Bounds on Finite Horizon QoS Metrics with 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

In: Proceedings of IEEE INFOCOM'96
San Francisco, CA, Mar. 1996, pp. 1338-1345

Abstract

In this paper we are concerned with a discrete time, single server system in which packets arrive from a finite population of sources. Under the assumption that arrivals from each source are modulated by a Markov process, we consider the following metrics *(i)* the fraction of an interval that the queue length exceeds a certain value, and *(ii)* the fraction of a group of packets from a single source that arrive to find the queue length above a certain value. For both metrics we derive upper and lower bounds on the probabilities that they exceed a threshold. These are important measures because they reflect more accurately the behavior perceived by applications such as networked audio and video. An application of these results to call admission is also given.

Keywords: Tail distribution; Exponential bound; Markov chain; Matrix analysis; Queues; Markov modulated process; Quality of service; Effective bandwidth; Call admission control.

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

1 Introduction

There exists a substantial body of work on the problem of providing guaranteed quality of service (QoS) to different service classes in BISDN's. Most of this work has been dedicated to guaranteeing that the packet loss probability seen by a random packet or the probability that the delay of a randomly chosen packet lies below some threshold, see [5, 8, 10, 17, 18] and references contained within for examples. However, this type of QoS metric appears inappropriate for envisaged audio and video services in BISDN's [3, 22, 4]. For example, metrics such as losses and delays in talkspurts [3] and losses within blocks of packets for packet video [3, 22, 4] are more appropriate.

The focus of this paper will be on two finite horizon QoS metrics, the *interval QoS* and the *block QoS*, introduced in [21]. These QoS criteria are defined over intervals of time and finite groups of packets from a single connection respectively. These metrics will be studied for a single network link modeled as a discrete time, single server system in which packets arrive from a finite population of sources. More specifically, we consider the following two metrics,

- the amount of time within an interval of time during which the queue length of the system exceeds a fixed value,
- the number of packets within a group from an individual source that arrive to find that the queue length exceeds a fixed value.

We develop upper and lower bounds on the probabilities that these quantities exceed a threshold for the case that the arrivals from the sources are modulated by a finite state Markov chain. These bounds are developed using bounds on the queue length distribution at an arbitrary time developed recently in [18]. Last, an application to call admission is also given.

Several papers have studied finite horizon metrics. However, they have focussed on very simple systems that usually contain a single source. For example, [20, 21] provides approximate analyses of these metrics for the case of an M/M/1/K queue and simulation results for a finite population of On-Off sources feeding a single server. Exact and asymptotic analyses of the block metric for a discrete time queue and the M/M/1/K queue when fed by a single source are given in [7] and [1, 6].

This paper is organized as follows. Section 2 contains a description of the model and the finite horizon metrics being considered along with a review of the results in [18] that will form the foundation for our analysis. Sections 3 and 4 contain the derivations of bounds on the interval

and block metrics, respectively. Numerical results and an application to call admission are given in Section 5. Finally, Section 6 summarizes the contributions of the paper.

2 Model and Preliminary Analysis

We model a statistical multiplexer as a single server serving an infinite capacity buffer in first in first out (FIFO) order in a discrete time system where the server can transmit up to c packets in one time unit. Assume that M sources, $M > c$, labeled $m = 1, \ldots, M$, feed packets to this multiplexer and denote by A_n^m the number of arrivals from source m during the n-th slot. Let $(Q_n)_n$ be the process describing the backlog in the queue at time n. It satisfies the following recursion,

$$Q_{n+1} = (Q_n + A_n - c)^+, \quad n = 0, 1, \ldots$$

where $A_n := \sum_{m=1}^M A_n^m$ is the total number of arrivals during the n-th slot.

We are interested in the following finite horizon performance measures ($l \geq -1, N \geq 1$):

$$P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right)$$

$$P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right), \quad m = 1, \ldots, M$$

where $(Q_n^m)_n$ is the queue-length process embedded at arrival epochs of packets from source m. The first of these will be referred to as the *interval metric* and the second one will be referred to as the *block metric*.

We will develop upper and lower bounds for these quantities under the assumption that the arrival processes $(A_n^m)_n$, $1 \leq m \leq M$, are modeled by M independent Markov Modulated Arrival Processes (MMAP's). That is we assume that $A_n^m = U_n^m(Y_n^m)$, where $(Y_n^m)_n$ is an irreducible, aperiodic, homogeneous Markov chain on the finite set $\mathcal{S}_m$ with transition matrix $\mathbf{P}_m$ and stationary distribution $\underline{\pi}_m$, and where $(U_n^m(k))_n$ is a renewal process for fixed m and k. We further assume that the Markov chains $(Y_n^m)_n$ $(1 \leq m \leq M)$ and the renewal processes $(U_n^m(k))$ $(k \in \mathcal{S}_m, 1 \leq m \leq M)$ are mutually independent processes. It is worth noting [9] that the aggregate arrival process $(A_n)_n$ is also a MMAP with state-space $\mathcal{S} = \prod_{m=1}^M \mathcal{S}_m$, underlying Markov chain $(Y_n)_n = (Y_n^1, \ldots, Y_n^M)_n$, transition matrix $\mathbf{P} = \otimes_{m=1}^M \mathbf{P}_m$, and stationary distribution $\underline{\pi} = \otimes_{m=1}^M \underline{\pi}_m m$, where $\otimes$ denotes

the Kronecker product. Last, define $U_n(k) = \sum_{m=1}^{M} U_n^m(k_m)$ for $k = (k_1, \ldots, k_M) \in \mathcal{S}$ so that $A_n = U_n(Y_n)$, and let $p_{i,j}$ be the (i,j)-entry of the transition matrix $\mathbf{P}$.

