
\tau_{\mathcal{S}_\uparrow} &:= \inf\{t > 0 : \varphi(t) \in \mathcal{S}_\uparrow\}.
\end{aligned}
$$

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:

$$
\begin{aligned}
\mathbb{E}[C_B] = f_2(0) = \mathbb{E}[\tau_{\mathcal{S}_\downarrow} \mid X(0) = 0, \varphi(0) = 2] &= -\frac{r_2 - r_1}{r_2 \alpha_1 + r_1 \alpha_2}, && (2.30) \\
\tau_{\mathcal{S}_\downarrow} : &= \inf\{t > 0 : X(t) = 0, \varphi(t) = 1\}.
\end{aligned}
$$

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 $1/\alpha_1$. Therefore:

$$\mathbb{E}[C_I] = \mathbb{E}[\tau_\uparrow \mid X(0) = 0, \varphi(0) = 1] = 1/\alpha_1.$$

By combining the expected busy period with the expected idle period we obtain an expression for the total expected busy cycle:

$$\mathbb{E}[C] = \mathbb{E}[C_B] + \mathbb{E}[C_I] = -\frac{r_2 - r_1}{r_2\alpha_1 + r_1\alpha_2} + \frac{1}{\alpha_1} = \left(\frac{r_1}{\alpha_1}\right) \cdot \frac{\alpha_1 + \alpha_2}{r_2\alpha_1 + r_1\alpha_2}. \tag{2.31}$$

In Equation (2.17) 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:

$$\max_{0 \leq k \leq c(t)} \{M_+(k) \leq x\} \leq M^*(t) \leq \max_{0 \leq k \leq c(t)+1} \{M_+(k) \leq x\}. \tag{2.32}$$

According to the weak law of large numbers we have:

$$\frac{c(t)}{t} \to \frac{1}{\mathbb{E}[C]}, \qquad t \to \infty. \tag{2.33}$$

Using Berman [13, Theorem 3.2] and Equation (2.17) the limiting distribution becomes:

$$\lim_{t\to\infty} \mathbb{P}\{\kappa M^*(t) - \log(bt) \leq x\} = \Lambda^{\frac{1}{\mathbb{E}[C]}}(x). \tag{2.34}$$

In Equation (2.34) the term $\frac{1}{\mathbb{E}[C]}$ from (2.33) represents the expected number of busy cycles per time unit (this corresponds to the c in Berman [13, Theorem 3.2]). The expression for the asymptotic distribution for the maximum of the two-state fluid queue

$$\mathbb{P}\{M^*(t) > b_{init}\} < p_{empty} \tag{2.35}$$

can now be expressed as:

$$\mathbb{P}\{\kappa M^*(t) - \log(bt) > x\} \approx 1 - \Lambda^{\frac{1}{\mathbb{E}[C]}}(x), \tag{2.36}$$

$$\mathbb{P}\{M^*(t) > b_{init}\} \approx 1 - \Lambda^{\frac{1}{\mathbb{E}[C]}}(\kappa b_{init} - \log(bt)), \tag{2.37}$$

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:

$$\mathbb{E}[M^*(t)] \to \frac{\log\left(\frac{bt}{\mathbb{E}[C]}\right) + \gamma}{\kappa}, \qquad t \to \infty, \tag{2.38}$$

where $\gamma \approx 0.577215665$ is the Euler-Mascheroni constant. The behavior with respect to the real process is illustrated in Figure 2.6. Observe that $\mathbb{E}[M^*(t)]$ grows logarithmically over time with logarithmic slope $\frac{1}{\kappa}$.

2.4 Dimensioning the initial buffer size

In Section 2.3 we showed that the probability of an empty play-out 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:

$$b_{init} > \frac{-\log[-\frac{\mathbb{E}[C]}{bT_{play}}\log(1-p_{empty})]}{\kappa}. \tag{2.39}$$

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

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

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).

