Tail probabilities of low-priority waiting times and queue lengths in $MAP/GI/1$ queues *

Vijay Subramanian [a,**] and R. Srikant [b]

[a] *Mathematics of Communication Networks, Motorola, 1501 W. Shure Drive, Arlington Heights, IL 60004, USA*
E-mail: VSUBRAM2@email.mot.com

[b] *Coordinated Science Laboratory and Department of General Engineering, University of Illinois, 1308 W. Main Street, Urbana, IL 61801, USA*
E-mail: rsrikant@uiuc.edu

Received 15 January 1999; revised 19 August 1999

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 a numerical investigation of the large-buffer asymptotic behavior of the the waiting-time distribution for low-priority sources shows that these asymptotics are often non-exponential.

Keywords: priority queues, tail probabilities, waiting times, Markovian arrival processes

1. Introduction

In this paper, we consider a $MAP/GI/1$ queue with two priority classes operating under a non-preemptive priority service discipline, and we are interested in the behavior of the waiting time of the low priority arrival. MAPs are very commonly used to describe 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 [24,25,27,31].

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-

* Research supported in part by NSF Grant NCR-9701525.
** The first author was previously with the University of Illinois.

priority waiting time. Our numerical results suggest that the tail of the low-priority waiting time distribution is often non-exponential. An inspiration for our work is the results of Abate and Whitt [4], who showed the non-exponential 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. While the results in [4] show the exact nature of the asymptote, we present strong numerical evidence to only demonstrate that the asymptote is non-exponential. In addition, also using numerical results, we show that the behavior of the asymptote is significantly different in the case of general MAP arrivals as compared to the case of Poisson arrivals studied in [4]. Specifically, in the case with Poisson arrivals, the low-priority waiting-time distribution has an exponential asymptote when the low-priority arrival rate is above a certain threshold. This is not necessarily true in the case with MAP arrivals. Other differences are pointed out in 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 prespecified *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, [16,23,36]. Thus, given a finite buffer size B, a natural approximation to the loss probability $P(X > B)$ is

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

Similar large deviations results and approximations are also available for waiting times and queue lengths [36]. In ATM networks, the cell size is constant. A common approximation to this is to use a Erlang service-time distribution with many phases, often making the problem amenable to a matrix–geometric type analysis. This approximation of deterministic service times by a phase-type distribution is discussed extensively in [11], but we do not address it here.

While there are several sources of error in using the approximation (2), it has been shown that the approximation can be refined as $P(X > B) \approx \alpha\,\mathrm{e}^{-\delta B}$, where α can be computed either numerically [11], or by capturing the gains of bufferless statistical multiplexing using the Chernoff bound [17] or appropriate large deviations scaling to account for a large number of sources [8,14]. 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, $P(X > B) \sim \alpha\,\mathrm{e}^{-\delta B}$ [2], where $f(x) \sim g(x)$ denotes $\lim_{x\to\infty} f(x)/g(x) = 1$. Assuming a preemptive resume model,

large deviations results of the form (2) can also be obtained for statistical multiplexers with priorities [6,10], based on earlier work characterizing the large-deviations rate function of the output process of a single-class, single-server queue [9,15]. A more general case has been considered in [7,29], where large-deviation asymptotics have been obtained for queue lengths and virtual waiting-times in two queues operating under the generalized processor-sharing (GPS) discipline. One of the contributions of our paper is to point out that for the nonpreemptive priority model (which should be very close to the preemptive resume model when the packet sizes are small), the asymptotic form $P(X > B) \sim \alpha\, \mathrm{e}^{-\delta B}$ does not hold, in general, for the low-priority sources.

Previously, Takine et al. [35] considered nonpreemptive priority queues with Markovian arrivals and independent and identically distributed service times for each packet. They derived expressions for the mean waiting time of each packet class and also presented algorithms for evaluating the mean waiting times. Takine et al. [34] derived expressions for the workload in a $MAP/GI/1$ queue with state-dependent arrivals. This was used to derive the Laplace–Stieltjes transform (LST) for the waiting time of packets in each class of a $MAP/GI/1$ queue with preemptive resume priority. Sugahara et al. [32] analyzed a two-priority (nonpreemptive priority) queue with the high class arrivals from a 2-state Markov Modulated Poisson process and the low class arrivals from a Poisson process. Using a supplementary variable technique for solving the partial differential equations for the system, they derived the joint probability generating functions of the stationary queue-length distributions and the LST of the stationary waiting time-distributions of high- and low-priority packets. Our derivation of these quantities use a different approach primarily relying on the properties of MAP arrival processes and the key renewal theorem for Markov renewal processes [20] as in the derivation of the results in [26]. Further, we also extend the results to the case where both the high- and low-priority classes are MAPs using Palm theory [5].

Elwalid and Mitra [18] consider a multi-service multiplexing system with Markov modulated fluid inputs and derive approximations to the tail probabilities of the queue-length distributions. The main idea there is an elegant Markovian approximation for the busy-period of the high-priority packets. The Markovian approximation then naturally provides a exponential tail for the low-priority buffer occupancy, which is used in an admission control scheme for such a multiplexer. Since our model is not a fluid model, the results in [18] do not directly apply. It may be possible to connect our results to the results in [18] using appropriate fluid limits as in [12]. However, we do not pursue this avenue of research in this paper. Using the results in [18], Presti et al. [30] have obtained approximations for the more general case of GPS service discipline.

The rest of this paper is organized as follows. In section 2, we introduce MAPs and present some known properties of MAPs and $MAP/GI/1$ queues that will be used later. Priority queues are considered in section 3 and an exact expression for the transform of the waiting-time distribution of the low-priority packets is derived. Section 4 presents numerical results and discusses the conclusions from these results.

2. MAPs and $MAP/GI/1$ queues

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

$$Q = \begin{bmatrix} D_0 & D_1 & \mathbf{0} & \mathbf{0} & \dots \\ & D_0 & D_1 & \mathbf{0} & \dots \\ & & D_0 & D_1 & \ddots \\ & & & \ddots & \ddots \end{bmatrix},$$

where $D \equiv D_0 + D_1$ with $D \neq D_0$, is also a generator. We assume that D is irreducible. 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 [27] and studied under various names [24,25,31], 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 [25] 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 [24–26,31].

