COMMUN. STATIST.-STOCHASTIC MODELS, 10(1), 99-143 (1994)

ASYMPTOTICS FOR STEADY-STATE TAIL PROBABILITIES IN STRUCTURED MARKOV QUEUEING MODELS

Joseph ABATE
900 Hammond Road
Ridgewood, NJ 07450-2908

Gagan L. CHOUDHURY
AT&T Bell Laboratories
Holmdel, NJ 07733-3030

Ward WHITT
AT&T Bell Laboratories
Murray Hill, NJ 07974-0636

Key Words: Markov chains, $M/G/1$-type Markov chain, batch Markovian arrival process, versatile Markovian point process, $BMAP/G/1$ queue, asymptotics for steady-state distributions, tail probabilities in queues, Perron-Frobenius eigenvalue, Tauberian theorem

ABSTRACT

We apply Tauberian theorems with known transforms to establish asymptotics for the basic steady-state distributions in the $BMAP/G/1$ queue. The batch Markovian arrival process ($BMAP$) is equivalent to the versatile Markovian point process or Neuts (N) process; it generalizes the Markovian arrival process (MAP) by allowing batch arrivals. We consider the waiting time, the workload (virtual waiting time) and the queue lengths at an arbitrary time, just before an arrival and just after a departure. We begin by establishing asymptotics for steady-state distributions of $M/G/1$-type Markov chains. Then we treat steady-state distributions in the $BMAP/G/1$ and $MAP/MSP/1$ queues. The MSP is a MAP independent of the arrival process generating service completions during

the time the server is busy. In great generality, but not always, the basic steady-state distributions in these models have asymptotically exponential tails. When they do, the asymptotic parameters of the different distributions are closely related.

1. Introduction

In this paper we establish asymptotics for steady-state distributions of *GI/M/*1-type and *M/G/*1-type Markov chains, as in Neuts [26,28], supplementing recent work by Asmussen and Perry [5], Baiocchi [6], Elwalid and Mitra [14] and Falkenberg [16,17], and earlier work by Neuts [27], Neuts and Takahashi [29] and Takahashi [32]. We apply Tauberian theorems for generating functions and Laplace transforms in Feller [17, pp. 445-7] to establish asymptotics for the steady-state distributions of *M/G/*1-type Markov chains and the basic processes of the *BMAP/G/*1 queue. The generating function of the steady-state distribution of an *M/G/*1-type Markov chain is given in Neuts [28, p. 143]. The transforms of the steady-state distributions in the *BMAP/G/*1 queue are given in Ramaswami [30], Neuts [28] and Lucantoni [23,24]. The *BMAP/G/*1 model has a single server, the first-come first-served service discipline, an unlimited waiting room and i.i.d. service times that are independent of a batch Markovian arrival process (*BMAP*).

To model service times that are not necessarily i.i.d., we also consider the *MAP/MSP/*1 queue, which has a Markovian service process (*MSP*) as well as a Markovian arrival process (*MAP*). The *MSP* is a *MAP* independent of the arrival process that generates service completions when the server is busy. The *MAP/MSP/*1 queue can be represented as a quasi-birth-and-death (*QBD*) process, so that the asymptotic behavior is determined by previous results of Neuts [26,27]. We can also treat the multi-server *MAP/MSP/m* queue, because it can be represented in a *GI/M/*1-type Markov chain. The *MAP/MSP/m* model is interesting because the arrival and service processes have the same structure. The spectral analysis of this model when the *MAP* and *MSPs* are Markov modulated Poisson processes (*MMPPs*) is discussed in Elwalid and Mitra [14]. The references contain related asymptotic results for other models and additional references; e.g., see [1,3,5,7,8,19].

Let W be the steady-state *waiting time* experienced by an arriving customer before beginning service; let L be the steady-state *workload* at an arbitrary time, or the virtual waiting time; and let Q, Q^a and Q^d be the steady-state *queue lengths,* the number in system at an arbitrary time, just before an arrival and just after a departure, respectively. Under quite general conditions, we show that there exist positive constants $\eta, \sigma, \alpha_L, \alpha_W, \beta$, β^a and β^d such that

$$e^{\eta x} P(L > x) \to \alpha_L \text{ and } e^{\eta x} P(W > x) \to \alpha_W \text{ as } x \to \infty \qquad (1)$$

and

$$\sigma^{-k}P(Q > k) \to \beta\,,\ \ \sigma^{-k}P(Q^a > k) \to \beta^a$$
$$\text{and } \sigma^{-k}P(Q^d > k) \to \beta^d \text{ as } k \to \infty\,. \tag{2}$$

We call η and σ the *asymptotic decay rates* and $\alpha_L, \alpha_W, \beta, \beta^a$ and β^d the associated *asymptotic constants.* Instead of (2), we actually establish related results for probability mass functions, such as

$$\sigma^{-k}P(Q = k) \to \beta(1-\sigma)/\sigma \text{ as } k \to \infty\,. \tag{3}$$

The convergence in (3) and (2) are equivalent, but (2) is usually of greater applied interest. We also establish limits for the joint distribution of these variables and the auxiliary phase state.