We shall assume throughput the paper that the Markov chain $(Y_n)_n$ begins in equilibrium, that is $P(Y_0 = k) = \pi(k)$ for all $k \in \mathcal{S}$. Under this assumption the sequence $(A_n)_n$ is a stationary sequence and the stability condition for this model is $E[A_n] < c$ [19]. We will assume from now on that $E[A_n] < c$ and will denote by Q the stationary regime of the process $(Q_n)_n$.

We conclude this section by introducing some additional notation and by reviewing several results from [18] pertaining to the tail distribution of the backlog, $P(Q_n \geq x)$.

Define $F_k(x) = P(U_n(k) \leq x)$ for $k \in \mathcal{S}$ and $\psi_k(\theta) = E[\exp(\theta U_n(k))]$ for $k \in \mathcal{S}$.

We will assume that the set $\Theta = \{\theta > 0 : \psi_k(\theta) < \infty, \forall k \in \mathcal{S}\}$ is non empty and open. These technical assumptions are satisfied in most cases of practical interest which includes r.v.'s with phase-type distributions.

Let us introduce further notation. Let $\mathbf{A}$ be an n-by-n matrix with real entries. $\mathbf{A}^T$ will denote its transpose, $\mathbf{A}^k$ its k-th power, and $r(\mathbf{A})$ its spectral radius. For any vector $\underline{a} = (a_1, \ldots, a_n)$, diag $(\underline{a})$ or diag $((a_i, i = 1, 2, \ldots, n))$ will denote the diagonal matrix with diagonal elements $a_1, \ldots, a_n$ and $|\underline{a}|$ will stand for $\sum_{k=1}^{n} a_k$.

For $\theta \in \Theta$, define the matrix

$$\mathbf{H}(\theta) = \left(\mathbf{P}_1^T \, \mathbf{\Psi}^1(\theta)\right) \otimes \cdots \otimes \left(\mathbf{P}_M^T \, \mathbf{\Psi}^M(\theta)\right)$$

where $\mathbf{\Psi}^m(\theta) := \text{diag } (E[\exp(\theta U_n^m(k))],\ k \in \mathcal{S}_m)$.

Since $\mathbf{H}(\theta)$ is nonnegative and irreducible we know from Perron-Frobenius theory [14] that the spectral radius $\tau(\theta) = r(\mathbf{H}(\theta))$ is an eigenvalue and that any right-eigenvector corresponding to this eigenvalue has strictly positive components. Denote by $\underline{z}(\theta) = (z_k(\theta),\ k \in \mathcal{S})$ the unique right-eigenvector such that $|\underline{z}(\theta)| = 1$.

In [18] we showed that

$$P(Q_n > x) \leq C(\theta)\, e^{-\theta x}, \quad x \geq 0,\ n = 1, 2, \ldots \tag{1}$$

for all $\theta \in \Theta$ such that $\tau(\theta) \leq \exp(\theta c)$, where

$$C(\theta) = \sup_{x \geq 0, j \in \mathcal{S}} \frac{\sum_{k \in \mathcal{S}} p_{k,j} \pi(k)(1 - F_k(x))}{\sum_{k \in \mathcal{S}} p_{k,j} z_k(\theta) \int_x^\infty e^{\theta(u-x)} dF_k(u)} \tag{2}$$

and

$$B\, e^{-\theta^* x} \leq P(Q_n > x), \quad x \geq 0,\ n = 1, 2, \ldots \tag{3}$$

where θ^* is the unique solution in $(0, \infty)$ of the equation $\tau(\theta) = \exp(\theta c)$, and B is given as

$$B = \inf_{x \geq 0, j \in \mathcal{S}} \frac{\sum_{k \in \mathcal{S}} p_{k,j} \pi(k)(1 - F_k(x))}{\sum_{k \in \mathcal{S}} p_{k,j} z_k(\theta) \int_x^\infty e^{\theta(u-x)} dF_k(u)}. \tag{4}$$

The upper (resp. lower) bound in (1) (resp. (3)) will hold if it holds for $n = 0$. For instance, (1) holds for $n = 0$ if the queue is initially empty. The same bounds hold for the tail of the stationary backlog distribution, $P(Q > x)$, without any additional conditions.

We will find it useful to use bounds on the backlog distribution conditioned on the state of the Markov chain. These are, see [18],

$$P(Q_n > x \,|\, Y_n = k) \leq C(\theta)\, (z_k(\theta)/\pi(k))\, e^{-\theta x} \tag{5}$$

and

$$P(Q_n > x \,|\, Y_n = k) \geq B\, (z_k(\theta^*)/\pi(k))\, e^{-\theta^* x} \tag{6}$$

for all $x \geq 0$, $k \in \mathcal{S}$, $n = 1, 2, \ldots$ and for all $\theta \in \Theta$ such that $\tau(\theta) \leq \exp(\theta c)$.

It has been shown elsewhere (e.g., [18]) that it is much easier to compute θ^*, $C(\theta)$, and B for the case of independent sources, than for the case of an arbitrary arrival process, even if the numbers of states in the underlying Markov chains are the same. In addition, one can introduce the notion of effective bandwidth [8, 11, 10, 13, 15, 16] for each source when the performance criterion is

$$P(Q > x) \leq e^{-\theta x} \tag{7}$$

as $x \to \infty$. The effective bandwidth, $c_m(\theta)$ for source m is

$$c_m(\theta) = \frac{1}{\theta} \log \tau_m(\theta)$$

where $\tau_m(\theta) = r\left(\mathbf{P}_m^T\, \mathbf{\Psi}^m(\theta)\right)$.

We have the following result (Proposition 3.1 in [18]).