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 $\{(n, j),\ n \geqslant 0,\ j = 1, 2, \dots, m\}$ with generator Q. Define $D(z)$ to be

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

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

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

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

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

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

$$\lambda = \pi D_1 e,$$

and since D is irreducible, this is also the long-term arrival rate.

[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.

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, $h(s) = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}H(t)$) and mean $1/\mu$ with $\mu > 0$. The traffic intensity at the $MAP/GI/1$ queue is given by $\rho = \lambda/\mu$.

Lemma 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

$$A(z, s) = h\big(sI - D(z)\big), \quad \text{where } \mathrm{Re}(s) \geqslant 0 \text{ and } |z| \leqslant 1.$$

In the previous lemma $h(sI - D(z))$ is a scalar function evaluated at a matrix argument. This is evaluated in the standard manner by substituting the matrix argument in the power series expansion of h. This same interpretation will be used in the rest of the paper whenever a scalar function is evaluated at a matrix argument. For more details, the reader is referred to [25, section 5.1].

Lemma 2. We have the following equation for the 2-dimensional transform of the number served during, and the duration of, the busy period

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

Lemma 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 at times of departures given that the system is empty.

Lemma 4. Define $A(z) \equiv A(z, 0) = h(-D(z))$. The queue-length generating function of a $MAP/GI/1$ queue at an arbitrary time is given by

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

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 $D_0^l = -\lambda_l$ and $D_1^l = \lambda_l$. Define $D_h(z) \equiv D_0^h + D_1^h z$ and $D_S(z) \equiv D_0^h + D_1^h z + z\lambda_l I - \lambda_l I$. Then, lemmas 1–4 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 [37] 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 (3) 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 $R(t) = \mu \int_0^t (1 - H(u))\,\mathrm{d}u$ 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 [13; 19, chapter XIV]. 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.[2]
- W_2: The $n-1$ high-priority busy periods generated by the $n-1$ packets ahead of it in the queue.[3]

The LST of W_1 can be easily seen to be

$$G_l(s) = r\big(sI - D_h\big[G_h(s)\big]\big).$$

[2] Note that we need to consider the phase of the arrival process at the beginning of the service and at the end. Thus, W_1 is actually a matrix-valued random variable.

[3] Again we need to take into account of the arrival process phases. Thus, W_2 is also a matrix-valued random variable.

From lemma 2, the high-priority busy period is given by

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

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

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

Now we are ready to state and prove the following theorem.

Theorem 5. 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

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

Proof. Let $y_S(n)$ be distribution of the total number of customers in the system, i.e., the inverse transform of $Y_S(z)$ given by (3) with $D(z)$ replaced by $D_S(z)$. In other words, $Y_S(z) = \sum_{n=0}^{\infty} y_S(n)z^n$. Thus, $y_S(n)$ is a vector such that each element of the vector corresponds to an arrival phase and gives the probability that the total number (sum of high-priority and low-priority packets) is n. Therefore, from the discussion prior to the theorem, we have that the LST of the waiting-time distribution of the low-priority packets $w_l(s)$ is given by

$$w_l(s) = (1-\rho) + \sum_{n>0} y_S(n)G_l(s)G_h^{n-1}(s)e, \tag{7}$$

where we have used the fact that the waiting time is zero when a packet arrives when the system is empty, which occurs with probability $1-\rho$. From equation (4) we deduce that $G_h(s)$ is a power series in $D_h[G_h(s)]$. From equation (5) we can deduce that even $G_l(s)$ is a power series in $D_h[G_h(s)]$. Thus, $G_l(s)$ and $G_h(s)$ commute. Using this we have the following form for $w_l(s)$:

$$\begin{aligned} w_l(s) &= (1-\rho) + \left\{\sum_{n>0} y_S(n)G_h^n(s)\right\}G_h^{-1}(s)G_l(s)e \\ &= (1-\rho) + \big\{Y_S\big(G_h(s)\big) - (1-\rho)g_S\big\}G_h^{-1}(s)G_l(s)e. \end{aligned} \tag{8}$$

For the last equation above, we have used the interpretation of g_S from lemma 3. □

The next result presents the transform of the stationary distribution of number of high-priority and low-priority packets in the system, the proof of which is provided in the appendix. 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 6. 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 i. 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

$$\begin{aligned} TY(z,w) &= P(0,0)\big(-\big(D_0^h - \lambda_l I\big)^{-1}\big)\big(zD_1^h + w\lambda_l I - I\big)\big(I - A(z,w)\big)\big(-D(z,w)^{-1}\big) \\ &\quad + P(0,0)\big[-\big(I - A(z,0)\big)\big(-D(z,0)^{-1}\big) + \big(-\big(D_0^h - \lambda_l\big)^{-1}\big)\big] \\ &\quad + P(z,w)\big(I - A(z,w)\big)\big(-D(z,w)^{-1}\big) \\ &\quad + P(z,0)\big(I - A(z,0)\big)\big(-D(z,0)^{-1}\big), \end{aligned} \tag{9}$$

where

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

$$\begin{aligned} P(z,w) &= \left\{ P(0,w)\left(\frac{z}{w} - 1\right) + P(0,0)\,\frac{z}{w}\big[-w\big(D_0^h - \lambda_l I\big)^{-1}\big(D_1^h + \lambda_l I\big) - I\big] \right\} \\ &\quad \times A(z,w)\big(zI - A(z,w)\big)^{-1}, \end{aligned} \tag{10}$$

and

$$A(z,w) = \int_0^{\infty} \mathrm{e}^{(D_0^h - \lambda_l I + D_1^h z + \lambda_l I w)x}\,\mathrm{d}H(x), \tag{11}$$

