Statistical Multiplexing With Priorities: Tail Probabilities of Queue Lengths, Workloads and Waiting Times

Vijay G. Subramanian and R. Srikant
Coordinated Science Laboratory
University of Illinois
vgsubram@uiuc.edu; rsrikant@uiuc.edu

Abstract

We consider the problem of estimating tail probabilities of waiting times in statistical multiplexing systems with two classes of sources – one with high priority and the other with low priority. The priority discipline is assumed to be nonpreemptive. Exact expressions for the transforms of these quantities are derived assuming that packet or cell streams are generated by Markovian Arrival Processes (MAPs). Then we numerically investigate the large-buffer asymptotic behavior of the the waiting-time distribution for low-priority sources and show that these asymptotics may be non-exponential.

1 Introduction

In this paper, we consider a $MAP/GI/1$ with non-preemptive priorities, restricting our attention to two priorities as the n priority system can be considered as a two-priority system by superposition of the arrival processes and the non-preemptive nature of the priorities. MAPs are very commonly used to described arrival processes in high-speed networks. We assume that the service-time distribution is the same for both priority classes, a special application of this would be in ATM networks where the cell size is a constant 53 bytes. Since the models we consider fall under the class of $M/G/1$ type models defined by Neuts, we would draw upon the substantial literature on such models in [13, 11, 8, 9].

Our main results are follows: We derive exact expressions for the transforms of the waiting time and queue length of the low-priority sources. The transforms could be used to obtain the exact tail probability of these quantities using Laplace transform inversion techniques. We also numerically investigate the asymptotic behavior of the tail probabilities. For our numerical studies, we focus on the tail behavior of the low-priority waiting time. Our principal conclusion is that $P(W > T) \sim \alpha T^{\beta} e^{-\delta T}$ as $T \to \infty$, where W is the waiting time and β could be positive, negative or zero. An inspiration for our work is the results of Abate and Whitt [3] who showed the above asymptotic behavior for the low-priority sources for the case when the arrival processes for both classes are Poisson. Our results generalize this to $MAP/G/1$ queues and provide exact expressions for the transforms of the low priority waiting times. In addition, using numerical results, we show that the behavior of β is significantly different in the case of general MAP arrivals as compared to the case of Poisson arrivals. Specifically, in the case with Poisson arrivals, $\beta = 0$ when the low-priority arrival rate is above a certain threshold. This is not necessarily true in the case with MAP arrivals. Further, β is always less than or equal to zero with Poisson arrivals whereas β can also be positive with general MAP arrivals (see Section 4).

An important application of tail probability computations is in the design of admission control schemes for high-speed networks, where admitted sources are guaranteed a pre-specified *Quality-of-Service* (QoS). QoS is typically specified in terms of quantities such as probability of cell loss, upper limits on waiting times, etc. A popular technique is to convert these QoS requirements into a single number called the *effective bandwidth* and use this quantity very much like a bandwidth requirement in circuit-switched networks. The effective bandwidth approximation relies on large-deviations results of the following form:

$$\lim_{B \to \infty} \frac{1}{B} \log P(X > B) = -\delta, \tag{1}$$

where X is the workload and $\delta > 0$; see for example [15]. Thus, given a finite buffer size B, a natural approximation to the loss probability, $P(X > B)$, is

$$P(X > B) \approx e^{-\delta B}. \tag{2}$$

While there are several sources of error in using the approximation (2), it has been shown that the approx-

imation can be refined as $P(X > B) \approx \alpha e^{-\delta B}$. Such a refinement is considered to be reasonably accurate for most practical applications. Further, for most well-behaved service-time distributions, if α is exactly calculated, $P(X > B) \sim \alpha e^{-\delta B}$ [2]. Assuming a preemptive resume model, large deviations results of the form (2) can also be obtained for statistical multiplexers with priorities [6, 4, 5, 12]. One of the contributions of our paper is to point out that for non-preemptive priority model (which should be very close to the preemptive resume model when the packet sizes are small), the asymptotic form $P(X > B) \sim \alpha e^{-\delta B}$ does not hold, in general, for the low-priority sources.