Proposition 2.1

$$\lim_{x\to\infty} \frac{\log P(Q > x)}{x} \leq -\theta \quad \textit{if and only if} \quad \sum_{m=1}^{M} c_m(\theta) \leq c.$$

This carries the implication that admission control can consist of simply checking if there is sufficient excess bandwidth at a server to cover the effective bandwidth requirement of a new source. This type of result has been shown in more generality (see [18]). It has also spawn considerable interest in developing practical call admission policies based on the idea; see [10] for one example.

3 Interval Metric

Our interest in this section is to develop, to the extent possible, an equivalent theory for the interval metric $P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right)$ as exists for the backlog distribution which was described in the previous section.

We begin by establishing an upper bound. An application of Chernoff's bound yields

$$\begin{aligned} P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right) &\leq \frac{1}{L} E\left[\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\}\right] \\ &\leq \frac{N}{L}\, C(\theta)\, e^{-\theta x} \end{aligned} \tag{8}$$

for all $x \geq 0$ and for all $\theta \in \Theta$ such that $\tau(\theta) \leq \exp(\theta c)$. The last inequality follows from (1).

We next obtain a lower bound. We have the following inequality

$$P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right) \geq P(Q_{l+1} > x, \cdots, Q_{l+L} > x). \tag{9}$$

We focus on the right-hand side of (9). It can be expressed as

$$P(Q_{l+1} > x, \cdots, Q_{l+L} > x)$$

$$
\begin{aligned}
&= \sum_{j_1,\cdots,j_{L-1}} P(Q_{l+1} > x, \cdots, Q_{l+L} > x, Y_{l+1} = j_1, \cdots, Y_{l+L-1} = j_{L-1}) \\
&\geq \sum_{j_1,\cdots,j_{L-1}} P(Q_{l+1} > x, U_{l+1}(Y_{l+1}) > c, \cdots, U_{l+L-1}(Y_{l+L-1}) > c, \qquad (10) \\
&\qquad\qquad Y_{l+1} = j_1, \cdots, Y_{l+L-1} = j_{L-1}), \\
&= \sum_{j_1,\cdots,j_{L-1}} P(Q_{l+1} > x \,|\, Y_{l+1} = j_1)(1 - F_{j_1}(c))\, \pi(j_1) \prod_{i=2}^{L-1} p_{j_{i-1},j_i}(1 - F_{j_i}(c)) \\
&\geq B\, e^{-\theta^* x} \sum_{j_1,\cdots,j_{L-1}} z_{j_1}(\theta^*)(1 - F_{j_1}(c)) \prod_{i=2}^{L-1} p_{j_{i-1},j_i}\,(1 - F_{j_i}(c)). \qquad (11)
\end{aligned}
$$

(10) is a consequence of the definition of Q_n whereas (11) follows from the application of the inequality (6). Combining (9) and (11) yields

$$
P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right) \geq \begin{cases} B e^{-\theta^* x}, & \text{for } L = 1 \\ \underline{z}(\theta^*)\, \mathbf{D}\, (\mathbf{P}\, \mathbf{D})^{L-2}\, \mathbf{1}^T\, B\, e^{-\theta^* x}, & \text{for } L \geq 2 \end{cases} \qquad (12)
$$

with $\mathbf{D} = \text{diag}\,(P\,(U_n(j) > c),\, j \in \mathcal{S})$ and $\mathbf{1}^T = (1, 1 \ldots, 1)$.

We turn our attention now to the notion of effective bandwidth.

Assume now that the performance criterion is

$$
P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right) \leq \exp(-\theta x) \qquad (13)
$$

as $x \to \infty$ in such a way that $\log(L/N)/x \to -\xi$ with $0 \leq \xi < \infty$.

The following result follows from (8):

Proposition 3.1

$$
\lim_{\substack{x\to\infty \\ \log(L/N)/x \to -\xi}} \frac{\log P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right)}{x} \leq -\theta \quad \textit{if} \quad \sum_{m=1}^{M} c_m(\theta + \xi) \leq c.
$$

We make the following observations. First, since $c_m(\theta)$ is nondecreasing in θ, we see from Proposition 3.1 that fewer sessions will be admitted when applying criterion (13) rather than the criterion (7).

Second, Proposition 3.1 is not as strong as Proposition 2.1 in that we have not established that $P\left(\sum_{n=l+1}^{l+N} \mathbf{1}\{Q_n > x\} \geq L\right) \leq e^{-\theta x}$ (as $x \to \infty$) implies $\sum_{m=1}^{M} c_m(\theta + \xi) \leq c$. We conjecture that this is, in fact, true. However, our lower bound (12) is not tight enough for us to establish the result.

4 Block Metric

We now turn our attention to the block metric $P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right)$. Again our objective is to derive upper and lower bounds and to address the existence of an effective bandwidth theory for the block metric.

Let T_n^m be the time of the $(n+1)$-st arrival from source m. It is easily observed from the statistical assumptions placed on the model that for every $m = 1, 2, \ldots, M$, $(Q_n^m, Y_{T_n^m})_n$ is a Markov chain, further ergodic under the stability condition $E[A_n] < c$. From now on we will assume that the $M+1$ Markov chains $(Q_n, Y_n)_n$, $(Q_n^m, Y_{T_n^m})_n$, $m = 1, 2, \ldots, M$, all begin in equilibrium (this assumption is made possible because of the property that any ergodic Markov chain on a countable state space couples with its stationary version after a time which is finite a.s. [2, 143-144]). This assumption implies, in particular, that

$$\left(Q_0^m, Y_{T_0^m}\right) \stackrel{\text{st}}{=} (Q_n^m, Y_{T_n^m}) \quad \text{and} \quad (Q_0, Y_0) \stackrel{\text{st}}{=} (Q_n, Y_n) \quad \forall n = 0, 1, \ldots \tag{14}$$