$$P(0,w) = Q(w)\big(P(0,1)e\big), \tag{12}$$

$$Q(0) = \frac{P(0,0)}{(P(0,1)e)}, \tag{13}$$

$$Q(w) = Q(0)\big\{-w\big(D_0^h - \lambda_l I\big)^{-1}\big(D_1^h + \lambda_l\big) - I\big\}G_H(w)\big(wI - G_H(w)\big)^{-1}, \tag{14}$$

$$G_H(w) = G_h(\lambda_l - \lambda_l w), \tag{15}$$

$$P(0,0) = \frac{1-\rho}{\lambda_h + \lambda_l I}\, g_{\mathrm{S}}\big(-D_0^h + \lambda_l I\big), \tag{16}$$

$$P(0,1)e = \frac{\lambda_l}{\lambda_h + \lambda_l} + \frac{1-\rho}{\lambda_h + \lambda_l}\, g_{\mathrm{S}} D_1^h e. \tag{17}$$

We can now write out the expression for $Y^C(w)$ using the expression for $Y(z,w)$ and the following relation:

$$Y^C(w) = \frac{1 - Y(1,w)e}{1-w}. \tag{18}$$

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 approaches standard in matrix-analytic theory, we have the following theorem which is the equivalent of theorem 6. The differences between the proofs of the two theorem is outlined in the appendix. Note that we only need to make the appropriate changes to the renewal function of the Markov process, the distribution of arrivals, and the distribution of customers at departure epochs in the proof of theorem 6 given in the appendix. While theorem 6 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 7. 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 i. 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 (9) 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) $D_h(z) = D_0^h \otimes I_l + D_1^h \otimes I_l z$,

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

(vii) $D_{\mathrm{S}}(z) = D(z,z)$, and

(viii) $G_H(w)$ solves

$$\begin{aligned} G_H(w) &= A\big(G_H(w), w\big) \\ &= \int_0^{\infty} \mathrm{e}^{(D_0^h \otimes I_l + I_h \otimes D_0^l + (D_1^h \otimes I_l) G_H(w) + I_h \otimes D_1^l w)x}\, \mathrm{d}H(x) \\ &= h\big(-D_0^h \otimes I_l - I_h \otimes D_0^l - \big(D_1^h \otimes I_l\big) G_H(w) - I_h \otimes D_1^l w\big). \end{aligned} \quad (19)$$

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 [26, theorem 9] where the relationship

between time averages and customer averages in [33] is generalized to $MAP/GI/1$ queues.

Theorem 8. 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

$$Y_S^a(z) = Y_S(z)\frac{(I_h \otimes D_0^l)e}{\pi_l D_0^l e_l}, \tag{20}$$

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

$$\widetilde{Y}_S^a(z) = \frac{1 - Y_S^a(z)}{1 - z}. \tag{21}$$

Theorem 9. 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

$$Y^a(z, w) = Y(z, w)\frac{(I_h \otimes D_0^l)e}{\pi_l D_0^l e_l}, \tag{22}$$

where $Y(z, w)$ is given in theorem 7. Therefore, the tail of the distribution of the number of low-priority packets seen by a low-priority arrival has a z-transform given by

$$\widetilde{Y}_a(w) = \frac{1 - Y_a(1, w)}{1 - w}. \tag{23}$$

Theorem 10. The virtual waiting-time distribution of the low-priority arrivals is given by

$$\begin{aligned} W_V^l(s) &= (1-\rho)g_S + \sum_{n>0} y_S(n)G_l(s)G_h^{n-1} \\ &= (1-\rho)g_S + \big\{Y_S\big(G_h(s)\big) - (1-\rho)g_S\big\}G_h^{-1}(s)G_l(s). \end{aligned} \tag{24}$$

Theorem 11. The waiting-time distribution of the low-priority arrivals is given by

$$w_l(s) = W_V^l(s)\frac{(I_h \otimes D_0^l)e}{\pi_l D_0^l e_l}. \tag{25}$$

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

$$w_l^C(s) = \frac{1 - w_l(s)}{s}. \tag{26}$$

4. Asymptotics of the tail probabilities

In this section, we present results to show that the low-priority tail asymptotics for waiting times are non-exponential. Assuming an asymptote of the form

$$P(W > T) \sim \alpha T^{-\beta} \, \mathrm{e}^{-\delta T}, \tag{27}$$

we present numerical evidence to show that often $\beta \neq 0$. 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]. However, to use these numerical results, one has to first hypothesize a form for the asymptotics and, then, proceed to obtain the parameters of the asymptotic form. Therefore, we first present a simple model which suggests asymptotics of the form (27). This example proving the asymptotic form is not for MAP processes but it is for fluid models which are closely related to MAPs when the service times are small, as shown in [12]. However, it should be noted that nonpreemptive and preemptive priorities are the same in fluid models, and thus, we lose the detailed structure of the MAP queueing system when considering the fluid model. Thus, this example only suggests the asymptotics, but is not a conlcusive proof that the asymptote is indeed of the assumed form. Nevertheless, it serves as a simple example to illustrate why one might suspect non-exponential asymptotic behavior for the low-priority tail probability.

Our principal conclusion from the numerical study *does not* depend on the assumption that the asymptotic form is given by (27). Our principal conclusion is simply that the asymptote is non-exponential, which is arrived at by showing that β converges to a nonzero value. Hence, the assumption of the asymptotic form is not important to reach this conclusion. But the fluid-model example and the results in [4] suggest the form (27), and, thus, it is interesting to study whether the parameter β behaves as it does for the case of Poisson arrivals studied in [4]. The numerical evidence will show that, in general, the behavior is different with general MAP arrivals, i.e., either the asymptotic form is not of the type given in (27), or, if it is, then the behavior of β is different.

