EXPONENTIAL BOUNDS WITH APPLICATIONS 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

February 14, 1997
To appear in *Journal of the ACM*

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 Markov additive inputs. Specifically, the bounds are on quantities such as backlog, queue length, and response time. Explicit or computable expressions for our bounds are given in the context of queueing theory and numerical comparisons with other bounds and exact results are presented. The paper concludes with two applications to admission control in multimedia systems.

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

*P. Nain was supported in part by NSF under grant NCR-9116183. 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 witnessing a phenomenal growth in the deployment and usage of networked multimedia applications. Numerous networked teleconferencing applications have recently been introduced, [40, 58, 30, 41]. In addition, there are plans to deploy large-scale multimedia servers in the not too distant future, [55]. All of these applications share 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 computable upper and lower bounds on the tail of the distributions of quantities such as backlog, delay and queue length 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 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

$$X_{n+1} = \max(0, X_n + U_n), \qquad \forall n \geq 0 \tag{1.1}$$

with $X_0 \geq 0$ a.s, where the real-valued increments $(U_n)_n$ are modulated by a Markov chain $(Y_n)_n$ such that $(Y_n, \sum_{m=0}^{n} U_m)_n$ is a Markov Additive (MA) process [39]. In our context, one application is when X_n represents the waiting time of the n-th customer in a First-In-First-Out (FIFO) G/G/1 single server queue, $U_n = \sigma_n - \tau_n$, where $(\sigma_n)_n$ and $(\tau_n)_n$ are the service requirement and interarrival time sequences, respectively.

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 strictly positive constants a, a_n, b, b_n and θ such that

$$\begin{aligned} a_n\, e^{-\theta x} &\leq P(X_n > x) \leq b_n\, e^{-\theta x} \\ a\, e^{-\theta x} &\leq P(X > x) \leq b\, e^{-\theta x} \end{aligned}$$

for all $x \geq 0$, $n \geq 0$.

In the particular case where $(\sigma_n)_n$ and $(\tau_n)_n$ are two mutually independent renewal sequences (GI/GI/1 queue), Kingman [44, 45] showed that $a \exp(-\eta x) \leq P(X > x) \leq \exp(-\eta x)$ for all $n \geq 0$ and $x \geq 0$, where η is the unique solution in $(0, \infty)$ of the equation $E[\exp(\theta\,(\sigma_n - \tau_n)] = 1$ under the stability condition $E[\sigma_n - \tau_n] < 0$ (a refinement of Kingman's upper bound was proposed by Ross [57]; see also [8, p. 139]). 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 Modulated Process (MMP) pioneered by Neuts [53] (see, in particular, the works by Regterschot and de Smit on the M/G/1 queue with Markov modulated arrivals and services [56] and by Lucantoni, et al. on the transient analysis of the BMAP/G/1 queue [51], as well as [29] for a recent survey of this area). These computations are often very expensive and do not easily yield the tail probability distribution. Consequently, there has been considerable interest in the development of approximations or asymptotics. These include methods which approximate the arrival processes by simple Markovian models, (e.g., [36]) or fluids, (e.g., [3]), are based on asymptotic properties of statistical multiplexers (e.g., [2]) or on diffusion processes, (e.g., [31]). The problem with these methods is that there is no way of knowing how accurate they are in any one application. This has motivated interest in the development of performance bounds for general arrival processes. This is exemplified by the works of Asmussen and Rolski [7], Chang [12], Cruz [18, 19], Duffield [22], Kurose [46], and Yaron and Sidi [62]. With the exception of the work of Asmussen and Duffield these papers make very few assumptions regarding the arrival processes and the resulting bounds are very loose.

Previous work most closely related to ours include those of Asmussen and Rolski [7] and Duffield [22]. Asmussen and Rolski derived bounds in the context of risk theory and Asmussen [6] showed how they can be mapped into bounds on the tail of the queue length distribution of an MMPP/G/1 queue. Our techniques apply to a larger class of systems. Moreover, as will be described later, our bounds are, in general, better than those in [7]. The mapping described in [6] can be used to apply our bounds to risk theory. Duffield uses a martingale approach (similar to [44] for the G/G/1 queue) to obtain upper bounds similar to ours for a Markovian environment. This approach does not appear easily to yield lower bounds. Neither of the two approaches reported in [7, 22] appear easily to yield bounds on the transient behavior.

We apply our bounds to several systems that have received considerable prior attention. These include the MMPP/E_N/1 queue, the MMPP/D/1 queue and the fixed rate discrete time queue fed by a homogeneous population of on/off sources. For the first two models we present easily computable bounds on the tail of the response time distribution and compare them with the bounds in [7, 22] and the exact distribution. We observe from a large number of examples (see Sections 3.4 and 3.5) that our bounds are usually better than those in [7]. We also observe that the difference between the upper and lower bounds is always smaller than that of [7]. For our examples with 25 homogeneous two-state MMPPs, the times to compute these bounds differ from the times to compute the distribution exactly by two or more orders of magnitude. For the discrete time model, we present easily computable bounds which are then used to address the call admission problem. Comparisons are made with the effective bandwidth approach [35] which illustrate the conservative nature of the latter.

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 of the bounds to queues operating in a Markovian environment are found in Section 3 along with comparisons to the bounds developed in [7, 22]. Applications to discrete time queues and to call admission in multimedia systems are found in Section 4.

2 Exponential Bounds

In this section we derive exponential upper bounds (Section 2.2) and lower bounds (Section 2.3) for the tail distribution of X_n as well as for the tail distribution of its stationary regime X (Section 2.4). We establish these results by extending the approach of Kingman [45] to the multidimensional case using matrix analysis techniques. Prior to deriving the bounds, we introduce some notation.

2.1 Notation and Assumptions

Throughout this paper we assume that the real-valued increments $(U_n)_n$ are modulated by a Markov chain $(Y_n)_n$ such that

(A) U_n and Y_{n+1} conditioned on $(X_0, Y_0, \ldots, Y_n, U_0, \ldots, U_{n-1})$ depend only on Y_n.

We shall assume for the sake of simplicity that the Markov chain $(Y_n)_n$ has a finite state-space $\mathcal{S} = \{1, 2, \ldots, K\}$. The extension of our results to general state-spaces can be found in [47] for the case when $(Y_n, \sum_{m=0}^n U_m)_n$ is an uncoupled MA process (i.e., when in addition to assumption (A), U_n and Y_{n+1} are conditionally independent given Y_n). The case of arbitrary MA processes and general state-spaces can easily be handled by combining the approach in [47] with that in the present paper.

For any Borel set Γ of $(-\infty, \infty)$, $i, j \in \mathcal{S}$, define

$$F_{ij}(\Gamma) = P(Y_{n+1} = j,\, U_n \in \Gamma \,|\, Y_n = i) \tag{2.1}$$

the kernel of the MA process $(S_n, Y_n)_n$, and its transform

$$F_{ij}^*(\theta) = \int_{-\infty}^{\infty} e^{\theta u}\, F_{ij}(du), \quad \theta \in (-\infty, \infty). \tag{2.2}$$

With a slight abuse of notation, $F_{ij}(x)$ will correspond to $F_{ij}((-\infty,x])$.

We assume the Markov chain $(Y_n)_n$ is homogeneous, aperiodic and irreducible, with transition matrix $\mathbf{P}=[p_{ij}]$ (note that $p_{ij}=F_{ij}(\infty)$). The irreducibility of $\mathbf{P}$ implies that Perron-Frobenious theory applies to $\mathbf{F}^*(\theta)$ for all $\theta\in\mathcal{D}$ [39, Section 7(ii)]. Here $\mathcal{D}$ is defined as

$$\mathcal{D}=\left\{\theta\,:\,F^*_{ij}(\theta)<\infty\quad\forall i,j\in\mathcal{S}\right\}$$

As a result, we know that the matrix $\mathbf{F}^*(\theta)$ has a unique left eigenvector $\underline{z}(\theta)=(z_1(\theta),\ldots,z_K(\theta))$, with strictly positive components, corresponding to its largest eigenvalue $\rho(\theta)$ and such that $\sum_{k\in\mathcal{S}} z_k(\theta)=1$ [37, Theorem 8.4.4] (throughout this paper upper case boldface will denote matrices and lower case underlined will denote vectors). In the sequel we will assume that $\theta\in\mathcal{D}$.

To avoid triviality we further assume that the set $\mathcal{M}\subset\mathcal{S}^2$ defined by

$$\mathcal{M}=\left\{(i,j)\in\mathcal{S}^2\,:\,F_{ij}(0)<p_{ij}\right\}\tag{2.3}$$

is nonempty, as otherwise $X_n\to 0$ a.s. as n goes to ∞.

Last, we denote by $\underline{\pi}_n=(\pi_n(0),\ldots,\pi_n(K))$ and $\underline{\pi}=(\pi(0),\ldots,\pi(K))$ the probability distribution vector at time n and the stationary probability distribution vector, respectively, of the Markov chain $(Y_n)_n$. Unless otherwise mentioned, the initial probability distribution vector $\underline{\pi}_0$ is arbitrary in the sense that we do not assume stationarity of the Markov chain $(Y_n)_n$.

2.2 Exponential Upper Bounds

In this section, we derive upper bounds for the tail distribution of X_n. Let $(\gamma_j^n, j\in\mathcal{S})$, $n=0,1,\ldots,$ $\gamma_j^n:[0,\infty)\to[0,\infty)$, be a set of functions such that

$$\sum_{k\in\mathcal{S}}\left[\int_{-\infty}^{x}\gamma_k^n(x-u)\,F_{kj}(du)+(p_{kj}-F_{kj}(x))\,\pi_n(k)\right]\le\gamma_j^{n+1}(x).\tag{2.4}$$

The following result holds:

Proposition 2.1 *Let P_m denote the property that*

$$P(X_m>x\,,\,Y_m=j)\le\gamma_j^m(x)\tag{2.5}$$

for all $x\ge 0$, $j\in\mathcal{S}$.

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

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.

We have for all $x \geq 0$, $j \in \mathcal{S}$,

$$
\begin{aligned}
& P(X_{n+1} > x,\, Y_{n+1} = j) \\
&= \sum_{k\in\mathcal{S}} \pi_n(k)\, P(X_n + U_n > x,\, Y_{n+1} = j \,|\, Y_n = k) \\
&= \sum_{k\in\mathcal{S}} \pi_n(k) \left[\int_{-\infty}^{x} P(X_n > x - u \,|\, U_n = u, Y_n = k, Y_{n+1} = j)\, F_{kj}(du) + p_{kj} - F_{kj}(x)\right] \\
&= \sum_{k\in\mathcal{S}} \left[\int_{-\infty}^{x} P(X_n > x - u, Y_n = k)\, F_{kj}(du) + (p_{kj} - F_{kj}(x))\, \pi_n(k)\right] \qquad (2.6) \\
&\leq \sum_{k\in\mathcal{S}} \left[\int_{-\infty}^{x} \gamma_k^n(x-u)\, F_{kj}(du) + (p_{kj} - F_{kj}(x))\, \pi_n(k)\right] \qquad (2.7) \\
&\leq \gamma_j^{n+1}(x).
\end{aligned}
$$

where (2.7) follows from the induction hypothesis, and where (2.6) is a consequence of assumption (A). ∎

The following result provides an upper bound for $P(X_n > x, Y_n = j)$.