We first establish an upper bound. Applying again Chernoff's bound yields

$$\begin{aligned} P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right) &\leq \frac{1}{L} \sum_{n=0}^{N-1} P(Q_n^m > x) \\ &= \frac{N}{L} P(Q_0^m > x) \quad \text{from (14).} \end{aligned} \tag{15}$$

On the other hand,

$$\begin{aligned} P(Q_0^m > x) &= P(Q_0 > x \,|\, A_0^m > 0) \\ &= \sum_{k \in \mathcal{S}} P(Q_0 > x, Y_0 = k \,|\, A_0^m > 0) \\ &= \frac{1}{P(A_0^m > 0)} \sum_{k \in \mathcal{S}} P(Q_0 > x \,|\, Y_0 = k)\, P(A_0^m > 0 \,|\, Y_0 = k)\, \pi(k) \end{aligned}$$

$$\leq \frac{C(\theta)}{P(A_0^m > 0)} \sum_{k \in \mathcal{S}} z_k(\theta)\ P(A_0^m > 0 \,|\, Y_0 = k)\, e^{-\theta x} \tag{16}$$

$$= \frac{C(\theta)}{P(A_0^m > 0)}\, \underline{z}(\theta)\, \mathbf{E}\, \mathbf{1}^T\, e^{-\theta x} \tag{17}$$

by using (5), where

$$\mathbf{E} = \text{diag}\ (P(U_n^m(i) > 0),\ i \in \mathcal{S}).$$

By combining (15) and (17) we finally obtain

$$P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right) \leq \left(\frac{N}{L}\right) \frac{C(\theta)}{P(A_0^m > 0)}\, \underline{z}(\theta)\, \mathbf{E}\, \mathbf{1}^T\, e^{-\theta x} \tag{18}$$

for all $x \geq 0$ and for all $\theta \in \Theta$ such that $\tau(\theta) \leq \exp(\theta c)$.

We next obtain a lower bound.

We have

$$P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right) \geq P(Q_0^m > x, Q_1^m > x, \ldots, Q_{L-1}^m > x)$$

$$= P(Q_0 > x, Q_1^m > x \ldots, Q_{L-1}^m > x \,|\, A_0^m > 0) \quad \text{from the stationarity assumption (14)}$$

$$= \sum_{\substack{i_j \in \mathcal{S} \\ j=0,1,\ldots,L-1}} P(Q_0 > x, Q_1^m > x, \ldots, Q_{L-1}^m > x, Y_0 = i_0, Y_{T_1^m} = i_1, \ldots, Y_{T_{L-1}^m} = i_{L-1} \,|\, A_0^m > 0)$$

$$\geq \sum_{\substack{i_j \in \mathcal{S} \\ j=0,1,\ldots,L-1}} P\Big(Q_0 > x, Q_1^m > x, \ldots, Q_{L-1}^m > x, Y_0 = i_0, Y_{T_1^m} = i_1, \ldots, Y_{T_{L-1}^m} = i_{L-1},$$

$$\text{at least c arrivals in each slot of the period } [T_j^m, T_{j+1}^m], j = 0, 1, \ldots, L-2 \,|\, A_0^m > 0\Big)$$

$$= \frac{1}{P(A_0^m > 0)} \sum_{\substack{i_j \in \mathcal{S} \\ j=0,1,\ldots,L-1}} P(Q_0 > x \,|\, Y_0 = y_0)\, P(A_0^m > 0 \,|\, Y_0 = i_0)\, \pi(i_0) \prod_{j=0}^{L-2} R_{i_j, i_{j+1}}$$

$$\geq \frac{B^*}{P(A_0^m > 0)}\, e^{-\theta^* x} \sum_{\substack{i_j \in \mathcal{S} \\ j=0,1,\ldots,L-1}} z_{i_0}(\theta^*)\, P(A_0^m > 0 \,|\, Y_0 = i_0) \prod_{j=0}^{L-2} R_{i_j, i_{j+1}} \tag{19}$$

where (19) follows from (6), and

$$R_{i,j} = P(Y_{T_1^m} = j,\ \text{at least c arrivals in each slot of the period } [T_0^m, T_1^m] \,|\, Y_{T_0^m} = i).$$

In order to compute the matrix $\mathbf{R} = [R_{i,j}]$ we introduce the matrix $\mathbf{R}' = [R'_{i,j}]$, where $R'_{i,j}$ is the joint probability that the next arrival of source m will occur when the Markov chain is in state j and that at least c arrivals will be generated in each slot between the current time (say t) and the arrival time of the next arrival from source m given that there is no arrival from source m at time t and that the Markov chain is in state i at time t, namely,

$$\begin{aligned} R'_{i,j} &= P(Y_{T_n^m} = j, \text{ at least c arrivals in each slot of the period } [t, T_n^m] \\ &\qquad \mid Y_t = i, A_k^m = 0, \ \forall k = t, t+1, \ldots, T_n^m - 1). \end{aligned}$$

We have

$$\begin{aligned} R'_{i,j} &= p_{i,j}\, P(A_t \geq c \,|\, A_t^m = 0, Y_t = i)\, P(A_{t+1}^m > 0 \,|\, Y_{t+1} = j) \\ &\quad + \sum_{l \in \mathcal{S}} p_{i,l}\, P(A_t \geq c \,|\, A_t^m = 0, Y_t = i)\, P(A_{t+1}^m = 0 \,|\, Y_{t+1} = l)\, R'_{l,j} \\ R_{i,j} &= p_{i,j}\, P(A_t \geq c \,|\, A_t^m > 0, Y_t = i)\, P(A_{t+1}^m > 0 \,|\, Y_{t+1} = j) \\ &\quad + \sum_{l \in \mathcal{S}} p_{i,l}\, P(A_t \geq c \,|\, A_t^m > 0, Y_t = i)\, P(A_{t+1}^m = 0 \,|\, Y_{t+1} = l)\, R'_{l,j} \end{aligned}$$