2 MAPs and MAP/GI/1 Queues

A MAP is a continuous-time Markov chain described by the following generator:

$$Q = \begin{bmatrix} D_0 & D_1 & 0 & 0 & \cdot \\ & D_0 & D_1 & 0 & \cdot \\ & & D_0 & D_1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix},$$

where $D \equiv D_0 + D_1$ with $D \neq D_0$, is also a generator. One can associate an arrival process with this Markov chain as follows: an arrival occurs whenever there is a state transition into a state corresponding to a D_1 block and there is no arrival otherwise. D_0 and D_1 are $m \times m$ matrices, where m is the number of phases of the arrival process. This is a special case of versatile Markovian point processes introduced by Neuts [11] and studied under various names [13, 8, 9], most commonly referred to as the Batch Markovian Arrival Processes (BMAPs). We have simply restricted our attention to the case where the batch size at each arrival point can be at most 1. We will not discuss all the properties of MAPs here, the reader is referred to [9] for an introduction. However, before we introduce priority queues, we present some well-known properties of MAPs and $MAP/GI/1$ queues which will be useful for our analysis later. These results can be found in [13, 10, 8, 9].

If we let $N(t)$ be the number of arrivals in $(0, t]$ and $J(t)$ be the phase of the system at time t, then $(N(t), J(t))$ is a Markov chain on the state space $\{(n, j), n \geq 0, j = 1, 2, \ldots, m\}$ with generator Q. Define $D(z)$ to be

$$D(z) \equiv D_0 + D_1 z \quad \text{for } |z| \leq 1,$$

and $P_a(n, t)$ be a $m \times m$ matrix denoting the number of arrivals in $(0, t]$ with the (i, j)th element defined as

$$[P_a(n,t)]_{ij} \equiv \text{Prob}(N(t)=n, J(t)=j | N(0)=0, J(0)=i).$$

Then, the matrix generating function $P^*(z, t)$ of the number of arrivals in $[0, t]$ [1] is given by

$$P^*(z, t) = e^{D(z)t}, \quad \text{for } |z| \leq 1,\ t \geq 0,$$

where $P^*(z, t) = \sum_{n=0}^{\infty} P_a(n, t) z^n$. Let π be the stationary distribution of D (assuming that it irreducible). Then the mean arrival rate is given by

$$\lambda = \pi D_1 e.$$

In the rest of this section, we state a few lemmas listing some properties of $MAP/GI/1$ queues that will be used later. Let H be the service-time distribution with LST h (i.e, $h(s) = \int_0^\infty e^{-st} dH(t)$) and mean $1/\mu$ with $\mu > 0$. The traffic intensity at the $MAP/GI/1$ queue is given by $\rho = \lambda/\mu$.

Lemma 2.1 *Let $A(z, s)$ be the 2-dimensional transform of the number of arrivals and time interval between successive departure epochs given that the first departure did not leave the system empty. $A(z, s)$ is given by*

$$A(z, s) = h(sI - D(z)), \text{ where } Re(s) \geq 0 \text{ and } |z| \leq 1.$$

Lemma 2.2 *Define $A(z) \equiv A(z, 0) = h(-D(z) - z\lambda_l I + \lambda_l I)$. We have the following equation for the 2-dimensional transform of the number served during, and the duration of, the busy period*

$$G(z, s) = zh(sI - D[G(z, s)]).$$

Lemma 2.3 *Define $G \equiv G(1, 0)$. The matrix G has the following properties:*

1. *G is stochastic whenever $\rho < 1$. Let g be the invariant probability vector of G.*
2. *The jth component of vector g is the stationary probability that the arrival process is in phase j given that the system is empty.*

Lemma 2.4 *The queue-length generating function of a $MAP/GI/1$ queue is given by*

$$Y(z) = (1 - \rho) g (z - 1) A(z) [zI - A(z)]^{-1}. \quad (3)$$

[1] Most quantities in matrix-analytic theory are matrices such as $P_a(n, t)$ and $P^*(z, t)$ with each element corresponding to a particular phase transition. However, for convenience, we will often refer to quantities such as "the number of arrivals in $[0, t]$" without referring to the phase transition, but this should not cause confusion.

3 Priority $MAP/GI/1$ Queues

In this section, we derive the transforms of various quantities of interest for $MAP/GI/1$ queues with two priority classes where the classes have separate buffers. Throughout, we assume that the service-time distribution is the same for both classes. For ease of exposition, we first consider the case where the low-priority arrivals are Poisson. Later, we generalize this to the case where both classes generate arrivals according to MAPs.