However, a word of caution is in order. *While the limits in (1)–(3) typically hold, they are by no means automatic.* Some conditions must be satisfied. Even for a stable *M/G/*1 queue where the service-time distribution has a finite moment generating function in a neighborhood of the origin, (1)–(3) need *not* hold. This is illustrated by Example 5 of [1]. In that example, the rightmost singularity of the Laplace transform is a branch point, not a simple pole.

We also establish important relations among the asymptotic parameters in (1)–(3) in the *BMAP/G/*1 model, extending previous results in that direction by Neuts [27]. For example, if V is a generic service time, the asymptotic decay rates η and σ are related by $Ee^{\eta V} = \sigma^{-1}$. There are also simple relations among the asymptotic constants: $\beta^a = \alpha_W$, $\beta = \alpha_L$ and $\alpha_L/\alpha_W = \rho(1-\sigma)/\eta\sigma$. Therefore, if we know the asymptotic decay rates and one asymptotic constant, then we know the other asymptotic constants. Indeed, assuming that we know the service-time transform Ee^{sV}, we only need to know one of the asymptotic decay rates. Additional results of this kind appear in [2] and [19].

In Section 5 we illustrate our results for the *BMAP/G/*1 queue by analyzing an $MMPP_2/D_2/1$ example. The $MMPP_2$ arrival process is a Markov modulated Poisson process (*MMPP*) with a two-state environment Markov chain, while D_2 is a two-point service-time distribution. As in [1,2], our exact numerical results are based on the algorithms in Lucantoni [23], using numerical transform inversion algorithms in Abate and Whitt [4], as implemented by Choudhury, Lucantoni and Whitt [11]. The asymptotic parameters are also calculated by a moment-based numerical inversion algorithm in Choudhury and Lucantoni [9]. The algorithm in [9] calculates moments of all desired orders.

The *BMAP/G/*1 and *MAP/MSP/*1 models are attractive since they include superposition arrival processes. Whitt [33] applies the asymptotic decay rates in the *BMAP/G/*1 and *MAP/MSP/*1 queues to develop effective bandwidths for independent sources to use for admission control in multi-service networks. This paper provides theoretical support for the procedures in [33]. For related work, see Chang [8], Elwalid and Mitra [13,14] and references therein. However, a further word of caution is in order, because we have found that the quality of the effective-bandwidth approximations can deteriorate dramatically with superposition arrival processes as the number of component arrival processes increases; see [10].

2. Structured Markov Chains

For Markov chains of *GI/M/*1 type, the asymptotics of steady-state distributions is discussed in Neuts [27]. The steady-state probability vector is matrix-geometric, i.e., of the form $x_0, x_1, x_1 R, x_1 R^2, \ldots$ where R is the rate matrix. It has the asymptotic form

$$x_1 R^i = \sigma^i (x_1 r) l + o(\sigma^i) \text{ as } i \to \infty , \tag{4}$$

where the decay rate σ is the Perron-Frobenius eigenvalue of R, and l and r are left and right eigenvectors associated with σ normalized so that $le = lr = 1$, with e being a vector of 1's; see [27, p. 224]. (Our σ and η are η and ξ in [27].) Neuts points out that the asymptotic decay rate σ is relatively easy to obtain, even for large models, but the asymptotic constant (vector) $(x_1 r) l$ usually is not.

We now investigate the asymptotics of steady-state distributions of *M/G/*1-type Markov chains. The transition probability matrix is of the form

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & B_4 \cdots \\ C_0 & A_1 & A_2 & A_3 & A_4 \cdots \\ 0 & A_0 & A_1 & A_2 & A_3 \cdots \\ 0 & 0 & A_0 & A_1 & A_2 \cdots \\ 0 & 0 & 0 & A_0 & A_1 \cdots \\ 0 & 0 & 0 & 0 & A_0 \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

with component $m \times m$ matrices A_k, $m_0 \times m_0$ matrix B_0, $m_0 \times m$ matrices B_k and $m \times m_0$ matrix C_0. Our matrix P corresponds to the matrix $Q(\infty)$ in (2.1.9) of Neuts [28, p. 76], with $A_k = A_k(\infty)$, $B_k = B_k(\infty)$ and $C_0 = C_0(\infty)$. We assume that the Markov chain with transition matrix P is irreducible and positive recurrent. In the queueing models, positive recurrence primarily corresponds to $\rho < 1$, where ρ is the traffic intensity. (We also need a moment condition on $\{B_k\}$.)