Proposition 2.2 (Exponential upper bound)

If $\rho(\theta) \leq 1$ and if

$$P(X_0 > x, Y_0 = j) \leq C_0(\theta)\, z_j(\theta)\, e^{-\theta x}, \qquad \forall x \geq 0,\ j \in \mathcal{S} \tag{2.8}$$

then, for all $n \geq 0$, $x \geq 0$, $j \in \mathcal{S}$,

$$P(X_n > x, Y_n = j) \quad \leq \quad C_n(\theta)\, z_j(\theta)\, e^{-\theta x} \tag{2.9}$$

with

$$C_n(\theta) = \sup_{\substack{(x,j)\in\mathcal{E} \\ 0\leq m\leq n}} \frac{\displaystyle\sum_{k\in\mathcal{S}} \pi_m(k)\,(p_{kj} - F_{kj}(x))}{\displaystyle\sum_{k\in\mathcal{S}} z_k(\theta) \int_x^{\infty} e^{\theta(u-x)}\, F_{kj}(du)} < \infty \tag{2.10}$$

where $\mathcal{E} = \{(x,j) \in [0,\infty) \times \mathcal{S} \,:\, F_{kj}(x) < p_{kj}$ for some $k \in \mathcal{S}\}$.

In particular,

$$P(X_n > x) \leq C_n(\theta)\, e^{-\theta x}, \qquad \forall x \geq 0,\ n \geq 0. \tag{2.11}$$

Proof. Define

$$\gamma_j^n(x) = C_n(\theta)\, z_j(\theta)\, e^{-\theta x}. \tag{2.12}$$

Thanks to Proposition 2.1 it suffices to prove that the functions in (2.12) satisfy (2.4) to establish (2.9).

We have

$$
\begin{aligned}
&\sum_{k\in\mathcal{S}} \left[\int_{-\infty}^{x} \gamma_k^n(x-u)\, F_{kj}(du) + (p_{kj} - F_{kj}(x))\, \pi_n(k)\right] \\
&= \sum_{k\in\mathcal{S}} \left[\int_{-\infty}^{\infty} C_n(\theta)\, z_k(\theta)\, e^{\theta(u-x)}\, F_{kj}(du) - \int_{x}^{\infty} C_n(\theta)\, z_k(\theta)\, e^{\theta(u-x)}\, F_{kj}(du) + (p_{kj} - F_{kj}(x))\, \pi_n(k)\right] \\
&= e^{-\theta x}\, C_n(\theta) \sum_{k\in\mathcal{S}} F^*_{kj}(\theta)\, z_k(\theta) - \sum_{k\in\mathcal{S}} \left[\int_x^\infty \left(C_n(\theta)\, z_k(\theta)\, e^{\theta(u-x)} - \pi_n(k)\right) F_{kj}(du)\right] \\
&\leq e^{-\theta x}\, C_n(\theta) \sum_{k\in\mathcal{S}} F^*_{kj}(\theta)\, z_k(\theta) \qquad (2.13) \\
&= e^{-\theta x}\, C_n(\theta)\, \rho(\theta)\, z_j(\theta) \qquad (2.14) \\
&\leq e^{-\theta x}\, C_{n+1}(\theta)\, z_j(\theta) = \gamma_j^{n+1}(x) \qquad (2.15)
\end{aligned}
$$

where (2.13), (2.14) and (2.15) follow from the definition of $C_n(\theta)$, the identity $\underline{z}(\theta)\, \mathbf{F}(\theta) = \rho(\theta)\, \underline{z}(\theta)$, and the inequalities $\rho(\theta) \leq 1$ and $C_n \leq C_{n+1}$, respectively. This proves (2.9).

Summing up over all j in $\mathcal{S}$ both sides of (2.9) and using the normalizing condition $\sum_{j=1}^K z_j(\theta) = 1$ yields (2.11).

We conclude this proof by showing that the constant $C_n(\theta)$ is always finite. This property follows from the obvious inequalities

$$
C_n(\theta) \leq \sup_{\substack{(x,j)\in\mathcal{E} \\ 0\leq m\leq n}} \frac{\sum_{k\in\mathcal{S}} \pi_m(k)\, (p_{kj} - F_{kj}(x))}{\sum_{k\in\mathcal{S}} z_k(\theta)\, (p_{kj} - F_{kj}(x))} \leq \sup_{\substack{0\leq m\leq n \\ j\in\mathcal{S}}} \frac{\pi_m(j)}{z_j(\theta)} < \infty \qquad (2.16)
$$

where the last inequality follows from the positiveness of the eigenvector vector $\underline{z}(\theta)$. ∎

Define $\theta^* = \sup\{\theta \in \mathcal{D} : \rho(\theta) \leq 1\}$. An interesting issue is to determine when $\theta^* > 0$ or, equivalently, when does an exponential upper bound exist for the tail distribution of X_n. In the case when the set $\mathcal{D}$ is open, the answer is provided by Duffield [22, Lemma 2] who showed that $\theta^* > 0$ if and only if the stability condition $E_\pi[U_0] < 0$ holds, where E_π denotes the expectation operator associated with a stationary Markov chain $(Y_n)_n$ (i.e., $\underline{\pi}_0 = \underline{\pi}$). This result in turn implies that an exponential upper bound exists for $P(X_n > x)$ if and only if the system is stable (see Remark 2.1). In that case, θ^* is the largest exponential decay rate among all positive decay rates such that $\rho(\theta) \leq 1$. However, this leaves open the question whether θ^* is the best possible decay rate over all $\theta \geq 0$. An affirmative answer to this question again follows from Duffield [22] (see also Glynn and Whitt [33, Theorem 1]) who established that

$$
\lim_{x\to\infty} \frac{\log P(X > x)}{x} = -\theta^* \qquad (2.17)
$$

when the set $\mathcal{D}$ is open. The results in [22] require that the recurrent condition (3.2) in [39] be satisfied. However, this condition is automatically fulfilled when the Markov chain $(Y_n)_n$ has a finite state-space, as observed by Iscoe et al. [39, Section 7(ii)], which therefore validates the use of Duffield's results here. In the case when the state-space is general, then condition (3.2) in [39] must be assumed.

As mentioned above, the results in [22] also require that the set $\mathcal{D}$ be open. While it is not difficult to contruct examples where this assumption is violated, it turns out that a large class of distributions yields an open set $\mathcal{D}$. This class includes the distributions with rational Laplace transforms (e.g. phase-type distributions).

Large deviation results for queues like (2.17) have also been obtained lately by Abate et al. [2], Chang [12], Courcoubetis and Weber [17], de Veciana et al. [21], Duffield and O'Connell [23], Elwalid and al. [27], Kesidis et al. [38], Parulekar and Makowski [54], Simonian and Guibert [59], among others.

Remark 2.1 *When the Markov chain (Y_n) is stationary, the stability condition $E_\pi[U_0] < 0$ follows from Loynes [50]. In the non-stationary case one may use a coupling argument due to Borovkov and Foss [9] to prove that $E_\pi[U_0] < 0$ is also the stability condition or, in other words, that there exists an almost finite r.v. X such that X_n converges in law to X as $n \to \infty$ independently of the joint distribution of X_0 and Y_0.*

2.3 Exponential Lower Bound

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

Proposition 2.3 (Exponential lower bound)

Assume that $\rho(\theta^*) = 1$. If

$$P(X_0 > x, Y_0 = j) \geq B_0\, z_j(\theta^*)\, e^{-\theta^* x}, \qquad \forall x \geq 0, j \in \mathcal{S} \tag{2.18}$$

then, for all $n \geq 0$, $x \geq 0$, $j \in \mathcal{S}$,

$$P(X_n > x, Y_n = j) \geq B_n\, z_j(\theta^*)\, e^{-\theta^* x} \tag{2.19}$$

where

$$B_n = \inf_{\substack{(x,j)\in\mathcal{E}\\ 0\leq m\leq n}} \frac{\sum_{k\in\mathcal{S}} \pi_m(k)\,(p_{kj} - F_k(x))}{\displaystyle\sum_{k\in\mathcal{S}} z_k(\theta^*) \int_x^\infty e^{\theta^*(u-x)}\, F_k(du)}. \tag{2.20}$$

In particular,

$$P(X_n > x) \geq B_n\, e^{-\theta^* x}, \qquad \forall x \geq 0,\ n \geq 0. \tag{2.21}$$

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

$$\sum_{k\in\mathcal{S}} \left[\int_{-\infty}^{x} \delta_k^n(x-u)\, F_{kj}(du) + (p_{kj} - F_k(x))\, \pi_n(k)\right] \geq \delta_j^{n+1}(x) \tag{2.22}$$

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

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

for all $x \geq 0$, $n \geq 0$, $j \in \mathcal{S}$. 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.

Define now the functions $\delta_j^n(x) = B_n\, z_j(\theta^*)\, \exp(-\theta^* x)$. By using the same arguments as in the proof of Proposition 2.2 and the identity $\rho(\theta^*) = 1$, it is easily checked that the functions $\delta_j^n(x)$ satisfy (2.22), from which (2.19) and (2.21) follow. ∎

The equation $\rho(\theta) = 1$ always has one and only one solution $\theta = \theta^*$ in $\mathcal{D} \cap (0,\infty)$ when the set $\mathcal{D}$ is open. This follows from the strict convexity of $\rho(\theta)$ on $\mathcal{D}$ (which itself is a consequence of the strict convexity of $\log \rho(\theta)$ [39, Lemma 3.4(i)]), of $\lim_{\theta \to \delta\mathcal{D}} \rho(\theta) = \infty$ [39, Corollary 3.1], of $\rho(0) = 1$, and of $\rho'(0) = E_\pi[U_0] < 0$.

2.4 Bounds for the Stationary Regime

In this section we determine upper and lower bounds for $P(X > x)$, the stationary tail distribution of X_n, and we discuss cases when the lower bound is not trivial.

Proposition 2.4 (Stationary lower and upper bounds)

Assume that the stability condition $E_\pi[U_0] < 0$ holds (see Remark 2.1). If $\rho(\theta) \leq 1$ then

$$P(X > x) \leq C(\theta)\, e^{-\theta x}, \qquad \forall x \geq 0 \tag{2.23}$$

for all $0 \leq \theta \leq \theta^*$, where

$$C(\theta) = \sup_{(x,j)\in\mathcal{E}} \frac{\sum_{k\in\mathcal{S}} \pi(k)\,(p_{kj} - F_{kj}(x))}{\sum_{k\in\mathcal{S}} z_k(\theta) \int_x^\infty e^{\theta(u-x)}\, F_{kj}(du)}. \tag{2.24}$$

Furthermore, if $\rho(\theta^*) = 1$ then

$$B\, e^{-\theta^* x} \leq P(X > x), \qquad \forall x \geq 0 \tag{2.25}$$

where

$$B = \inf_{(x,j)\in\mathcal{E}} \frac{\sum_{k\in\mathcal{S}} \pi(k)\,(p_{kj} - F_{kj}(x))}{\sum_{k\in\mathcal{S}} z_k(\theta^*) \int_x^\infty e^{\theta^*(u-x)}\, F_{kj}(du)}. \tag{2.26}$$

The proof of Proposition 2.4 follows from Propositions 2.2 and 2.3 and from the result that $P(X_n > x) \to_n P(X > x)$ independently of the joint distribution of X_0 and Y_0 whenever the stability condition $E_\pi[U_0] < 0$ is satisfied, as already discussed in Remark 2.1.

It is simple to construct examples where the constant B in (2.25) is equal to 0. However, we expect $B > 0$ in practice. In the rest of this section we discuss two cases where $B > 0$: the case when the increments $(U_n)_n$ are bounded from above, and the case when they have phase-type distributions.

Our discussion will be based on the following technical lemma whose proof is given in Appendix A.

Lemma 2.1 *For $(j,k) \in \mathcal{M}$, let $\Delta_{kj} = \inf\{x \geq 0,\ F_{kj}(x) = p_{kj}\}$ (see (2.3) for the definition of $\mathcal{M}$). If for every pair of states $(j,k) \in \mathcal{M}$ such that $\Delta_{kj} = \infty$ the constant ξ_{kj} defined by*

$$\xi_{kj} = \liminf_{x \to \infty} \frac{p_{kj} - F_{kj}(x)}{\displaystyle\int_x^\infty e^{\theta^*(u-x)}\, F_{kj}(du)} \tag{2.27}$$

is strictly positive, then $B > 0$.

An immediate corollary of this lemma is that $B > 0$ when the increments $(U_n)_n$ are bounded from above, that is, when $\Delta_{kj} < \infty$ for all $j,k \in \mathcal{M}$.