or, in matrix form,

$$\mathbf{R}' = \mathbf{G}_1\mathbf{P}\mathbf{E} + \mathbf{G}_1\mathbf{P}\overline{\mathbf{E}}\mathbf{R}' \tag{20}$$

$$\mathbf{R} = \mathbf{G}_2\mathbf{P}\mathbf{E} + \mathbf{G}_2\mathbf{P}\overline{\mathbf{E}}\mathbf{R}' \tag{21}$$

with

$$\begin{aligned} \overline{\mathbf{E}} &= \mathbf{I} - \mathbf{E} \\ \mathbf{G}_1 &= \text{diag } (P(A_n \geq c \,|\, A_n^m = 0, Y_n = i),\ i \in \mathcal{S}) \\ \mathbf{G}_2 &= \text{diag } (P(A_n \geq c \,|\, A_n^m > 0, Y_n = i),\ i \in \mathcal{S}) \end{aligned}$$

where the matrix $\mathbf{E}$ has been defined earlier.

Solving for $\mathbf{R}'$ in (20) and substituting the obtained matrix for $\mathbf{R}'$ in (21) finally gives

$$\mathbf{R} = \mathbf{G}_2\mathbf{P}\left[\mathbf{I} + \overline{\mathbf{E}}\left(\mathbf{I} - \mathbf{G}_1\mathbf{P}\overline{\mathbf{E}}\right)^{-1}\mathbf{G}_1\mathbf{P}\right]\mathbf{E}. \tag{22}$$

In summary, we have shown (cf. (19) and (22)) that

$$P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right) \geq \frac{B^*}{P(A_0^m > 0)}\, \underline{z}(\theta^*)\, \mathbf{E}\, \mathbf{R}^{L-1}\, \mathbf{1}^T\, e^{-\theta^* x}, \quad \forall x > 0. \tag{23}$$

Assume now that the performance criterion is

$$P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right) \leq \exp(-\theta x) \tag{24}$$

as $x \to \infty$ in such a way that $\log(L/N)/x \to -\xi$ with $0 \leq \xi < \infty$. The following effective bandwidth-type result is a direct consequence of (18):

Proposition 4.1

$$\lim_{\substack{x\to\infty \\ \log(L/N)/x\to-\xi}} \frac{\log P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right)}{x} \leq -\theta \quad \textit{if} \quad \sum_{m=1}^{M} c_m(\theta+\xi) \leq c.$$

Again, the discussion following Proposition 3.1 for the interval metric applies equally as well to the block metric.

5 Call Admission

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

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

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 355 ms and 640 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 32 Kb/s. With this data, it is easily seen that the size of a packet is 512 bits and that the available bandwidth is 1,536 Mb/s, which in turn implies that the time needed to serve a packet is 1/3 ms.

Observe that there is no contention if the number of sources M is less than 49 and that the system is unstable whenever $M > 134$.

Define D_n and D_n^m as the delay at time n and as the delay at the n-th arrival epoch of a packet from source m, respectively.

We ask ourselves the following questions:

(1) What is the maximum number $M_{\rm im}$ of voice sessions that can be supported by the channel such that $P\left(\sum_{n=l+1}^{l+22} \mathbf{1}\{D_n > b\} \geq 1\right) \leq q$?

(2) For fixed m, what is the maximum number $M_{\rm bm}$ of voice sessions that can be supported by the channel such that $P\left(\sum_{n=0}^{21} \mathbf{1}\{D_n^m > b\} \geq 1\right) \leq q$?

Here b represents the maximum tolerable delay (in ms) and q a tolerance. Note that $N = 22$ corresponds to the average duration of an on period. Hence we are interested in the probability that the delay exceeds b at least once during an on period ($L = 1$) and in particular, the number of sessions that can be supported while ensuring that this probability lies below the tolerance q.

We shall only concentrate here on determining a lower bound on $M_{\rm im}$ (resp. $M_{\rm bm}$) which we will denote as $M_{\rm im}^{\rm lower}$ (resp. $M_{\rm bm}^{\rm lower}$). Since $D_n = Q_n/3$ ms and $D_n^m = Q_n^m/3$ ms from the definition of the model, the distribution bounds in (8) and in (18) can be used to obtain these lower bounds – namely

$$M_{\rm im} \geq \text{argmax}_{49 \leq M \leq 134} \{M \,:\, \ln(22\, C(\theta^*)/q) - 3b\,\theta^* \leq 0\} = M_{\rm im}^{\rm lower}$$
$$M_{\rm bm} \geq \text{argmax}_{49 \leq M \leq 134} \{M \,:\, \ln(22\, D/q) - 3b\,\theta^* \leq 0\} = M_{\rm bm}^{\rm lower}$$

where $D := C(\theta^*)\, \underline{z}(\theta^*)\, \mathbf{E}\, \mathbf{1}^T / P(A_0^m > 0)$.

Hints for the computation of θ^*, $C(\theta^*)$ and D are given in Appendix A.

Figures 1 and 2 give lower bounds on $M_{\rm im}$ and on $M_{\rm bm}$, respectively, as a function of the tolerable delay, b and for tolerances of 1%, 5% and 10%. Also included are approximations for the lower bounds on $M_{\rm im}$ and $M_{\rm bm}$ based on the effective bandwidth approach (cf. Propositions 3.1 and 4.1), where we let $\xi = \ln(1/22)/b$ for every fixed $b \in (0, 1000]$. We observe, as in[10], that the effective bandwidth approach is very conservative for small values of b (i.e. for $b < 100$).