4.1. Fluid model example

As mentioned earlier, non-exponential tails for waiting times of low-priority waiting times was shown in [4] for $M/GI/1$ models. The key idea was to use a geometric random sum representation for the transform of the low-priority waiting times where the stationary excess of the "service time" of the low-priority packets includes the effect of the busy period due to the high-priority sources. The example we provide here complements the results in [4] by considering a high-priority source that is an on–off Markov-modulated rate process which is commonly used in modeling arrival processes. However, as we shall see, our example which considers the tail of the low-priority workload has a simpler direct interpretation in terms of the busy period of

the $M/GI/1$ queue. Since we are considering a fluid model, it is simpler to consider workload, instead of waiting times, and hence, we do so.

Consider a multiplexer with two buffers. Buffer 1 is fed by source 1 that alternates between on and off states. It spends an exponentially distributed amount of time in each state with the mean *on* and *off* times given by $1/q_d$ and $1/q_u$, respectively. When source 1 is *on*, it produces fluid at rate λ_1. Source 2 is a constant rate source producing fluid at rate λ_2. The server rate is c and its service discipline is generalized processor sharing (GPS) with weights ϕ_1 and ϕ_2 for sources 1 and 2, respectively [28]. Comparing this to our MAP models earlier, this would be the fluid-equivalent of a two-state Markov-modulated Poisson process (MMPP) for source 1 and a Poisson process for source 2 with the service times being a Erlang-k distribution for large k. Note that the choice of $\phi_2 = 0$ leads to a strict priority scheme with lower priority for source 2 as we had been considering in the previous sections. But here we let ϕ_2 be possibly nonzero to allow us to study the more general case of GPS.

We assume that

$$\lambda_1 > \frac{\phi_1 c}{\phi_1 + \phi_2} \quad \text{and} \quad \lambda_2 > \frac{\phi_2 c}{\phi_1 + \phi_2}.$$

We will denote service rate guaranteed to source i as g_i where $g_i = c\phi_i/(\phi_1 + \phi_2)$. Clearly, when $\lambda_2 < g_2$, buffer 2 is always empty. In what follows, we show that one can exactly determine the asymptotic form of the workload in buffer 2 and when $\lambda_2 > g_2$, this asymptotic form is non-exponential. For stability, we assume that

$$\left(\frac{q_u}{q_u + q_d}\right)\lambda_1 + \lambda_2 < c.$$

We first note the following two facts that we will use:

- Due to our assumptions on the arrival rates λ_1 and λ_2, whenever source 1 is either *on* or back-logged, it receives service at rate g_1, and source 2 receives service at rate g_2.
- When buffer 1 is empty, source 2 receives service at rate c when it is back-logged and at rate λ_2 when its buffer is empty.

A busy period of source 1 begins in the *on* state and ends when its buffer is empty. We denote the LST of its busy period as $B(s)$. We derive an expression for $B(s)$ by exploiting a connection between the fluid model and an appropriately defined $M/GI/1$ queue as in [21,22].

Let $X_1(t)$ be $X_2(t)$ be the workload in buffers 1 and 2, respectively. Define $Y_1(t) \equiv X_1(t)/g_1$. Let t_n be the begin time of the nth *off* period of source 1. Then,

$$Y_1(t_n) = \big[Y_1(t_{n-1}) - a_n\big]^+ + \frac{\lambda_1 - g_1}{g_1}\, b_n, \tag{28}$$

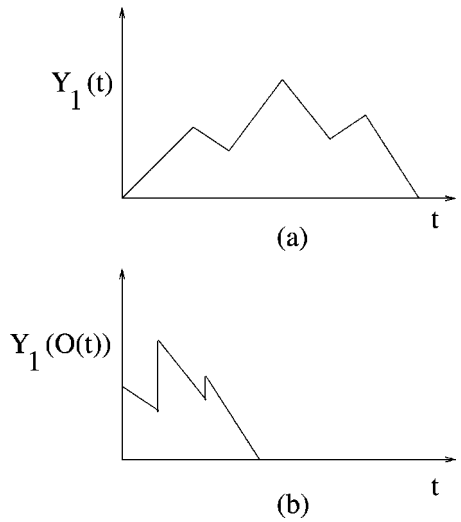

Figure 1. (a) Actual busy period due to source 1. (b) Busy period considering only the off-times of source 1.

where a_n is the duration of the $(n-1)$th *off* period and b_n is the duration of the *on* period following the $(n-1)$th *off* period. Further, for $0 \leqslant \delta < a_{n+1}$,

$$Y_1(t_n + \delta) = \big[Y_1(t_n) - \delta\big]^+. \tag{29}$$

Compare (28) and (29) to Lindley's difference equations for the workload in a $G/GI/1$ queue, with the beginning of an *on*-time corresponding to an arrival epoch. We immediately recognize that, by restricting our attention to *off* times alone, $Y_1(\mathrm{O}(t))$ is the workload in an $M/M/1$ queue with arrival rate q_u and mean service-time $(\lambda_1 - g_1)/g_1 q_d$, where $\mathrm{O}(t)$ is the total *off* time in $[0, t]$. From figure 1, it is also immediate that the busy-period of source 1 is $\lambda_1/(\lambda_1 - g_1)$ times the busy-period of the $M/M/1$ queue. If we denote the LST of the busy-period of the $M/M/1$ queue, then the busy-period LST of source 1 is given by

$$B(s) = B_M\left(\frac{\lambda_1 s}{\lambda_1 - g_1}\right). \tag{30}$$

Having characterized the busy-period of source 1, we now turn our attention to buffer 2 whose asymptotic behavior we would like to characterize. Based on our analysis of the busy-period of source 1, source 2 can be thought of being served by a server that alternates between an *up* and a *down* state as follows:

- *Up* state: Server stays in this state for an exponentially distributed duration $1/q_u$ and the server capacity is c in this state.
- *Down* state: Duration in this state has an LST $B(s)$, and the server capacity is g_2 in this state.