We now address the case where $F_{kj}(x)$ has a polynomial-exponential density function. A probability density function $f(x)$ of a $(0,\infty)$-valued r.v. is polynomial-exponential if it has the form

$$f(x) = \sum_{i=1}^{n} a_i\, x^{m_i}\, e^{-\beta_i x}, \quad \forall x > 0$$

where a_i's are nonzero real numbers, m_i's nonnegative integers and β_i's are strictly positive real numbers. The set of r.v.'s with polynomial-exponential density functions is quite large and includes, in particular, the set of r.v.'s with phase-type distributions (e.g., Coxian distributions – see [5, pp. 74-75]). Recall that the latter set is dense in the set of probability distributions on $(0,\infty)$ [5, Theorem 6.2, p.76]. The following result, proven in Appendix A, holds:

Corollary 2.1 *If for every $(k,j) \in \mathcal{M}$ either $\Delta_{kj} < \infty$ or $F_{jk}(x)$ has a polynomial-exponential density function, then $B > 0$.*

Instances where F_{kj} has a polynomial-exponential density function and $\Delta_{kj} < \infty$ may be found in Section 3.2 and 3.3, respectively.

3 Application to Queues and Comparison with Other Bounds

In this section we specialize the recursion (1.1) to the case when the increments $(U_n)_n$ are in the form $U_n = \sigma_n - \tau_n$ with $\sigma_n \geq 0$ and $\tau_n \geq 0$. In this setting $(X_n)_n$ typically represents the waiting time process in a FIFO G/G/1 queue with interarrival times $(\tau_n)_n$ and service requirements $(\sigma_n)_n$, and equation (1.1) is called the Lyndley's equation. Our aim is to give explicit formulae for the coefficients $C(\theta)$ and B that appear in the upper bound (2.24) and in the lower bound (2.26), respectively, and to numerically compare these bounds with bounds that have been recently proposed in the literature. This section is organized as follows: in Section 3.1 we derive lower and upper bounds for the tail distribution of the stationary waiting time for queues in Markovian environment; in Sections 3.2 and 3.3 these bounds are specialized to the case of MMPP/E_N/1 and MMPP/D/1 queues, respectively; in Section 3.4 we review bounds proposed by Asmussen and Rolski [7] and Duffield [22] and place them into the context of the queueing models introduced in Section 3.1; Section 3.5 concludes with numerical results and a discussion on the tightness of the various bounds presented in Sections 3.2-3.4.

3.1 Bounds for Queues in Markovian Environment

We assume that customers arrive at a FIFO single server queue according to a Markov modulated Poisson process $(t_n)_n$ [29]. More precisely, we assume that the arrival process is a doubly stochastic Poisson process with arrival rate $\lambda_{Z(t)}$ at time t, where $(Z(t),\, t \geq 0)$ is an irreducible Markov process on the set $\mathcal{S} = \{1, 2, \ldots, K\}$, with infinitesimal generator $\mathbf{Q} = [q_{ij}]$, rate matrix $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_K)$, and invariant measure $\underline{q} = (q(1), \ldots, q(K))$.

Service requirements $(\sigma_n)_n$ are also modulated according to the Markov process $(Z(t),\, t \geq 0)$ in the sense that the probability distribution of the service requirement of the n-th customer, σ_n, may depend on $Z(t_n-)$. We denote by $H_i(x) = P(\sigma_n \leq x \,|\, Z(t_n-) = i)$ and $H_i^*(\theta) = E[\exp(\theta\, \sigma_n) \,|\, Z(t_n-) = i]$ the probability distribution and the Laplace transform of σ_n, respectively, given that $Z(t_n-) = i$. We also assume that the service requirements are mutually independent r.v.'s, and that the service requirement σ_n is independent of the state $Z(t_{n+1}^-)$ and interarrival time τ_n given $Z(t_n^-)$. Last, we will assume that the queue is stable [29].

In order to apply the results in Section 2 we need to identify the Markov chain $(Y_n)_n$, the kernel (2.1) and its transform (2.2). In this setting, it is easy to see that the Markov chain $(Y_n)_n$ must be chosen as the Markov chain embedded in $(Z(t),\, t \geq 0)$ at arrival instants, that is $Y_n = Z(t_n-)$. Its transition matrix $\mathbf{P}$ is given by (see [29])

$$\mathbf{P} = (\mathbf{\Lambda} - \mathbf{Q})^{-1}\, \mathbf{\Lambda}. \tag{3.1}$$

Let us determine the kernel $F_{ij}(x)$. We first observe that $F_{ij}(x)$ needs only to be determined for $x \geq 0$ as the supremum and the infimum in (2.24) and (2.26), respectively, are only taken over nonnegative values of x. We have, for $x \geq 0$,

$$\begin{aligned} F_{ij}(x) &= p_{ij} - P(Y_{n+1} = j,\, \tau_n < \sigma_n - x \,|\, Y_n = i) \\ &= p_{ij} - \int_x^\infty P(Y_{n+1} = j,\, \tau_n < y - x \,|\, Y_n = i)\, dH_i(y) \end{aligned} \tag{3.2}$$

or, in matrix notation,

$$\mathbf{F}(x) = \mathbf{P} - \int_x^\infty d\mathbf{H}(y)\, \mathbf{G}(y-x), \qquad \forall x \ge 0 \tag{3.3}$$

with $\mathbf{F}(x) = [F_{ij}(x)]$, $\mathbf{H}(x) = \text{diag}\,(H_i(x),\, i \in \mathcal{S})$ and $\mathbf{G}(x) = [P(Y_{n+1} = j,\, \tau_n < x \,|\, Y_n = i)]$. It is known (see [29, formula (5)] for instance) that

$$\mathbf{G}(x) = \mathbf{P} - \exp((\mathbf{Q} - \mathbf{\Lambda})\, x)\, \mathbf{P}, \qquad \forall x \ge 0 \tag{3.4}$$

so that, from (3.3),

$$\mathbf{F}(x) = \mathbf{P} - \int_x^\infty d\mathbf{H}(y)\; (\mathbf{I} - \exp((\mathbf{Q} - \mathbf{\Lambda})\,(y-x)))\, \mathbf{P}, \qquad \forall x \ge 0 \tag{3.5}$$

where $\mathbf{I}$ stands for the identity matrix. This, in turn, implies that

$$d\mathbf{F}(x) = \int_x^\infty d\mathbf{H}(y)\; \exp((\mathbf{Q} - \mathbf{\Lambda})\,(y-x))\, \mathbf{\Lambda}\; dx, \qquad \forall x \ge 0 \tag{3.6}$$

by using the identity $(\mathbf{\Lambda} - \mathbf{Q})\, \mathbf{P} = \mathbf{\Lambda}$.

We are now in position to write down the coefficients $C(\theta)$ and B (see Proposition 2.4). In matrix form, these coefficients become by using (3.5) and (3.6)

$$C(\theta) = \sup_{(x,j)\in\mathcal{E}} g_j(x,\theta), \qquad B = \inf_{(x,j)\in\mathcal{E}} g_j(x,\theta^*) \tag{3.7}$$

with

$$g_j(x,\theta) = \frac{\underline{\pi}\,\left(\int_x^\infty d\mathbf{H}(u)\,(\mathbf{I} - \exp((\mathbf{Q}-\mathbf{\Lambda})(u-x))\right)\mathbf{P}\,\underline{e}_j}{\underline{z}(\theta)\,\left(\int_x^\infty e^{\theta\,(u-x)}\int_u^\infty d\mathbf{H}(y)\;\exp((\mathbf{Q}-\mathbf{\Lambda})(y-u))\,du\right)\mathbf{\Lambda}\,\underline{e}_j} \tag{3.8}$$

for $0 \le \theta \le \theta^*$, where $\underline{e}_j$ is the vector whose components are 0 except the j-th one which is equal to 1.

Let us now determine the matrix $\mathbf{F}^*(\theta)$ for $\theta \in \mathcal{D} \cap [0,\infty)$. Since, for all $n \ge 0$, σ_n is independent of τ_n, given Y_n, we clearly have

$$\begin{aligned} \mathbf{F}^*(\theta) &= \mathbf{H}^*(\theta) \int_0^\infty e^{-\theta\,x}\, d\mathbf{G}(x) \\ &= \mathbf{H}^*(\theta)\,(\theta\,\mathbf{I} + \mathbf{\Lambda} - \mathbf{Q})^{-1}\,\mathbf{\Lambda}, \qquad \forall \theta \in \mathcal{D}\cap[0,\infty) \end{aligned} \tag{3.9}$$

with $\mathbf{H}^*(\theta) = \text{diag}\,(H_i^*(\theta),\, i \in \mathcal{S})$, where (3.9) follows from (3.4).

>From (3.9) we may compute the left-eigenvector $\underline{z}(\theta)$ of $\mathbf{F}^*(\theta)$ corresponding to the largest eigenvalue $\rho(\theta)$, and derive the optimal exponential decay rate θ^* as the unique solution in $(0,\infty)$ of the equation $\rho(\theta) = 1$.

We conclude this subsection by briefly discussing the case when the interarrival and customer requirement sequences are mutually independent renewal sequences (GI/GI/1 queue). In this case, our lower bound (2.25) reduces to the lower bound found by Kingman [45] and the upper bounds (2.23) reduce to the upper bounds derived by Ross [57] (see also [60]). In particular, the lower bound and the upper bound in (2.25) are equal when the service times are exponentially distributed (GI/M/1 queue).

3.2 Bounds for the MMPP/E_N/1 Queue

We consider the queueing model defined in Section 3.1 but we now assume that the service requirements $(\sigma_n)_n$ form a renewal sequence, independent of the arrival process, with common distribution function $H(x)$ given by an N-stage Erlang probability distribution (MMPP/E_N/1 queue), namely, $H(x) = 1 - e^{-\mu x} \sum_{l=0}^{N-1} (\mu x)^l / l!$.

This assumption implies, in particular, that (cf. (3.9)) $\mathbf{F}^*(\theta) = (\mu/(\mu-\theta))^N \, (\theta\, \mathbf{I} + \mathbf{\Lambda} - \mathbf{Q})^{-1}\, \mathbf{\Lambda}$ for all $\theta \in \mathcal{D} \cap [0,\infty) = [0,\mu)$.

Recall the definition of $g_j(x,\theta)$ (see (3.8)). Straightforward algebra yields

$$g_j(x,\theta) = \left(\frac{\mu-\theta}{\mu}\right)^N \frac{\underline{\pi}\, \mathbf{\Psi}(x,\mu,N) \mathbf{\Delta}\, \underline{e}_j}{\underline{z}(\theta)\, \mathbf{\Psi}(x,\mu-\theta,N) \mathbf{\Delta}\, \underline{e}_j} \tag{3.10}$$

where

$$\mathbf{\Psi}(x,\mu,N) = \sum_{l=0}^{N-1} \sum_{r=0}^{l} \frac{(x\,\mu)^r}{r!} \,(\mu\, \mathbf{\Delta})^{N-1-l}, \quad x \ge 0,\ n = 1, \ldots$$

and $\mathbf{\Delta} = (\mu\, \mathbf{I} + \mathbf{\Lambda} - \mathbf{Q})^{-1}$ (hint: use the identity $(\mathbf{I} - \mu\, \mathbf{\Delta})\, \mathbf{P} = \mathbf{\Delta}\, \mathbf{\Lambda}$).

The complexity of computing $C(\theta)$ (resp. $B(\theta^*)$) is dominated by the search for the value of x that yields the supremum (resp. infimum) in the expression (3.7). In the case of the Erlang distribution, it is easily shown that no more than $N-1$ values of x (possibly including $x=0$ and $x=\infty$) need to be checked and that, except for $x = 0, \infty$, they are the positive real roots of the polynomial

$$\phi_j(x) := \underline{\pi} \left(\frac{d\mathbf{\Psi}(x,\mu,N)}{dx} \mathbf{\Delta} \underline{e}_j \underline{z}\, \mathbf{\Psi}(x,\mu-\theta,N) - \mathbf{\Psi}(x,\mu,N) \mathbf{\Delta} \underline{e}_j \underline{z}\, \frac{d\mathbf{\Psi}(x,\mu-\theta,N)}{dx} \right) \mathbf{\Delta} \underline{e}_j$$

which can be shown to be of degree $2(N-2)$.

