EXPONENTIAL BOUNDS WITH AN APPLICATION TO CALL ADMISSION

Zhen LIU[1]* Philippe NAIN[1†] and Don TOWSLEY[2‡]

[1]INRIA, B.P. 93, 06902, Sophia Antipolis Cedex, France

[2]Department of Computer Science
University of Massachusetts, Amherst, MA 01003, USA

Revised version — October 4, 1994

Abstract

In this paper we develop a framework for computing upper and lower bounds of an exponential form for a large class of single resource systems with non-renewal inputs. Specifically, the bounds are on quantities such as backlog, queue length, and response time. The exponential upper bound is shown to exist if the process describing the system is stable. The optimum decay rate for the bound is obtained by establishing a large deviation result for this process. The paper concludes with several applications including admission control to multimedia systems such as high speed networks and video servers.

Keywords: Tail distribution; Exponential bound; Large deviation principle; Ergodicity; Markov chain; Matrix analysis; Queues; Markov modulated process; Effective bandwidth; Call admission control.

*Z. Liu was supported in part by the CEC DG XIII under the ESPRIT BRA grant QMIPS.

†P. Nain was supported in part by NSF under grant NCR-9116183 and by the CEC DG XIII under the ESPRIT BRA grant QMIPS. This work was done when this author was visiting the University of Massachusetts in Amherst during the academic year 1993-94.

‡D. Towsley was supported in part by NSF under grant NCR-9116183.

1 Introduction

We are about to witness a phenomenal growth in the introduction and usage of multimedia applications. Numerous teleconferencing applications that run over networks have recently been introduced (e.g., vat [21] and NeVoT [34] for voice, nv [15] for video, and wb [22] and shdr [32] for shared whiteboard). In addition, there are plans to deploy largescale multimedia servers in the not too distant future, [33]. All of these applications have in common the need for a minimal *quality of service (QOS)* guarantee in the form of either an end-to-end delay constraint or a maximum tolerable fraction of loss. Providing QOS guarantees to these applications poses one of the most challenging problems facing designers of multimedia systems and applications.

In this paper we focus on a *single resource* and develop a framework within which to obtain easily computable upper and lower bounds on the tail of the distributions of quantities such as backlog, delay, queue length, etc. at that resource. These bounds are exponential in nature when the combined arrival and service processes (to be made precise) can be described by a Markov chain and the system is stable. In addition to obtaining distributional bounds, we also obtain the effective bandwidths of sessions for a rich and large class of traffic sources and service processes. Last, we apply these results to the problem of call admission in a network and in a multimedia server setting.

More precisely, we consider the behavior of a single server as described by the recursion

$$X_n = [X_{n-1} + U_{n-1}(Y_{n-1})]^+, \qquad \forall n \geq 1 \tag{1.1}$$

with $[a]^+ := \max(0, a)$, where $(U_n(Y_n))_n$ is a Markov modulated process, with $(Y_n)_n$ being a finite-state, irreducible, aperiodic, and homogeneous Markov chain on the set $\mathcal{S} := \{1, 2, \ldots, K\}$, and $(U_n(i))_n$, $1 \leq i \leq K$, being K mutually independent renewal sequences of $(-\infty, \infty)$-valued random variables (r.v.'s), further independent of the Markov chain $(Y_n)_n$. In our context, one application is where X_n represents the waiting time of the n-th customer in a first-in-first-out G/G/1 single server queue, $U_n(k) = S_n(k) - T_n(k)$ for every $n \geq 0$, $k \in \mathcal{S}$ where $(S_n(k))_n$ and $(T_n(k))_n$ are the service requirement and interarrival time sequences in state k. We will assume that X_0 is a nonnegative and almost surely (a.s.) finite r.v.

Our primary objective is to compute exponential upper and lower bounds for the tail distribution of X_n, both for every $n \geq 0$ and for the stationary regime X of X_n (when it exists), namely, to find $a > 0$, $b > 0$, and $\theta > 0$ such that

$$a\, e^{(-\theta x)} \leq P(X_n \geq x), P(X \geq x) \leq b\, e^{(-\theta x)}$$

for all $x > 0$, $n \geq 0$. We also want the coefficients a, b and θ to be easily computable so that on-line computations may be carried out.

In the particular case where $(S_n)_n$ and $(T_n)_n$ are two mutually independent renewal sequences (GI/GI/1 queue), Kingman [26, 27] showed that $P(X \geq x) \leq \exp(-\eta x)$ for all $n \geq 0$ and $x > 0$, where η is the unique solution in $(0, \infty)$ of the equation $E[\exp(\theta U_n(1))] = 1$ under the stability

condition $E[U_n(1)] < 0$. Our results can be considered as an extension of Kingman's result to stochastic recursions of the form (1.1) where $(X_n)_n$ is no longer a Markov chain.

As mentioned before, our work is motivated by the need to characterize the response time distribution and/or backlog distributions in multimedia systems. Many multimedia applications have real-time constraints (e.g., voice, video) for which it is important to characterize the response time distribution at a single resource, whether it is a hop in a network or the I/O system at a server. Although such applications have real-time constraints, they are able to tolerate a small fraction of packets missing their deadlines (approx. 1% for voice). Bounds on the tail distribution of quantities such as buffer occupancy and response times can be used by designers to size systems. Furthermore, bounds can be used to develop policies for controlling the admission of new applications (sessions) to the network.

Previous work in this area falls into three categories. First, a considerable amount of work has focussed on the development of algorithms for computing the response time distribution of a statistical multiplexer being fed by a Markovian Arrival Process (MAP) pioneered by Neuts [31] (see[14] for a recent survey of this area). Unfortunately, these computations are typically very expensive. Consequently, there has been considerable interest in the development of bounds on performance for very general arrival processes. This is exemplified by the works of Cruz [6, 7], Kurose [28], Duffield [9], Chang [3], and Yaron and Sidi [36]. The last two papers are of particular interest as they deal with systems in which response times are finite but unbounded. Because these papers make very few assumptions regarding the arrival processes, the resultant bounds are very loose. We will observe that where there is an overlap between our model and those given in the references above, our bounds are demonstrably tighter and easier to calculate.

The previous work most closely related to ours is that of Duffield [9] which uses a martingale approach (similar to [26] for the G/G/1 queue) to obtain upper bounds similar to ours for the case of a Markovian environment. This approach does not appear easily to yield lower bounds or transient results, nor considers the problem of call admission.

Third, there has been developed a *theory of effective bandwidths* for the purpose of controlling the admission of new sessions to a network. Briefly, it has been noted that it is possible to associate an easily calculated quantity with each session, referred to as the *effective bandwidth*, that captures the behavior of the tail of the response time distribution at a multiplexer. The call admission problem is then solved by checking whether the effective bandwidth of the aggregate user population, including the new user, exceeds the service capacity. The reader is referred to [16, 23, 17, 12, 5] for treatments of this problem. As an application of our model we will obtain results for the source models commonly encountered in the study of network and multimedia server performance.

We apply our results to several systems that have received considerable prior attention. These include: *i)* a discrete-time queue with a fixed rate server, *ii)* a fixed rate server being fed by a Markov modulated Poisson process (MMPP), and *iii)* a fixed rate server being fed by a fluid process. For the first of these models, we present a more easily computable upper bound on buffer occupancy than previously reported in [3, 9]. These bounds can then be used to address the call

admission problem. For all of these cases, we also develop simple expressions for the effective bandwidths of a population of sources feeding the server. In the case of the MMPP, our results generalize those previously obtained in [12, 24].

The organization of the paper is as follows. Upper and lower bounds are derived in Section 2. This section includes a derivation of the largest exponential decay rate and a treatment of both transient and stationary regimes. It concludes with a demonstration of the tightness of the bounds. Applications to the problem of call admission in a multimedia system is found in Section 3. A generalization of the results to the case where the Markovian environment contains a countable number of states is given in section 4.

2 Exponential Bounds

In this section, we derive exponential upper bounds (Section 2.2) and lower bounds (Section 2.4) for the tail distribution of X_n. We establish these results by extending the approach of Kingman [27] to multidimensional case using matrix analysis techniques. In Section 2.3, we provide simpler (but looser) upper bounds based on the main result in Section 2.2. In Section 2.5, we provide upper and lower bounds for the stationary regime. The tightness of the bounds is addressed in Section 2.6. Last, a discussion of the implications of one of our technical assumptions required to determine the largest exponential decay rate is found in Section 2.7. Prior to deriving the bounds, we introduce some notation.

2.1 Notation

Let $P = [p_{ij}]$ be the transition matrix of the Markov chain $(Y_n)_n$ and let $\pi = (\pi(1), \ldots, \pi(K))$ be its invariant measure. Define $\pi_n(i) = P_{\pi_0}(Y_n = i)$ for $i \in \mathcal{S}$, $n \geq 1$, and let $\pi_0 = (\pi_0(1), \ldots, \pi_0(K))$ be the initial probability distribution. In the following, we will drop the subscript π_0 in P_{π_0} except when $\pi_0 = \pi$ to emphasize the fact that the Markov chain $(Y_n)_n$ is stationary.

In order to avoid triviality, we will assume that there exists at least one state $i \in \mathcal{S}$ for which $U_n(i)$ takes strictly positive value, since otherwise X_n decreases to 0 a.s. We also assume that $E[U_n(i)]$ exists for all $i \in \mathcal{S}$ and is finite. Let

$$\mu := \min_{i \in \mathcal{S}} E[U_n(i)]. \tag{2.1}$$

Define $F_i(x) = P(U_n(i) < x)$ and let $\phi_i(\theta) = E[\exp(\theta U_n(i))]$ for $i \in \mathcal{S}$, $\theta \in (-\infty, \infty)$. Let $\theta_+ = \inf\{\theta > 0 : \phi_i(\theta) = \infty \text{ for some } i \in \mathcal{S}\}$ and $\theta_- = \sup\{\theta < 0 : \phi_i(\theta) = \infty \text{ for some } i \in \mathcal{S}\}$. We will assume that $\theta_+ > 0$ and $\theta_- < 0$.

Introduce the sets

$$\begin{aligned} \Theta &= \{\theta_- \leq \theta \leq \theta_+ : \phi_i(\theta) < \infty, \quad \forall i \in \mathcal{S}\} \\ \Theta_+ &= \{\theta \in \Theta : \theta \geq 0\}. \end{aligned}$$

Some of our results require either that Θ be open or that $\theta_+ \notin \Theta$. A discussion of the implications of this assumption is deferred to Section 2.7.

2.2 Exponential Upper Bounds

Let $(\gamma_j^n)_{n,j}$, $\gamma_j^n : [0,\infty) \to [0,\infty)$, be a set of functions such that

$$\sum_{k \in \mathcal{S}} p_{kj}\, \pi_n(k) \left[\int_{-\infty}^{x} \gamma_k^n(x-u)\, dF_k(u) + 1 - F_k(x) \right] \leq \pi_{n+1}(j)\, \gamma_j^{n+1}(x). \tag{2.2}$$

The following result holds:

Proposition 2.1 *Let P_m denote the property that*

$$P(X_m \geq x \,|\, Y_m = j) \leq \gamma_j^m(x) \tag{2.3}$$

for all $x > 0$ and for all $j \in \mathcal{S}$ such that $\pi_m(j) \neq 0$.

If P_0 is true, then P_m is true for all $m \geq 1$.

Proof. We use an induction argument on m. Assume that P_m is true for $m = 0, 1, \ldots, n$ and let us show that P_{n+1} is true.

Assume that $\pi_{n+1}(j) \neq 0$. We have

$$\begin{aligned}
&P(X_{n+1} \geq x \,|\, Y_{n+1} = j) \\
&= \sum_{k \in \mathcal{S}} P(X_n + U_n(k) \geq x \,|\, Y_n = k, Y_{n+1} = j) P(Y_n = k \,|\, Y_{n+1} = j) \\
&= \sum_{k \in \mathcal{S}} p_{kj}\, \frac{\pi_n(k)}{\pi_{n+1}(j)}\, P(X_n + U_n(k) \geq x \,|\, Y_n = k, Y_{n+1} = j) \\
&= \sum_{k \in \mathcal{S}} p_{kj}\, \frac{\pi_n(k)}{\pi_{n+1}(j)}\, P(X_n + U_n(k) \geq x \,|\, Y_n = k) \\
&= \sum_{k \in \mathcal{S}} p_{kj}\, \frac{\pi_n(k)}{\pi_{n+1}(j)} \left[\int_{-\infty}^{x} P(X_n \geq x - u \,|\, Y_n = k)\, dF_k(u) + 1 - F_k(x) \right] \\
&\leq \sum_{k \in \mathcal{S}} p_{kj}\, \frac{\pi_n(k)}{\pi_{n+1}(j)} \left[\int_{-\infty}^{x} \gamma_k^n(x-u)\, dF_k(u) + 1 - F_k(x) \right] \qquad (2.4) \\
&\leq \gamma_j^{n+1}(x)
\end{aligned}$$

where (2.4) follows from the induction hypothesis. This concludes the proof. ♣

From Proposition 2.1 we deduce the following

Corollary 2.1 *If P_0 is true then*

$$P(X_n \geq x) \leq \sum_{j \in \mathcal{S}} \pi_n(j)\, \gamma_j^n(x), \qquad \forall x > 0,\, n \geq 0. \tag{2.5}$$

The next step consists in finding functions $\gamma_j^n(x)$ satisfying (2.2). We first introduce some additional notation and recall some basic results of matrix analysis.

Let $\mathcal{M}_n$ be the set of all n-by-n matrices with real entries. Recall that a matrix/vector is nonnegative (resp. positive) if all its entries are real and nonnegative (resp. strictly positive). For any matrix $A \in \mathcal{M}_n$, we will denote by A^T its transpose matrix and by $r(A)$ its spectral radius. We shall denote by $\tilde{r}(A)$ the largest real eigenvalue of A, if any.

For any matrix $A = [a_{ij}] \in \mathcal{M}_n$ the matrix $A^k = [a_{ij}^{(k)}] \in \mathcal{M}_n$ will denote the k-th power of A, and A^T the transpose of A. For any vector $\mathbf{v} = (v_1, \ldots, v_K)$, $\mathbf{v}^T$ will denote the transpose of $\mathbf{v}$, and $|\mathbf{v}|$ will stand for $\sum_{i=1}^K v_i$.

For $\theta \in \Theta$, define the matrix $\Phi(\theta) = \mathrm{diag}(\phi_1(\theta), \phi_2(\theta), \ldots, \phi_K(\theta))$. Observe that all entries of $\Phi(\theta)$ are finite.

The following theorem is due to Perron and Frobenius [19, Theorem 8.4.4, Theorem 8.5.1]:

Theorem 2.1 (Perron-Frobenius) *Let $A \in \mathcal{M}_n$ be a nonnegative irreducible. Then,*

(i) $r(A) = \tilde{r}(A) > 0$, $r(A)$ is a simple eigenvalue of A, and there exists a positive right eigenvector of the matrix A corresponding to the eigenvalue $r(A)$;

(ii) there exists a constant matrix $L \in \mathcal{M}_n$ such that $\lim_{m\to\infty}[A/r(A)]^m = L$.