6 Summary

In this paper, we have derived lower and upper bounds of an exponential form on two QoS metrics, the interval metric and the block metric. In addition, based on these bounds, we have partially developed a theory of effective bandwidths for there two metrics. An application to call admission has been presented to show the applicability of the bounds. Future work will focus on completing

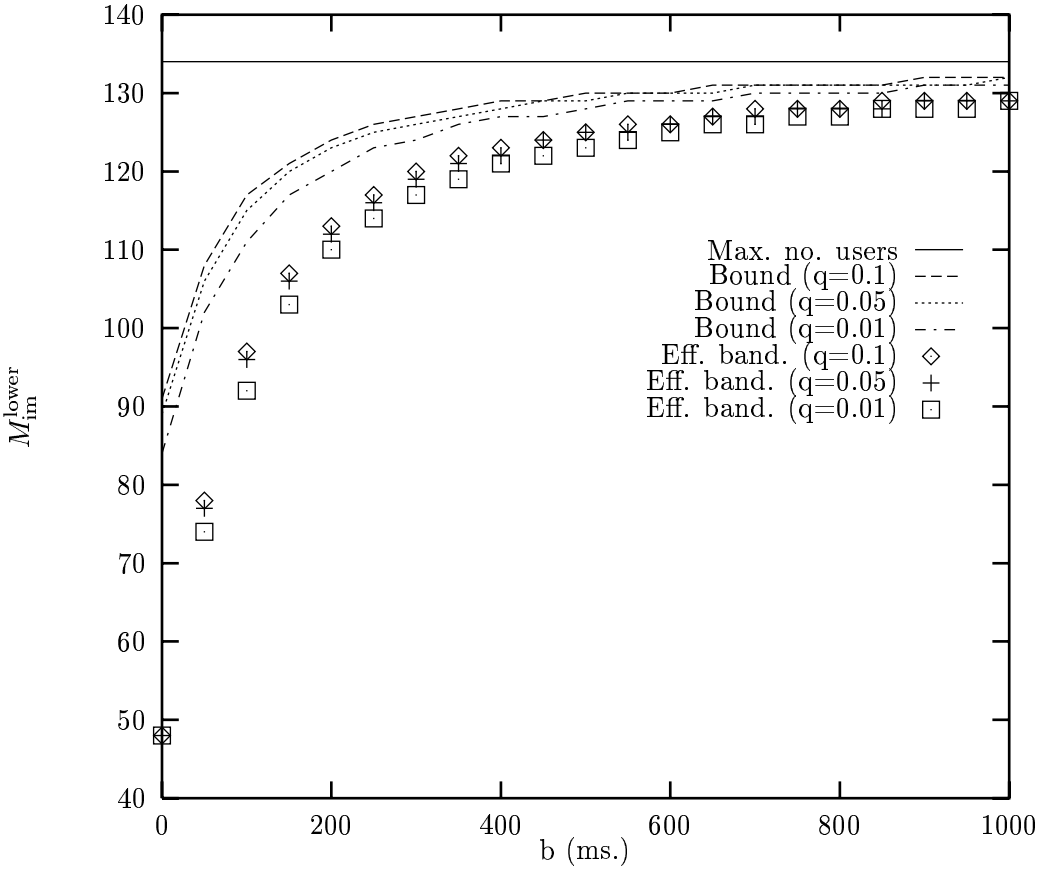

Figure 1: Supportable number of voice sessions for the Interval Metric ($N = 22, L = 1$)

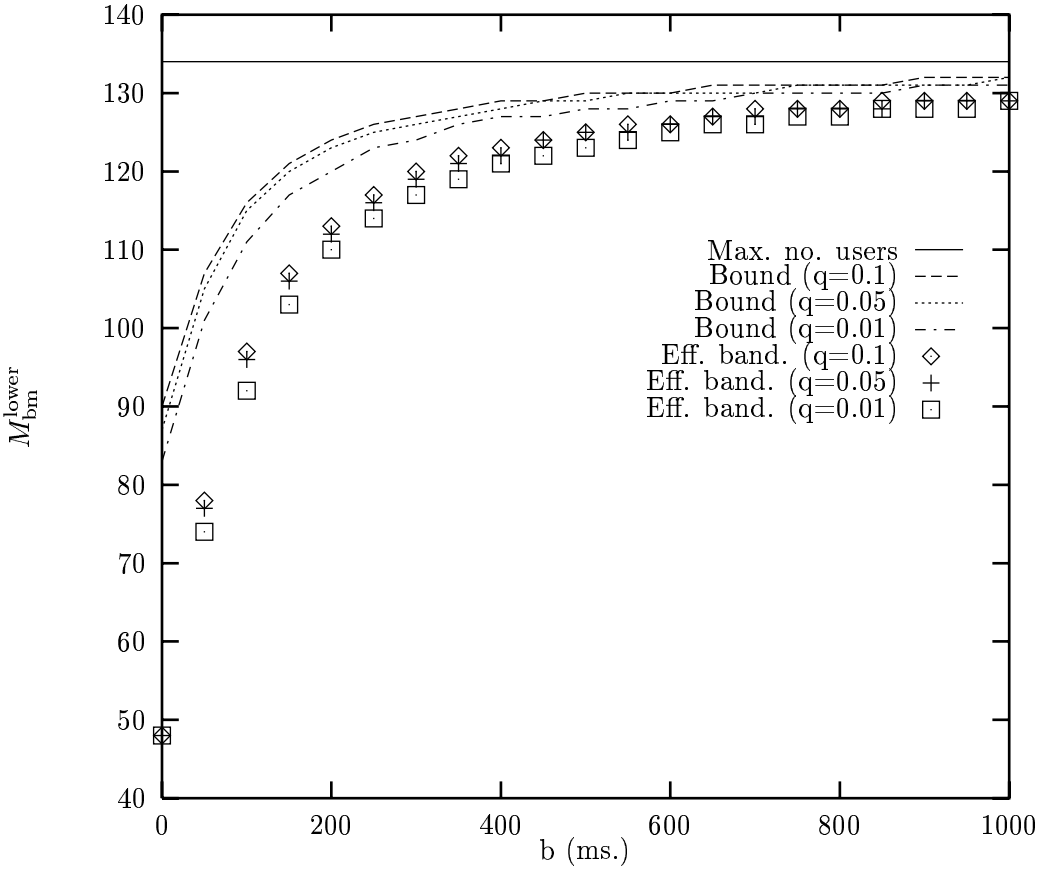

Figure 2: Supportable number of voice sessions for the Block Metric ($N = 22, L = 1$)

the theory of effective bandwidths for these two metrics and on tightening the bounds presented in this paper.