For the case $N = 1$ (MMPP/M/1 queue), $g_j(x,\theta)$ does not depend on x and we have

$$C(\theta) = \left(\frac{\mu-\theta}{\mu}\right) \max_{j\in\mathcal{S}} \frac{\underline{\pi}\, \mathbf{\Delta}\, \underline{e}_j}{\underline{z}(\theta)\, \mathbf{\Delta}\, \underline{e}_j}, \qquad B = \left(\frac{\mu-\theta^*}{\mu}\right) \min_{j\in\mathcal{S}} \frac{\underline{\pi}\, \mathbf{\Delta}\, \underline{e}_j}{\underline{z}(\theta^*)\, \mathbf{\Delta}\, \underline{e}_j}. \tag{3.11}$$

For the case $N = 2$, $\phi_j(x)$ is a constant and we have been able to establish that the supremum (resp. infimum) of $g_j(x,\theta^*)$ over x in $[0,\infty)$ is always achieved at $x = \infty$ (resp. $x = 0$). For the case $N > 2$ (MMPP/E$_N$/1 queue) we conjecture that $\phi_j(x)$ has no positive real roots. We further conjecture that the supremum (resp. infimum) of $g_j(x, \theta^*)$ over x in $[0, \infty)$ is always reached at $x = \infty$ (resp. $x = 0$). These conjectures have always checked true in all the numerical experiments we have performed using the Maple V[1] software for symbolic computation. When the last conjecture holds, then

$$\begin{aligned} C(\theta^*) &= \left(\frac{\mu - \theta^*}{\mu}\right) \max_{j \in \mathcal{S}} \frac{\underline{\pi}\, \mathbf{\Delta}\, \underline{e}_j}{\underline{z}\,(\theta^*)\, \mathbf{\Delta}\, \underline{e}_j} & (3.12) \\ B &= \left(\frac{\mu - \theta^*}{\mu}\right)^N \min_{j \in \mathcal{S}} \frac{\underline{\pi}\, (\mathbf{I} - (\mu \mathbf{\Delta})^N)\, (\mathbf{I} - \mu\, \mathbf{\Delta})^{-1}\, \mathbf{\Delta}\, \underline{e}_j}{\underline{z}\,(\theta^*)\, (\mathbf{I} - ((\mu - \theta^*)\, \mathbf{\Delta})^N)\, (\mathbf{I} - (\mu - \theta^*)\, \mathbf{\Delta})^{-1}\, \mathbf{\Delta}\, \underline{e}_j}. & (3.13) \end{aligned}$$

3.3 Bounds for the MMPP/D/1 Queue

We now specialize the queueing model in Section 3.1 to the case when the service requirements $(\sigma_n)_n$ are all equal to the same constant s (MMPP/D/1 queue). Then, cf. (3.9), $\mathbf{F}^*(\theta) = e^{\theta\, s}\, (\theta\, \mathbf{I} + \mathbf{\Lambda} - \mathbf{Q})^{-1}\, \mathbf{\Lambda}$ for all $\theta \in \mathcal{D} \cap [0, \infty) = [0, \infty)$.

In this case, $C(\theta) = \sup_{0 \le x < s,\, j \in \mathcal{S}} g_j(x, \theta)$ and $B = \inf_{0 \le x < s,\, j \in \mathcal{S}} g_j(x, \theta^*)$, and it is not difficult to show that

$$g_j(x, \theta) = \frac{\underline{\pi}\, (\mathbf{I} - \exp((\mathbf{Q} - \mathbf{\Lambda})\, (s - x))\, (\mathbf{\Lambda} - \mathbf{Q})^{-1}\, \underline{e}_j}{\underline{z}\,(\theta)\, (\mathbf{I} \exp(-\theta(s - x) - \exp((\mathbf{Q} - \mathbf{\Lambda})\, (s - x)))\, (\theta\, \mathbf{I} + \mathbf{\Lambda} - \mathbf{Q})^{-1}\, \underline{e}_j} \qquad (3.14)$$

for $0 \le x < s$.

Again, we conjecture that the supremum (resp. infimum) in $g_j(x, \theta^*)$ is always reached for $x = s$ (resp. $x = 0$) as this has always been observed through our experiments. In particular, it is true for the M/D/1 queue. When this is true, then $C(\theta^*)$ takes the simple form

$$C(\theta^*) = \max_{j \in \mathcal{S}} \frac{\pi(j)}{z_j(\theta^*)}.$$

3.4 Other Bounds for Queues in a Markovian Environment

In this section we review bounds recently proposed by Asmussen and Rolski [6] and Duffield [22].

The bounds proposed by Asmussen and Rolski [7] have been derived in the context of risk theory. In the queueing setting of Section 3.1 Asmussen and Rolski's bounds read [7, Corollary 4.1], [6, Theorem 3.8]:

$$\sum_{k \in \mathcal{S}} q(k)\, h_k(\gamma^*)\, C_-(k)\, e^{-\gamma^*\, x} \le P(X \ge x) \le \sum_{k \in \mathcal{S}} q(k)\, h_k(\gamma^*)\, C_+(k)\, e^{-\gamma^*\, x}, \quad \forall x \ge 0 \qquad (3.15)$$

[1] Maple V is a registered trademark of Waterloo Maple Software.

with

$$C_{+}(k) = \max_{j\in\mathcal{S}} \frac{1}{h_j(\gamma^*)} \sup_{x\geq 0} \frac{1-H_k(x)}{\displaystyle\int_x^\infty e^{\gamma^*(u-x)}\, dH_k(u)} \tag{3.16}$$

$$C_{-}(k) = \min_{j\in\mathcal{S}} \frac{1}{h_j(\gamma^*)} \inf_{x\geq 0} \frac{1-H_k(x)}{\displaystyle\int_x^\infty e^{\gamma^*(u-x)}\, dH_k(u)}. \tag{3.17}$$

Note that bounds in [5] are only available for the stationary regime and for Markov chains $(Y_n)_n$ with a finite state-space. To define the unknown quantities γ^* and $h_k(\gamma^*)$ in (3.15)-(3.17), introduce the matrix $\mathbf{M}(\gamma)=\mathbf{S}(\gamma)+\mathbf{Q}^*-\gamma\,\mathbf{I}$, where $\mathbf{S}(\gamma)=\mathrm{diag}\,(\lambda_i\,(H_i^*(\gamma)-1),\ i\in\mathcal{S})$ and where $\mathbf{Q}^*=[q^*_{ij}]$ with $q^*_{ij}=q(j)\,q_{ji}/q(i)$ for $i\neq j$, is the infinitesimal generator of the reversed Markov process $(Z(t),\ t\geq 0)$. Let $h(\gamma)=(h_1(\gamma),\ldots,h_K(\gamma))^T$ be the right-eigenvector of the matrix $\mathbf{M}(\gamma)$ corresponding to the eigenvalue $\kappa(\gamma)$ with the largest real part. Then, when the queue is stable, γ^* is the unique solution in $(0,\infty)$ of the equation $\kappa(\gamma)=0$. It can be shown that $\theta^*=\gamma^*$. Note that this result directly follows from inequalities (2.25) and (3.15) whenever $B>0$ and $\sum_{k\in\mathcal{S}} q(k)\,h_k(\gamma^*)\,C_-(k)>0$. Last, when the service times are either deterministic or taken from an Erlang distribution, it can be shown that the sup and inf in (3.16) and (3.17) are achieved at $x=\infty$ and $x=0$ respectively.

In [22] Duffield derived a set of upper bounds for the tail distribution of the stationary regime X of a stochastic process (X_n) defined by the recursion (1.1) under the assumptions that $(Y_n,\sum_{m=0}^n U_m)$ is a MA process (same assumption as ours) and that the Markov chain $(Y_n)_n$ is stationary (we do not impose this condition). Specializing Duffield's bounds to the queueing model of Section 3.1 yields, for $\lambda_i>0$ for all $i\in\mathcal{S}$,

$$P(X\geq x)\leq D(\theta)\,e^{-\theta\,x}, \qquad \forall x\geq 0 \tag{3.18}$$

for all $\theta\in[0,\theta^*]$, with

$$D(\theta)=\sup_{j\in\mathcal{S}} \frac{1}{r_j(\theta)}, \tag{3.19}$$

where $r_j(\theta)$ is the j-th component of the unique vector $\underline{r}\,(\theta)$ satisfying the relations $F^*(\theta)\,\underline{r}\,(\theta)=\rho(\theta)\,\underline{r}\,(\theta)$ and $\sum_{i\in\mathcal{S}}\pi(i)\,r_i(\theta)=1$.

Observe that $D(\theta)\geq 1$ (since $\sum_{i\in\mathcal{S}}\pi(i)\,r_i(\theta)=1$) as opposed to the coefficients in Asmussen and Rolski's upper bound and in ours which may be smaller than one (see numerical results in Section 3.5 for $x=0$).

However, it is difficult in general to analytically compare the bounds in [7] and in [22] to ours since they appear in very different forms (see (2.25) where B and $C(\theta)$ are given in (3.7)-(3.8), (3.15), and (3.18)), which is a consequence of the fact that they have been derived using very different techniques: risk theory and Lundberg's inequalities for Asmussen and Rolski, martingales and large deviations for Duffield, and the extension of Kingman's method for our bounds. A comparison based on numerical results is presented in the next section.

Other (upper) bounds have also been recently obtained by Chang [12], and Yaron and Sidi [62], for queues with very general arrival patterns. These bounds, based on Chernoff's inequality, are in general not as tight as our bounds.

3.5 Numerical Results and Discussion

In this section we report numerical experiments performed for various queueing models with MMPP arrival processes. More precisely, we assume that the arrival process is the superposition of M ($M = 25$ in the tables) independent, homogeneous, two-state MMPP's. Observe that the superposition of independent MMPP's is again an MMPP (see [29] for instance) so that the results obtained so far in this section apply.

Tables 3.1–3.4 display our lower and upper bounds (LNT l.b/u.b.; see Sections 3.2, 3.3) and Asmussen and Rolski's lower and upper bounds (AR l.b./u.b.; see (3.15)) for the tail distribution, $P(X > x)$, of the stationary waiting time for MMPP/M/1, MMPP/E$_2$/1, MMPP/E$_5$/1, and MMPP/D/1 queues respectively. These bounds have been computed for different values of the traffic intensity ρ ($\rho \in \{0.4, 0.75, 0.95\}$) and for various values of x. In each case, the mean service time is 1.

Before commenting on the numerical results we first describe an efficient way to compute θ^* and $\underline{z}(\theta^*)$ as a brute force approach may not be applicable for large values of M. The optimal decay rate θ^* was computed by using the "effective bandwidth decomposition" for MMPP/GI/1 queues fed by M independent MMPP's, that is, when $(\mathbf{Q}, \mathbf{\Lambda}) = \left(\oplus_{m=1}^{M} \mathbf{Q}_m, \oplus_{m=1}^{M} \mathbf{\Lambda}_m\right)$. In this case the optimal decay rate θ^* satisfies the equation

$$\theta^* = \sum_{m=1}^{M} pf\left((H^*(-\theta^*) - 1)\,\mathbf{\Lambda}_m + \mathbf{Q}_m\right) \tag{3.20}$$

where $H^*(\theta)$ is the Laplace-Stieltjes transform of the service times (as usual $pf(\mathbf{A})$ denotes the Perron-Frobenius eigenvalue of the matrix $\mathbf{A}$). Equation (3.20) follows (for instance) from [61, eq. (6.22) and Proposition 14] (see also [26] and [43] among others).