The nonnegative and irreducible matrix $H(\theta) := P^T \Phi(\theta)$ will play an important role in the following. Let $\rho(\theta) := r(H(\theta))$ be its spectral radius and let $\mathbf{z}(\theta) = (z_1(\theta), \ldots, z_K(\theta))^T$ be its unique right eigenvector corresponding to the eigenvalue $\rho(\theta)$ such that $|\mathbf{z}(\theta)| = 1$.

The following result holds:

Proposition 2.2 (Exponential upper bound)

Assume that $\theta \in \Theta_+$. If $\rho(\theta) \leq 1$, and if

$$P(X_0 \geq x) \leq C(\theta)\, e^{-\theta x}, \qquad \forall x > 0 \tag{2.6}$$

then, for all $n \geq 0$, $x > 0$,

$$P(X_n \geq x) \;\; \leq \;\; C(\theta)\, e^{-\theta x} \tag{2.7}$$

where

$$C(\theta) = \sup_{\substack{x \geq 0 \\ n \geq 0 \\ j \in \mathcal{S}}} \frac{\sum_{k \in \mathcal{S}} p_{kj}\, \pi_n(k)\, (1 - F_k(x))}{\sum_{k \in \mathcal{S}} p_{kj}\, z_k(\theta) \int_x^\infty e^{\theta(u-x)}\, dF_k(u)} < \infty. \tag{2.8}$$

If $X_0 = 0$ *a.s. then*

$$P(X_n \geq x) \leq \inf_{\{\theta \in \Theta_+ \,:\, \rho(\theta) \leq 1\}} C(\theta)\, e^{-\theta x}, \qquad \forall n \geq 0,\, \forall x > 0. \tag{2.9}$$

It is worth observing that the upper bound in (2.9) is always smaller than or equal to 1 when the Markov chain $(Y_n)_n$ is stationary since $C(0) = 1$ thanks to the identity $\mathbf{z}(0) = \pi$.

Proof of Proposition 2.2. Define

$$\gamma_j^n(x) = \begin{cases} C(\theta)\, (z_j(\theta)/\pi_n(j))\, e^{-\theta x} & \text{if } \pi_n(j) \neq 0 \\ 1 & \text{if } \pi_n(j) = 0. \end{cases} \tag{2.10}$$

Let us show that these functions satisfy (2.2). First, observe from the identities $\sum_{k \in \mathcal{S}} p_{kj}\, \pi_n(k) = \pi_{n+1}(j)$, $j \in \mathcal{S}$, that (2.2) trivially holds if $\pi_{n+1}(j) = 0$ since the left-hand side of (2.2) also vanishes in this case.

Assume now that $\pi_{n+1}(j) \neq 0$. We have

$$\begin{aligned}
&\sum_{k \in \mathcal{S}} p_{kj}\, \pi_n(k) \left[\int_{-\infty}^x \gamma_k^n(x-u)\, dF_k(u) + 1 - F_k(x) \right] \\
&= \sum_{k \in \mathcal{S}} p_{kj}\, \pi_n(k) \left[\int_{-\infty}^\infty C(\theta)\, \frac{z_k(\theta)}{\pi_n(k)}\, e^{\theta(u-x)}\, dF_k(u) - \int_x^\infty C(\theta)\, \frac{z_k(\theta)}{\pi_n(k)}\, e^{\theta(u-x)}\, dF_k(u) + 1 - F_k(x) \right] \\
&= e^{-\theta x}\, C(\theta) \sum_{k \in \mathcal{S}} p_{kj}\, \phi_k(\theta)\, z_k(\theta) - \sum_{k \in \mathcal{S}} p_{kj}\, \pi_n(k) \left[\int_x^\infty \left(C(\theta)\, \frac{z_k(\theta)}{\pi_n(k)}\, e^{\theta(u-x)} - 1 \right) dF_k(u) \right] \\
&\leq e^{-\theta x}\, C(\theta) \sum_{k \in \mathcal{S}} p_{kj}\, \phi_k(\theta)\, z_k(\theta), \quad \text{from (2.8)} \\
&= e^{-\theta x}\, C(\theta)\, \rho(\theta)\, z_j(\theta), \quad \text{since } H(\theta)\, \mathbf{z}(\theta) = \rho(\theta)\, \mathbf{z}(\theta) \\
&\leq e^{-\theta x}\, C(\theta)\, z_j(\theta) = \pi_{n+1}(j)\, \gamma_j^{n+1}(x)
\end{aligned}$$

from the assumption that $\rho(\theta) \leq 1$. The proof of (2.7) now follows from Corollary (2.1).

Since (2.6) is automatically satisfied when $X_0 = 0$ a.s., the inequality (2.7) is seen to hold for all $\theta \in \Theta_+$ such that $\rho(\theta) \leq 1$, which yields (2.9). The finiteness of $C(\theta)$ is a consequence of (2.16) ♣

In Section 2.3 below, we shall derive various upper bounds of $C(\theta)$ which have simpler forms than the right-hand side of (2.8).

The next issue to be addressed is the existence of θ in Θ_+ such that $\rho(\theta) \leq 1$. We will need the following lemma whose proof is forwarded to Appendix A.

Lemma 2.1 *The spectral radius $\rho(\theta)$ is log convex and is strictly convex for $\theta \in \Theta$. Moreover, if $\theta_+ \notin \Theta$, then $\rho(\theta)$ goes to ∞ when θ goes to θ_+.*

Proposition 2.3 *Define $\theta^* = \sup\{\theta \in \Theta : \rho(\theta) \leq 1\}$. Then,*

$$\theta^* > 0 \qquad \text{if and only if} \qquad \sum_{k \in \mathcal{S}} \pi(k)\, E[U_n(k)] < 0. \tag{2.11}$$

Moreover, if $\sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)] < 0$ and if $\theta_+ \notin \Theta$ then $\rho(\theta^*) = 1$ and $\rho(\theta) < 1$ for $0 < \theta < \theta^*$.

Proof. Since the matrix $H(\theta)$ is differentiable at $\theta = 0$ and since $\rho(\theta)$ is a simple eigenvalue of this matrix, Theorem 6.3.12 in [19, p. 372] implies that the derivative of $\rho(\theta)$ at 0 is given by $\rho'(0) = \sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)]$. Hence, (2.11) follows from the strict convexity of $\rho(\theta)$ (see Lemma 2.1) and from $\rho(0) = 1$.

The proof of the last statement is a direct consequence of the identities $\rho(0) = 1$ and $\rho'(0) = \sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)]$, together with Lemma 2.1. ♣

We will show in Section 2.5 (see Lemma 2.2) that condition $\sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)] < 0$ in Proposition 2.3 is the stability condition of this system. Therefore, Proposition 2.3 says that an exponential upper bound exists for $P(X_n \geq x)$ when the system is stable. In that case, θ^* is the largest exponential decay among all positive θ such that $\rho(\theta) \leq 1$. However, this leaves open the question whether θ^* the best possible exponential decay over all $\theta \geq 0$.

The following result whose proof is given in Appendix B readily implies that θ^* is indeed the best exponential decay.

Proposition 2.4 *Assume that Θ is an open set, namely $\Theta = (\theta_-, \theta_+)$. If $\sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)] < 0$ then*

$$\lim_{x\to\infty} \frac{\log P(X \geq x)}{x} = -\theta^*. \tag{2.12}$$

2.3 Simpler Upper Bounds

Various upper bounds can be derived from Proposition 2.2. Let $\tilde{\mathcal{S}}(x) = \{k \mid k \in \mathcal{S},\ F_k(x) < 1\}$, and $\tilde{\mathcal{S}} = \tilde{\mathcal{S}}(0)$.

It is simple to check from (2.8) that

$$C(\theta) \;=\; \sup_{\substack{x\geq 0\\ n\geq 0\\ j\in\mathcal{S}}} \frac{\sum_{k\in\tilde{\mathcal{S}}(x)} p_{kj}\, \pi_n(k)\, (1 - F_k(x))}{\sum_{k\in\tilde{\mathcal{S}}(x)} p_{kj}\, z_k(\theta) \int_x^\infty e^{\theta(u-x)}\, dF_k(u)} \tag{2.13}$$

$$\leq \sup_{\substack{x\geq 0\\ n\geq 0}} \frac{\sum_{k\in\widetilde{\mathcal{S}}(x)} \pi_n(k)\,(1-F_k(x))}{\sum_{k\in\widetilde{\mathcal{S}}(x)} z_k(\theta)\int_x^\infty e^{\theta(u-x)}\,dF_k(u)} \tag{2.14}$$

$$\leq \sup_{\substack{x\geq 0\\ n\geq 0}} \max_{k\in\widetilde{\mathcal{S}}(x)} \frac{\pi_n(k)}{z_k(\theta)} \tag{2.15}$$

$$= \sup_{n\geq 0} \max_{k\in\widetilde{\mathcal{S}}} \frac{\pi_n(k)}{z_k(\theta)} \tag{2.16}$$

where (2.15) follows from the inequality $1-F_k(x)\leq \int_x^\infty e^{\theta(u-x)}\,dF_k(u)$.

In the case where $X_0=0$ a.s. all quantities in (2.13)—(2.16) yield, in combination with Proposition 2.2, upper bounds for the tail distribution of X_n. The next result provides an easily computable upper bound.

Corollary 2.2 *If $\rho(\theta)\leq 1$, and if*

$$P(X_0\geq x)\leq |\boldsymbol{y}(\theta)|\,e^{-\theta x}, \qquad \forall x>0 \tag{2.17}$$

then, for all $n\geq 0$,

$$P(X_n\geq x)\leq |\boldsymbol{y}(\theta)|\,e^{-\theta x}, \qquad \forall x>0 \tag{2.18}$$

where $\boldsymbol{y}(\theta)=(y_1(\theta),\ldots,y_K(\theta))^T$ is any right eigenvector of $H(\theta)$ corresponding to the eigenvalue $\rho(\theta)$ such that

$$y_j(\theta)\geq \sup_{n\geq 0}\pi_n(j), \qquad \forall j\in\mathcal{S}. \tag{2.19}$$

If $X_0=0$ a.s. then for all $n\geq 0$, $x>0$,

$$P(X_n\geq x)\leq \inf_{\{\theta\in\Theta_+:\,\rho(\theta)\leq 1\}} |\boldsymbol{y}(\theta)|\,e^{-\theta x}. \tag{2.20}$$

Proof. The proof is analogous to the proof of Proposition 2.2 except that $C(\theta)$ in (2.10) must now be replaced by $|\mathbf{y}(\theta)|$. ♣

Clearly, $C(\theta)\leq |\mathbf{y}(\theta)|$ under condition (2.19) (hint: replace $z_k(\theta)$ by $y_k(\theta)/|\mathbf{y}(\theta)|$ in (2.16) and use (2.19)). We also observe that $|\mathbf{y}(\theta)|$ is actually equal to the right-hand side of (2.16) when the equality in (2.19) takes place for the smallest component of $\mathbf{y}(\theta)$.

Condition (2.19) will hold, in particular, if $\mathbf{y}(\theta)$ is chosen to be the unique right eigenvector of $H(\theta)$ corresponding to the eigenvalue $\rho(\theta)$ such that $\min_{j\in\mathcal{S}} y_j(\theta)=1$ Also observe that conditions (2.19) reduce to $y_j(\theta)\geq\pi(j)$ for all $j\in\mathcal{S}$ when the Markov chain $\mathbf{Y}$ is stationary (i.e. when $\pi_0=\pi$).

The bound in (2.18) reduces to Kingman's exponential upper bound [27] when X_n represents the waiting time of the n-th customer in a first-in-first-out GI/GI/1 single server queue (i.e. when $\mathcal{S}=\{1\}$ and $U_n(1)=S_n-T_n$, where $(S_n)_n$ and $(T_n)_n$ are independent i.i.d. sequences of positive r.v.'s.).

2.4 Exponential Lower Bound

In this section we address the problem of computing an exponential lower bound for the tail distribution of X_n.

Proposition 2.5 (Exponential lower bound)

Assume that $\rho(\theta^*) = 1$ *and that* $\sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)] < 0$. *If*

$$P(X_0 \geq x) \geq B(\theta^*)\, e^{-\theta^* x}, \qquad \forall x > 0 \tag{2.21}$$

then, for all $n \geq 0$,

$$P(X_n \geq x) \geq B(\theta^*)\, e^{-\theta^* x}, \qquad \forall x > 0 \tag{2.22}$$

where

$$B(\theta^*) = \inf_{\substack{x>0 \\ n\geq 0 \\ j\in\mathcal{S}}} \frac{\sum_{k\in\mathcal{S}} p_{kj}\, \pi_n(k)\, (1 - F_k(x))}{\sum_{k\in\mathcal{S}} p_{kj}\, z_k(\theta^*) \int_x^\infty e^{\theta^*(u-x)}\, dF_k(u)}. \tag{2.23}$$

Proof. Let $(\delta_j^n)_{n,j}$, $\delta_j^n : [0,\infty) \to [0,\infty)$ be a set of functions such that

$$\sum_{k\in\mathcal{S}} p_{kj}\, \pi_n(k) \left[\int_{-\infty}^{x} \delta_k^n(x-u)\, dF_k(u) + 1 - F_k(x) \right] \geq \pi_{n+1}(j)\, \delta_j^{n+1}(x) \tag{2.24}$$

for $j \in \mathcal{S}$, $n \geq 0$. Let Q_n be the property that

$$P(X_n \geq x \,|\, Y_n = j) \geq \delta_j^n(x)$$

for all $x > 0$ and for all $j \in \mathcal{S}$ such that $\pi_n(j) \neq 0$. Mimicking the proof of Proposition 2.1 we can prove that Q_n is true for all $n \geq 0$ if Q_0 is true. In direct analogy with Corollary 2.1 we then deduce that, if X_0 satisfies

$$P(X_n \geq x) \geq \sum_{j\in\mathcal{S}} \pi_n(j)\, \delta_j^n(x), \qquad \forall x > 0 \tag{2.25}$$

for $n = 0$, then (2.25) holds for all $n \geq 0$.

Define the functions $\delta_j^n(x) = B(\theta^*)\, z_j(\theta^*)\, \exp(-\theta^* x)/\pi_n(j)$ if $\pi_n(j) \neq 0$ and $\delta_j^n(x) = 1$ (for instance) if $\pi_n(j) = 0$. Using the same arguments as in the proof of Proposition 2.2 together with $\rho(\theta^*) = 1$ it is easily checked that the functions $(\delta_j^n)_{n,j}$ satisfy (2.24), from which the result follows.