A Appendix

This section contains simple formulas for the numerical computation of the upper bounds both for the interval metric and for the block metric in the case where the offered traffic is the superposition of M independent, identical, on-off sources as described in Section 5.

More precisely, we assume that each on-off source is modulated by a Markov chain $(Y_n^m)_n$ with state space $\mathcal{S}_m = \{0, 1\}$, where 0 (resp. 1) corresponds to the off (resp. on) state, with transition matrix

$$\mathbf{P}_m = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

and where $U_n^m(Y_n^m) = \lambda$ if $Y_n^m = 1$ and 0 if $Y_n^m = 0$ ($\lambda = 1$ in Section 5). In this case, it is easily seen that the stationary distribution $\underline{\pi}_m = (\pi_0, \pi_1)$ of the Markov chain $(Y_n^m)_n$ is given by $\underline{\pi}_m = (q/(p+q), p/(p+q))$. We now examine the computation of the different quantities involved in the upper bounds reported in (8) and in (18).

Computation of θ^*.

Let $\nu(\theta)$ be the spectral radius of $\mathbf{P}_m^T \, \Psi^m(\theta)$ (note that $\nu(\theta)$ is independent of m since the sources are all identical). From an elementary result from the theory of Kronecker products [12, p. 27] we have

$$\tau(\theta) = \nu(\theta)^M$$

so that θ^* (cf. Section 2) is simply the unique position solution of the equation $M \, \log(\nu(\theta)) = \theta c$, where

$$\nu(\theta) = \frac{(1-p) + (1-q)\, e^{\lambda\theta} + \sqrt{((1-p) + (1-q)\, e^{\lambda\theta})^2 - 4\,(1-p-q)\, e^{\lambda\theta}}}{2}.$$

Computation of $C(\theta)$.

By observing that $1 - F_k(x) = \mathbf{1}\{\lambda\ e|k| > x + c\}$ for all $k \in \mathcal{S} = \{0,1\}^M$, we get (cf. (2))

$$\begin{aligned} C(\theta) &= \sup_{\substack{x\geq 0 \\ j\in\mathcal{S}}} \frac{\sum_{k\in\mathcal{S}} p_{k,j}\pi(k)(1-F_k(x))}{\sum_{k\in\mathcal{S}} p_{k,j} z_k(\theta) \int_x^\infty e^{\theta(u-x)} dF_k(u)} \\ &= \max_{0\leq r\leq M} \left\{ \max_{\substack{l_0\leq l\leq M \\ j\in\mathcal{S},|j|=r}} \frac{\sum_{i=l}^{M} \sum_{k\in\mathcal{S},|k|=i} p_{k,j}\pi(k)}{\sum_{i=l}^{M} \sum_{k\in\mathcal{S},|k|=i} p_{k,j} z_k(\theta)\, e^{\lambda\theta(i-l)}} \right\} \end{aligned} \tag{25}$$

where $l_0 := \inf\{l = 1, 2, \ldots : l\lambda > c\}$. The right-hand side of (25) can be further simplified by noting that $\pi(k) = \pi_1^i\, \pi_0^{M-i}$ for all $k \in \mathcal{S}$ such that $|k| = i$. Similarly, we get that $z_k(\theta) = v_1(\theta)^i\, v_0(\theta)^{M-i}$ for all $k \in \mathcal{S}$ such that $|k| = i$, where $\underline{v}(\theta) = (v_0(\theta), v_1(\theta))$ is the unique right-eigenvector of the matrix $\mathbf{P}_m^T\, \Psi^m(\theta)$ corresponding to the eigenvalue $\nu(\theta)$ such that $|\underline{v}(\theta)| = 1$ (here we use the result that $\underline{z}(\theta) = \otimes_{m=1}^M \underline{v}(\theta)$ [12, p. 27]). By a simple algebraic computation we obtain $v_0(\theta) = (e^{\theta\lambda} - \nu(\theta))/(e^{\theta\lambda} - 1)$ and $v_1(\theta) = (\nu(\theta) - 1)/(e^{\theta\lambda} - 1)$.

Reporting these simplifications in (25) yields

$$C(\theta) = \left(\frac{\pi_0}{v_0(\theta)}\right)^M \max_{0\leq r\leq M} \left\{ \max_{\substack{l_0\leq l\leq M \\ j\in\mathcal{S},|j|=r}} \frac{e^{\lambda l\theta} \sum_{i=l}^{M} (\pi_1/\pi_0)^i \sum_{k\in\mathcal{S},|k|=i} p_{k,j}}{\sum_{i=l}^{M} \left(v_1(\theta)\, e^{\lambda\theta}/v_0(\theta)\right)^i \sum_{k\in\mathcal{S},|k|=i} p_{k,j}} \right\}. \tag{26}$$