In the case that the arrival process is the superposition of M independent *homogeneous two-state* MMPP's each with infinitesimal generator and rate matrix given by

$$\mathbf{Q}_m = \begin{pmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{pmatrix}, \qquad \mathbf{\Lambda}_m = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

respectively, then (3.20) reduces to

$$\theta^* = M\,\frac{-(q + \lambda\,\overline{H}^*(\theta^*)) + \sqrt{(q + \lambda\,\overline{H}^*(\theta^*))^2 - 4\left(\prod_{i=1}^{2}(q_i + \lambda_i\,\overline{H}^*(\theta^*)) - q_1 q_2\right)}}{2} \tag{3.21}$$

with $\overline{H}^*(\theta) = 1 - H^*(-\theta)$, $q = q_1 + q_2$ and $\lambda = \lambda_1 + \lambda_2$. Various efficient numerical procedures can be used for computing θ^* from (3.21).

Let us now turn to the computation of $\underline{z}\,(\theta^*)$. Depending on the dimension of the matrix $\mathbf{F}^*(\theta)$ (see (3.9); here $\mathbf{F}^*(\theta) = H^*(-\theta)\,(\theta\,\mathbf{I} + \mathbf{\Lambda} - \mathbf{Q})^{-1}\,\mathbf{\Lambda}$ with our notation) various approaches (including a brute force approach and an iterative approach) may be used for computing $\underline{z}\,(\theta^*)$, the (normalized) left-eigenvector of $\mathbf{F}^*(\theta^*)$ associated with the eigenvalue 1. In the case that the input process is the superposition of M independent homogeneous two-state MMPP's then the input process can be modeled as a $(M+1)$-state MMPP with infinitesimal generator $\mathbf{Q} = [Q_{i,j}]$ and rate matrix $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda^1, \ldots, \lambda^{M+1})$, where $\mathbf{Q}$ is a tridiagonal matrix with $Q_{i,i-1} = (i-1)\,q_2$ for $2 \le i \le M$, $Q_{i,i+1} = (M+1-i)\,q_1$ for $1 \le i \le M$, and where $\lambda^i = (M-i)\,\lambda_1 + i\,\lambda_2$ for $1 \le i \le M+1$.

This corresponds to aggregating all MMPP's into a single MMPP with arrival rates depending on the number of MMPP's in state 2 at any time in the initial system. In this case, $\underline{z}\,(\theta^*)$ can be computed explicitly as briefly discussed below.

We first observe from the definition of $\underline{z}\,(\theta^*)$ that $\underline{z}\,(\theta^*) = \underline{\phi}^T\,\mathbf{\Lambda}$ where $\underline{\phi}$ satisfies the equation $H^*(-\theta^*)\,\mathbf{\Lambda}\,\underline{\phi} = \mathbf{M}\underline{\phi}$, with $\mathbf{M} = \theta^*\,\mathbf{I} + \mathbf{\Lambda} - \mathbf{Q}^T$. By noting that $\mathbf{M}$ as the same structure as the matrix $\mathbf{M}$ in [3, p. 1875] a similar analysis to that in [3, Section 2.1] allows us to obtain the j-th component $(1 \le j \le M+1)$ of the vector $\underline{\phi}$ as the coefficient of x^{j-1} in the polynomial $(x - x_1)^k\,(x - x_2)^{M-k}$ (assuming that $\phi_{M+1} = 1$), where $r_2 < 0 < r_1$ are the roots of the polynomial $q_1\,x^2 + b\,x - q_2$ with $b = q_2 + \lambda_2 - (q_1 + \lambda_1) + (\lambda_1 - \lambda_2)\,\theta^*\,H^*(-\theta^*)$. The integer k (see [3]) is given by $k = \left(q_1 M(x_1 - 1) - \lambda_1 M\overline{H}^*(\theta^*) - \theta^*\right)/(q_1\,(x_1 - x_2))$. The vector $\underline{z}\,(\theta^*)$ is now obtained by normalizing the vector $\underline{\phi}^T\,\mathbf{\Lambda}$ so that its components sum up to one.

In general, we observe that the tightness of the bounds increases as the traffic intensity and the variability of the service times increase and, in particular, our bounds appear to be very tight for the MMPP/M/1 queue. Although our lower bound is always better than Asmussen and Rolski's, our upper bound is sometimes looser (in the case of deterministic service times). We observe that the gap between our lower and upper bounds is always (occasionally considerably) smaller than the corresponding gap for Asmussen and Rolski's bounds. Last, we omitted Duffield's upper bound because as we have always observed it to be worse than ours and that of Asmussen and Rolski.

We also include exact results obtained by inverting the Laplace transform of the tail of the wait time distribution [29, 15]. The inversion is performed using the EULER algorithm described in [1] using parameter values $A = 19.1$ and $m = 11$. The third parameter of the algorithm, n was set to 100 in the case of exponential and Erlang distributions. Because we had difficulty determining a set of parameter values with which to generate the exact solution for a deterministic service time, we approximate it by an Erlang random variable with 8192 stages. For this case, we set the parameter n to be 200. We found the our bounds valuable in helping us to set these values (see [1] for a further discussion on the use of bounds for computing exact distributions). Our experience has been that the time required to calculate $P(X > x)$ exactly for 8-12 values of x is 100 to 600 times more expensive than calculating the bounds for the systems presented in Tables 3.1–3.4. The exact calculation of a larger number of values is even more expensive as the cost is linear in the number of values whereas it is essentially constant (the cost of calculating B and $C(\theta^*)$) for the bounding methods.

Similar observations regarding the quality of the bounds hold for other values of the number of sources. We have not compared the costs of computing our bounds with those of computing the distribution exactly in the case of a heterogeneous population of sources. However, if it consists of independent two-state sources, the computation requirements will not increase significantly as it differs from that described above in the determination of the eigenvector $\underline{z}(\theta^*)$; it must now be obtained by solving a set of linear equations. Although the computation costs for an exact solution will not change, they will still remain significantly greater than for the bounds.

We conclude this section by noting that even more striking differences in computation costs have been observed when our bounding approach was adapted to a heterogeneous population of Markov on/off fluid sources [4].

x	0	100	200	300	400
LNT u.b.	0.951	$0.654 \ 10^{-2}$	$0.449 \ 10^{-4}$	$0.308 \ 10^{-6}$	$0.212 \ 10^{-8}$
AR u.b	0.955	$0.656 \ 10^{-2}$	$0.450 \ 10^{-4}$	$0.309 \ 10^{-6}$	$0.213 \ 10^{-8}$
LNT l.b.	0.945	$0.649 \ 10^{-2}$	$0.446 \ 10^{-4}$	$0.306 \ 10^{-6}$	$0.210 \ 10^{-8}$
AR l.b.	0.937	$0.644 \ 10^{-2}$	$0.442 \ 10^{-4}$	$0.304 \ 10^{-6}$	$0.209 \ 10^{-8}$
Exact	0.950	$0.653 \ 10^{-2}$	$0.448 \ 10^{-4}$	$0.308 \ 10^{-6}$	$0.211 \ 10^{-8}$

(a) $\rho = 0.95$ ($\lambda_1 = 0.6,\ \lambda_2 = 2,\ q_{12} = 1,\ q_{21} = 3$)

x	0	12	36	48	72
LNT u.b.	0.753	$3.759 \ 10^{-2}$	$0.937 \ 10^{-4}$	$4.676 \ 10^{-6}$	$1.165 \ 10^{-8}$
AR u.b.	0.760	$3.792 \ 10^{-2}$	$0.945 \ 10^{-4}$	$4.718 \ 10^{-6}$	$1.176 \ 10^{-8}$
LNT l.b.	0.740	$3.695 \ 10^{-2}$	$0.921 \ 10^{-4}$	$4.597 \ 10^{-6}$	$1.146 \ 10^{-8}$
AR l.b.	0.723	$3.608 \ 10^{-2}$	$0.899 \ 10^{-4}$	$4.488 \ 10^{-6}$	$1.119 \ 10^{-8}$
Exact	0.750	$3.745 \ 10^{-2}$	$0.933 \ 10^{-4}$	$4.659 \ 10^{-6}$	$1.161 \ 10^{-8}$

(b) $\rho = 0.75$ ($\lambda_1 = 0.6,\ \lambda_2 = 1.2,\ q_{12} = 1,\ q_{21} = 3$)

x	0	3	12	24	30
LNT u.b.	0.403	$6.670 \ 10^{-2}$	$3.026 \ 10^{-4}$	$0.227 \ 10^{-6}$	$0.623 \ 10^{-8}$
AR u.b.	0.413	$6.832 \ 10^{-2}$	$3.099 \ 10^{-4}$	$0.233 \ 10^{-6}$	$0.638 \ 10^{-8}$
LNT l.b.	0.387	$6.401 \ 10^{-2}$	$2.904 \ 10^{-4}$	$0.218 \ 10^{-6}$	$0.597 \ 10^{-8}$
AR l.b.	0.355	$5.880 \ 10^{-2}$	$2.667 \ 10^{-4}$	$0.200 \ 10^{-6}$	$0.549 \ 10^{-8}$
Exact	0.400	$6.695 \ 10^{-2}$	$3.001 \ 10^{-4}$	$0.226 \ 10^{-6}$	$0.619 \ 10^{-8}$

(c) $\rho = 0.4$ ($\lambda_1 = 0.3,\ \lambda_2 = 0.8,\ q_{12} = 1,\ q_{21} = 4$)

Table 3.1: Bounds and exact value for $P(X > x)$ for MMPP/M/1 queue with 25 homogenous two state sources; (a) $\rho = 0.95$, (b) $\rho = 0.75$, (c) $\rho = 0.4$.

x	0	50	150	250	300
LNT u.b.	0.969	$3.451 \ 10^{-2}$	$0.438 \ 10^{-4}$	$0.555 \ 10^{-7}$	$0.198 \ 10^{-8}$
AR u.b .	0.973	$3.464 \ 10^{-2}$	$0.439 \ 10^{-4}$	$0.557 \ 10^{-7}$	$0.199 \ 10^{-8}$
LNT l.b.	0.938	$3.340 \ 10^{-2}$	$0.424 \ 10^{-4}$	$0.537 \ 10^{-7}$	$0.191 \ 10^{-8}$
AR l.b.	0.917	$3.267 \ 10^{-2}$	$0.414 \ 10^{-4}$	$0.525 \ 10^{-7}$	$0.187 \ 10^{-8}$
Exact	0.950	$3.404 \ 10^{-2}$	$0.432 \ 10^{-4}$	$0.548 \ 10^{-7}$	$0.195 \ 10^{-8}$

(a) $\rho = 0.95$ ($\lambda_1 = 0.6$, $\lambda_2 = 2$, $q_{12} = 1$, $q_{21} = 3$)

x	0	18	30	42	54
LNT u.b.	0.834	$0.172 \ 10^{-2}$	$0.277 \ 10^{-4}$	$0.449 \ 10^{-6}$	$0.726 \ 10^{-8}$
AR u.b.	0.843	$0.173 \ 10^{-2}$	$0.280 \ 10^{-4}$	$0.453 \ 10^{-6}$	$0.734 \ 10^{-8}$
LNT l.b.	0.720	$0.148 \ 10^{-2}$	$0.240 \ 10^{-4}$	$0.388 \ 10^{-6}$	$0.627 \ 10^{-8}$
AR l.b.	0.651	$0.134 \ 10^{-2}$	$0.217 \ 10^{-4}$	$0.351 \ 10^{-6}$	$0.567 \ 10^{-8}$
Exact	0.750	$0.160 \ 10^{-2}$	$0.258 \ 10^{-4}$	$0.419 \ 10^{-6}$	$0.677 \ 10^{-8}$