3.1 Poisson Low-Priority Arrivals

Let the low-priority arrivals be Poisson with rate λ_l. The high priority arrival process is a MAP with D_0^h and D_1^h as the relevant parameters and the rate is denoted as λ_h. The traffic intensity is given by $\rho_l \equiv \lambda_l/\mu$ and $\rho_h \equiv \lambda_h/\mu$ with $\rho \equiv \rho_h + \rho_l < 1$ for stability. Note that the low-priority class can be considered as a MAP with parameters $D_0^l = -\lambda$ and $D_1^l = \lambda$. Define $D_S(z) \equiv D_0^h + D_1^h z - z\lambda_l I - \lambda_l I$. Then, Lemmas 2.1 through 2.4 can be applied to the single class $MAP/GI/1$ obtained by considering the two buffers (high and low priority) together, and the arrival process being the superposition of the high and low-priority arrival processes. As in $D_S(z)$, throughout we will use the subscript S to denote this single-priority-class $MAP/GI/1$.

We first derive an expression for the LST of the waiting time distribution of the lower priority packets. Since the low-priority and high-priority packets have the same service-time distributions, we only need to monitor the total number of packets in the system. We first note two useful facts:

- The low-priority arrival process is Poisson and thus, we use the PASTA property to claim that upon arrival a low-priority packet sees $Y_S(z)e$ as the distribution of the number of packets in service, where $Y_S(z)$ is given by (2.4) with $D(z)$ replaced by $D_S(z)$, and e is a vector with all elements equal to 1.
- The remaining service of the packet in service (if any) is given by the stationary excess distribution associated with H (again by PASTA). Denote the residual service-time distribution by R and its LST by r, where R is given by $R(t) = \mu \int_0^t (1-H(u))du$, and thus, $r(s) = (1-h(s))\mu/s$.

Now we proceed along the lines of the derivation of the Takàcs equation and the Kendall functional equation for the busy period of the $M/G/1$ queue. Specifically, we note that if a lower-priority packet sees n packets (of both classes together) ahead of it, then its waiting time is independent of the order in which all the packets ahead of it (and any high-priority arrivals before it begins service) are served. Thus, the waiting time is composed of

- W_1 : The service times of the packet in service and all high-priority packets that arrive during the service of the packet currently in service.
- W_2 : The $n-1$ high-priority busy periods generated by the $n-1$ packets ahead of it in the queue.

The LST of W_1 can be easily seen to be

$$G_l(s) = r(sI - D_S[G_h(s)]).$$

From Lemma 2.2, the high-priority busy period is given by

$$G_h(s) = h(sI - D_S[G_h(s)]). \tag{4}$$

Therefore, using the relationship between $r(s)$ and $h(s)$, $G_l(s)$ can be rewritten as

$$G_l(s) = \mu(1 - G_h(s))[sI - D_S[G_h(s)]]^{-1}. \tag{5}$$

Now we are ready to state the following theorem.

Theorem 3.1 *The LST of the complementary cumulative distribution function (ccdf) of the low-priority waiting time is given by* $w_l^C(s) = \frac{1-w_l(s)}{s}$, *where*

$$\begin{aligned} w_l(s) &= (1-\rho) + Y_S(G_h(s))G_h^{-1}(s)G_l(s)e \\ &- (1-\rho)g_S G_h^{-1}(s)G_l(s)e. \end{aligned} \tag{6}$$

Proof: See [14].

The next result presents the transform of the stationary distribution of number of high-priority and low-priority packets in the system, the proof can be found in a longer version of this paper [14]. By PASTA, this is also the same as the distribution of number of high-priority and low-priority packets seen by a low-priority packet arrival.

Theorem 3.2 *Let $y(n,m)$ be a vector such that its ith component is the probability that there are n high-priority and m low-priority packets in the system, and the arrival process is in phase j. Its 2-dimensional transform* $Y(z,w) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} y(n,m)z^n w^m$ *is given by*

$$\begin{aligned} TY(z,w) &= P(0,0)(-(D_0^h - \lambda_l I)^{-1})(zD_1^h + w\lambda_l I - I)(I - A(z,w))(-D(z,w)^{-1}) \\ &+ P(0,0)[-(I - A(z,0))(-D(z,0)^{-1}) + (-(D_0^h - \lambda_l)^{-1})] \\ &+ P(z,w)(I - A(z,w))(-D(z,w)^{-1}) \\ &+ P(z,0)(I - A(z,0))(-D(z,0)^{-1}), \end{aligned} \tag{7}$$