Now, define $Y_2(t) \equiv X_2(t)/(c - \lambda_2)$. Let $F(y)$ be the steady-state cdf of of Y_2, $\widehat{F}_u(y)$ be the steady-state $P(Y_2 < y \mid \text{server is up})$ and $\widehat{F}_d(y)$ be the steady-state $P(Y_2 < y \mid \text{server is down})$. In a manner similar to the busy-period analysis for buffer 1, it is easy to see that $\widehat{F}_u^{\mathrm{c}}(y) = F_w^{\mathrm{c}}(y)$, where $F_w(y)$ is the waiting-time cdf in an $M/GI/1$ queue with arrival rate q_u, and service-time distribution LST $\widehat{B}(s)$ given by

$$\widehat{B}(s) = B\left(\frac{\lambda_2 - g_2}{c - \lambda_2}\, s\right), \tag{31}$$

and the superscript c denotes ccdf. Thus, the LST of $\widehat{F}_u^{\mathrm{c}}(y)$, denoted by $\hat{f}_u^{\mathrm{c}}(s)$, can be written as $\hat{f}_u^{\mathrm{c}}(s) = (1 - \hat{f}_u(s))/s$, where $\hat{f}_u(s)$ is given by the Kendall functional equation for the busy-period of an $M/GI/1$ queue

$$\hat{f}_u(s) = \widehat{B}\big(s + q_u - \hat{f}_u(s)\big). \tag{32}$$

Now, let $F_u^{\mathrm{c}}(y) = P(Y_2 > y \text{ and server is up})$ and $F_d^{\mathrm{c}}(y) = P(Y_2 > y \text{ and server is down})$ in steady-state. Equating probability drifts, which can be justified rigorously using level-crossing analysis, it is easy to see that

$$F_u^{\mathrm{c}}(y) = \frac{\lambda_2 - g_2}{c - \lambda_2}\, F_d^{\mathrm{c}}(y).$$

Therefore,

$$F^{\mathrm{c}}(y) = F_u^{\mathrm{c}}(y) + F_d^{\mathrm{c}}(y) = \frac{c - g_2}{\lambda_2 - g_2}\, F_w^{\mathrm{c}}(y) = \left(\frac{c - g_2}{\lambda_2 - g_2}\right)\gamma \widehat{F}_u^{\mathrm{c}}(y), \tag{33}$$

where

$$\gamma \equiv 1 - \frac{q_u}{q_u + q_d}\,\frac{\lambda_1}{g_1}$$

is the probability that the server is *up*.

Now it is easy to see that the $F^{\mathrm{c}}(y) \sim \alpha y^{-3/2}\, \mathrm{e}^{-\delta y}$ as follows: The asymptotic form of $F^{\mathrm{c}}(y)$ is the same as that of $F_u^{\mathrm{c}}(y)$ except for a constant. Since $F_u^{\mathrm{c}}(y)$ has the same LST as that of the busy-period of an $M/M/1$ queue, from [3], its asymptotic form is as desired. The exact values of α and δ can be calculated from [3, equation (4.1)] using (31)–(33).

4.2. Numerical results

We generate high and low MAP sources using a superposition of *on–off* sources each with an average rate of 0.0125 as in [11]. As in [11], 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

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

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

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	$1.3 \cdot 10^{-6}$
0.05	0.4197	4.85	$9 \cdot 10^{-10}$

the previous sections to calculate the exact tail probabilities and use the moment-based technique in [1] to compute the asymptote. To ensure the accuracy of the numerical examples, we use a large number of moments ($\geqslant$40) and also stop the computation only when there is less than a 10% change in the computed parameters. For instance, if δ is computed to be 0.4, then the number of moments used in the computation would be large enough to ensure that the value of δ changed by less than 0.01 in successive steps of the iteration given in [1].

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.

In the $M/G, G/1$ model studied in [4], it was observed that, as the low-priority arrivals 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. From table 2, we see that the behavior of the asymptote is different with general MAP arrivals than with Poisson arrivals: either the asymptotic form of the tail distribution is different or the behavior of β is different. As an example, if the asymptotic form that has been hypothesized is correct, then β can be positive with MAP arrival processes which is different from the behavior with Poisson arrivals.

While the form of the asymptote is yet to be proven, due to the fluid example and the results in [4], it is interesting to verify how well the expression (27) performs as an approximation to the exact tail probability. The analysis of $M/G/1$ queues with priorities in [4] suggests that, when $\beta \neq 0$, the asymptote may not lead to a good approximation of the exact tail probabilities. This is illustrated in figures 3, 4. The

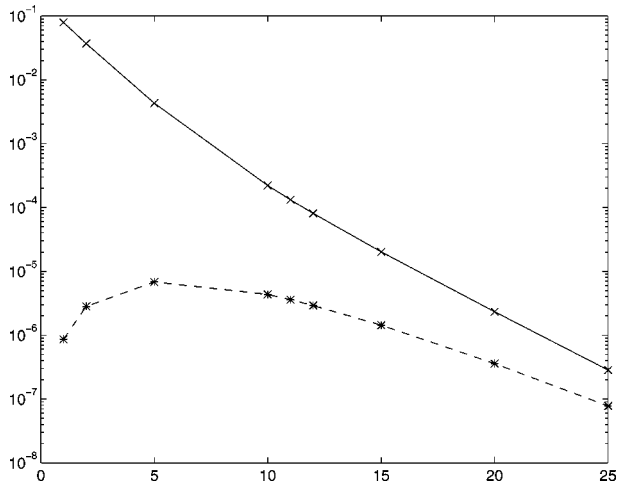

Figure 2. Non-exponential asymptote (dashed) and exact low-priority waiting-time tail probability (solid) for $\rho_h = 0.0125$, $\rho_l = 0.0375$.