(b) $\rho = 0.75$ ($\lambda_1 = 0.6$, $\lambda_2 = 1.2$, $q_{12} = 1$, $q_{21} = 3$)

x	0	6	12	18	21
LNT u.b.	0.567	$0.285 \ 10^{-2}$	$0.143 \ 10^{-4}$	$0.718 \ 10^{-7}$	$0.509 \ 10^{-8}$
AR u.b.	0.584	$0.293 \ 10^{-2}$	$0.147 \ 10^{-4}$	$0.740 \ 10^{-7}$	$0.525 \ 10^{-8}$
LNT l.b.	0.359	$0.180 \ 10^{-2}$	$0.905 \ 10^{-5}$	$0.454 \ 10^{-7}$	$0.322 \ 10^{-8}$
AR l.b.	0.262	$0.132 \ 10^{-2}$	$0.661 \ 10^{-5}$	$0.332 \ 10^{-7}$	$0.235 \ 10^{-8}$
Exact	0.400	$0.232 \ 10^{-2}$	$0.117 \ 10^{-4}$	$0.585 \ 10^{-7}$	$0.415 \ 10^{-8}$

(c) $\rho = 0.4$ ($\lambda_1 = 0.3$, $\lambda_2 = 0.8$, $q_{12} = 1$, $q_{21} = 4$)

Table 3.2: Bounds and exact values for $P(X > x)$ for MMPP/E_2/1 queue with 25 homogenous 2 state sources; (a) $\rho = 0.95$, (b) $\rho = 0.75$, (c) $\rho = 0.4$.

x	0	90	150	210	270
LNT u.b.	0.988	0.527 10^{-3}	0.346 10^{-5}	0.228 10^{-7}	0.150 10^{-9}
AR u.b.	0.991	0.529 10^{-3}	0.348 10^{-5}	0.229 10^{-7}	0.150 10^{-9}
LNT l.b.	0.931	0.497 10^{-3}	0.327 10^{-5}	0.215 10^{-7}	0.141 10^{-9}
AR l.b.	0.898	0.479 10^{-3}	0.315 10^{-5}	0.207 10^{-7}	0.136 10^{-9}
Exact	0.950	0.513 10^{-3}	0.337 10^{-5}	0.222 10^{-7}	0.146 10^{-9}

(a) $\rho = 0.95$ ($\lambda_1 = 0.6,\ \lambda_2 = 2,\ q_{12} = 1,\ q_{21} = 3$)

x	0	12	24	36	48
LNT u.b.	0.923	0.452 10^{-2}	0.221 10^{-4}	1.081 10^{-7}	0.529 10^{-9}
AR u.b.	0.932	0.456 10^{-2}	0.223 10^{-4}	1.091 10^{-7}	0.534 10^{-9}
LNT l.b.	0.703	0.344 10^{-2}	0.168 10^{-4}	0.823 10^{-7}	0.403 10^{-9}
AR l.b.	0.588	0.288 10^{-2}	0.141 10^{-4}	0.689 10^{-7}	0.337 10^{-9}
Exact	0.750	0.394 10^{-2}	0.193 10^{-4}	0.943 10^{-7}	0.461 10^{-9}

(b) $\rho = 0.75$ ($\lambda_1 = 0.6,\ \lambda_2 = 1.2,\ q_{12} = 1,\ q_{21} = 3$)

x	0	4	8	12	16
LNT u.b.	0.780	0.599 10^{-2}	0.459 10^{-4}	0.352 10^{-6}	0.270 10^{-8}
AR u.b.	0.804	0.617 10^{-2}	0.473 10^{-4}	0.363 10^{-6}	0.278 10^{-8}
LNT l.b.	0.331	0.254 10^{-2}	0.195 10^{-4}	0.150 10^{-6}	0.115 10^{-8}
AR l.b.	0.194	0.149 10^{-2}	0.114 10^{-4}	0.088 10^{-6}	0.067 10^{-8}
Exact	0.400	0.403 10^{-2}	0.309 10^{-4}	0.237 10^{-6}	0.182 10^{-8}

(c) $\rho = 0.4$ ($\lambda_1 = 0.3,\ \lambda_2 = 0.8,\ q_{12} = 1,\ q_{21} = 4$)

Table 3.3: Bounds and exact values for $P(X > x)$ for MMPP/E_5/1 queue with 25 homogeneous 2 state sources; (a) $\rho = 0.95$, (b) $\rho = 0.75$, (c) $\rho = 0.4$.

x	0	50	100	150	200
LNT u.b.	1.010	$0.650\ \ 10^{-2}$	$0.418\ \ 10^{-4}$	$0.269\ \ 10^{-6}$	$0.173\ \ 10^{-8}$
AR u.b	1.009	$0.650\ \ 10^{-2}$	$0.418\ \ 10^{-4}$	$0.269\ \ 10^{-6}$	$0.173\ \ 10^{-8}$
LNT l.b.	0.926	$0.596\ \ 10^{-2}$	$0.383\ \ 10^{-4}$	$0.247\ \ 10^{-6}$	$0.159\ \ 10^{-8}$
AR l.b.	0.879	$0.566\ \ 10^{-2}$	$0.364\ \ 10^{-4}$	$0.234\ \ 10^{-6}$	$0.151\ \ 10^{-8}$
Exact	0.950	$0.623\ \ 10^{-2}$	$0.401\ \ 10^{-4}$	$0.258\ \ 10^{-6}$	$0.166\ \ 10^{-8}$

(a) $\rho = 0.95$ ($\lambda_1 = 0.6,\ \lambda_2 = 2,\ q_{12} = 1,\ q_{21} = 3$)

x	0	8	16	24	32
LNT u.b.	1.029	$1.273\ \ 10^{-2}$	$1.576\ \ 10^{-4}$	$1.950\ \ 10^{-6}$	$2.413\ \ 10^{-8}$
AR u.b	1.028	$1.272\ \ 10^{-2}$	$1.575\ \ 10^{-4}$	$1.949\ \ 10^{-6}$	$2.412\ \ 10^{-8}$
LNT l.b.	0.688	$0.852\ \ 10^{-2}$	$1.054\ \ 10^{-4}$	$1.305\ \ 10^{-6}$	$1.616\ \ 10^{-8}$
AR l.b.	0.532	$0.658\ \ 10^{-2}$	$0.815\ \ 10^{-4}$	$1.008\ \ 10^{-6}$	$1.248\ \ 10^{-8}$
Exact	0.750	$1.031\ \ 10^{-2}$	$1.277\ \ 10^{-4}$	$1.582\ \ 10^{-6}$	$1.959\ \ 10^{-8}$

(b) $\rho = 0.75$ ($\lambda_1 = 0.6,\ \lambda_2 = 1.2,\ q_{12} = 1,\ q_{21} = 3$)

x	0	3	6	9	12
LNT u.b.	1.088	$0.860\ \ 10^{-2}$	$0.679\ \ 10^{-4}$	$5.370\ \ 10^{-7}$	$4.240\ \ 10^{-8}$
AR u.b	1.085	$0.857\ \ 10^{-2}$	$0.677\ \ 10^{-4}$	$5.350\ \ 10^{-7}$	$4.220\ \ 10^{-8}$
LNT l.b.	0.307	$0.243\ \ 10^{-2}$	$1.918\ \ 10^{-5}$	$1.515\ \ 10^{-7}$	$1.197\ \ 10^{-9}$
AR l.b.	0.144	$0.114\ \ 10^{-2}$	$0.900\ \ 10^{-5}$	$0.711\ \ 10^{-7}$	$0.562\ \ 10^{-9}$
Exact	0.400	$0.470\ \ 10^{-2}$	$3.680\ \ 10^{-5}$	$2.910\ \ 10^{-7}$	$2.296\ \ 10^{-9}$

(c) $\rho = 0.4$ ($\lambda_1 = 0.3,\ \lambda_2 = 0.8,\ q_{12} = 1,\ q_{21} = 4$)

Table 3.4: Bounds, approximations and exact values for $P(X > x)$ for MMPP/D/1 queue with 25 homogeneous 2 state sources; (a) $\rho = 0.95$, (b) $\rho = 0.75$, (c) $\rho = 0.4$.

4 Applications to Call Admission in Multimedia Systems

The aim of this section is to present various applications of our results to the problem 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., [35, 28, 52]). 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 [20].

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 $P(X > b) \leq q$ 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 $P(X > b) \leq q$ implies that, for the case of a buffer with finite capacity b, the cell loss probability does not exceed q.

Using the bounds established in Section 2, we obtain bounds on the number of calls that can be admitted to a single resource system. We will observe that use of the upper bound on the tail of the backlog distribution for the purpose of call admission results in a larger number of admitted calls than the popular effective bandwidth approach [35].

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.

4.1 Markov Arrival Process

Consider an irreducible, aperiodic Markov chain $(Y_n)_n$ with state space $\mathcal{S} := \{1, \ldots, K\}$ and transition matrix $\mathbf{P}$. Let $(A_n)_n$ be a sequence of $\{0, 1, 2, \ldots\}$-valued r.v.'s such that $(A_n, Y_n)_n$ is a Markov chain with transition kernel $G_{kj}(x) = P(Y_{n+1} = j, A_n \leq x \,|\, Y_n = k)$. Then, the process $(A_n)_n$ is called a Markov Arrival Process (MAP). In the following, a MAP will be represented by the 4-tuple $(A_n, Y_n, \mathcal{S}, \mathbf{P})$ whenever there is a need to specify the Markov environment associated with it; otherwise, we will simply say that $(A_n)_n$ is a MAP.

Assume now that the increments $(U_n)_n$ in (1.1) are given by $U_n = A_n - c$, where is $(A_n)_n$ a MAP and c is a nonnegative constant. From the definition of a MAP it is seen that the sequence $(U_n)_n$ satisfies assumption (A) in Section 1 so that all of the results obtained in Section 2 will apply to $(X_n)_n$.

Consider now N independent MAP's, $(A_n^i, Y_n^i, \mathcal{S}_i, \mathbf{P}_i)$, $1 \leq i \leq N$, and let $(A_n)_n$ be the process resulting from the superposition of these MAP's, namely, $A_n = \sum_{i=1}^N A_n^i$. It is known that $(A_n, (Y_n^1, \ldots, Y_n^N), \times_{i=1}^N \mathcal{S}_i, \otimes_{i=1}^N \mathbf{P}_i)$ is a MAP ($\otimes$ denotes the Knonecker product of matrices). By using elementary properties of Kronecker product of matrices [11, 34] together with the independence assumption of MAP's $(A_n^i)_n$ it is easily seen that the spectral radius $\rho(\theta)$ of the matrix $\mathbf{F}^*(\theta)$ is given by

$$\rho(\theta) = e^{-\theta c} \prod_{i=1}^N \tau_i(\theta) \tag{4.1}$$

where $\tau_i(\theta)$ is the spectral radius of the matrix with (k,j)-entry given by $E[\exp(\theta\, A_n^i)\,(Y_{n+1}^i = j) \,|\, Y_n^i = k)$. Therefore, we deduce from Proposition 2.4 that

$$P(X > x) \leq C(\theta)\, e^{-\theta\, x}, \quad \forall x \geq 0 \qquad \text{if} \qquad \sum_{i=1}^N \frac{\log(\tau_i(\theta))}{\theta} \leq c. \tag{4.2}$$

The quantity $c_i(\theta) = \log(\tau_i(\theta))/\theta$ is called the *effective bandwidth* of the process $(A_n^i)_n$ [16, 26, 32, 35, 42, 32].

A similar result was presented by Chang and Cheng in [13, Example 3.4] but with a different coefficient $C(\theta)$. The coefficient in [13], denoted as $\Gamma(\theta)$, is given by $\Gamma(\theta) = \max_{i,j} r_i(\theta)/r_j(\theta)$, where $(r_1(\theta), \ldots, r_K(\theta))^T$ is the (positive) right eigenvector of the matrix $\mathbf{F}^*(\theta)$ associated with its spectral radius $\rho(\theta)$. In general, the bound in [13] appears to be looser that ours. In particular, $\Gamma(\theta)$ is always larger than 1 for $\theta > 0$ unlike $C(\theta)$ which maybe smaller than 1 (see Section 4.1.1).