where $T = \frac{1}{\lambda_h+\lambda_l}$, $D(z,w) = D_0^h - \lambda_l I + D_1^h z + \lambda_l I w$,

$$\begin{aligned} P(z,w) &= \{P(0,w)(z/w-1)+P(0,0)(z/w)[-w(D_0^h-\lambda_l I)^{-1}(D_1^h+\lambda_l I)-I]\} \\ &\times A(z,w)(zI-A(z,w))^{-1}, \textit{and} \end{aligned} \quad (8)$$

$$A(z,w) = \int_0^\infty e^{(D_0^h-\lambda_l I+D_1^h z+\lambda_l I w)x}\, dH(x), \quad (9)$$

$$P(0,w) = Q(w)(P(0,1)e), \quad (10)$$

$$Q(0) = P(0,0)/(P(0,1)e), \quad (11)$$

$$\begin{aligned} Q(w) &= Q(0)\{-w(D_0^h-\lambda_l I)^{-1}(D_1^h+\lambda_l)-I\} \\ &\times G_H(w)(wI-G_H(w))^{-1}, \end{aligned} \quad (12)$$

$$G_H(w) = G_h(\lambda_l-\lambda_l w), \quad (13)$$

$$P(0,0) = \frac{1-\rho}{\lambda_h+\lambda_l I} g_S(-D_0^h+\lambda_l I), \quad (14)$$

$$P(0,1)e = \frac{\lambda_l}{\lambda_h+\lambda_l} + \frac{1-\rho}{\lambda_h+\lambda_l} g_S D_1^h e. \quad (15)$$

The following corollary immediately follows from the definition of $Y(z,w)$.

Corollary 3.1 *The z-transform of ccdf of the number of low-priority packets in the system seen by a low priority arrival is given by* $Y^C(w) = \frac{1-Y(1,w)e}{1-w}$.

3.2 MAP Arrival Process for the Low-Priority Class

Here we let the high-priority class have D_0^h, D_1^h as the matrices associated with the arrival process and similarly, let D_0^l, D_1^l be the matrix parameters of the low-priority arrival process. We still assume that the service-times are distributed identically for the two classes. Let I_h, I_l denote identity matrices of the same order as D_1^h, D_1^l respectively.

From standard matrix-analytic theory, we have the following theorem which is the equivalent of Theorem 3.2. While Theorem 3.2 also gives the transform of the distribution seen by low priority arrivals when they are Poisson, the following theorem gives only the stationary distribution.

Theorem 3.3 *Let* $y(n,m)$ *be a vector such that its ith component is the probability that there are n high-priority and m low-priority packets in the system, and the arrival process is in phase j. Its 2-dimensional transform* $Y(z,w) = \sum_{n=0}^\infty \sum_{m=0}^\infty y(n,m) z^n w^m$ *is given by (7) with the following modifications:*

(i) D_0^h *replaced by* $D_0^h \otimes I_l$,

(ii) $-\lambda_l I$ *replaced by* $I_h \otimes D_0^l$,

(iii) D_1^h *replaced by* $D_1^h \otimes I_l$,

(iv) $\lambda_l I$ *replaced by* $I_h \otimes D_1^l$,

(v) $D_S(z) = D_0^h \otimes I_l + D_1^h \otimes I_l z$.

(vi) $D(z,w) = D_0^h \oplus D_0^l + D_1^h \otimes I_l z + I_h \otimes D_1^l w$, *and*

(vii) $G_H(w)$ *solves*

$$\begin{aligned} G_H(w) &= A(G_H(w),w) \\ &= \int_0^\infty e^{(D_0^h\otimes I_l + I_h\otimes D_0^l + (D_1^h\otimes I_l)G_H(w) + I_h\otimes D_1^l w)x}\, dH(x) \\ &= h(-D_0^h\otimes I_l - I_h\otimes D_0^l - (D_1^h\otimes I_l)G_H(w) - I_h\otimes D_1^l w). \end{aligned} \quad (16)$$

In the rest of this subsection we provide expressions for the distribution of the number of packets seen by a low-priority arrival and the waiting time distribution of the low-priority packets, i.e., $w_l(s)$. Since the stationary distributions remain the same as before, we only need to derive the relevant customer (packet) averages. Let π_l be the stationary distribution of the state of the low-priority arrival process. The proofs of these are essentially along the lines of [10, Theorem 9].