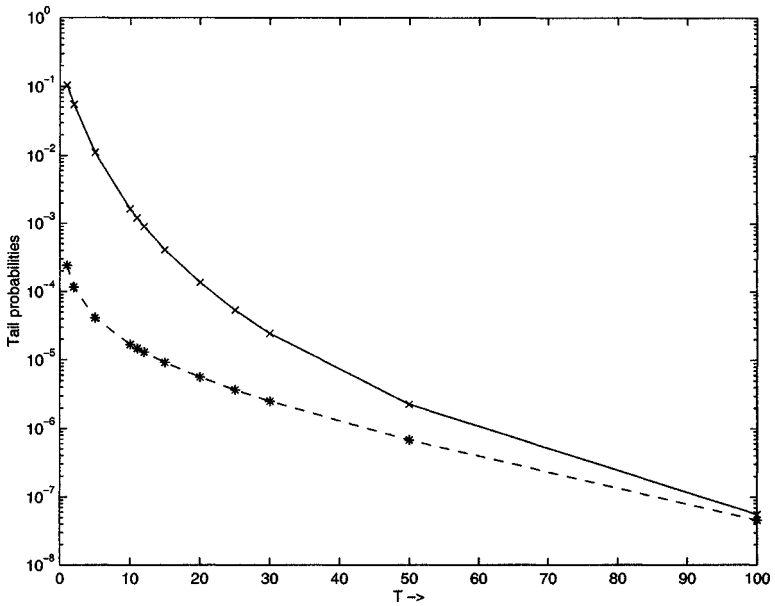

Figure 3. Non-exponential asymptote (dashed) and exact low-priority waiting-time tail probability (solid) for $\rho_h = 0.0625$, $\rho_l = 0.025$.

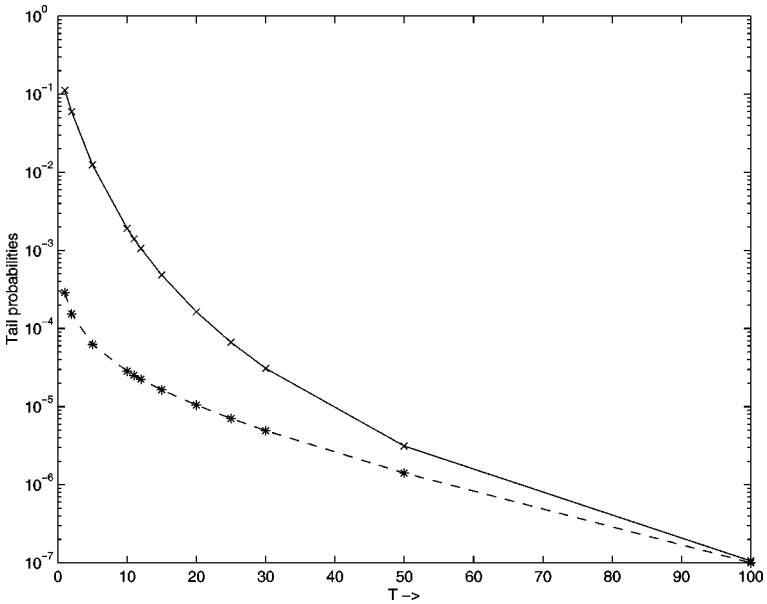

Figure 4. Non-exponential asymptote (dashed) and exact low-priority waiting-time tail probability (solid) for $\rho_h = 0.0625$, $\rho_l = 0.0375$.

figures also show another important feature: even if the asymptote is valid, it need not be an upper bound for the exact tail probability. However, in general, it is not always optimistic either. For instance, in the case of Poisson arrivals for both high-priority and low-priority sources, the non-exponential asymptote in [4, table 14.2] provides a conservative approximation to the exact tail probability.

Figure 2 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 2 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 proving the exact form of the asymptote, and exploring other approximations (other than the asymptote) which are both accurate as well as faster than computing the exact tail probability.

Appendix

Proof of theorem 6

Our proof is a generalization of a similar proof for vacation models in [26]. Consider the Markov renewal process obtained by sampling the system at departure epochs. Let $p(n, m)$ be the stationary probability vector at departure epochs, i.e., the jth component of $p(n, m)$ gives the stationary probability that there are n high-priority

and m low-priority customers, and the arrival process is in phase j at a departure epoch. From [35], the 2-dimensional transform of $p(n, m)$, denoted by $P(z, w)$, is given by (10), where

- $A(z, w)$ is the transform of the number of arrivals of each class (z denotes the high priority and w the low priority) in a service-time,
- $Q(w)$ is the stationary distribution of the number of low-priority packets in the system obtained by sampling the system just after departures which leave no high-priority packets in the system, and
- $G_H(w)$ is the transform of the number of low-priority packet arrivals in a high-priority busy-period.

From the stationary distribution obtained by sampling the system just after departures, we need to derive the stationary distribution of the queueing process. Using PASTA we would then have the distribution of packets including the number of each class, seen by a low priority arrival.

Let $Y(z, w) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} y(n, m) z^n w^m$ be the transform of the stationary distribution of the number of packets of each class in the queue.

Let $y(n, m, t)$ be a a vector whose ith component is the probability that there are n packets of high priority, m packets of low priority, and the phase of the input process is i, at time t, starting at some initial state at time 0. The initial state is irrelevant because we will be considering only steady-state quantities and the key renewal theorem helps us remove the dependence on the initial state. Thus, we suppress the initial state in the analysis that follows. Let $M^{n,m}(t)$ denote the renewal function (here a vector) associated with the Markov regenerative process (with the above initial state), i.e., the ith component of the vector $M^{n,m}(t)$ is the mean number of visits to the state (n, m) and phase i in $[0, t)$. The following equations hold for $y(n, m, t)$:

(i) $n > 0,\ m > 0,$

$$
\begin{aligned}
&y(n, m, t) \\
&\quad = \int_0^t \mathrm{d}M^{0,0}(u) \int_0^{t-u} \mathrm{d}P_a(1, 0, v) P_a(n-1, m, t-u-v)\big(1 - H(t-u-v)\big) \\
&\qquad + \int_0^t \mathrm{d}M^{0,0}(u) \int_0^{t-u} \mathrm{d}P_a(0, 1, v) P_a(n, m-1, t-u-v)\big(1 - H(t-u-v)\big) \\
&\qquad + \sum_{k=1}^{n} \sum_{l=0}^{m} \int_0^t \mathrm{d}M^{k,l}(u) P_a(n-k, m-l, t-u)\big(1 - H(t-u)\big) \\
&\qquad + \sum_{l=1}^{m} \int_0^t \mathrm{d}M^{0,l}(u) P_a(n, m-l, t-u)\big(1 - H(t-u)\big); \qquad \text{(A.1)}
\end{aligned}
$$