♣

The condition $\rho(\theta^*) = 1$ holds if $\theta_+ \notin \Theta$ (cf. Proposition 2.3). It is simple to construct examples when $B(\theta^*) = 0$. However, we expect $B(\theta^*) > 0$ in practice, especially when $\pi_0 = \pi$. This can be

proved if the r.v.'s $U_n(k)$ are uniformly bounded from above by a constant, that is, if there exists $\Delta > 0$ such that $P(U_n(k) \leq \Delta) = 1$ for all $k \in \mathcal{S}$. In that case, we have (with $\pi_0 = \pi$)

$$\begin{aligned} B(\theta^*) &= \inf_{\substack{x>0 \\ j \in \mathcal{S}}} \frac{\sum_{k \in \mathcal{S}} p_{kj}\, \pi(k) \int_x^\Delta dF_k(u)}{\sum_{k \in \mathcal{S}} p_{kj}\, z_k \int_x^\Delta e^{\theta^*(u-x)}\, dF_k(u)} \\ &\geq \left(\frac{\pi_{\min}}{z_{\max}}\right) \inf_{\substack{x>0 \\ j \in \mathcal{S}}} \frac{\sum_{k \in \mathcal{S}} p_{kj} \int_x^\Delta dF_k(u)}{\sum_{k \in \mathcal{S}} p_{kj} \int_x^\Delta e^{\theta^*(u-x)}\, dF_k(u)} \\ &\geq \left(\frac{\pi_{\min}}{z_{\max}}\right) e^{-\theta^* \Delta} > 0 \end{aligned} \tag{2.26}$$

where $\pi_{\min} = \min_{j \in \mathcal{S}} \pi(j)$, $z_{\max} = \max_{j \in \mathcal{S}} z_j(\theta^*)$.

Another situation where $B(\theta^*) > 0$ is discussed in Section 2.6.

2.5 Exponential Upper and Lower Bounds for the Stationary Regime

In this section we will discuss the stationary version of the results we have obtained so far.

The first result addresses the condition of existence of a stationary regime for the process $(X_n)_n$.

Lemma 2.2 (Stationary regime) *Let $(Y_n)_n$ be an irreducible, aperiodic, ergodic, homogeneous Markov chain with invariant measure π (we do not assume that the state-space is finite). Let $(X_n)_n$ be a sequence of r.v.'s satisfying the recursion (1.1).*

If $\sum_k \pi(k)\, E[U_n(k)] < 0$ then there exists an almost surely finite r.v. X such that for, all $x \geq 0$,

$$P(X_n < x) \to_n P(X < x)$$

independently of the distributions of X_0 and Y_0.

If $\sum_k \pi(k)\, E[U_n(k)] > 0$ then $X_n \to_n \infty$ a.s.

Proof. We know (cf. [1, Theorem 1]) that there exists a stationary Markov chain $(Y^n)_n$, defined on the same probability space as $(Y_n)_n$, independent of Y_0, and such that

$$\lim_{n \to \infty} P(Y_k = Y^k, \forall k \geq n) = 1. \tag{2.27}$$

Since the sequences $(U_n(i))_n$, $i \in \mathcal{S}$, are independent i.i.d. sequences, we may conclude from (2.27) that there exists a stationary and ergodic sequence $(\xi_n)_n$, independent of Y_0, such that

$$\lim_{n \to \infty} P(\xi_k = U_k(Y^k), \forall k \geq n) = 1. \tag{2.28}$$

Let $(V_n)_n$ be a sequence satisfying the recursion $V_{n+1} = \max(0, V_n + \xi_n)$ for $n \geq 0$, where $V_0 \geq 0$ is a finite random variable. Since $(\xi_n)_n$ is stationary and ergodic, it is known (see [29]) that if $E[\xi] < 0$ then there exists a stationary sequence $(V^n)_n$ with generic element denoted as X, such that $X < \infty$ a.s. and such that $\lim_{n\to\infty} P(V_k = V^k,\, k \geq n) = 1$ for any (possibly random) initial value V_0.

Therefore, Theorem 5 in [1] applies to the sequence $(X_n)_n$, which entails that if $E[\xi] < 0$ then

$$\lim_{n\to\infty} P(X_k = V^k,\, \forall k \geq n) = 1$$

for any (possibly random) initial values X_0 and Y_0. This proves the first part of the lemma since clearly $E[\xi] = \sum_k \pi(k)\, E[U_n(k)]$.

Define $M = \min\{n \geq 0 : Y_k = Y^k,\ \forall k \geq n\}$. We have

$$P(X_n > x) \geq P(X_n > x,\, n \geq M) \geq P\left(\sum_{i=M}^{n} \xi_i > x,\, n \geq M\right) = P\left(\sum_{i=M}^{n} \xi_i > x\right) - P(n < M). \tag{2.29}$$

Since $M < \infty$ a.s. from (2.27) and since $\lim_{n\to\infty} \sum_{i=M}^{n} \xi_i = +\infty$ a.s. when $(\xi_n)_n$ is ergodic and $E[\xi] > 0$, we get from (2.29) that $\lim_{n\to\infty} P(X_n > x) = 1$ for all $x \geq 0$ if $E[\xi] > 0$, which completes the proof. ♣

Recall the definition of θ^* in Proposition 2.3. We have the following result:

Proposition 2.6 (Exponential bounds for the stationary tail distribution)

If $\rho(\theta^*) = 1$ and if $\sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)] < 0$, then X, the stationary variable of the process $(X_n)_n$, satisfies

$$B^*\, e^{-\theta^* x} \leq P(X \geq x) \leq C^*\, e^{-\theta^* x}, \qquad \forall x > 0 \tag{2.30}$$

where

$$B^* = \inf_{\substack{x>0 \\ j\in\mathcal{S}}} \frac{\sum_{k\in\mathcal{S}} p_{kj}\, \pi(k)\, (1 - F_k(x))}{\sum_{k\in\mathcal{S}} p_{kj}\, z_k(\theta^*) \int_x^\infty e^{\theta^*(u-x)}\, dF_k(u)} \tag{2.31}$$

$$C^* = \sup_{\substack{x>0 \\ j\in\mathcal{S}}} \frac{\sum_{k\in\mathcal{S}} p_{kj}\, \pi(k)\, (1 - F_k(x))}{\sum_{k\in\mathcal{S}} p_{kj}\, z_k(\theta^*) \int_x^\infty e^{\theta^*(u-x)}\, dF_k(u)}. \tag{2.32}$$

Proof. The existence of a stationary regime for $(X_n)_n$ under the condition $\sum_{k\in\mathcal{S}} \pi(k)\, E[U_n(k)] < 0$ has been established in Lemma 2.2.

Set $X_0 = 0$ and assume that the Markov chain $(Y_n)_n$ is stationary (i.e. $\pi_0 = \pi$). From Proposition 2.2 we have that $P_\pi(X_n \geq x) \leq C^* \exp(-\theta^* x)$ for all $x > 0$, $n \geq 0$. Letting n go to infinity in this inequality, and using the result that $P(X_n \geq x) \to_n P(X \geq x)$ independently of the distribution of X_0 and Y_0 from Lemma 2.2, yields

$$P(X \geq x) \leq C^*\, e^{-\theta^* x}, \qquad \forall x > 0. \tag{2.33}$$

This establishes the upper bound.

Let now X_0 be a r.v. satisfying (2.21). Hence, $P_\pi(X_n \geq x) \geq B^* \exp(-\theta^* x)$ for all $x > 0$, $n \geq 0$, from (2.22). Letting n go to infinity, and applying again Lemma 2.2 gives the lower bound. ♣

Similarly, the following upper bounds can be obtained from Proposition 2.2 and Corollary 2.2.

Proposition 2.7 *Assume that $\theta \in \Theta_+$. If $\rho(\theta) \leq 1$, then for all $x > 0$*

$$P(X \geq x) \leq C(\theta)\, e^{-\theta x} \tag{2.34}$$

where $C(\theta)$ is defined in (2.8) with $\pi_n(k)$ substituted for $\pi(k)$, and

$$P(X \geq x) \leq |y(\theta)|\, e^{-\theta x} \tag{2.35}$$

where $y(\theta) = (y_1(\theta), \ldots, y_K(\theta))^T$ is any right eigenvector of $H(\theta)$ corresponding to the eigenvalue $\rho(\theta)$ such that $y_j(\theta) \geq \pi(j)$ for all $j \in \mathcal{S}$. Also,

$$P(X \geq x) \leq \inf_{\{\theta \in \Theta_+ : \rho(\theta) \leq 1\}} C(\theta)\, e^{-\theta x} \leq \inf_{\{\theta \in \Theta_+ : \rho(\theta) \leq 1\}} |y(\theta)|\, e^{-\theta x}. \tag{2.36}$$

2.6 Tightness of the Upper and Lower Bounds

We now show that the upper and lower bounds in Proposition 2.6 are tight.

Consider the case $U_n(k) = S_n(k) - T_n(k)$, where $S_n(k)$ and $T_n(k)$ are nonnegative and independent random variables. Assume that $S_n(k)$ is an exponential random variable with parameter β. Then, simple computations yield, for all $x > 0$, $0 \leq \theta < \theta_+ = \beta$,

$$1 - F_k(x) = e^{-\beta x} E\left[e^{-\beta T_n(k)}\right] \tag{2.37}$$

$$\int_x^\infty e^{\theta(u-x)} dF_k(u) = \frac{\beta}{\beta - \theta}\, e^{-\beta x} E\left[e^{-\beta T_n(k)}\right]. \tag{2.38}$$

Note that $\theta_+ = \beta \notin \Theta$. Introducing (2.37) and (2.38) in (2.31) and (2.32) yields the following

Corollary 2.3 *Assume that $U_n(k) = S_n(k) - T_n(k)$, where $S_n(k)$ and $T_n(k)$ are nonnegative and independent random variables. Assume further that $S_n(k)$ is an exponential random variable with parameter β. If $E_\pi[U_n(Y_n)] < 0$ (stability condition) then*

$$\frac{1}{\zeta}\left(\frac{\beta - \theta^*}{\beta}\right) e^{-\theta^* x} \leq P(X \geq x) \leq \zeta \left(\frac{\beta - \theta^*}{\beta}\right) e^{-\theta^* x} \tag{2.39}$$

with

$$\zeta = \frac{\max_{j,k \in \mathcal{S}}\ p_{jk}}{\min_{j,k \in \mathcal{S}}\ p_{jk}}.$$

In particular, when $\zeta = 1$, then

$$P(X \geq x) = \left(\frac{\beta - \theta^*}{\beta}\right) e^{-\theta^* x}. \tag{2.40}$$

It is interesting to note that (2.40) provides the exact solution for the tail distribution of the customer waiting time in a GI/M/1 queue.

2.7 Discussion

Proposition 2.4 requires the assumption $\Theta = (\theta_-, \theta_+)$. This assumption is necessary in order to be able to apply the Gartnër-Ellis theorem [10] which is the main ingredient in the proof that θ^* is the largest exponential decay rate. Also, Propositions 2.5 and 2.6 require the assumption $\rho(\theta^*) = 1$. Thanks to Proposition 2.3 this assumption holds, in particular, when $\theta_+ \notin \Theta$.

Fortunately, many interarrival time and service time distributions satisfy the assumption that $\Theta = (\theta_-, \theta_+)$. These include distributions with rational Laplace transforms (e.g., phase type distributions). On the other hand, it is not difficult to construct an example where Θ is not an open set. Consider the case where the interarrival times and service times are two independent renewal sequences of nonnegative random variables. Let $f_1(t) = (C_1/(1+t^2)) \exp(-\gamma_1 t)\mathbf{1}(t \geq 0)$ and $f_2(t) = (C_2/(1+t^2)) \exp(-\gamma_2 t)\mathbf{1}(t \geq 0)$ be the density functions of the service times and interarrival times, respectively, with $\gamma_1 > 0$ and $\gamma_2 > 0$. Here, C_1 and C_2 are normalizing constants. It is then easily seen that $\theta_- = -\gamma_2$, $\theta_+ = \gamma_1$ so that $\Theta = [-\gamma_2, \gamma_1]$.

3 Applications to Call Admission in Multimedia Systems

In this section we will present various applications of our results to the problems of call admission in a multimedia system such as a network or a server. A call admission algorithm aims at admitting a new multimedia application (session) into a network or a server only if it can be guaranteed a minimal quality of service (QOS) without violating the QOS of other applications already in the system. In the case of a network, there is the additional constraint that the algorithm must be simple enough so that the decision to accept or to reject a new session can be carried out on-line.

Consider the network setting. A call admission algorithm must typically be concerned with guaranteeing an *end-to-end* QOS over a path that may contain two or more hops. This is a difficult problem and one approach taken is to divide the end-to-end QOS requirement among all of the hops and perform call admission at each hop (e.g., [16, 13, 30]). Thus, if any one hop decides not to admit the call, the call is not admitted end-to-end. Under this approach, it suffices to consider the call admission problem for a single channel. Note that, in the case of call admission to a multimedia server, the server can also be modeled as a single resource [8].

Consider a communication channel equipped with a buffer of finite or infinite size, that can transmit up to c units of information (e.g., c ATM cells) per unit of time. When the buffer is of infinite size a typical performance criterion is $P(X \geq b) \leq q$ where X may represent either the buffer content at arrival epochs in steady state or the packet delay in steady state. Observe that if X is the steady-state content of a buffer of infinite size, then $P(X \geq b) \leq q$ implies that, for the case of a buffer with finite capacity b, the cell loss probability does not exceed q.

Of particular practical interest is the notion of associating an *effective bandwidth* to an application such that the test of whether a set of applications can obtain their desired QOS consists of comparing the sum of their individual effective bandwidths to the channel capacity. This can easily be applied to the problem of call admission; accept a new application if its effective bandwidth is less than the excess capacity available at the channel, i.e., the difference between the channel capacity and the sum of the effective bandwidths of the active applications. This has been explored in a number of papers (see [17, 23, 12, 16]) where the notion has been formalized and established rigorously for several different QOS measures for a number of different application traffic models.

In the following, we will only consider a buffer of infinite size. The resource (communication channel in a network, I/O system in a server) will be modeled as a single server queueing system with service capacity c. The call admission problem will be addressed for the performance criterion $P(X > b) \leq \exp(-\theta b)$ when $b \to \infty$. We will consider three different traffic models commonly encountered in the literature: a discrete-time queueing model fed by independent Markov Modulated Arrival Processes (MMAP's; Section 3.1), a continuous-time queueing model fed by independent Markov Modulated Poisson Processes (MMPP's; Section 3.2) and a continuous-time queueing model fed by independent Markov Modulated Fluid Processes (MMFP's; Section 3.3). Last, applications to call admission are given in Section 3.4.

3.1 Markov Modulated Arrivals

We first establish preliminary results in the case when the discrete-time queueing system is fed by a single MMAP (Section 3.1.1). Then, we address the case the queue is fed by N independent MMAP's and derive the effective bandwidth of each source (Section 3.1.2).

3.1.1 Discrete-Time Queue Fed by a Single MMAP

Consider a discrete-time single server queue with one class of customers and an infinite waiting room. For simplicity we will assume that the queue is empty at time 0. Let $(Y_n)_n$ be an irreducible, aperiodic, homogeneous Markov chain on the set finite $\mathcal{S} = \{1, 2, \ldots, K\}$ with transition matrix P. Let $(A_n(i))_n$, $i = 1, 2, \ldots, K$, be K mutually independent sequences of i.i.d. random variables. The process $(A_n(Y_n))_n$ is called a Markov Modulated Arrival Process (MMAP), and $A_n(i)$ gives the number of customers that arrive in the interval of time $[n, n+1)$ given that $Y_n = i$.

Define $U_n(i) = A_n(i) - c$, where $c > 0$ is the capacity of the server (number of customers served by

unit of time). Hence, it is seen that the process $(X_n)_n$ in (1.1) represents the queue-length process of this queueing system.