2.5 Numerical experiments

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

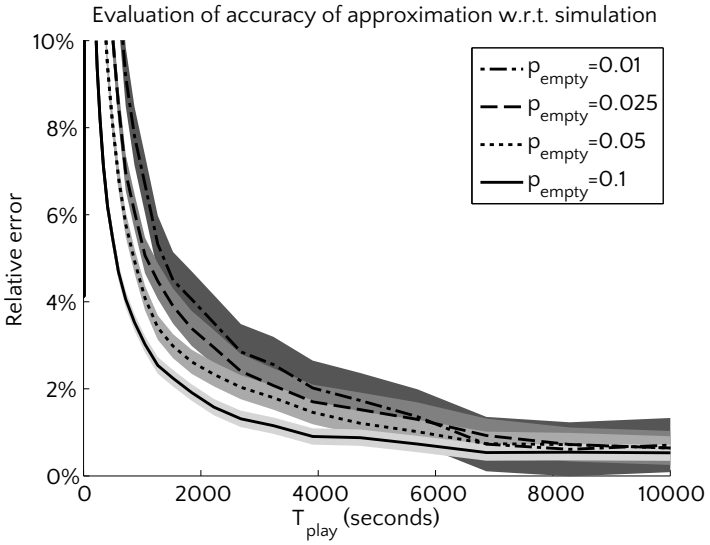

Figure 2.4: *Relative difference of buffer under-run probability to simulation. The gray bands around the lines are the 95% confidence intervals of the simulation.*

simulations in order to evaluate the accuracy of our mapping. Our parameter setting is as follows: $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, $s_1 = 8$*Mbps*, $s_2 = 2$*Mbps*, $R_{play} = 4$*Mbps*, $r_1 = -4$, $r_2 = 2$, $R = \text{diag}\left(\begin{bmatrix} r_1 & r_2 \end{bmatrix}\right)$ and

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

The simulation consists of $10,000,000$ sample paths. Figure 2.4 represents the relative difference between target tail probability p_{empty} and the actual fraction of sample paths that exceed the buffer level approximation. We define the relative difference of approximation (app) and simulation (sim) by:

$$\text{diff}_{relative}(app, sim) = \left| \frac{app - sim}{sim} \right|. \tag{2.40}$$

From Figure 2.4 it can be observed that for the the tail probability $\mathbb{P}\{M(T_{play}) > b_{init}\}$ the error quickly approaches the region below 5%. Figure 2.5 represents the actual fraction of sample paths that exceeds the theoretical asymptotic percentiles. The theoretical percentiles are based on Equation (2.39). The straight thin dashed lines represent the desired tail probability.

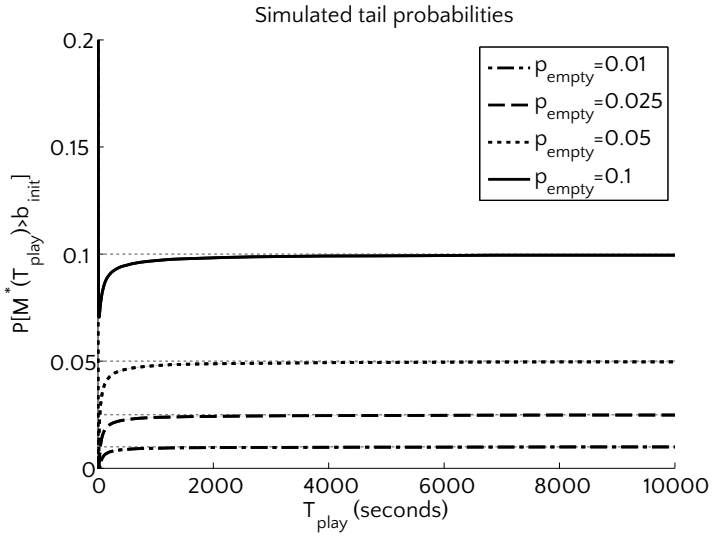

Figure 2.5: *Tail probabilities using empirical distribution based on simulation, evaluated on theoretical asymptotic percentiles.*

The tail probabilities in Figure 2.5 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 2.3.2 we derived the asymptotic mean in Equation (2.38). We compare the asymptotic mean to the simulation results in Figure 2.6. 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 2.6 we observe that this is indeed the case.

In Figure 2.7 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 (2.39). On a linear time scale, simulation and asymptotic percentiles coincide quite closely.

Figure 2.8 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 possible growth rate of $X(t) + Y(t)$. The "notch" corresponds to the upper bound (which is curved due to logarithmic time scale).