Example 4.1 (Computation of $C(\theta)$ for discrete time on/off sources)

Consider the case when $(A_n)_n$ is the superposition of N independent and identical 2-state MAP's $(A_n^i, Y_n^i, \{1,2\}, \mathbf{P}_i\}$ such that

$$P(Y_{n+1}^i = j, A_n^i \leq x \,|\, Y_n^i = k) = p_{kj}\, F_k(x)$$

where p_{kj} is the (k,j)-entry of the transition matrix $\mathbf{P}_i$, and $F_k(x) = P(A_n^i \leq x \,|\, Y_n^i = k)$ for $k, j = 1, 2,$. Assume that $F_1(x) = 1$ for all $x \geq 0$ and that $F_2(x) = 1$ for all $x < \lambda$. In other words, each MAP $(A_n^i)_n$ is a discrete time on/off source which emits packets at rate λ in state 2 and does *not emit any packet in state* 1. *Then, it can be shown [49] that*

$$C(\theta) = \left(\frac{\pi_1}{z_1(\theta)}\right)^N \max_{\substack{1\le r\le N\\ l_0\le l\le N}} \frac{e^{\lambda l\theta} \sum_{i=l}^{N} \binom{N}{i} (\pi_2/\pi_1)^i \, \alpha_{ir}}{\sum_{i=l}^{N} \binom{N}{i} \left(z_2(\theta)\, e^{\lambda\theta}/z_1(\theta)\right)^i \, \alpha_{ir}} \tag{4.3}$$

with $\pi_1 = p_{21}/(p_{12}+p_{21})$, $\pi_2 = 1-\pi_1$, $l_0 = \inf\{l = 1,2,\ldots : l\lambda > c\}$, $z_1(\theta) = (\exp(\theta\lambda) - \nu(\theta))/(\exp(\theta\lambda)-1)$, $z_2(\theta) = 1 - z_1(\theta)$,

$$\nu(\theta) = \frac{(1-p_{12}) + (1-p_{21})\, e^{\lambda\theta} + \sqrt{((1-p_{12}) + (1-p_{21})\, e^{\lambda\theta})^2 - 4\,(1-p_{12}-p_{21})\, e^{\lambda\theta}}}{2}$$

and where α_{ir}, *the probability that* r *sources are on at time* n *given that* i *sources were on at time* $n-1$, *is given by*

$$\alpha_{ir} = \sum_{s=\max(0,i-r)}^{\min(i,N-r)} \binom{i}{l} p_{21}^{l}\,(1-p_{21})^{i-l} \binom{N-i}{r-(i-l)} p_{12}^{r-(i-l)}\,(1-p_{12})^{N-r-l}.$$

□

Consider now the performance criterion $P(X > x) \le \exp(-\theta\, x)$ for $x \to \infty$. The following holds:

Proposition 4.1 *If the stability condition* $E_\pi[A_0] < c$ *is satisfied, and if the set* $\mathcal{D}$ *is open, then, for all* $\theta \in \mathcal{D} \cap (0,\infty)$

$$\lim_{x\to\infty} \frac{\log P(X > x)}{x} \le -\theta \qquad \text{if and only if} \qquad \sum_{i=1}^{N} c_i(\theta) \le c.$$

Proof. Assume that $E_\pi[A_0] < c$ and that the set $\mathcal{D}$ is open. Therefore, $\theta^* > 0$ (cf. discussion after the proof of Proposition 2.2), which in turn implies from the strict convexity of $\rho(\theta)$, the identity $\rho(0) = 1$ and $\rho'(0) < 0$ (see the discussion at the end of Section 2.3) that the condition $\rho(\theta) \le 1$, or, equivalently, $\sum_{i=1}^{N} c_i(\theta) \le c$ from (4.1), holds if and only if $0 \le \theta \le \theta^*$. From this and (2.17) we conclude that $\sum_{i=1}^{N} c_i(\theta) \le c$ if and only if $\lim_{x\to\infty}(1/x)\,\log\, P(X > x) \le -\theta$. ∎

Proposition 4.1 is not new, as the same result was announced by Kesidis et al. [43] (but proved through a heuristic argument). This proposition was mainly stated for future reference (see Section 4.1.1) and the proof we gave was presented for the sake of completeness. The same result can also be obtained in an even more general context (see Assumptions (C1)-(C3) in [12]) from the work of Chang [12, Proposition 3.9] by using the same arguments as ours. In particular, Chang showed that $c_i(\theta) = (1/\theta)\, \lim_{n\to\infty}(1/n)\, \log\, E[\exp(\theta \sum_{m=0}^{n-1} A_m^i)]$ [12, Example 3.3], which provides a nice interpretation of the effective bandwidth of a source.

We now 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.

4.1.1 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 discrete time on/off source as defined in Example 4.1. We assume that these sources are mutually independent and all identical, with common transition matrix

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

The mean of on and off periods correspond to 352ms and 650 ms, respectively. The service rate of the 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 $P(X > b) \leq q$? 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 (4.2) and (2.25) can be used to obtain bounds on N_{max} – namely

$$N_{lb} \leq N_{max} \leq N_{ub}$$

where

$$\begin{aligned} N_{lb} &= \max_{49 \leq N \leq 134} \{N \,:\, \ln(C(\theta^*)/q) - \theta^* \, b \leq 0\} \\ N_{ub} &= \max_{49 \leq N \leq 134} \{N \,:\, \ln(B^*/q) - \theta^* \, b > 0\} \end{aligned}$$

where for each $N = 49, \ldots, 134$, θ^* is the unique solution in $(0, \infty)$ of the equation $\sum_{i=1}^{N} c_i(\theta) = c$. The coefficient $C(\theta^*)$ has been computed from (4.3) for various values of b, q and r. It is worth noting that $C(\theta)$ has always been found smaller than 1, ranging from $1.03\,10^{-20}$ for 49 sources to 0.9995 for 134 sources. The coefficient B has also been computed from (4.3) after substituting "max" for "min" and θ for θ^*.

Table 4.5 reports N_{lb} and N_{ub} as a function of the tolerable delay, b, for tolerances of 0.1%, 1% and 5%, Also included are the number of sessions N_{eb} that can be supported based on the effective bandwidth approach, namely, $N_{eb} = \max\{N \,:\, \sum_{i=1}^{N} c_i(\theta) \leq c\}$ (cf. Proposition 4.1). We observe that the quality of the bounds increases as b and/or q increase. In particular, the relative error $r_e := (N_{ub} - N_{lb})/N_{ub}$ is such that $r_e \leq 0.25$ for $b \geq 20$, $r_e \leq 0.2$ for $b \geq 50$ and $r_e \leq 0.05$ for $b \geq 200$. In addition, the effective bandwidth approach turns out to be very conservative for small delay constraints (say, for $b \leq 100$) and lies between the bounds only for large b ($b \geq 500$). The fact that the effective bandwidth yields conservative admission controls has been observed elsewhere as well (see [35]) where enhancements have been proposed.

4.1.2 Call admission in a video server

We consider requests to a video server for movies. Sources are homogeneous, independent, 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 (s) segments. Each source is modeled as a discrete time on/off source as in Example 4.1, with common transition matrix

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

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 (stability condition).

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 (N_{ub}) and lower (N_{lb}) bounds for N_{max} for $.5s \leq b \leq 60s$ for tolerances of 1, 5, and 10%. For the range given above, the bounds obtained on N_{max} do not depend on b and is presented in Table 4.6. Also included are the number of sessions that can be supported as predicted by the effective bandwidth approach (N_{eb}). Observe that the effective bandwidth approach yields the same number of sessions as can be supported through a peak rate allocation.

5 Concluding Remarks

In this paper we have presented upper and lower bounds of an exponential form for the tail distribution of both X_n and of its stationary regime X, in the case where $(X_n)_n$ is defined by the stochastic recursion (1.1). Applications to queues have been discussed and our bounds have been numerically compared to other bounds and to the exact distribution. Last, we have provided an application of our results in the setting of admission control. Our work has been lately extended in several directions including more general stochastic recursions in the max-plus framework [48] and more general admission control criteria related to the probability that k or more customers within a group of n arrive to find the buffer occupancy greater than some level, [49]. Also, it has been used to derive upper and lower bounds on the tail distribution of the stationary backlog in a multiplexer fed by independent and nonhomogeneous Markov on/off fluid sources [4].

Acknowledgments: The authors would like to thank Alain Jean-Marie for useful discussions during the course of this work, and Zhi-Li Zhang for the numerical calculations in Section 4.1.1. The authors are also very grateful to the reviewers for their comments - especially for pointing out the generalization of the early results to Markov additive processes and for bringing [7] and [33] to our attention.

b	0	10	20	30	40	50	100	200	500	1000
q=0.001										
N_{lb}	86	91	95	98	101	104	113	122	129	131
N_{ub}	127	127	127	127	127	127	128	129	131	132
N_{eb}	48	58	69	78	86	93	110	121	129	131
q=0.01										
N_{lb}	94	98	102	105	108	111	119	125	130	132
N_{ub}	129	129	129	129	129	129	130	131	132	133
N_{eb}	48	63	78	90	98	103	117	125	130	132
q=0.05										
N_{lb}	100	105	109	112	115	117	123	128	132	133
N_{ub}	131	131	131	131	131	131	131	132	133	133
N_{eb}	48	72	90	101	108	113	123	128	132	133

Table 4.5: Supportable number of voice sessions

q	0.001	0.01	0.05
N_{lb}	105	107	108
N_{ub}	111	113	114
N_{eb}	100	100	100

Table 4.6: Supportable numbers of video sessions

A Proofs of Lemma 2.1 and Corollary 2.1

Proof of Lemma 2.1: Define the set $\mathcal{G} = \{(x,j,k) \in [0,\infty) \times \mathcal{S}^2 \, : \, F_{kj}(x) < p_{kj}\}$. Observe that $\mathcal{G}$ is a nonempty set thanks to the assumption that the set $\mathcal{M}$ (see (2.3)) is nonempty.

$>$From the definition of B (see (2.26)) it is easily seen that

$$\begin{aligned} B &\geq \inf_{(x,j,k) \in \mathcal{G}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) g_{kj}(x) \\ &= \min \left\{ \inf_{\substack{(x,j,k) \in \mathcal{G} \\ \Delta_{kj} < \infty}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) g_{kj}(x) \, ; \, \inf_{\substack{(x,j,k) \in \mathcal{G} \\ \Delta_{kj} = \infty}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) g_{kj}(x) \right\} \\ &\geq \min \left\{ \min_{\substack{(j,k) \in \mathcal{M} \\ \Delta_{kj} < \infty}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) e^{-\theta^* \Delta_{kj}} \, ; \, \inf_{\substack{x \geq 0, j,k \in \mathcal{S} \\ \Delta_{kj} = \infty}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) g_{kj}(x) \right\} \end{aligned} \tag{A.1}$$

where $g_{kj}(x) = (p_{kj} - F_{kj}(x)) / \int_x^\infty \exp(\theta^*(u-x)) \, F_{kj}(du)$.

On the other hand, we deduce from assumption (2.27) that when $\Delta_{kj} = \infty$ then there exist constants $\delta_{kj} < \infty$ and $\epsilon > 0$ such that $g_{kj}(x) \geq \epsilon$ for all $x \geq \delta_{kj}$. This observation readily implies that

$$\begin{aligned} \inf_{\substack{x \geq 0, j,k \in \mathcal{S} \\ \Delta_{kj} = \infty}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) g_{kj}(x) &= \min_{\substack{j,k \in \mathcal{S} \\ \Delta_{kj} = \infty}} \left\{ \inf_{0 \leq x < \delta_{kj}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) g_{kj}(x) \, ; \, \inf_{x \geq \delta_{kj}} \left(\frac{\pi(k)}{z_k(\theta^*)} \right) g_{kj}(x) \right\} \\ &\geq \min_{\substack{j,k \in \mathcal{S} \\ \Delta_{kj} = \infty}} \left\{ \left(\frac{\pi(k)}{z_k(\theta^*)} \right) e^{-\theta^* \delta_{kj}} \, ; \, \left(\frac{\pi(k)}{z_k(\theta^*)} \right) \epsilon \right\} > 0. \end{aligned} \tag{A.2}$$

Combining (A.1) and (A.2) yields $B > 0$. ∎

Proof of Corollary 2.1: Thanks to Lemma 2.1 it suffices to show that $\xi_{kj} > 0$ when $F_{jk}(x)$ has a polynomial-exponential density function for $(k,j) \in \mathcal{M}$ since in this case $\Delta_{kj} = \infty$.