Let $\Psi(\theta) = \text{diag}\,(E[\exp(\theta\, A_n(1))], \ldots, E[\exp(\theta\, A_n(K))])$ and define $\tau(\theta) = r(P^T\Psi(\theta))$. Notice that $\tau(\theta) = \rho(\theta)\,\exp(\theta c)$ since $r(\alpha A) = \alpha\, r(A)$ for any scalar number α.

By Proposition 2.2 and Corollary 2.2 we have

$$P(X_n \geq x) \leq C(\theta)\, e^{-\theta x} \leq |\mathbf{y}(\theta)|\, e^{-\theta x} \quad x > 0,\, n \geq 0, \tag{3.1}$$

for all θ such that $\tau(\theta) \leq \exp(\theta c)$ or, equivalently, for all θ such that $a^*(\theta) \leq c$ with $a^*(\theta) := (1/\theta)\,\log(\tau(\theta))$, where $\mathbf{y}(\theta)$ is any eigenvector of the matrix $P^T\Psi(\theta)$ corresponding to the eigenvalue $\tau(\theta)$ such that (2.19) holds.

In [3] $a^*(\theta)$ is referred to as the Minimum Envelope Rate (MER) of the arrival process with respect to θ. Specializing Chang's model to ours yields (cf. [3, Theorem 3.7])

$$P(X_n \geq x) \leq \beta(\theta)\, e^{-\theta x} \quad x > 0,\, n \geq 0, \tag{3.2}$$

provided that $a^*(\theta) < c$, with

$$\beta(\theta) \geq \frac{e^{\theta\hat{\sigma}(\theta)}}{1 - e^{-\theta(a^*(\theta)-c)}} \tag{3.3}$$

where $\hat{\sigma}(\theta) = \left[\sup_{n\geq 0}\left((1/\theta)\sup_{m\geq 0}\log\left(E\left[\exp(\theta\sum_{i=m}^{n+m-1} A_i(Y_i))\right]\right) - a^*(\theta)\, n\right)\right]^+$.

Note that both coefficients $C(\theta)$ and $|\mathbf{y}(\theta)|$ in (3.1) remain bounded for all values of θ such that $a^*(\theta) \leq c$ (i.e. for $0 \leq \theta \leq \theta^*$ from Proposition 2.3) unlike the coefficient $\beta(\theta)$ in (3.2) that goes to infinity when $a^*(\theta) \to c$. We also believe that $\mathbf{y}(\theta)$ should be easier to compute than $\beta(\theta)$. Last, observe that in (3.1) we allow $a^*(\theta)$ to be equal to c (which occurs when $\theta = \theta^*$ by Proposition 2.3), whereas in (3.2) the condition $a^*(\theta) < c$ is required.

3.1.2 Discrete-Time Queue Fed by N Independent MMAP's

We consider the model in Section 3.1.1 but we now assume that the queue is fed by N independent MMAP's $(A_n^j(Y_n^j))_n$, $j = 1, 2, \ldots, N$. Let $\mathcal{S}_j$ be the (finite) state-space of the aperiodic, irreducible, homogeneous Markov chain $(Y_n^j)_n$, and denote by $P_j = [{}_jp_{kl}]$ its transition matrix and by π_j its invariant measure.

It is known (cf. [14] for instance) that the superposition of N such independent Markov modulated processes is again a Markov modulated process $(A_n(Y_n))_n$, where Y_n is an aperiodic, irreducible, homogeneous Markov chain on the states $S = \prod_{j=1}^N \mathcal{S}_j$ with transition matrix P and invariant measure π given by

$$P = P_1 \otimes \cdots \otimes P_N \tag{3.4}$$

$$\pi = \pi_1 \otimes \cdots \otimes \pi_N \tag{3.5}$$

where $\otimes$ denotes the Kronecker product [2, 18].

Define $U_n(\mathbf{i}) = \sum_{j=1}^{N} A_n^j(i_j) - c$ with $\mathbf{i} = (i_1, \ldots, i_N) \in S$, and let

$$\Phi(\theta) = e^{-\theta c} \left(\Psi^1(\theta) \otimes \cdots \otimes \Psi^N(\theta) \right) \tag{3.6}$$

where $\Psi^j(\theta) = \operatorname{diag}\left(E\left[\exp(\theta A_n^j(i)) \right],\ i \in \mathcal{S}_j \right)$.

Thus, by using the identities $(A \otimes B)^T = A^T \otimes B^T$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ (cf. [2]), the matrix $H(\theta) = P^T \Phi(\theta)$ is given by

$$e^{-\theta c} \left(P_1 \otimes \cdots \otimes P_N \right)^T \left(\Psi^1(\theta) \otimes \cdots \otimes \Psi^N(\theta) \right) = e^{-\theta c} \left(P_1^T \, \Psi^1(\theta) \right) \otimes \cdots \otimes \left(P_N^T \, \Psi^N(\theta) \right). \tag{3.7}$$

The process $(X_n)_n$ defined in (1.1) now corresponds to the queue-length process in this multiclass queueing system.

The following lemma will have interesting consequences.

Lemma 3.1 *Let $A \in \mathcal{M}_m$ (resp. $B \in \mathcal{M}_n$) be a matrix with eigenvalues $(\alpha_i)_{i=1}^m$ (resp. $(\beta_j)_{j=1}^n$) and corresponding eigenvectors $(\boldsymbol{v}_i)_{i=1}^m$ (resp. $(\boldsymbol{w}_j)_{j=1}^m$). Then, the eigenvalues of the matrix $A \otimes B \in \mathcal{M}_{mn}$ are $\alpha_i \beta_j$, $1 \le i \le m$, $1 \le j \le n$, with corresponding eigenvectors $\boldsymbol{v}_i \otimes \boldsymbol{w}_j$, $1 \le i \le m$, $1 \le j \le n$. Moreover, $|\boldsymbol{v}_i \otimes \boldsymbol{w}_j| = |\boldsymbol{v}_i| \times |\boldsymbol{w}_j|$ for $1 \le i \le m$, $1 \le j \le n$.*

Proof. The proof of all the statements, but the last one, can be found in [18, p. 27]. For the last statement, let $\mathbf{v}_i = (v_{i1}, \ldots, v_{im})$ and $\mathbf{w}_j = (w_{j1}, \ldots, w_{jn})$. It is easy to see that

$$|\mathbf{v}_i \otimes \mathbf{w}_j| = \sum_{k=1}^{m} \sum_{l=1}^{n} v_{ik} w_{jl} = \left(\sum_{k=1}^{m} v_{ik} \right) \times \left(\sum_{l=1}^{n} w_{jl} \right) = |\mathbf{v}_i| \times |\mathbf{w}_j|.$$

♣

From (3.7) and Lemma 3.1 we get that

$$\rho(\theta) = e^{-\theta c} \prod_{j=1}^{N} \tau_j(\theta) \tag{3.8}$$

where $\tau_j(\theta) := r\left(P_j^T \Psi^j(\theta) \right)$ for $j = 1, 2, \ldots, N$.

Therefore, if there exists $\theta > 0$ such that $\prod_{j=1}^{N} \tau_j(\theta) \le \exp(\theta c)$, then by (3.8), Corollary 2.2 and Lemma 3.1, we will have

$$P(X_n \ge \boldsymbol{x}) \le |\mathbf{y}^1(\theta) \otimes \cdots \otimes \mathbf{y}^N(\theta)| \, e^{-\theta x} = \left(\prod_{j=1}^{N} |\mathbf{y}^j(\theta)| \right) e^{-\theta x}, \qquad \forall n \ge 0, \forall \boldsymbol{x} > 0 \tag{3.9}$$

when $X_0 = 0$ a.s., where $\mathbf{y}^j(\theta) = (y_i^j(\theta),\, i \in \mathcal{S}_j)$ is any eigenvector of the matrix $P_j^T \Psi^j(\theta)$ corresponding to the eigenvalue $\tau_j(\theta)$ such that $y_i^j(\theta) \geq \sup_{n \geq 0} P(Y_n^j = i)$ for $i \in \mathcal{S}_j$, $j = 1, 2, \ldots, N$ (here, we use the property that $\mathbf{v} \otimes \mathbf{w} \geq \mathbf{a} \otimes \mathbf{b}$ if $\mathbf{v} \geq \mathbf{a}$ and $\mathbf{w} \geq \mathbf{b}$). Observe from Proposition 2.7 that the bound in (3.9) also holds for $P(X \geq x)$.

Tighter bounds can be obtained by using directly Proposition 2.2 in the transient case and Proposition 2.7 in the stationary case.

We now give an example where the coefficient in front of the exponential in (3.9) is bounded from above by 1 for any number of sources.

Consider N independent high/low rate sources such that $\mathcal{S}_j = \{1, 2\}$ and $A_n^j(1) = r_{j1}$ and $A_n^j(2) = r_{j2}$, with $0 \leq r_{j1} \leq r_{j2}$, $j = 1, 2, \ldots, N$. The transition probabilities are ${}_jp_{12} = q_j$ and ${}_jp_{21} = p_j$. Without loss of generality, we assume that the Markov chains $(Y_n^j)_n$, $j = 1, 2, \ldots, N$ are all stationary and that the system is initially empty. Let $\bar{q}_j = 1 - q_j$ and $\bar{p}_j = 1 - p_j$. By a simple algebraic computation we obtain for $j = 1, 2, \ldots, N$,

$$\pi^j(1) = \frac{p_j}{p_j + q_j}, \qquad \pi^j(2) = \frac{q_j}{p_j + q_j} \tag{3.10}$$

$$z_1^j(\theta) = \frac{\tau_j(\theta) - \bar{p}_j e^{\theta r_{j2}}}{\tau_j(\theta) + q_j e^{\theta r_{j1}} - \bar{p}_j e^{\theta r_{j2}}}, \qquad z_2^j(\theta) = \frac{q_j e^{\theta r_{j1}}}{\tau_j(\theta) + q_j e^{\theta r_{j1}} - \bar{p}_j e^{\theta r_{j2}}} \tag{3.11}$$

where

$$\tau_j(\theta) = \frac{1}{2}\left(\bar{q}_j e^{\theta r_{j1}} + \bar{p}_j e^{\theta r_{j2}} + \sqrt{\left(\bar{q}_j e^{\theta r_{j1}} + \bar{p}_j e^{\theta r_{j2}}\right)^2 - 4(\bar{p}_j - q_j) e^{\theta(r_{j1} + r_{j2})}}\right) \tag{3.12}$$

is the spectral radius of

$$P_j^T \Psi^j = \begin{pmatrix} \bar{q}_j e^{\theta r_{j1}} & p_j e^{\theta r_{j2}} \\ q_j e^{\theta r_{j1}} & \bar{p}_j e^{\theta r_{j2}} \end{pmatrix}$$

and $(z_1^j(\theta), z_2^j(\theta))$ is the right eigenvector corresponding to $\tau_j(\theta)$ such that $z_1^j(\theta) + z_2^j(\theta) = 1$.

Therefore, Proposition 2.2 implies that if there exists $\theta > 0$ such that $\prod_{1 \leq j \leq N} \tau_j(\theta) \leq \exp(\theta c)$, then

$$P(X_n \geq x) \leq C(\theta)\, e^{-\theta x}, \qquad \forall x > 0,\, \forall n \geq 0 \tag{3.13}$$

where $C(\theta)$ is defined by (2.8) with π and $z(\theta)$ being the Kronecker products of the vectors of $(\pi^j)_{j=1}^N$ and $(z^j(\theta))_{j=1}^N$, see Lemma 3.1 and (3.5). We now show that $C(\theta) \leq 1$ when $p_j + q_j \leq 1$, $j = 1, 2, \ldots, N$.

For $j = 1, 2, \ldots, N$, let $Y_j, Z_j \in \{r_{j1}, r_{j2}\}$ be random variables such that $P(Y_j = r_{j1}) = \pi^j(1)$ and $P(Y_j = r_{j2}) = \pi^j(2)$, $P(Z_j = r_{j1}) = z_1^j(\theta)$ and $P(Z_j = r_{j2}) = z_2^j(\theta)$. If $p_j + q_j \leq 1$, then one can verify from (3.10), (3.11) and (3.12) that $z_2^j(\theta) \geq \pi^j(2)$. Therefore, $Y_j \leq_{st} Z_j$, $j = 1, 2, \ldots, N$, where the symbol $\leq_{st}$ stands for "stochastically larger". A real r.v. Z is stochastically larger than a real r.v. Y, if and only if $P(Y \geq x) \leq P(Z \geq x)$ for all $x \in I\!R$. Assume that $Y_1, \ldots, Y_N$ (resp.

$Z_1, \ldots, Z_N$) are mutually independent r.v.'s. Define $Y = \sum_{j=1}^N Y_j$ and $Z = \sum_{j=1}^N Z_j$. It then follows that $Y \leq_{st} Z$ if $p_j + q_j \leq 1$, $j = 1, 2, \ldots, N$.

Let V be the finite set of possible values that Y can take. By using the bound in (2.14) after substituting $\pi_n(k)$ for $\pi(k)$, we obtain

$$C(\theta) \leq \sup_{v \in V} \frac{\sum_{\substack{(u_1,\ldots,u_N)\in\{1,2\}^N,\\ r_{1u_1}+\cdots+r_{Nu_N}\geq v}} \prod_{1\leq j\leq N} \pi^j(u_j)}{\sum_{\substack{(u_1,\ldots,u_N)\in\{1,2\}^N,\\ r_{1u_1}+\cdots+r_{Nu_N}\geq v}} \prod_{1\leq j\leq N} z^j_{u_j}(\theta) \cdot e^{\theta((r_{1u_1}+\cdots+r_{Nu_N})-v)}} \tag{3.14}$$

$$\leq \sup_{v \in V} \frac{\sum_{\substack{(u_1,\ldots,u_N)\in\{1,2\}^N,\\ r_{1u_1}+\cdots+r_{Nu_N}\geq v}} \prod_{1\leq j\leq N} \pi^j(u_j)}{\sum_{\substack{(u_1,\ldots,u_N)\in\{1,2\}^N,\\ r_{1u_1}+\cdots+r_{Nu_N}\geq v}} \prod_{1\leq j\leq N} z^j_{u_j}(\theta)} \tag{3.15}$$

$$= \sup_{v \in V} \frac{P(Y \geq v)}{P(Z \geq v)}. \tag{3.16}$$

Thus, if $p_j + q_j \leq 1$ for $j = 1, 2, \ldots, N$, then $C(\theta) \leq 1$ in view of (3.16). Consequently,

Corollary 3.1 *Consider a single server queue with service capacity c, fed by N independent high/low rate sources with $\mathcal{S}_j = \{1, 2\}$, $A^j_n(1) = r_{j1}$ and $A^j_n(2) = r_{j2}$, such that $0 \leq r_{j1} \leq r_{j2}$, $j = 1, 2, \ldots, N$. The transition probabilities are ${}_jp_{12} = q_j$ and ${}_jp_{21} = p_j$. Assume that the Markov chains $(Y^j_n)_n$, $j = 1, 2, \ldots, N$, are stationary and that the system is initially empty. If $p_j + q_j \leq 1$ for all $1 \leq j \leq N$, and if $\prod_{1\leq j\leq N} \tau_j(\theta) \leq \exp(\theta c)$, then*

$$P(X_n \geq x) \leq e^{-\theta x}, \qquad \forall n \geq 0, \forall x > 0. \tag{3.17}$$

Remark. This result is consistent with observations made in [16, 4] that the constant in front of the exponential is less than one for bursty sources.