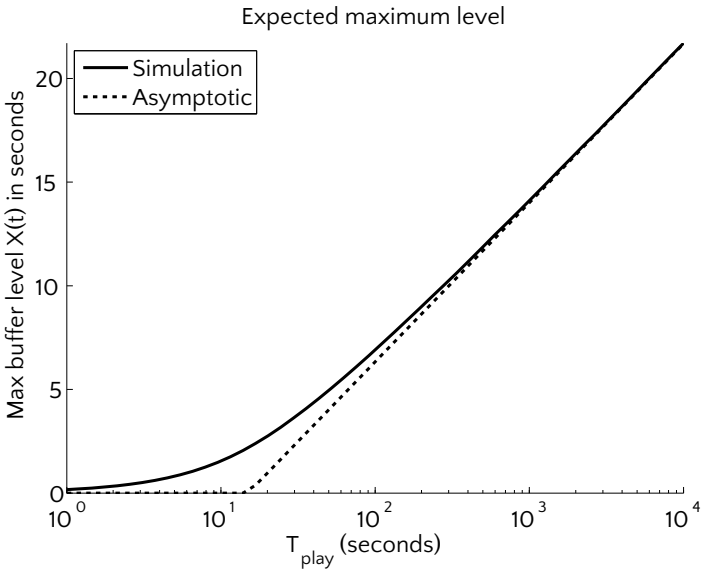

Figure 2.6: *Simulated and theoretical (asymptotic, see Equation (2.38)) expectation of the maximum level $M^*(t)$ on logarithmic time-scale.*

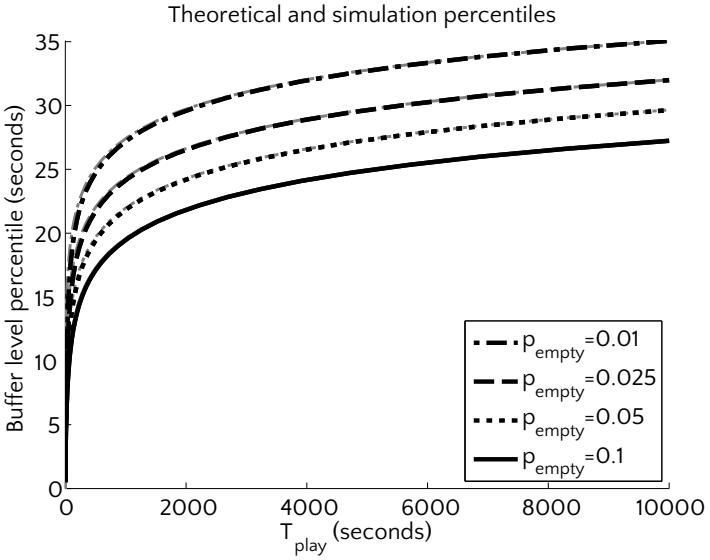

Figure 2.7: *Black lines represent simulation percentiles, gray lines represent theoretical asymptotic percentiles.*

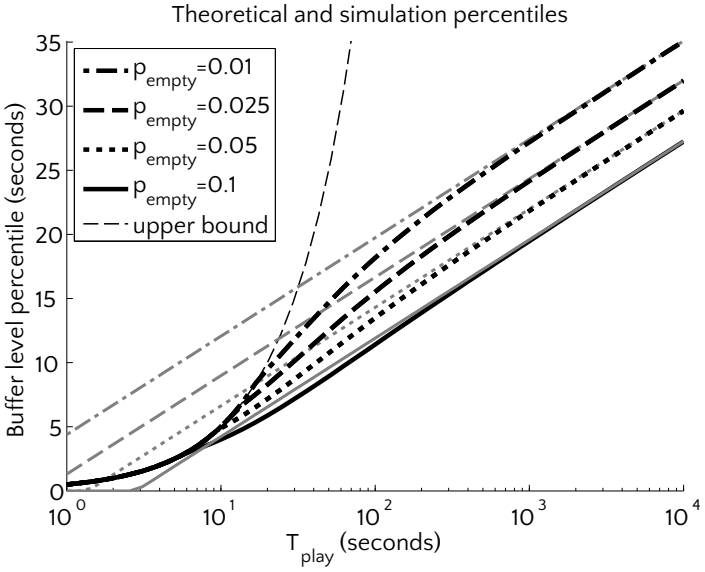

Figure 2.8: *Percentiles on logarithmic time scale. Black lines represent percentiles from simulation, gray lines represent theoretical asymptotic percentiles.*

2.6 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}.

2.6.1. Remark (Variable bitrate)**.** 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 of the maximum congestion level $M(t)$ exceeding the initial buffer level b_{init}. 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 fluctuating available bandwidth. Our analysis facilitates proper parameter selection with respect to the altered network parameters.

2.6.2. Remark (Transition probabilities)**.** 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 [86, 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 2.39. 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.