(ii) $n > 0,\ m = 0,$

$$y(n,0,t) = \int_0^t \mathrm{d}M^{0,0}(u) \int_0^{t-u} \mathrm{d}P_a(1,0,v) P_a(n-1,0,t-u-v)\big(1-H(t-u-v)\big) + \sum_{k=1}^{n} \int_0^t \mathrm{d}M^{k,0}(u) P_a(n-k,0,t-u)\big(1-H(t-u)\big); \tag{A.2}$$

(iii) $n = 0,\ m > 0,$

$$y(0,m,t) = \int_0^t \mathrm{d}M^{0,0}(u) \int_0^{t-u} \mathrm{d}P_a(0,1,v) P_a(0,m-1,t-u-v)\big(1-H(t-u-v)\big) + \sum_{l=1}^{m} \int_0^t \mathrm{d}M^{0,l}(u) P_a(0,m-l,t-u)\big(1-H(t-u)\big); \tag{A.3}$$

(iv) $n = 0,\ m = 0,$

$$y(0,0,t) = \int_0^t \mathrm{d}M^{0,0}(u) P_a(0,0,t-u), \tag{A.4}$$

where $P_a(k,l,t)$ is the matrix of probabilities of k high-priority arrivals, and l low-priority arrivals in $(0,t)$. We will explain equation (A.1); the other equations are special cases. We implicitly take into account information of the state of the arrival process by considering all quantities to be vectors or matrices. Equation (A.1) is obtained by conditioning on the last epoch of the Markov renewal process as follows:

(i) The last epoch of the Markov renewal process is the vector $(0,0)$ and one of the two following events occur so that the state at time t is (n,m):

  (a) An arrival of high priority occurs first; this packet gets into service and remains in service. During the service of this packet $n-1$ high-priority packets and m low-priority customers arrive.

  (b) An arrival of low priority occurs first; this packet gets into service and remains in service. During the service of this packet n high-priority packets and $m-1$ low-priority customers arrive.

(ii) The last epoch of the Markov renewal process is (k,l) with $0 < k \leqslant n,\ 0 \leqslant l \leqslant m$. One of the high-priority packets (the first actually) gets into service and stays in service. During the service of this packet $n-k$ high-priority customers and $m-l$ low-priority customers arrive.

(iii) The last epoch of the Markov renewal process is $(0,l)$ with $0 < l \leqslant m$. The first low-priority packet gets into service and stays in service. During the service of this packet n high-priority customers and $m-l$ low-priority packets arrive.

As in [26], taking limits as $t \to \infty$, and using the key renewal theorem, we get the following equations for $y(n,m) = \lim_{t\to\infty} y(n,m,t)$:

(i) $n > 0,\ m > 0,$

$$\begin{aligned} y(n,m) = {} & P(0,0)/T\big(-\big(D_0^h - \lambda_l I\big)^{-1}\big) D_1^h \int_0^\infty P_a(n-1,m,t)\big(1-H(t)\big)\,\mathrm{d}t \\ & + P(0,0)/T\big(-\big(D_0^h - \lambda_l I\big)^{-1}\big)\lambda_l I \int_0^\infty P_a(n,m-1,t)\big(1-H(t)\big)\,\mathrm{d}t \\ & + \sum_{k=1}^{n}\sum_{l=0}^{m} P(k,l)/T \int_0^\infty P_a(n-k,m-l,t)\big(1-H(t)\big)\,\mathrm{d}t \\ & + \sum_{l=1}^{m} P(0,l)/T \int_0^\infty P_a(n,m-l,t)\big(1-H(t)\big)\,\mathrm{d}t; \end{aligned} \tag{A.5}$$

(ii) $n > 0,\ m = 0,$

$$\begin{aligned} y(n,0) = {} & P(0,0)/T\big(-\big(D_0^h - \lambda_l I\big)^{-1}\big) D_1^h \int_0^\infty P_a(n-1,0,t)\big(1-H(t)\big)\,\mathrm{d}t \\ & + \sum_{k=1}^{n} P(k,0)/T \int_0^\infty P_a(n-k,0,t)\big(1-H(t)\big)\,\mathrm{d}t; \end{aligned} \tag{A.6}$$

(iii) $n = 0,\ m > 0,$

$$\begin{aligned} y(0,m) = {} & P(0,0)/T\big(-\big(D_0^h - \lambda_l I\big)^{-1}\big)\lambda_l I \int_0^\infty P_a(0,m-1,t)\big(1-H(t)\big)\,\mathrm{d}t \\ & + \sum_{l=1}^{m} P(0,l)/T \int_0^\infty P_a(0,m-l,t)\big(1-H(t)\big)\,\mathrm{d}t; \end{aligned} \tag{A.7}$$

(iv) $n = 0,\ m = 0,$

$$y(0,0) = P(0,0)/T\big(-\big(D_0^h - \lambda_l I\big)^{-1}\big). \tag{A.8}$$

The 2-dimensional transform of $y(n,m)$ yields the desired expression for $Y(z,w)$.

Proof of theorem 7

As in the proof of theorem 6 we can follow a detailed cookie-cutting argument for the evolution of the distribution of customers in the queue. The resulting equations are exactly like equations (A.1)–(A.4) in the proof of theorem 6 with the definitions of $P_a(k,l,t)$ and $M^{n,m}(t)$ modified to take into account the joint phase of the high- and low-priority arrival processes. Defining $P_a(z,w,t)$ to be the (double) z-transform of $P_a(k,l,t)$, we have that $P_a(z,w,t) = \mathrm{e}^{D(z,w)t}$. The Key renewal theorem gives a limit similar to that obtained in equations (A.5)–(A.8) with the changes given in the statement of theorem 7. We only have to take into account the changed $P_a(z,w,t)$. For the renewal function, the dependence on the initial state disappears in the limit

as a consequence of the Blackwell renewal theorem. The case of lower traffic being Poisson is obtained as a special case with $D(z,w) = D_0^h - \lambda_l I_h + D_1^h z + \lambda_l I_h w$.