We conclude this section by computing the effective bandwidth for each source in the case when the performance criterion is $P(X \geq b) \leq \exp(-\theta b)$ for $b \to \infty$.

Proposition 3.1 *Define*

$$c_j(\theta) = \frac{1}{\theta} \log \tau_j(\theta), \qquad \forall j = 1, 2, \ldots, N. \tag{3.18}$$

If the system is stable and if $\theta_+ \notin \Theta$ then, for all $\theta \in \Theta_+$,

$$\lim_{b\to\infty} \frac{\log P(X \geq b)}{b} \leq -\theta \qquad \textit{if and only if} \qquad \sum_{j=1}^N c_j(\theta) \leq c.$$

Proof. Assume that the system is stable, namely, $E_\pi[U_n(Y_n)] < 0$, so that $\theta^* > 0$ by Proposition 2.3. Therefore, $\rho(\theta) \leq 1$ or, equivalently, $\sum_{j=1}^N c_j(\theta) \leq c$ from (3.8), if and only if $0 \leq \theta \leq \theta^*$. This result follows from the definition of θ^*, the convexity of $\rho(\theta)$ (see Lemma 2.1) and the identity $\rho(0) = 1$.

On the other, since $U_n(i)$ is of the form $U_n(i) = S_n(i) - c$ with $S_n(i) \geq 0$ it is seen that $\theta_- = -\infty$, so that $\Theta = (-\infty, \theta_+)$ owing to the assumption that $\theta_+ \notin \Theta$. Therefore, Θ is an open set and Proposition 2.4 applies to give $\lim_{b\to\infty} \log P(X \geq b)/b \leq -\theta$ for $\theta \in \Theta_+$ if and only if $0 \leq \theta \leq \theta^*$.

Combining both results yields the proposition. ♣

This result has been derived by Kesidis et al. [24] through a heuristic argument. The same result can also be derived in an even more general context (see Assumptions (C1)-(C3) in [3]) from the work by Chang [3, Proposition 3.9] by using the same arguments as ours.

3.2 Markov Modulated Poisson Processes

We now apply our results to the case when the traffic is modeled as the superposition of N independent MMPP's. When $N = 1$, we will refer to this queueing system as the single class MMPP/G/1 queue and, when $N \geq 2$, as the multiclass MMPP/G/1 queue.

This section is organized as follows: single class and multiclass MMPP/G/1 queueing models are introduced in Section 3.2.1. Exponential upper bounds for the tail stationary distribution of the workload in a single class MMPP/G/1 queue are derived in Section (3.2.2). These bounds are obtained by considering a discretized version of the model. The CAC problem is then addressed in Section (3.2.3).

3.2.1 Bounding the Customer Waiting Time in a MMPP/G/1 Queue

Consider a single-server first-in-first-out queue with an infinite buffer where customers arrive according to a Markov Modulated Poisson Process (see [14] and references therein) with arrival rate $\lambda_{Z(t)}$ at time t, where $(Z(t),\, t \geq 0)$ is an irreducible Markov process on $\mathcal{S} = \{1, 2, \ldots, K\}$ with infinitesimal generator Q and invariant measure $q = (q(1), q(2), \ldots, q(K))$. Let $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_K)$ denote the rate matrix. We assume that the service times of customers generated in state $i \in \mathcal{S}$ form a renewal sequence denoted by $(S_n(i))_n$ and that all service times and interarrival times are assumed to be mutually independent r.v.'s. Last, we assume that the Markov process $(Z(t),\, t \geq 0)$ is stationary.

Define the matrix $P = [p_{ij}] \in \mathcal{M}_K$ as

$$P = (\Lambda - Q)^{-1}\Lambda \tag{3.19}$$

where the existence of $(\Lambda - Q)^{-1}$ follows from [19, Corollary 6.2.27]. If we further assume that

$\lambda_i > 0$ for $i \in \mathcal{S}$, then the matrix P is irreducible and aperiodic, and p_{ij} gives the probability that a customer arrives in state j given the previous customer arrived in state i (see [14]).

Define $\phi_i(\theta) = E\left[\exp(\theta(S(i) - T(i))\right]$, where $S(i)$ and $T(i)$ are generic r.v.'s for the service times of customers generated in state i and for the interarrival between two customers (say C_1 and C_2) given C_1 was generated in state i. It is easily seen (see [14]) that

$$\phi_i(\theta) = E\left[e^{\theta S(i)}\right] \left((\theta I_K + \Lambda - Q)^{-1} \underline{\lambda}\right)_i \tag{3.20}$$

where $\underline{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$. Here $(\mathbf{v})_i$ denotes the i-th component of any vector $\mathbf{v}$ and I_K stands for the identity matrix in $\mathcal{M}_K$. The process $(X_n)_n$ in (1.1) now corresponds to the waiting time process in this queueing system. It is a simple exercise to show that the stability condition is $\sum_{i \in \mathcal{S}} q(i)\, \lambda_i E[S(i)] < 1$.

The case when $\lambda_i = 0$ for some i can be handled in the same way (in that case P and $\Phi(\theta)$ are n-by-n matrices where n is the number of nonzero arrival rates) and is left to the reader.

Consider now the case when the MMPP/G/1 queueing system is fed by N independent MMPP's $\left(\lambda^k_{Z_k(t)},\, t \geq 0\right)$, $k = 1, 2, \ldots, N$. Let Q_k and $\mathcal{S}_k$ be the infinitesimal generator and the state-space, respectively, of the irreducible Markov process $\mathbf{Z}_k = (Z_k(t), t \geq 0)$, and denote by Λ_k its rate matrix. Let $K_k < \infty$ be the cardinality of the state-space $\mathcal{S}_k$. It is known (e.g., see [14]) that the superposition of these N independent MMPP's is again a MMPP with $Z(t) \in \prod_{k=1}^N \mathcal{S}_k$, infinitesimal generator $Q' = Q_1 \oplus \cdots \oplus Q_N$ and rate matrix $\Lambda' = \Lambda_1 \oplus \cdots \oplus \Lambda_N$, where $\oplus$ denotes the Kronecker sum [2]. For k fixed in $\{1, 2, \ldots, K\}$, let $(S^k_n(i))_n$ be K_k renewal sequences where $S^k_n(i)$ represents the service times of the n-th customer of stream k generated when $\mathbf{Z}_k$ is in state $i \in \mathcal{S}_k$. We further assume that the r.v.'s $(S^k_n(i))_{i,k,n}$ are all mutually independent.

Thus, the computation of exponential bounds for the waiting time process in this multiclass MMPP/G/1 queue reduces to the computation of exponential bounds for a single class MMPP/G/1 queue with infinitesimal generator Q', rate matrix Λ', and moment generating functions $\phi_{\mathbf{i}}(\theta) = \sum_{k=1}^N (\lambda^k_{i_k}/(\lambda^1_{i_1} + \cdots + \lambda^N_{i_N})) E[\exp(\theta S^k_n(i_k))]\ \left((\theta I_K + \Lambda' - Q')^{-1}\, \underline{\lambda}'\right)_{\mathbf{i}}$ for all $\mathbf{i} := (i_1, \ldots, i_N) \in \prod_{k=1}^N \mathcal{S}_k$ with $K := \sum_{k=1}^N K_k$, $\underline{\lambda}' = \underline{\lambda}^1 \oplus \cdots \oplus \underline{\lambda}^N$ and $\underline{\lambda}^k = (\lambda^k_i,\, i \in \mathcal{S}_k)^T$ for $k = 1, 2, \ldots, N$.

3.2.2 Bounding the Workload Process in a MMPP/G/1 queue

We now analyze the workload process of the MMPP/G/1 queue. We will assume that the queue is empty at time 0 and that the server has capacity c (note that the notion of server capacity is not important in Section 3.2.1 as customer service times are specified in the model). We will further assume that the Markov process $\mathbf{Z} = (Z(t), t \geq 0)$ is stationary. Introduce $(B_n(i))_n$ to be the sequence of amounts of work generated when $\mathbf{Z}$ is in state i. Observe that $(B_n(i))_n$ is an i.i.d. sequence of r.v.'s and let $B(i)$ be a generic element of this sequence. Define $\upsilon_i(\theta) = E[\exp(\theta\, B(i))]$

for $i \in \mathcal{S}$ and let $\Upsilon(\theta) = \text{diag}\ (\upsilon_i(\theta),\, i \in \mathcal{S})$. Let $V(t)$ be the amount of work generated in $[0,t)$ and let $G_i(t,\theta) = E[\exp(\theta V(t)) \,|\, Z(0) = i]$ for $i \in \mathcal{S}$. Also define $G(t,\theta) = \text{diag}\ (G_i(t,\theta),\, i \in \mathcal{S})^T$.

We will appeal to the following technical lemma whose proof is given in Appendix C.

Lemma 3.2 *For all $t \geq 0$, $\theta \in \Theta_+$,*

$$G(t,\theta) = e^{(Q+\Upsilon(\theta)\Lambda-\Lambda)t}\, \mathbf{1}^T \tag{3.21}$$

with $\mathbf{1}^T := (1,1,\ldots,1)^T$, and

$$\lim_{t\to\infty} \frac{1}{t} \log E\left[e^{\theta V(t)}\right] = \tilde{r}\left(Q + \Upsilon(\theta)\Lambda - \Lambda\right). \tag{3.22}$$

We now discretize the MMPP model in the following way. Let $\delta > 0$ be an arbitrarily fixed real number. For $n \geq 1$ the amount of work ${}^\delta A_{n+1}$ arrived during the time slot $[n\delta, (n+1)\delta)$ is placed into the buffer at time $n\delta$ for service. The random variable ${}^\delta A_n$ is determined by the Markov chain $(Z(n\delta))_n$ whose transition matrix is given by

$${}^\delta P = e^{Q\delta}. \tag{3.23}$$

Moreover,

$$E\left[e^{\theta\, {}^\delta A_n} | Z(n\delta) = i\right] = G_i(\delta,\theta)$$

so that, by (3.21),

$${}^\delta\Psi(\theta) := \text{diag}\ (G(\delta,\theta)) = \text{diag}\ \left(e^{(Q+\Upsilon(\theta)\Lambda-\Lambda)\delta}\mathbf{1}^T\right) \tag{3.24}$$

where diag $(\mathbf{v})$ stands for diag $(v_1, v_2 \ldots, v_n)$ for any vector $\mathbf{v} = (v_1, v_2 \ldots, v_n)$.

Let ${}^\delta X_n$ be the workload just before time $n\delta$ with ${}^\delta X_0 = 0$. It is clear that

$${}^\delta X_{n+1} = \left[\, {}^\delta X_n + {}^\delta A_{n+1} - \delta c\right]^+, \qquad \forall n \geq 0. \tag{3.25}$$

Therefore, all the results obtained in Sections 2 and 3 apply to the process $({}^\delta X_n)_n$.

Let ${}^\delta X$ be the stationary regime of $({}^\delta X_n)_n$ under the stability condition $\sum_{i\in\mathcal{S}} q(i)\,\lambda_i\, E[B_i] < c$. Then, we know by Proposition 2.7 that an exponential upper bound with decay $\theta > 0$ exists for $P({}^\delta X \geq x)$ if

$$r({}^\delta\Psi(\theta)\, {}^\delta P) \leq e^{\theta\delta c}. \tag{3.26}$$

Mimicking the argument used in the proof of Proposition 2.4 (see Appendix B), one can show that

$$\log r({}^\delta\Psi(\theta)\, {}^\delta P) = \lim_{n\to\infty} \frac{1}{n} \log E\left[e^{\theta \sum_{i=1}^n {}^\delta A_i}\right] = \lim_{n\to\infty} \frac{1}{n} \log E\left[e^{\theta V(n\delta)}\right]. \tag{3.27}$$

Combining this result with (3.22) yields

$$\log r({}^{\delta}\Psi(\theta)\,{}^{\delta}P) = \delta\,\tilde{r}(A(\theta)) \tag{3.28}$$

with $A(\theta) := Q + \Upsilon(\theta)\Lambda - \Lambda$.

Thus, condition (3.26) simply becomes

$$\tilde{r}(A(\theta)) \leq \theta c. \tag{3.29}$$

Let X be the stationary workload of the original MMPP/G/1 queue. The following result relates X to ${}^{\delta}X$ and will be used in the next section.

Lemma 3.3 *For all $\delta > 0$,*

$$P({}^{\delta}X \geq x) \leq P(X \geq x) \leq P({}^{\delta}X + {}^{\delta}A \geq x) \qquad \forall x > 0, \tag{3.30}$$

where ${}^{\delta}A$ is the stationary regime of the sequence $({}^{\delta}A_n)_n$.

Proof. Fix $\delta > 0$. Let $X(t)$ be the workload at time t of the original MMPP/G/1 queue. We will assume that the sample paths of $(X(t), t \geq 0)$ are all left-continuous and that $X(0-) = 0$. We claim that, for every sample path,

$${}^{\delta}X_m \leq X(m\delta) \leq {}^{\delta}X_{m-1} + {}^{\delta}A_m, \qquad \forall m \geq 1. \tag{3.31}$$

We use an induction argument to establish this result. ¿From the obvious inequalities

$${}^{\delta}X_1 = [{}^{\delta}A_1 - \delta c]^+ \leq X(\delta) \leq {}^{\delta}A_1 = {}^{\delta}X_0 + {}^{\delta}A_1$$

we see that (3.31) holds for $m = 1$. Assume that (3.31) holds for $m = 1, 2, \ldots, n$ and let us show that it still holds for $m = n + 1$. Clearly,

$$[X(n\delta) + {}^{\delta}A_{n+1} - \delta c]^+ \leq X((n+1)\delta) \leq [X(n\delta) - \delta c]^+ + {}^{\delta}A_{n+1}$$

so that,

$${}^{\delta}X_{n+1} = [{}^{\delta}X_n + {}^{\delta}A_{n+1} - \delta c]^+ \leq X((n+1)\delta) \leq [{}^{\delta}X_{n-1} + {}^{\delta}A_n - \delta c]^+ + {}^{\delta}A_{n+1} = {}^{\delta}X_n + {}^{\delta}A_{n+1} \tag{3.32}$$

by using the induction hypothesis and (3.25). Hence, for all $x \geq 0$, $n \geq 1$,

$$P({}^{\delta}X_{n+1} \geq x) \leq P(X(n\delta) \geq x) \leq P({}^{\delta}X_n + {}^{\delta}A_n \geq x), \qquad \forall n \geq 1. \tag{3.33}$$

Letting $n \to \infty$ in (3.33) yields (3.30). ♣

3.2.3 Call Admission Control and Effective Bandwidth for MMPP's

Consider now the communication channel with bandwidth c. The traffic is generated by N independent MMPP sources $\left(Q_k = [q_{ij}^k],\ \mathcal{S}_k,\ \Lambda_k = \text{diag}\,(\lambda_{1k},\ k \in \mathcal{S}_k)\right)$ for $k = 1, 2, \ldots, N$, When the Markov process $\mathbf{Z}_k = (Z_k(t), t \geq 0)$ is in state $i \in \mathcal{S}_k$, customers arrive in the queue according to a Poisson process with parameter λ_{ik}. Let $(B_n(i,k))_n$ be the sequence of amounts of work generated by source k when $\mathbf{Z}_k$ is in state i, for $i \in \mathcal{S}_k$, $k = 1, 2, \ldots, N$. We assume that all these sequences constitute mutually independent sequences of i.i.d. random variables. Define $\Upsilon_k(\theta) = \text{diag}\,(E[\exp(\theta B_n(i,k))], i \in \mathcal{S}_k)$.