Theorem 3.4 *Let* $Y_S^a(z)$ *be the z-transform of the distribution of the total number of packets in the system seen by a low-priority arrival.* $Y_S^a(z)$ *is given by*

$$Y_S^a(z) = Y_S(z)(I_h \otimes D_0^l)e/(\pi_l D_0^l e_l), \quad (17)$$

where e is a column vector of ones of size given by the size of $I_h \otimes D_0^l$ *and* e_l *is a column vector of ones with size the same as* D_0^l. *Thus, the ccdf has a z-transform given by*

$$\tilde{Y}_S^a(z) = \frac{1-Y_S^a(z)}{1-z}. \quad (18)$$

Theorem 3.5 *The 2-dimensional transform of the number of high-priority and low-priority packets seen by a low-priority arrival, denoted by* $Y_a(z,w)$ *is given by*

$$Y^a(z,w) = Y(z,w)(I_h \otimes D_0^l)e/(\pi_l D_0^l e_l), \quad (19)$$

where $Y(z,w)$ *is given in Theorem 3.3. Therefore, the tail of the distribution of the number of low-priority packets seen be a low-priority arrival has a z-transform given by*

$$\tilde{Y}_a(w) = \frac{1-Y_a(1,w)}{1-w}. \quad (20)$$

Theorem 3.6 *The virtual waiting time distribution of the low-priority arrivals is given by*

$$\begin{aligned} W_V^l(s) &= (1-\rho)g_S + \sum_{n>0} y_S(n) G_l(s) G_h^{n-1} \\ &= (1-\rho)g_S + Y_S(G_h(s))G_h^{-1}(s)G_l(s) \\ &- (1-\rho)g_S G_h^{-1}(s) G_l(s). \end{aligned} \quad (21)$$

Theorem 3.7 *The waiting-time distribution of the low-priority arrivals is given by*

$$w_l(s) = W_V^l(s)(I_h \otimes D_0^l)e/(\pi_l D_0^l e_l). \quad (22)$$

Thus, the LST of the low-priority waiting time ccdf is given by

$$w_l^C(s) = \frac{1 - w_l(s)}{s}. \quad (23)$$

4 Asymptotics of the Tail Probabilities

In this section, we present results to show that the low-priority tail asymptotics for waiting times are of the form

$$P(W > T) \sim \alpha T^{-\beta} e^{-\delta T}. \quad (24)$$

For general MAPs, it is difficult to obtain closed-form expressions for α, β and δ. Therefore, we compute these numerically using the transform expressions obtained in the earlier section. For this purpose, we use the techniques in [1].

We generate high and low MAP sources using a superposition of *on-off* sources each with an average rate of 0.0125 as in [7]. As in [7], the mean *on* and *off* times are 436.36 and 4363.63, with the arrival rate during an *on* period being 0.1375. Thus, a high-priority mean arrival rate of 0.05 would mean that there are four independent high-priority MMPP sources. The service times are exponential with mean 1. Of course, the arrival processes and service times could be more general, but the model considered here is sufficient to illustrate the asymptotic behavior. We use the results in the previous sections to calculate the exact tail probabilities and use the moment-based technique in [1] to compute the asymptote.

The parameters of the asymptote, α, β and δ are shown in Table 1 for various values of ρ_h and ρ_l. As can be seen from the table, β can be non-negligible, thus leading to non-exponential asymptotics, in general. It is also interesting to note that the so-called *effective bandwidth approximation*, $e^{-\delta T}$, does not change very much as ρ_l is changed with $\rho_h = 0.0625$. However, the true asymptote changes significantly as revealed by the different values for α and β.

ρ_h	ρ_l	δ	β	$\alpha \times 10^3$
0.025	0.025	0.2322	-0.7170	5.9
0.025	0.0375	0.2356	-0.3710	3.2
0.0625	0.025	0.04	-1	2.5
0.0625	0.0375	0.0411	-0.845	3
0.0625	0.05	0.0425	-0.5	1.4

Table 1: Parameters of the tail asymptote for low-priority waiting times for various values of arrival rates for high and low priority sources

In the $M/G, G/1$ model studied in [3], it was observed that, as the low priority arrival increased, the value of β exhibited the following behavior: β is equal to $-3/2$ till a threshold value of the low-priority arrival rate is reached. At this threshold, β is equal to $-1/2$ and above this threshold, β is equal to zero. We numerically study if this behavior holds with MAP arrival processes. We consider an example with the same type of *on-off* sources as before. We set the high-priority arrival rate $\lambda_h = 0.0125$, and increase the low-priority arrival rate. The results are presented in Table 2.