Let $(x_0, x_1, \ldots)$ be the steady-state vector, with $x_0 = (x_{01}, \ldots, x_{0m_0})$ and $x_i = (x_{i1}, \ldots, x_{im})$, and let $X(z) = \sum_{i=1}^{\infty} x_i z^i$ be the (vector) generating function. By (3.3.2) of [28, p. 143], the generating function satisfies the equation

$$X(z)[zI - A(z)] = zx_0 B(z) - zx_1 A_0 , \tag{5}$$

for all complex z with $|z| \le 1$, where $A(z) = \sum_{k=0}^{\infty} A_k z^k$ and $B(z) = \sum_{k=1}^{\infty} B_k z^k$.

We now want to establish (3). One way is to use a "probabilistic" or "large-deviations" approach, exploiting change of measure, as in [5,8,19]. However, since the transform is available in (5), it is natural to use it. One way to use the transform is to study the singularities of $X(z)$, as discussed in Wilf [34, Section 5.2]. With this approach, we need to identify the *radius of convergence* and show that the only singularity on the radius of convergence is a *simple pole on the real axis.* This approach is used by Falkenberg [15,16]; for related work see Gail, Hantler and Taylor [18].

What we do is apply the Tauberian theorem for generating functions in Feller [17, p. 447]. However, Tauberian theorems require extra conditions. Under regularity conditions, we prove that $(1-z\sigma)X(z)$ converges to a specified limit ξ as $z \to \sigma^{-1}$. This almost implies that $\sigma^{-i}x_i$ converges to the same limit ξ as $i \to \infty$, but not quite. By [17], the transform limit implies *Cesàro convergence* of $\sigma^{-i}x_i$, i.e., $n^{-1}\sum_{i=1}^{n}\sigma^{-i}x_i \to \xi$ as $n \to \infty$. Thus, we assume that $\sigma^{-i}x_i$ converges to something, possibly 0 or ∞ in each coordinate. Then the Cesàro limit implies that $\sigma^{-i}x_i$ actually must converge to ξ.

Our analysis shows that σ^{-1} is the radius of convergence of the generating function $X(z)$ and that there is a singularity on the real axis at σ^{-1}. Our extra assumption is tantamount to *assuming* that there are no other singularities on the circle $|z| = \sigma^{-1}$. Falkenberg [15,16] makes a similar assumption. A virtue of our approach is that we obtain explicit expressions for both asymptotic parameters ξ and σ.

Example 1. To see that in general a transform limit does not imply ordinary convergence, consider the probability density function

$$f(x) = e^{-2x}(\frac{8}{3} + \sin x - \frac{4}{3}\cos x) , x \geq 0 ,$$

with Laplace transform

$$\hat{f}(s) \equiv \int_0^\infty e^{-sx} f(x)\,dx = \frac{4}{3}\left[\frac{2}{2+s}\right] + \frac{1-(4/3)(s+2)}{1+(s+2)^2} .$$

It is easy to see that $(s+2)\hat{f}(s) \to 8/3$ as $s \to -2$, but $e^{2x}f(x)$ oscillates as $x \to \infty$; i.e., $e^{2x}f(x)$ converges to 8/3 with Cesàro convergence but not ordinary convergence. (Similar examples can be constructed with generating functions.) The transform $\hat{f}(s)$ has three singularities for s such that $\mathrm{Re}(s) = -2$, namely, -2 and $-2 \pm i$. In this example, $e^{sx}f(x)$ is periodic, but it is not difficult to construct examples where $e^{sx}f(x)$ is not periodic, e.g., it could be asymptotically periodic, as with

$$f(x) = e^{-2x}\left[\frac{8}{3} + \sin(x^2/(x+1)) - \frac{4}{3}\cos(x^2/(x+1))\right] .$$

Nevertheless, when one of these densities is used as the service-time density in an $M/G/1$ queue, the limits in (1)–(3) are valid for all arrival rates such that $\rho < 1$. Hence, it may be possible to eliminate the convergence assumptions we make, e.g., on $\sigma^{-i} x_i$. ■

We first show that the asymptotic decay rate is independent of the auxiliary phase state. Motivated by Example 5 in [1], we allow asymptotic behavior of the form $\beta i^{-p} \sigma^i$ as $i \to \infty$.