As noticed earlier the process resulting from the superposition of these N MMPP's is itself a MMPP. Let X be the workload of this MMPP/G/1 queue in the stationary regime under the stability condition $\sum_{(i_1,\ldots,i_N)\in\times_{k=1}^N \mathcal{S}_k} \prod_{k=1}^N q_k(i_k) \sum_{k=1}^N \lambda_{ik}\, E[B_n(i,k)] < c$ where $(q_k(i), i \in \mathcal{S}_k)$ is the invariant measure of the Markov process $\mathbf{Z}_k$.

We now compute the effective bandwidth for each source in the case when the performance criterion is $(\log(P(X \geq b)))/b \leq -\theta$ with $b \to \infty$.

Proposition 3.2 *Define*

$$d_j(\theta) = \frac{1}{\theta}\,\tilde{r}\left(Q_j + \Upsilon_j(\theta)\Lambda_j - \Lambda_j\right), \qquad \forall j = 1, 2, \ldots, N. \tag{3.34}$$

If the system is stable then, for all $\theta \geq 0$,

$$\lim_{b\to\infty} \frac{\log P(X \geq b)}{b} \leq -\theta \qquad \textit{if and only if} \qquad \sum_{j=1}^N d_j(\theta) \leq c.$$

Proof. Define the process $({}^\delta X_n)_n$ like in (3.25) with ${}^\delta A_n = \sum_{i=1}^N {}^\delta A_n^i$, where ${}^\delta A_n^i$ is the workload generated by source i in $[(n-1)\delta, n\delta)$, and let ${}^\delta X_n$ be its stationary regime. By repeating the argument in (3.27) and by using the independence of the r.v.'s ${}^\delta A_n^i$ and ${}^\delta A_n^j$ for $i \neq j$, we see that the condition (3.26) reduces to

$$\sum_{i=1}^N d_j(\theta) \leq c. \tag{3.35}$$

In the following, the notation ${}^\delta\theta_+$, ${}^\delta\Theta$, ${}^\delta\Theta_+$, and ${}^\delta C(\theta)$ will stand for θ_+, Θ, Θ_+, and $C(\theta)$, respectively, in the case when $\phi_i(\theta)$ in Section 2 is given by $\phi_i(\theta) = \exp(-\theta\delta c)\, G_i(\delta, \theta)$ for all $i \in \mathcal{S}$, $\delta > 0$.

Fix $\delta > 0$. Assume that $\lim_{b\to\infty}(\log P(X \geq b))/b \leq -\theta$. Then, by Lemma 3.3 $\lim_{b\to\infty}(\log P({}^\delta X \geq b))/b \leq -\theta$.

By using the same arguments as in the proof of Proposition 3.1, we obtain that if ${}^{\delta}\theta_+ \notin {}^{\delta}\Theta$, then, $\forall \theta \in {}^{\delta}\Theta_+$,

$$\lim_{b\to\infty} \frac{\log P({}^{\delta}X \geq b)}{b} \leq -\theta \qquad \text{if and only if} \qquad \sum_{j=1}^{N} d_j(\theta) \leq c.$$

This shows that for all $\theta \in {}^{\delta}\Theta_+$, $\sum_{j=1}^{N} d_j(\theta) \leq c$ if $\lim_{b\to\infty}(\log P(X \geq b))/b \leq -\theta$. Letting now $\delta \to 0$ we get that for all $\theta \geq 0$, $\sum_{j=1}^{N} d_j(\theta) \leq c$ if $\lim_{b\to\infty}(\log P(X \geq b))/b \leq -\theta$ since it is easily seen that the conditions ${}^{\delta}\theta_+ \notin {}^{\delta}\Theta$ and $\theta \in {}^{\delta}\Theta_+$ reduce to $\theta \geq 0$ as δ goes to 0.

Assume now that (3.35) is satisfied. For all $\delta > 0$, $a \in (0, b)$, $\theta \in {}^{\delta}\Theta_+$, we have

$$\begin{aligned} P(X \geq b) &\leq P({}^{\delta}X + {}^{\delta}A \geq b) && (3.36)\\ &\leq P({}^{\delta}X \geq b-a) + P({}^{\delta}A \geq a) \\ &\leq {}^{\delta}C(\theta)\, e^{\theta(b-a)} + P({}^{\delta}A \geq a) && (3.37) \end{aligned}$$

where (3.36) and (3.37) follow from Lemma 3.3 and from Proposition 2.7, respectively.

It is shown in Appendix D that

$$\limsup_{\delta\to 0} {}^{\delta}C(\theta) < \infty. \tag{3.38}$$

On the other hand, we have $\lim_{\delta\to 0} P({}^{\delta}A \geq a) = 0$ for all $a > 0$. Therefore, letting $\delta \to 0$ in (3.37), taking the logarithm, dividing by b, and letting $b \to \infty$ implies that $\lim_{b\to\infty}(\log P(X \geq b))/b \leq -\theta$ if $\sum_{j=1}^{N} d_j(\theta) \leq c$, which concludes the proof. ♣

The above result has been obtained by Mitra and Elwalid [12] for the case that all service times are identically distributed exponential r.v.s, and also by Kesidis et al. [24] through an heuristic argument for a simpler traffic pattern.

3.3 Markov Modulated Fluid Processes

Consider a single MMFP [12] with finite state-space $\mathcal{S} = \{1, 2, \ldots, K\}$ and irreducible infinitesimal generator $Q = [q_{ij}]$, serviced by a channel of constant capacity c which is equipped with an infinite buffer. We represent this source by $(Q, \mathcal{S}, \underline{\lambda})$ with $\underline{\lambda} = (\lambda_1, \ldots, \lambda_K)$, where λ_i is the rate at which traffic is generated when the source is in state $i \in \mathcal{S}$. It is a simple exercise to show that the process $(X_n)_n$ in (1.1) represents the buffer content at jump times of the Markov process when $P = D^{-1}(D - Q)$ with $D = \text{diag}\,(q_{11}, \ldots, q_{KK})$ and $\Phi(\theta) = \text{diag}\,(q_{kk}/(q_{kk} + \theta(\lambda_k - c)), k = 1, 2, \ldots, K)$ (hint: set $U_n(i) = T_n(i)(\lambda_i - c)$ where $T_n(i)$ is the sojourn time of the Markov process in state i).

Assume now that there are N independent Markov modulated fluid sources $(Q_k = [q_{ij}^k], \mathcal{S}_k, \underline{\lambda}^k = (\lambda_i^k, i \in \mathcal{S}_k))$, $k = 1, 2, \ldots, N$. It is known [12] that the aggregate source is itself a Markov-modulated fluid source $(Q, \mathcal{S}, \underline{\lambda})$ with $Q = Q_1 \oplus \cdots \oplus Q_N$, $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$, and $(\underline{\lambda}')^T =$

$\left(\underline{\lambda}^1\right)^T \oplus \cdots \oplus \left(\underline{\lambda}^N\right)^T$. Therefore, the computation of exponential bounds for this model reduces to considering a single Markov modulated fluid source.

Let $V(t)$ be the amount of fluid generated by a single Markov modulated fluid source by time t. In direct analogy with the derivation of (C.4) we can show that

$$\lim_{t\to\infty} \frac{1}{t} \log E\left[e^{\theta V(t)}\right] = \tilde{r}(Q + \theta\Lambda). \tag{3.39}$$

By discretizing the model as in Section 3.2.2 and using the line of arguments in Section 3.2.3, we obtain

Proposition 3.3 *Define*

$$f_j(\theta) = \frac{1}{\theta}\,\tilde{r}\left(Q_j + \theta\Lambda_j\right), \qquad \forall j = 1, 2, \ldots, N. \tag{3.40}$$

If the system is stable then, for all $\theta \geq 0$,

$$\lim_{b\to\infty} \frac{\log P(X \geq b)}{b} \leq -\theta \qquad \textit{if and only if} \qquad \sum_{j=1}^{N} f_j(\theta) \leq c.$$

It is worth observing from the identity $U_n(i) = T_n(i)(\lambda_i - c)$ that $\Theta = (-\infty, \theta_+)$ so that unlike in Propositions 3.1 and 3.2 the extra condition $\theta_+ \notin \Theta$ is automatically satisfied.

This result was first obtained by Gibbens and Hunt [17] in a special case, and later by Elwalid and Mitra [12]. It was also derived by Kesidis, Walrand and Chang [24] through an heuristic argument.

3.4 Call Admission

We consider two applications of the above analysis to call admission in multimedia systems. The first is to the admission of voice calls to a single T1 (1.536Mbs) channel. The second is to the admission of viewers to a video server.

3.4.1 Call admission in a network

Consider a single T1 channel serving a population of voice sessions. For simplicity we discretize time into 16ms. segments and model each voice source as an on-off source with transition matrix

$$P_j = \begin{bmatrix} .975 & .025 \\ .045 & .955 \end{bmatrix}$$

where the number of arrivals in a time unit is 0 when the source is in state 0 and 1 otherwise. The mean on and off periods correspond to 352ms and 650 ms, respectively. The service rate of the

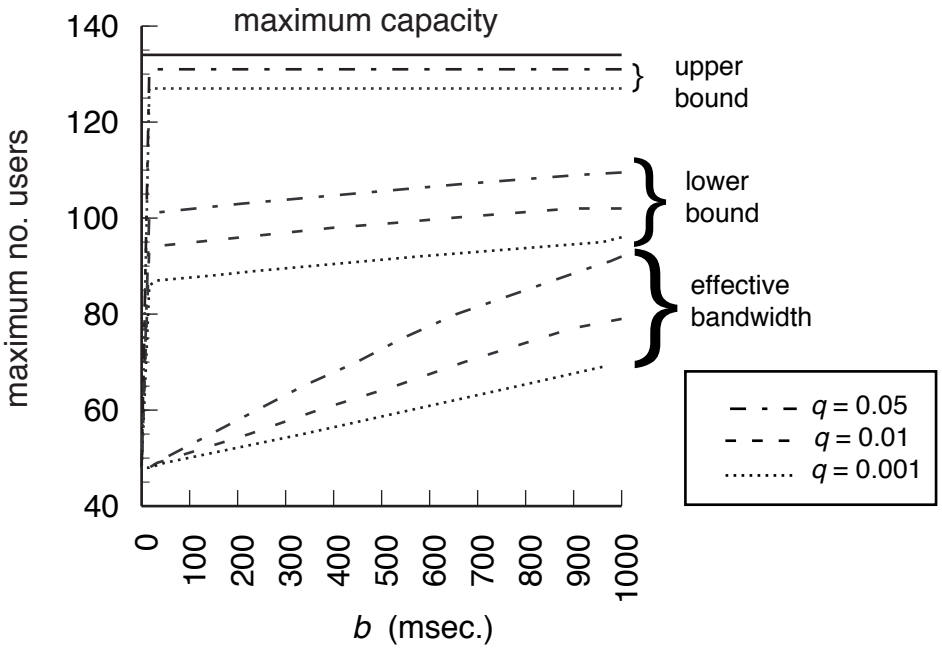

Figure 1: Supportable number of voice sessions.

channel is taken to be $c = 48$ which corresponds to each source generating data at a peak rate of 32Kbs. Observe that there is no contention if the number of sources N is less than 49 and that the system is unstable whenever $N > 134$.

We ask ourselves the following question: what is the number of voice sessions that can be supported by the channel such that $P(X \geq b) \leq q$? Here X is the backlog (measured in ms. of data), b the tolerable delay and q a tolerance. Let N_{max} denote this number. The distribution bounds in Proposition 2.6 can be used to obtain bounds on N_{max} - namely

$$\max_{49 \leq N \leq 134} \{N : -\ln(q/C^*)/b\} \} \leq N_{max} \leq \max_{49 \leq N \leq 134} \{N : -\ln(q/B^*)/b\}\}.$$

Figure 1 gives lower bounds on N_{max} as a function of the tolerable delay, b, for tolerances of 0.1%, 1% and 5%. Also included are the nmber of sessions that can be supported based on the effective bandwidth approach (cf., Proposition 3.1). We observe that there is a large gap between the lower and upper bounds. In addition, the effective bandwidth approach is very conservative, especially for very tight delay constraints. This has been observed elsewhere as well (see [16]) where enhancements have been proposed.

q	l.b.	u.b.	eff. bw
.01	105	111	100
.05	107	113	100
.1	108	114	100

Table 1: Supportable numbers of video sessions.

3.4.2 Call admission in a video server

We consider requests to a video server for movies. Sources are homogeneous and behave as follows. Each source cycles between playback of a movie during which it requires 1 resource unit and pause during which it releases its resource. For simplicity, time is divided into 1/2 second segments. The source is modeled as an on/off source with transition matrix

$$P_j = \begin{bmatrix} .9996667 & .0003333 \\ .9999444 & .0000556 \end{bmatrix}.$$

The playback period has an average length of 30 minutes and the pause period has average length of 5 minutes. Last, we assume that the video server has 100 resource units. Hence it can handle a minimum of 100 and a maximum of 116 viewers.

We again consider the question - how many viewers can this system handle such that the start of playback is not delayed beyond b time units with probability that exceeds q. Using the same approach as with the voice application, we have determined upper and lower bounds for N_{max} for $.5sec \leq b \leq 60sec$ for tolerances of 1, 5, and 10%. For the range given above, the bounds obtained on N_{max} does not depend on b and is presented in Table 1. Also included are the number of sessions that can be supported as predicted by the effective bandwidth approach. Observe that, the effective bandwidth approach yields the same number of sessions as can be supported through a peak rate allocation.

4 Extension to Countable Markov Chains

In this section we extend the results in Section 2 to the case where the state-space $\mathcal{S}$ of the Markov chain $(Y_n)_n$ is countable. This extension may be used to establish exponential bounds in Markovian feedforward queueing networks.

We assume that the Markov chain $(Y_n)_n$ is aperiodic, irreducible and positive recurrent. As before, we assume that $E[U_n(i)]$ is finite for all $i \in \mathcal{S}$.

The following condition will be needed (see condition (3.1) in [20]): there exists a probability measure Q on $\mathcal{S} \times I\!R$, an integer m_0, and real numbers $0 < a \leq b < \infty$, such that

$$a\,Q(j,x) \leq \left(\mathrm{diag}\left(F_k(x)\,, k \in \mathcal{S}\right) P\right)^{m_0}_{ij} \leq b\,Q(j,x) \tag{4.1}$$

for all $x \in I\!R$, $i, j \in \mathcal{S}$. Let $\hat{Q}(\theta) := \int_{-\infty}^{\infty} \exp(\theta x)\, Q(\mathcal{S}, dx)$ and define $\mathcal{D} := \{\theta : \hat{Q}(\theta) < \infty\}$.