For all $j,k \in \mathcal{S}$, let

$$f_{kj}(x) = \sum_{i=1}^{n_{kj}} a_{kj,i} \, x^{m_{kj,i}} \, e^{-\beta_{kj,i} x}$$

be the density function of $F_{kj}(x)$, where $a_{kj,i}$'s are nonzero real numbers, $m_{kj,i}$'s are nonnegative integers and $\beta_{kj,i}$'s are strictly positive real numbers. Assume without loss of generality that for all $j,k \in \mathcal{S}$, $\beta_{kj,1} \leq \beta_{kj,2} \leq \cdots \leq \beta_{kj,n_{kj}}$, and that if $\beta_{kj,i} = \beta_{kj,i+1}$, then $m_{kj,i} > m_{kj,i+1}$. As $f_{kj}(x) \geq 0$ for all $x > 0$, it is easy to see (by letting x go to infinity) that $a_{kj,1} > 0$ for all $j,k \in \mathcal{S}$.

It then follows that for all $x \geq 0$, $j,k \in \mathcal{S}$,

$$\frac{\displaystyle\int_x^\infty F_{kj}(du)}{\displaystyle\int_x^\infty e^{\theta^*(u-x)}\,dF_{kj}(u)} = \frac{\displaystyle\sum_{i=1}^{n_{kj}} a_{kj,i}\int_x^\infty u^{m_{kj,i}}e^{-\beta_{kj,i}\,u}du}{\displaystyle\sum_{i=1}^{n_{kj}} a_{kj,i}\int_x^\infty e^{\theta^*(u-x)}\,u^{m_{kj,i}}e^{-\beta_{kj,i}\,u}du}$$

$$= \frac{\displaystyle\sum_{i=1}^{n_{kj}} e^{-\beta_{kj,i}\,x}\,a_{kj,i}\,m_{kj,i}!\sum_{l=0}^{m_{kj,i}}\frac{x^l}{l!}\,\frac{1}{\beta_{kj,i}^{m_{kj,i}+1-l}}}{\displaystyle\sum_{i=1}^{n_{kj}} e^{-\beta_{kj,i}\,x}\,a_{kj,i}\,m_{kj,i}!\sum_{l=0}^{m_{kj,i}}\frac{x^l}{l!}\,\frac{1}{(\beta_{kj,i}-\theta^*)^{m_{kj,i}+1-l}}}. \qquad \text{(A.3)}$$

Dividing both the numerator and the denominator in the r.h.s. of (A.3) by $a_{kj,1}x^{m_{kj,1}}e^{-\beta_{kj,1}x}$ and using the fact that the couple $(\beta_{kj,1}, -m_{kj,1})$ is the smallest in the lexicographic order among all couples $(\beta_{kj,i}, -m_{kj,i})$, we obtain that

$$\lim_{x\to\infty}\frac{p_{kj}-F_{kj}(x)}{\int_x^\infty e^{\theta^*(u-x)}\,dF_{kj}(u)} = \frac{\beta_{kj,1}-\theta^*}{\beta_{kj,1}} > 0$$

where the strict positiveness is a consequence of the fact that $\mathcal{D} = \{\theta\,:\,\theta < \min_{k,j}\,\beta_{kj,1}\}$ together with the definition of θ^*. The proof is thus completed. ∎

References

[1] J. Abate, W. Whitt, "The Fourier-series Method for Inverting Transforms of Probability Distributions", *Queueing Systems*, **10**, pp. 5-88, 1992.

[2] J. Abate, G. L. Choudhury, W. Whitt, "Asymptotics for Steady-State Tail Probabilities in Structured Markov Queueing Models", *Commun. Statist. Stochastic Models*, **10**, 1, pp. 99-143, 1994.

[3] D. Anick, D. Mitra, M. M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources", *Bell Systems Technical Journal*, **61**, 1871-1894, 1982.

[4] D. Artiges and P. Nain, "Upper and Lower Bounds for the Multiplexing of Multiclass Markovian On/Off Sources", *Perf. Evaluation*, **27**&**28**, pp. 673–698, Oct. 1996.

[5] S. Asmussen, *Applied Probability and Queues*. J. Wiley & Sons, New York, 1987.

[6] S. Asmussen, "Stationary Distributions via First Passage Times". Preprint. Jan 1995.

[7] S. Asmussen and T. Rolski, "Risk Theory in a Periodic Environment: the Cramer-Lundberg Approximation and Lundberg's Inequality", *Mathematics of Operations Research*, **2**, 2, pp. 410-433, May 1994.

[8] A. A. Borovkov, *Stochastic Processes in Queueing Theory*. Springer-Verlag, new York, 1976.

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

[10] D. D. Botvich and N. G. Duffield, "Large Deviations, the Shape of the Loss Curve and Economies of Scale in Large Multiplexers", *Queueing Systems*, **20**, pp. 293-320, 1995.

[11] J. W. Brewer, "Kronecker Products and Matrix Calculus in System Theory", *IEEE Trans. on Circuit and Syst.*, **25**, 9, pp. 772-781, 1978.

[12] 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.

[13] C.-S. Chang and J. Cheng, "Computable Exponential Bounds for Intree Networks with Routing". *Proc. INFOCOM'95*, pp. 197–204, 1995.

[14] C.-S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, "The Application of Effective Bandwidth to Fast Simulation of Communication Networks", *Perf. Evaluation*, **20**, pp. 45–66, 1994.

[15] G. L. Choudhury, D. M. Lucantoni, W. Whitt, "Squeezing the Most out of ATM", *IEEE Trans. Commun.*, **44**, 2, pp. 203–217, Feb. 1996.

[16] 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.

[17] C. Courcoubetis and R. Weber, "Buffer Overflow Asymptotics for a Switch Handling Many Traffic Sources". *J. Appl. Prob.* **33**, pp. 886-903, Sept. 1996.

[18] R. L. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation", *IEEE Trans. Inf. Theory*, **37**, 1, pp. 114-131, Jan. 1991.

[19] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis", *IEEE Trans. Inf. Theory*, **37**, 1, pp. 132-141, Jan. 1991.

[20] 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.

[21] G. de Veciana, C. Courcoubetis and J. Walrand, "Decoupling Bandwidth for Networks", *UCB/ERL Technical Report M93/50*, Univ. California at Berkeley, Jun. 1993.

[22] N. G. Duffield, "Exponential Bounds for Queues with Markovian Arrivals", *Queueing Systems*, **17**, pp. 413-430, 1994.

[23] N. G. Duffield and N. O'Connell, "Large Deviations and Overflow Probabilities for the General Single-Server Queue", *Math. Proc. Phil. Soc.*, **118**, pp. 363-374, 1995.

[24] R. S. Ellis, "Large Deviations for a General Class of Random Vectors", *The Annals of Prob.*, **12**, 1, pp. 1-12, 1984.

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

[26] 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**, 3, pp. 329-343, Jun. 1993.

[27] A. I. Elwalid, D. Heyman, T. V. Laksjman, D. Mitra and A. Weiss, "Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing", *IEEE J. on Selected Areas Commun.*, **13**, 6, pp. 1004–1016, 1995.

[28] D. Ferrari and D. Verma. "A Scheme for Real-Time Channel Establishment in Wide-Area Networks", *IEEE J. Selected Areas Commun.*, **8**, pp. 368–379, Apr. 1990.

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

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

[31] E. Gelenbe, X. Mang, R.Önvural. "Diffusion based statistical call admission in ATM networks'," *Performance Evaluation*, **27**&**28**, pp. 411–436, Oct. 1996.

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

[33] P. W. Glynn and W. Whitt, "Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single-Server Queue", *Studies in Appl. Prob.*, J. Galambos and J. Gani eds, pp. 131–156, 1994.

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

[35] 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.

[36] H. Heffes, D.M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance", *IEEE J. Select. Areas Commun.*, **6**, 856-868, 1986.

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

[38] G. Kesidis, J. Walrand and C.-S. Chang, "Effective Bandwidth for Multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Trans. on Networking*, **1**, 4, pp. 424–428, 1993.

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

[40] V. Jacobson and S. McCanne, “vat”, Manual Pages, Lawrence Berkeley Laboratory, Berkeley, CA.

[41] V. Jacobson and S. McCanne, “Using the LBL Network Whiteboard”, Lawrence Berkeley Laboratory, Berkeley, CA.

[42] F. P. Kelly, “Effective Bandwidths at Multi-Class Queues”, *Queueing Systems*, **9**, pp. 5–16, 1991.

[43] G. Kesidis, J. Walrand and C.-S. Chang, “Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources”, *IEEE/ACM Trans. Networking*, **1**, 4, pp. 424–428, Aug. 1993.

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

[45] J. F. C. Kingman, “Inequalities in the Theory of Queues”, *J. Roy. Stat. Soc.*, Series B, **32**, pp. 102–110, 1970.

[46] 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.

[47] Z. Liu, P. Nain and D. Towsley, “On a Generalization of Kingman's Bounds”, *Technical Report No. 2423*, INRIA, Dec. 1994.

[48] Z. Liu, P. Nain and D. Towsley, “Bounds on the Tail Distribution of Markov-Modulated Stochastic Max-Plus Systems”, *Proc. of the 34th IEEE Conf. on Decision and Control*, New Orleans, pp. 1395-1399, Dec. 1995.

[49] Z. Liu, P. Nain and D. Towsley, “ Bounds on Finite Horizon QoS Metrics with Application to Call Admission”, *Proc. INFOCOM'96*, San Francisco, CA, pp. 1338-1345, Mar. 1996.

[50] 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,

[51] D.M. Lucantoni, G.L. Choudhury, W. Whitt, “The transient BMAP/G/1 queue,” *Stochastic Models,* **10**, pp. 145-182, 1994.

[52] R. Nagarajan, J. Kurose, D. Towsley, “Local allocation of end-to-end quality-of-service in high-speed networks”, *IFIP Transactions C-15: Modelling and Performance Evaluation of ATM Technology,* (ed. H. Perros, G. Pujolle, Y. Takahashi), pp. 99-118, 1993.

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

[54] M. Parulekar and A. M. Makowski, “Buffer Overflow Probabilities for a Multiplexer with Self-Similar Traffic”, *Proc. INFOCOM'96*, San Francisco, CA, pp. 1452-1459, Mar. 1996.

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

[56] G. J. K. Regterschot and J. H. A. de Smit, "The Queue with Markov Modulated Arrival and Services", *Math. of Opns. Research*, **11**, 3, pp. 465–483, 1986.

[57] S. M. Ross, "Bounds on the Delay Distribution in $GI/G/1$ Queues", *J. of Appl. Probab.*, **11**, pp. 417–421, 1974.

[58] 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)

[59] A. Simonian and J. Guibert, "Large Deviations Approximation for Fluid Queues Fed by a Large Number of On/Off Sources", *IEEE J. on Selected Areas Commun.*, **13**, 6, pp. 1017–1027, 1995.

[60] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*. Berlin: J. Wiley & Sons, 1983.

[61] W, Whitt, "Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-Class Queues", *Telecommunication Systems*, **2**, pp. 71–107, 1993.

[62] O. Yaron and M. Sidi, "Performance and Stability of Communication Networks Via Robust Exponential Bounds", *IEEE/ACM Trans. Networking*, **1**, 3, pp. 372–385, Jun. 1993.