ρ_l	δ	β	α
0.0125	0.399	-1.18	0.148
0.025	0.4088	-0.38	0.0388
0.0375	0.4089	2.3	$1.3 * 10^{-6}$
0.05	0.4197	4.85	$9 * 10^{-10}$

Table 2: Parameters of the low-priority tail asymptote with $\rho_h = 0.0125$

From Table 2, we see that the β can be positive with MAP arrival processes which is different from the behavior with Poisson arrivals. Figure 1 shows the plots of the exact tail and the asymptote for the case with $\rho_l = 0.0375$. The asymptote increases slightly for small values of T and then decreases since $\beta > 0$. Figure 1 again illustrates that the asymptote may not be very accurate even when T is large enough such that the tail probability is small (around 10^{-7}). Further, the tail probability estimate is optimistic, which may not be suitable for applications. This points to a need for further work on exploring other approximations (other than the asymptote) which are both accurate as well as faster than computing the exact tail probability.

References

[1] J. Abate, G. L. Choudhury, D. M. Lucantoni, and W. Whitt. Asymptotic analysis of tail probabilities based on the computation of moments. *Annals of Applied Probability*, 5:983–1007, 1995.

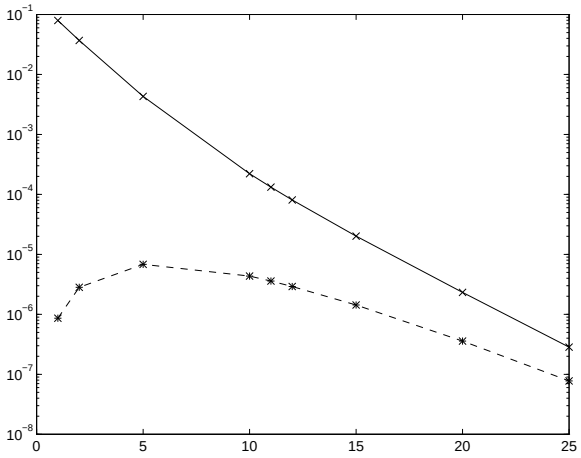

Figure 1: Non-exponential asymptote (dashed) and exact low-priority waiting-time tail probability (solid) for $\rho_h = 0.0125$, $\rho_l = 0.0375$

[2] J. Abate, G. L. Choudhury, and W. Whitt. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Commun. Statist.-Stochastic Models*, 10(1):99–143, 1994.

[3] J. Abate and W. Whitt. Asymptotics for $M/G/1$ low-priority waiting-time tail probabilities. *Queueing Systems*, 25:173–233, 1997.

[4] A. Berger and W. Whitt. Effective bandwidths with priorities, 1997. preprint.

[5] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis. Asymptotic buffer overflow probabilities in multiclass multiplexers, part I: The GPS policy, 1996. MIT LIDS Technical Report, LIDS-P-2341.

[6] C. S. Chang and T. Zajic. Effective bandwidths of departure processes from queues with time varying capacities. In *Proc. INFOCOM '95*, pages 1001–1009, Boston, Massachusetts, USA, 1995.

[7] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44:203–217, 1995.

[8] D.M. Lucantoni. New results on the single server queue with a batch markovian arrival process. *Comm. Statist.-Stochastic Models*, 7(1):1–46, 1991.

[9] D.M. Lucantoni. The $BMAP/G/1$ queue: A tutorial. In L. Donatiello and R. Nelson, editors, *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, pages 330–358. Springer-Verlag, 1993.

[10] D.M. Lucantoni, K.S. Meier-Hellstern, and M.F. Neuts. A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22:676–705, 1990.

[11] M. F. Neuts. *Structured Stochastic Matrices of the $M/G/1$ Type and Their Applications*. Marcel Dekker, Inc., 1989.

[12] I. Ch. Paschalidis. Performance analysis and admission control in multimedia communication networks. In *Proceedings of the IEEE Conference on Decision and Control*, 1997. To appear.

[13] V. Ramaswami. The $N/G/1$ queue and its detailed analysis. *Advances in Applied Probability*, 12:222–261, 1980.

[14] V. Subramanian and R. Srikant. Tail probabilties of low-priority waiting times and queue lengths in $MAP/G/1$ queues, 1997. http://tesla.csl.uiuc.edu/ srikant/pub.html.

[15] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems*, 2:71–107, 1993.