Recall the definition of the matrix $H(\theta)$ (cf. Section 2.2). Let $\partial\mathcal{D}$ denote the boundary of the set $\mathcal{D}$. The following lemma can be seen as an extension of Perron-Frobenius theory to infinite nonnegative and irreducible matrices.

Lemma 4.1 *Assume that (4.1) holds and that the set $\mathcal{D}$ is open. For each $\theta \in \mathcal{D}$, the matrix $H(\theta)$ has a maximal simple eigenvalue $\rho(\theta)$ with associate right-eigenvector $\mathbf{y}(\theta)$ such that*

(i) *$\mathbf{y}(\theta)$ is positive;*

(ii) *there exists a matrix L with finite elements such that $(H^T(\theta)/\rho(\theta))^n \to L$ as $n \to \infty$;*

(iii) *$\rho(\theta)$ is analytic and strictly convex on $\mathcal{D}$;*

(iv) *$\rho(\theta) \to \infty$ as $\theta \to \theta_0 \in \partial\mathcal{D}$.*

Proof. Define $S_n = S_0 + \sum_{k=0}^{n-1} U_k(Y_k)$ for all $n \geq 1$. The process $(Y_n, S_n)_n$ is a Markov-additive process (see [20]) with kernel of the generating functions of the additive components (i.e., the matrix denoted as $\hat{P}(\cdot)$ in [20]) given by $H^T(\cdot)$. Therefore, statements (i)-(ii) follow from [20, Lemma 3.1], statement (iii) follows from [20, Lemma 3.4], and statement (iv) follows [20, Corollary 3.1]. ♣

Lemma 4.1 allows us to extend all the results in Theorem 2.1 and in Lemma 2.1 to countable Markov chains, which in turn allows us to extend all the results in Section 2 to countable Markov chains.

Appendices

A Proof of Lemma 2.1

We already know from Kingman [25] that $\log(\rho(\theta))$ is convex in Θ, so that $\rho(\theta)$ is convex Θ. We show by contradiction that $\rho(\theta)$ is actually strictly convex in θ.

Assume that $\rho(\theta)$ is not strictly convex in Θ. This means that there exist $\theta_1, \theta_2 \in \Theta$ with $\theta_1 \neq \theta_2$, such that

$$\rho\left(\frac{\theta_1 + \theta_2}{2}\right) = \frac{\rho(\theta_1) + \rho(\theta_2)}{2}.$$

Therefore,

$$\log \rho\left(\frac{\theta_1 + \theta_2}{2}\right) = \log\left(\frac{\rho(\theta_1) + \rho(\theta_2)}{2}\right) > \frac{\log \rho(\theta_1) + \log \rho(\theta_2)}{2} \tag{A.1}$$

where the strict inequality comes from the fact that for $x_1 > 0$, $x_2 > 0$ such that $x_1 \neq x_2$ one has $x_1 + x_2 > 2\sqrt{x_1 x_2}$. Since inequality (A.1) contradicts the fact that $\log(\rho(\theta))$ is convex in Θ, we obtain that necessarily $\rho(\theta)$ is strictly convex in Θ.

We now turn to the proof of statement that $\rho(\theta)$ goes to ∞ when θ goes to θ_+. We need first to recall some results of matrix analysis. Let $A = [a_{ij}] \in \mathcal{M}_K$ with eigenvalues $(\lambda_i)_{i=1}^K$. It is known that (see [19, p. 43]) trace $(A^k) = \sum_{i=1}^K \lambda_i^k$ for all $k \geq 1$, which yields

$$|\text{trace } (A^k)| \leq K \, (r(A))^k \quad \forall k \geq 1. \tag{A.2}$$

On the other hand, because the matrix P is irreducible and nonnegative, we know from Theorem 2.1 and from [19, p. 516] that there exists some integer $m \geq 1$ such that

$$0 < P^m := [p_{ij}^{(m)}]. \tag{A.3}$$

Set $A = H(\theta)^T$ for $\theta \in \Theta$, and observe that $\rho(\theta)$ is also an eigenvalue of A. From (A.2) and the definition of the matrix $H(\theta)$ it is easily seen that

$$K \, (\rho(\theta))^m \geq \left(\min_{j \in \mathcal{S}} \phi_j(\theta) \right)^{m-1} \sum_{i \in \mathcal{S}} \phi_i(\theta) \, p_{ii}^{(m)} \geq e^{\theta \mu (m-1)} \sum_{i \in \mathcal{S}} \phi_i(\theta) \, p_{ii}^{(m)}, \quad \theta \in \Theta \tag{A.4}$$

where the last inequality follows from Jensen's inequality together with the definition of μ (see (2.1)). Since $P^m > 0$ and since $\liminf_{\theta \to \theta_+} \phi_i(\theta) \geq \phi_i(\theta_+)$ for all $i \in \mathcal{S}$ by Fatou's lemma, we see that the right-hand side of (A.4) converges to $+\infty$ when $\theta \to \theta_+$ if $\phi_i(\theta_+) = \infty$ for some $i \in \mathcal{S}$ or, equivalently, if $\theta_+ \notin \Theta$. This implies from (A.4) that $\lim_{\theta \to \theta_+} \rho(\theta) = \infty$ if $\theta_+ \notin \Theta$, which completes the proof. ♣

B Proof of Proposition 2.4

The proof will follow the line of arguments in [3]. We assume that $\Theta = (\theta_-, \theta_+)$ and that the system is stable.

According to Proposition 2.6 we have

$$\limsup_{x \to \infty} \frac{\log \, P(X \geq x)}{x} \leq -\theta^*. \tag{B.1}$$

Let us show that

$$\liminf_{x \to \infty} \frac{\log \, P(X \geq x)}{x} \geq -\theta^* \tag{B.2}$$

which will complete the proof.

Let us assume for the time being that $\pi_0 = \pi$. Introduce

$$J(\theta) = \lim_{n \to \infty} (1/n) \log E_\pi \left[e^{\theta Z_n} \right]$$

for $\theta \in (-\infty, \infty)$, where $Z_n := \sum_{i=0}^{n-1} U_i(Y_i)$. Let us determine $J(\theta)$.

We have

$$E_\pi \left[e^{\theta Z_n} \,|\, Y_0 = i \right] = \phi_i(\theta) \sum_{j \in \mathcal{S}} E_\pi \left[e^{\theta Z_{n-1}} \,|\, Y_0 = j \right] p_{ij}$$

which yields

$$E_\pi \left[e^{\theta Z_n} \right] = \pi \left(\Phi(\theta) P \right)^{n-1} \Phi(\theta) \, \mathbf{1}^T \tag{B.3}$$

with $\mathbf{1} = (1, 1, \ldots, 1)$. Since $\phi_i(\theta) = \infty$ for $\theta \notin \Theta$ for some $i \in \mathcal{S}$, and since $\pi(i) > 0$ for all $i \in \mathcal{S}$, we see from (B.3) that $E_\pi[\exp(\theta Z_n)] = \infty$ for all $\theta \notin \Theta$ and for all $n \geq 1$. Hence, $J(\theta) = \infty$ for $\theta \notin \Theta$.

Assume now that $\theta \in \Theta$. By applying the Perron-Frobenius theorem (Theorem 2.1, statement (ii)) to (B.3) it is easily seen that

$$J(\theta) = \log(\rho(\theta)), \quad \forall \theta \in \Theta. \tag{B.4}$$

Define $D_J = \{\theta : J(\theta) < \infty\}$ and observe from the above discussion that $D_J = \Theta$.

When $\pi_0 = \pi$ the sequence $(U_n(Y_n))_n$ is stationary so that, for all $x \geq 0$, $(P_\pi(X_n \geq x))_n$ is stochastically increasing in n if $X_0 = 0$ a.s. (see [29, Lemma 1]). Hence,

$$P(X \geq x) = \lim_{n \to \infty} P_\pi(X_n \geq x) \geq P_\pi(X_n \geq x) \geq P_\pi(Z_n \geq x), \qquad \forall n \geq 0, \, x > 0 \tag{B.5}$$

if $X_0 = 0$ a.s., where the first equality follows from Lemma 2.2 and the last inequality comes from the standard relation

$$X_n = \max \left(0, \left(\max_{j=0,1,\ldots,n-1} \sum_{i=j}^{n-1} U_i(Y_i) \right) \right), \qquad \forall n \geq 1 \tag{B.6}$$

which is easily derived from (1.1) if $X_0 = 0$.

Let $0 < \alpha < \beta < 1$. Setting $x = \alpha v n$ with $v > 0$ together with (B.5)-(B.6), yields

$$\begin{aligned} \liminf_{x \to \infty} \frac{\log P(X \geq x)}{x} &\geq \frac{1}{\alpha v} \liminf_{n \to \infty} \frac{1}{n} \log P_\pi \left(\frac{Z_n}{n} \geq \alpha v \right) \\ &\geq \frac{1}{\alpha v} \liminf_{n \to \infty} \frac{1}{n} \log P_\pi \left(\frac{Z_n}{n} > \beta v \right). \end{aligned} \tag{B.7}$$

The idea is now to use the lower bound of the Gartnër-Ellis theorem [10, Theorem II.2] to obtain

$$\liminf_{n \to \infty} \frac{1}{n} \log P_\pi \left(\frac{Z_n}{n} > \beta v \right) \geq -I(v) \tag{B.8}$$

with $I(v) = \sup_\theta \left(\theta v - J(\theta) \right)$.

The lower bound in (B.8) will hold (see [10, Theorem II.2]) if (a) $J(\theta)$ is differentiable in Θ, (b) $0 \in \Theta$, (c) $J(\theta)$ is a closed convex function, and (d) $J(\theta)$ is steep. Conditions (a) and (b) are

satisfied here. Since Θ is an open set by assumption, condition (d) will automatically hold if (c) holds (see [10, p. 2]). Let us show that the convex function $J(\theta)$ is closed, namely that, (e) $\lim_{\theta \to y,\, \theta \in \Theta} J(\theta) = +\infty$ for $y = \theta_-$ when θ_- is finite (resp. for $y = \theta_+$ when θ_+ is finite), see [11, pp. 213-214].

Fix $y \in \{\theta_-, \theta_+\}$. By Fatou's lemma, we have

$$\liminf_{\theta \to y,\, \theta \in \Theta} \phi_i(\theta) \geq \phi_i(y), \quad \forall i \in \mathcal{S}. \tag{B.9}$$

By combining now (A.4), (B.4), and (B.9) we readily deduce that (e) holds.

Since v is arbitrary in $(0, \infty)$, we deduce from (B.7) and (B.8) that

$$\liminf_{x \to \infty} \frac{\log P(X \geq x)}{x} \geq -\frac{1}{\alpha} \inf_{v > 0} \frac{I(v)}{v}, \quad \forall \alpha \in (0,1). \tag{B.10}$$

Letting now $\alpha \to 1$ in (B.10) yields

$$\liminf_{x \to \infty} \frac{\log P(X \geq x)}{x} \geq -\inf_{v > 0} \frac{I(v)}{v}. \tag{B.11}$$

The proof is completed if the right-hand side of (B.11) is shown to be $-\theta^*$.

This result can be obtained by noting that $J(\theta)$ is *strictly* convex in D_J [20, Lemma 3.4] (provided that condition (R) in [20] holds), and then by mimicking the proof of Theorem 3.9 in [3]. For the sake of completeness, we give below a simple proof that the right-hand side of (B.11) is equal to $-\theta^*$. Our proof neither requires the strict convexity of $J(\theta)$ nor condition (R) in [20].

Recall that $\rho(\theta)$ is strictly convex in Θ (cf. Appendix A) and that $\rho(0) = \rho(\theta^*) = 1$ (cf. Proposition 2.3). Therefore, $J(\theta) > 0$ for $\theta < 0$ or $\theta > \theta^*$, $J(\theta) < 0$ for $0 < \theta < \theta^*$ and $J(0) = J(\theta^*) = 0$, so that

$$\xi := -\inf_{\theta > 0} J(\theta) = -\inf_{0 < \theta < \theta^*} J(\theta) > 0.$$

It then follows from the definition of $I(v)$ that for all $v \geq 0$

$$I(v) \geq \sup_{\theta > 0} (\theta v - J(\theta)) \geq -\inf_{\theta > 0} J(\theta) = \xi > 0. \tag{B.12}$$

On the other hand, $I(v) \geq \sup_{\theta < 0}(\theta v - J(\theta)) \geq -J(0) = 0$ for all $v \leq 0$ and $I(0) = \sup_\theta(-J(\theta)) = \sup_{0 < \theta < \theta^*}(-J(\theta)) > 0$. Therefore,

$$\theta^* v - I(v) < 0, \qquad \forall v \leq 0. \tag{B.13}$$

From $J(\theta) = \sup_v (v\theta - I(v))$ (see [10, Theorem V.1]) and (B.13) we get

$$\begin{aligned} 0 &= J(\theta^*) = \max\left(\sup_{v \leq 0}(v\theta^* - I(v)), \sup_{v > 0}(v\theta^* - I(v))\right) \\ &= \sup_{v > 0}(v\theta^* - I(v)) \\ &= \sup_{v \geq \epsilon}(\theta^* v - I(v)) \end{aligned}$$

with $\epsilon := \xi/(2\theta^*)$, where the last equality comes from the fact (cf. (B.12)) that $v\theta^* - I(v) < 0$ for all $0 < v < \epsilon$. Thus,

$$0 = \sup_{v \geq \epsilon}(\theta^* v - I(v)) = \sup_{v \geq \epsilon}\left(\theta^* - \frac{I(v)}{v}\right) = \theta^* - \inf_{v \geq \epsilon}\frac{I(v)}{v} = \theta^* - \inf_{v > 0}\frac{I(v)}{v}$$

where the last equality comes from the fact (cf. (B.12)) that $I(v) > \theta^* v$ for all $0 < v < \epsilon$.