References

[1] J. Abate, G.L. Choudhury, D.M. Lucantoni and W. Whitt, Asymptotic analysis of tail probabilities based on the computation of moments, Ann. Appl. Probab. 5 (1995) 983–1007.

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

[3] J. Abate and W. Whitt, Approximations for the $M/M/1$ busy period, in: *Queueing Theory and Its Applications. Liber Amicorium for J.W. Cohen*, eds. O.J. Boxma and R. Syski (North-Holland, Amsterdam, 1988) pp. 149–191.

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

[5] F. Baccelli and P. Brémaud, *Elements of Queueing Theory* (Springer, New York, 1991).

[6] A. Berger and W. Whitt, Effective bandwidths with prioritie, IEEE/ACM Trans. Networking 6 (1998) 447–460.

[7] D. Bertsimas, I.Ch. Paschalidis and J.N. Tsitsiklis, Asymptotic buffer overflow probabilities in multiclass multiplexers, part I: The GPS policy, IEEE Trans. Automat. Control 43 (1996) 315–335.

[8] D.D. Botvich and N. Duffield, Large deviations, the shape of the loss curve and economies of scale in large multiplexers, Queueing Systems 20 (1995) 293–320.

[9] C.S. Chang, Sample path large deviations and intree networks, Queueing Systems 20 (1995) 7–36.

[10] C.S. Chang and T. Zajic, Effective bandwidths of departure processes from queues with time varying capacities, in: *Proc. of IEEE INFOCOM*, Boston, MA (1995) pp. 1001–1009.

[11] G.L. Choudhury, D.M. Lucantoni and W. Whitt, Squeezing the most out of ATM, IEEE Trans. Commun. 44 (1995) 203–217.

[12] G.L. Choudhury, A. Mandelbaum, M. Reiman and W. Whitt, Fluid and diffusion limits for queues in slowly changing environments, Stochastic Models 13(1) (1996).

[13] R.B. Cooper, *Queueing Theory* (North-Holland, New York, 1981).

[14] C. Courcoubetis and R.R. Weber, Buffer overflow asymptotics for a switch handling many traffic sources, J. Appl. Probab. (September 1996).

[15] G. de Veciana, C. Courcoubetis and J. Walrand, Decoupling bandwidths for networks: A decomposition approach to resource management, in: *Proc. of IEEE INFOCOM* (1994) pp. 466–473.

[16] G. de Veciana, G. Kesidis and J. Walrand, Resource management in wide-area ATM networks using effective bandwidths, IEEE J. Selected Areas Commun. 13 (1995) 1081–1090.

[17] A. Elwalid, D. Heyman, T.V. Lakshman, D. Mitra and A. Weiss, Fundamental results on the performance of ATM multiplexers with applications to video teleconferencing, IEEE J. Selected Areas Commun. (1995) 1004–1016.

[18] A.I. Elwalid and D. Mitra, Analysis, approximations and admission control of a multi-service multiplexing system with priorities, in: *Proc. of IEEE INFOCOM* (1995) pp. 463–472.

[19] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II (Wiley, New York, 1966).

[20] D. Heyman and M. Sobel, *Stochastic Models in Operations Research*, Vol. I (McGraw-Hill, New York, 1982).

[21] J.Q. Hu and D. Xiang, The queueing equivalance to a manufacturing system with failures, IEEE Trans. Automat. Control 38 (1993) 499–502.

[22] O. Kella and W. Whitt, A storage model with a two-state random environment, Oper. Res. 40 (1992) S257–S262.

[23] F.P. Kelly, Notes on effective bandwidths, in: *Stochastic Networks: Theory and Applications*, eds. F.P. Kelly, S. Zachary and I.B. Ziedins (1996) pp. 141–168.
[24] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Comm. Statist. Stochastic Models 7(1) (1991) 1–46.
[25] D.M. Lucantoni, The $BMAP/G/1$ queue: A tutorial, in: *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello and R. Nelson (Springer, Berlin, 1993) pp. 330–358.
[26] 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, Adv. in Appl. Probab. 22 (1990) 676–705.
[27] M.F. Neuts, *Structured Stochastic Matrices of the* $M/G/1$ *Type and Their Applications* (Marcel Dekker, New York, 1989).
[28] A. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The single node case, IEEE/ACM Trans. Networking (1993).
[29] I.Ch. Paschalidis, Performance analysis and admission control in multimedia communication networks, in: *Proc. of the IEEE Conf. on Decision and Control* (1997).
[30] F.L. Presti, Z.-L. Zhang and D. Towsley, Bounds, approximations and applications for a two-queue GPS system, in: *Proc. of IEEE INFOCOM* (1996).
[31] V. Ramaswami, The $N/G/1$ queue and its detailed analysis, Adv. Appl. Probab. 12 (1980) 222–261.
[32] A. Sugahara, T. Takine, Y. Takahashi and T. Hasegawa, Analysis of a nonpreemptive priority queue with SPP arrivals of high class, Performance Evaluation 21 (1995) 215–238.
[33] L. Takàcs, The limiting distribution of the virtual waiting time and the queue size for a single-server queue with recurrent input and general service times, Sankhyā A 25 (1963).
[34] T. Takine and T. Hasegawa, The workload in the $MAP/G/1$ queue with state-dependent services: its application to a queue with preemptive resume priority, Comm. Statist. Stochastic Models 10 (1994) 183–204.
[35] T. Takine, Y. Matsumoto, T. Suda and T. Hasegawa, Mean waiting times in non-preemptive priority queues with Markovian arrival and i.i.d. service processes, Performance Evaluation 20 (1994) 131–149.
[36] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues, Telecommunication Systems 2 (1993) 71–107.
[37] R.W. Wolff, *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, NJ, 1989).