Define $q_{i,r} = P(|Y_n| = r \,|\, |Y_{n=1}| = i)$ for all $i, r = 1, 2, \ldots, M$. In words, $q_{i,r}$ is the probability that there are r sources active at the beginning of a time-slot given that there were i sources active at the beginning of the previous time-slot. It is not difficult to show that for all $j \in \mathcal{S}$ such that $|j| = r$,

$$\sum_{k\in\mathcal{S},|k|=i} p_{k,j} = \frac{\binom{M}{i}}{\binom{M}{r}} q_{i,r} \tag{27}$$

and

$$q_{i,r} = \sum_{s=\max(0,i-r)}^{\min(i,M-r)} \binom{i}{l} q^l\, (1-q)^{i-l} \binom{M-i}{r-(i-l)} p^{r-(i-l)}\, (1-p)^{M-r-l}.$$

Combining (26) and (27) finally yields

$$C(\theta) = \left(\frac{\pi_0}{v_0(\theta)}\right)^M \max_{\substack{0\le r\le M \\ l_0\le l\le M}} \frac{e^{\lambda l\theta} \sum_{i=l}^{M} \binom{M}{i} (\pi_1/\pi_0)^i \, q_{i,r}}{\sum_{i=l}^{M} \binom{M}{i} \left(v_1(\theta)\, e^{\lambda\theta}/v_0(\theta)\right)^i \, q_{i,r}}.$$

It can be shown that the maximun is always reached for $l = M$ if $(\pi_1/\pi_0)/(v_1(\theta)\, \exp(\lambda\theta)/v_0(\theta)) \geq 1$. In this case $C(\theta) = (\pi_1/v_1(\theta)^M$.

Computation of the upper bound for the block metric.

Fix $\theta \in \Theta$ such that $\tau(\theta) \leq \exp(\theta c)$. From (15)-(16) we have

$$P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right) \leq \left(\frac{N}{L}\right) \frac{C(\theta)}{P(A_0^m > 0)} \sum_{k\in\mathcal{S}} z_k(\theta)\; P(A_0^m > 0 \,|\, Y_0 = k)\, e^{-\theta x}. \tag{28}$$

Since $P(A_0^m > 0 \,|\, Y_0 = k) = \mathbf{1}\{k_m = 1\}$ for all $k = (k_1, \ldots, k_M) \in \mathcal{S}$ from the definition of the model, we have

$$\begin{aligned}
\sum_{k\in\mathcal{S}} z_k(\theta)\; P(A_0^m > 0 \,|\, Y_0 = k) &= \sum_{k\in\mathcal{S}, k_m = 1} z_k(\theta) \\
&= v_1(\theta) \sum_{\substack{i_l\in\{0,1\} \\ l=1,2,\ldots,M-1}} \prod_{l=1}^{M-1} v_{i_l}(\theta) \qquad (29) \\
&= v_1(\theta) \qquad (30)
\end{aligned}$$

where (29) and (30) are direct consequences of the identities $\underline{z}(\theta) = \otimes_{m=1}^{M} \underline{v}(\theta)$ [12, p. 27] and $|\underline{v}(\theta)| = 1$, respectively.

By combining (28) and (30) and by noting that $P(A_0^m > 0) = \pi_1$ from the definition of the model, we finally obtain

$$P\left(\sum_{n=0}^{N-1} \mathbf{1}\{Q_n^m > x\} \geq L\right) \leq C(\theta) \left(\frac{N}{L}\right) \left(\frac{v_1(\theta)}{\pi_1}\right) e^{-\theta x}. \tag{31}$$

Acknowledgments: The authors are grateful to Mr. Zhi-Li Zhang for his help in generating the numerical results presented in this paper.

References

[1] E. Altman, A. Jean-Marie. "The Loss Process of Messages in an M/M/1/K Queue", *Proc. INFOCOM'94*, 1191 – 1198, 1994.

[2] S. Asmussen, *Applied Probability and Queues*. John Wiley & Sons, 1987.

[3] H.S. Bradlow. "Performance Measures for Real-Time Continuous Bit-Stream Oriented Services: Application to packet reassembly", *Computer Networks and ISDN Systems*, **20**, 15 – 26, 1990.

[4] E. Biersack. "Error Recovery in High-Speed Networks", *Proc. 2-nd Int. Workshop on Network and Operating System Support for Digital Audio and Video*, p. 222, 1991.

[5] C.-S. Chang, "Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks", *IEEE Trans. Aut. Contr.*, **39**, 5, 913 – 931, May 1994.

[6] I. Cidon, A. Khamisy, M. Sidi. "Analysis of Packet Loss Processes in High-Speed Networks", *IEEE Trans. Info. Theory*, **39**, 1, 98 –108, 1993.

[7] I. Cidon, A. Khamisy, M. Sidi. "Dispersed Messages in Discrete Time Queues: Delay, Jitter and Threshold Crossing", *Proc. INFOCOM'94*, 218 – 223, 1994.

[8] 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, 329 – 343, Jun. 1993.

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

[10] 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**, 968 – 981, 1991.

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

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

[13] J. Y. Hui, "Resource Allocation for Broadband Networks", *IEEE J. Select. Areas Commun.*, **6**, 1598 – 1608, 1988.

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

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

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

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

[18] Z. Liu, P. Nain and D. Towsley, "Exponential Bounds with an Application to Call Admission", Technical Report, University of Massachusetts, CMPSCI 94-63, Oct. 1994. Submitted to *JACM*

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

[20] R. Nagarajan. *Quality-of-Service Issues in High-Speed Networks*, PhD thesis, Univ. of Massachusetts, Amherst, Sept. 1993.

[21] R. Nagarajan, J. Kurose, D. Towsley. "Finite-Horizon Statistical Quality-of-Service Measures for High Speed Networks", to appear in *J. High Speed Networks*, 1995.

[22] N. Shacham. "Packet Recovery in High-Speed Networks Using Coding and Buffer Management", *Proc. INFOCOM'90*, 124 – 131, 1990.

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