Therefore,

$$\theta^* = \inf_{v>0}\frac{I(v)}{v} \tag{B.14}$$

which along with (B.11) implies that (B.2) holds. This completes the proof. ♣

C Proof of Lemma 3.2

Since the Markov process $\mathbf{Z}$ is stationary, we have

$$\begin{aligned}
G_i(t,\theta) = {} & e^{q_{ii}t}\sum_{n=0}^{\infty}\frac{(\lambda_i t)^n}{n!}e^{-\lambda_i t}E\left[e^{\theta\sum_{m=1}^{n}B_m(i)}\right] \\
& + \int_0^t (-q_{ii})\, e^{q_{ii}s}\left(\sum_{n=0}^{\infty}\frac{(\lambda_i s)^n}{n!}e^{-\lambda_i s}E\left[e^{\theta\sum_{m=1}^{n}B_m(i)}\right]\right)\left(\sum_{j\in\mathcal{S}-\{i\}}\frac{q_{ij}}{-q_{ii}}G_j(t-s,\theta)\right)ds \\
= {} & e^{q_{ii}t}\sum_{n=0}^{\infty}\frac{(\lambda_i t\, \upsilon_i(\theta))^n}{n!}e^{-\lambda_i t} + \int_0^t e^{(q_{ii}-\lambda_i)s}\left(\sum_{n=0}^{\infty}\frac{(\lambda_i s\, \upsilon_i(\theta))^n}{n!}\right)\left(\sum_{j\in\mathcal{S}-\{i\}}q_{ij}\, G_j(t-s,\theta)\right)ds \\
= {} & e^{(q_{ii}+(\upsilon_i(\theta)-1)\lambda_i)t} + \int_0^t e^{(q_{ii}+(\upsilon_i(\theta)-1)\lambda_i)s}\sum_{j\in\mathcal{S}-\{i\}}q_{ij}\, G_j(t-s)\, ds. \qquad \text{(C.1)}
\end{aligned}$$

We obtain, by a change of variable ($u = t - s$) in (C.1), that

$$G_i(t,\theta) = e^{(q_{ii}+(\upsilon_i(\theta)-1)\lambda_i)t}\left(1 + \int_0^t e^{-(q_{ii}+(\upsilon_i(\theta)-1)\lambda_i)u}\sum_{j\in\mathcal{S}-\{i\}}q_{ij}G_j(u,\theta)du\right).$$

Taking the derivative with respect to t yields

$$\begin{aligned}
\frac{d}{dt}G_i(t,\theta) &= (q_{ii} + (\upsilon_i(\theta)-1)\,\lambda_i)\,G_i(t,\theta) + \sum_{j\in\mathcal{S}-\{i\}}q_{ij}G_j(t,\theta) \\
&= (\upsilon_i(\theta)-1)\lambda_i\, G_i(t,\theta) + \sum_{j\in\mathcal{S}}q_{ij}G_j(t,\theta)
\end{aligned}$$

or, in vector form,

$$\frac{d}{dt}G(t,\theta) = (Q + \Upsilon(\theta)\Lambda - \Lambda)G(t,\theta)$$

where $G(t,\theta) = (G_i(t,\theta),\ i \in \mathcal{S})^T$ is a colomn vector. Thus, as $G(0,\theta) = \mathbf{1}^T := (1,1,\ldots,1)^T$, we obtain

$$G(t,\theta) = e^{(Q+\Upsilon(\theta)\Lambda-\Lambda)t}\,\mathbf{1}^T \tag{C.2}$$

which establishes (3.21).

Let us now prove (3.22). Let $A(\theta) = Q + \Upsilon(\theta)\Lambda - \Lambda$. Define $\sigma(\theta)$ to be such that $\sigma(\theta) > [\max_{i\in\mathcal{S}}(-q_{ii} - (v_i(\theta)-1)\,\lambda_i)]^+$. Therefore, the matrix $A(\theta)+\sigma(\theta)\,I$ is nonnegative and irreducible (since $A(\theta)$ is irreducible) and the Perron-Frobenious theory applies. By noting that the eigenvalues of $A(\theta)+\sigma(\theta)\,I$ are the eigenvalues of $A(\theta)$ shifted by $\sigma(\theta)$, we get that $r(A(\theta)+\sigma(\theta)I) = \tilde{r}(A(\theta))+\sigma(\theta)$. Recall here that $\tilde{r}(A)$ denotes the largest real eigenvalue, if any, of a matrix A.

We have from (C.2)

$$\begin{aligned}
\lim_{t\to\infty}\frac{1}{t}\log E\left[e^{\theta V(t)}\right] &= \lim_{t\to\infty}\frac{1}{t}\log\left(\pi\, e^{A(\theta)t}\,\mathbf{1}^T\right)\\
&= \log r\left(e^{A(\theta)+\sigma(\theta)I}\right) + \log r\left(e^{-\sigma(\theta)I}\right)\\
&\quad + \lim_{t\to\infty}\frac{1}{t}\log\left(\pi\,\left[\frac{e^{A(\theta)+\sigma(\theta)I}}{r\left(e^{A(\theta)+\sigma(\theta)I}\right)}\right]^t\left[\frac{e^{-\sigma(\theta)I}}{r\left(e^{-\sigma(\theta)I}\right)}\right]^t\mathbf{1}^T\right)\\
&= \log r\left(e^{A(\theta)+\sigma(\theta)I}\right) + \log r\left(e^{-\sigma(\theta)I}\right) \qquad \text{(C.3)}\\
&= \tilde{r}(A(\theta)) \qquad \text{(C.4)}
\end{aligned}$$

where (C.3) follows from Theorem 2.1 (statement (ii)) and (C.4) follows from $r(\exp(A)) = \exp(r(A))$ that holds for any nonnegative and irreducible matrix A and from $r(\exp(-\sigma(\theta)I)) = \exp(-\sigma(\theta))$.

♣

D Proof of (3.38)

We know from (2.16) that ${}^{\delta}C(\theta) \le \max_{k\in S}(1/{}^{\delta}z_k(\theta))$. Here, ${}^{\delta}\mathbf{z}(\theta)$ is the unique right-eigenvector of the matrix $({}^{\delta}P)^T\,{}^{\delta}\Psi(\theta)$ associated with the eigenvalue $r({}^{\delta}\Psi(\theta)\,{}^{\delta}P)$ such that $|{}^{\delta}\mathbf{z}(\theta)| = 1$, i.e.,

$$\begin{aligned}
({}^{\delta}P)^T\,{}^{\delta}\Psi(\theta)\;{}^{\delta}\mathbf{z}(\theta) &= r({}^{\delta}\Psi(\theta)\,{}^{\delta}P)\;{}^{\delta}\mathbf{z}(\theta)\\
&= e^{\delta\tilde{r}(A(\theta))}\mathbf{z}(\theta) \qquad \text{(D.1)}
\end{aligned}$$

where (D.1) follows from (3.28).

Assume first that $\lim_{\delta\to 0}{}^{\delta}\mathbf{z}(\theta)$ exists and is equal to some (componentwise finite) vector $\mathbf{z}(\theta)$.

Rewriting (D.1) as

$$\left(\frac{({}^{\delta}P)^T\,{}^{\delta}\Psi(\theta) - I}{\delta}\right){}^{\delta}\mathbf{z}(\theta) = \left(\frac{e^{\delta\tilde{r}(A(\theta))}-1}{\delta}\right){}^{\delta}\mathbf{z}(\theta) \tag{D.2}$$

and then letting δ go to 0 yields (hint: ${}^{\delta}P = I + \delta Q + \cdots$ and ${}^{\delta}\Psi(\theta) = I + \delta(\Upsilon(\theta)\Lambda - \Lambda) + \cdots$)

$$A^T(\theta)\mathbf{z}(\theta) = \tilde{r}(A(\theta))\mathbf{z}(\theta). \tag{D.3}$$

Recall the definition of $\sigma(\theta)$ in Appendix C and the property that $r(A^T(\theta) + \sigma(\theta)I) = \tilde{r}(A(\theta)) + \sigma(\theta)I$. Together with (D.3) this gives

$$\left(A^T(\theta) + \sigma(\theta)I\right)\mathbf{z}(\theta) = r(A(\theta) + \sigma(\theta)I)\,\mathbf{z}(\theta). \tag{D.4}$$

Since the matrix $A^T(\theta) + \sigma(\theta)I$ is irreducible and nonnegative from the very definition of $\sigma(\theta)$, Perron-Frobenius theory and (D.4) implies that the limit $\mathbf{z}(\theta)$ is positive.

Let us now show that ${}^{\delta}\mathbf{z}(\theta)$ has a finite limit as δ goes to 0. Let $(x_n^i)_n$, $i = 1, 2$, be two sequences of nonnegative real numbers such that $x_n^i \to 0$ as $n \to \infty$, and such that

$$\begin{aligned}
\mathbf{z}^1(\theta) &:= \lim_{n\to\infty} {}^{x_n^1}\mathbf{z}(\theta) = \liminf_{\delta\to 0} {}^{\delta}\mathbf{z}(\theta) \\
\mathbf{z}^2(\theta) &:= \lim_{n\to\infty} {}^{x_n^2}\mathbf{z}(\theta) = \limsup_{\delta\to 0} {}^{\delta}\mathbf{z}(\theta).
\end{aligned}$$

Observe that $\mathbf{z}^i(\theta)$, $i = 1, 2$, are componentwise finite vectors since ${}^{\delta}\mathbf{z}(\theta)$ lies in the compact set $[0,1]^K$ for all $\delta > 0$. It remains to show that $\mathbf{z}^1(\theta) = \mathbf{z}^2(\theta)$.

By repeating the above analysis with ${}^{\delta}\mathbf{z}(\theta)$ substituted for ${}^{x_n^i}\mathbf{z}(\theta)$, $i = 1, 2$, we get that $\mathbf{z}^1(\theta)$ and $\mathbf{z}^2(\theta)$ both satisfy the equation (D.4) or, equivalently, that both vectors are right-eigenvectors of the matrix $A^T(\theta) + \sigma(\theta)I$ associated with the eigenvalue $r(A(\theta) + \sigma(\theta)I)$. Such right-eigenvectors being all equal up to a multiplicative constant from Perron-Frobenius theory, we deduce that necessarily $\mathbf{z}^1(\theta) = \mathbf{z}^2(\theta)$, since clearly $|\mathbf{z}^i(\theta)| = 1$ for $i = 1, 2$. This concludes the proof. ♣

Acknowledgments: The authors would like to thank Alain Jean-Marie for useful discussions on the proof of Lemma 3.2 and Zhi-Li Zhang for the numerical calculations in Section 3.4.

References

[1] A. A. Borovkov, S. G. Foss, "Stochastically Recursive Sequences and Their Generalizations", *Siberian Advances in Mathematics*, **2**, pp. 16–81, 1992.

[2] J. W. Brewer, "Kronecker Products and Matrix Calculus in System Theory", *IEEE Trans. on Circuit and Syst.*, **25**, No. 9, pp. 772-781, 1978.

[3] C.-S. Chang, "Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks", *IEEE Trans. Aut. Contr.*, **39**, No. 5, pp. 913-931, May 1994.

[4] G. L. Choudhury, D. M. Lucantoni, W. Whitt, "Squeezing the Most out of ATM", preprint, Mar. 1993.

[5] C. Courcoubetis, G. Fouskas, R. Weber, "On the Performance of an Effective Bandwidth Formula", *Proc. of ITC'14*, pp. 201-212, Antibes, June 1994, Elsevier, Amsterdam, Eds. J. Labetoulle, J. Roberts.

[6] R. L. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation", *IEEE Trans. Inf. Theory*, **37**, No. 1, pp. 114-131, Jan. 1991.

[7] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis", *IEEE Trans. Inf. Theory*, **37**, No. 1, pp. 132-141, Jan. 1991.

[8] A. Dan, D. Sitaram, P. Shahabuddin, "Scheduling Policies for an On-Demand Video Server with Batching", *IBM Research Report, RC 19381*, Yorktown Heights, NY, 1994.

[9] N. G. Duffield, "Exponential Bounds for Queues with Markovian Arrivals", Technical Report DIAS-APG-93-01, 1993.

[10] R. S. Ellis, "Large Deviations for a General Class of Random Vectors", *The Annals of Prob.*, **12**, No. 1, pp. 1-12, 1984.

[11] R. S. Ellis, "Entropy, Large Deviations, and Statistical Mechanics". Springer-Verlag, New York, 1985.

[12] A. I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks", *IEEE/ACM Trans. on Networking*, **1**, No. 3, pp. 329-343, Jun. 1993.

[13] D. Ferrari, D. Verma. "A Scheme for Real-Time Channel Establishment in Wide-Area Networks", *IEEE J. Selected Areas in Communications*, **8**, pp. 368-379, April 1990.

[14] W. Fischer and K. Meier-Hellstern, "The Markov-Modulated Poisson Process (MMPP) Cookbook", *Perf. Evaluation*, **18**, pp. 149-172, 1992.

[15] R. Frederick, "nv", Manual Pages, Xerox Palo Alto Research Center.

[16] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks", *IEEE J. Select. Areas Commun.*, **9**, pp. 968-981, 1991.

[17] R. J. Gibbens and P. J. Hunt, "Effective Bandwidths for the Multi-Type UAS Channel", *Queueing Systems*, **9**, pp. 17-28, 1991.

[18] A. Graham, *Kronecker Products and Matrix Calculus with Applications*. Chischester: Ellis Horwood, 1981.

[19] R. A. Horn, and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[20] I. Iscoe, P. Ney and E. Nummelin, "Large Deviations of Uniformly Recurrent Markov Additive Processes", *Adv. in Appl. Math.*, **6**, pp. 373-412, 1985.

[21] V. Jacobson and S. McCanne, "vat", Manual Pages, Lawrence Berkeley Laboratory, Berkeley, CA.

[22] V. Jacobson and S. McCanne, "Using the LBL Network Whiteboard", Lawrence Berkeley Laboratory, Berkeley, CA.

[23] F. P. Kelly, "Effective Bandwidths at Multi-Class Queues", *Queueing Systems*, **9**, pp. 5-16, 1991.

[24] G. Kesidis, J. Walrand and C.-S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Trans. Networking*, **1**, No. 4, pp. 424-428, Aug. 1993.

[25] J. F. C. Kingman, "A Convexity Property of Positive Matrices", *Quart. J. Math. Oxford*, **12**, pp. 283-284, 1961.

[26] J. F. C. Kingman, "A Martingale Inequality in the Theory of Queues", *Camb. Phil. Soc.*, **59**, pp. 359–361, 1964.

[27] J. F. C. Kingman, "Inequalities in the Theory of Queues", *J. Roy. Stat. Soc.*, Series B, **32**, pp. 102-110, 1970.

[28] J. F. Kurose, "On Computing per-Session Performance Bounds in High-Speed Multi-Hop Computer Networks", *Proc. ACM SIGMETRICS and PERFORMANCE'92*, Newport, RI, pp. 128-139, Jun. 1992.

[29] R. M. Loynes, "The Stability of a Queue with Non-Independent Inter-Arrival and Service Times", *Proc. Cambridge Philos. Soc.*, **58**, pp. 497-520, 1962,

[30] R. Nagarajan, J. Kurose, D. Towsley. "Local allocation of end-to-end quality-of-service in high-speed networks", *Proc. 1993 IFIP Workshop on Perf. analysis of ATM Systems*, (H. Perros, ed.), North Holland.

[31] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* John Hopkins University Press, 1981.

[32] L. Padmanabhan, "Design and Implementation of a Shared White-Board", M.S. Project, Dept. of Computer Science, UMass, Amherst, MA 01003, May 1993.

[33] L. Press, "The Internet and Interactive Television", *Communications of the ACM*, **36**, 12, Dec. 1993.

[34] H. Schulzrinne, "Voice Communication Across the Internet: a Network Voice Terminal," Technical Report, Dept. of Computer Science, U. Massachusetts, Amherst MA, July 1992. (Available via anonymous ftp to gaia.cs.umass.edu in pub/nevot/nevot.ps.Z)

[35] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. Oxford, U.K. : Clarendon, 1965.

[36] O. Yaron and M. Sidi, "Performance and Stability of Communication Networks Via Robust Exponential Bounds", *IEEE/ACM Trans. Networking*, **1**, No. 3, pp. 372-385, Jun. 1993.