Theorem 1. *Consider an irreducible positive-recurrent Markov chain of M/G/1 type. Suppose that* $\underline{\lim}_{i \to \infty} \sigma^{-i} i^p x_{ij} = l_j$ *and* $\overline{\lim}_{i \to \infty} \sigma^{-i} i^p x_{ij} = u_j$ *for all j, where* $0 \le l_j \le u_j \le \infty$ *and* $-\infty < p < \infty$. *If* $A \equiv \sum_{k=0}^{\infty} A_k$ *is irreducible, then* $l_j > 0$ $(u_j < \infty)$ *holds for one j if and only if it holds for all j.*

Proof. Consider an initial state (i,j) where i is the level and j is the phase, with $i \ge m+1$. Let j' be a designated alternative phase state. Since A is irreducible and there are only m phase states, it is possible to go from (i,j) to $(i+k, j')$ in at most m steps for some k. Since the chain can go down at most one level at each transition, we can have $k \ge -m$. In particular, there is a finite product of at most m of the $m \times m$ submatrices A_l that produce this transition. It is significant that this bounding probability is independent of i, provided that $i \ge m+1$, since it is impossible to reach the lower boundary level from level above level $m+1$ in m steps. As a consequence, there is a constant ε as well as the constant k such that

$$x_{ij} \geq \varepsilon\, x_{i+k,j'} \tag{6}$$

for all i. Since the states j and j' are arbitrary, formula (6) implies that

$$\begin{aligned} \underline{\lim}_{i \to \infty} \sigma^{-i} i^p x_{ij} &\geq \varepsilon\, \underline{\lim}_{i \to \infty} \sigma^{-i} i^p x_{i+k,j'} \\ &\geq \varepsilon \sigma^k\, \underline{\lim}_{i \to \infty} \sigma^{-(i+k)} (i+k)^p x_{i+k,j'} \end{aligned} \tag{7}$$

and

$$\begin{aligned} \overline{\lim}_{i \to \infty} \sigma^{-(i+k)} (i+k)^p x_{i+k,j'} &\leq \varepsilon^{-1}\, \overline{\lim}_{i \to \infty} \sigma^{-(i+k)} (i+k)^p x_{ij} \\ &\leq \varepsilon^{-1} \sigma^{-k}\, \overline{\lim}_{i \to \infty} \sigma^{-i} i^p x_{ij}\ . \end{aligned} \tag{8}$$

The inequalities (7) and (8) imply the desired conclusion. ■

From (5) it is evident that the radius of convergence of the generating function $X(z)$ *should be* the minimal $z > 1$ such that $p(z) = z$ where $p(z) \equiv pf(A(z))$ is the *Perron-Frobenius eigenvalue* of $A(z)$. (It is understood that $p(z) = \infty$ if any elements of $A(z)$ are not finite.) Properties of $p(z)$ are discussed in Chapter 1 and Section 2.3 of Seneta [31] as well as in Section 2.3 and the Appendix of Neuts [28]. We will need the following basic convexity result due to Kingman [22].

Theorem 2. *If the elements of an irreducible nonnegative square matrix $A(s)$ are log-convex functions of s, then the Perron-Frobenius eigenvalue $pf(A(s))$ is log-convex.*

We apply Theorem 2 to transforms and related functions, using the following consequence of Hölder's inequality; see [22, p. 284].

Lemma 1. *For any random variable X, the transform Ee^{sX} is log-convex in s.*

Theorem 3. *Let $p(z) \equiv pf(A(z))$ be the Perron-Frobenius eigenvalue of $A(z)$ for $z > 0$, with $p(z) = \infty$ if $A(z)$ is not finite. Then $\log p(e^s)$ is an increasing convex function of s, so that when $A(1)$ is irreducible the equation $p(z) = z$ has at most one root for $z \neq 1$, and, if there is a root, this root must satisfy $z > 1$.*

Proof. The argument is a variant of the proof of Lemma 2.3.4 in Neuts [28, p. 94]. Make the change of variables $z = e^s$. Then the elements of the matrix $A(e^s)$ are log-convex functions of s by Lemma 1. (The elements are constant multiples of transforms.) Consequently, $\log p(e^s)$ is a convex function of s on $(-\infty,\infty)$ by Theorem 2. (The relevant domain for s in [28] is $(-\infty,0)$, but we are interested in $(0,\infty)$; we can conclude that $\log p(e^s)$ is convex on the interval it is finite.) Moreover, $pf(A_1) \leq pf(A_2)$ if A_1 and A_2 are nonnegative irreducible matrices with $A_1 \leq A_2$ (elementwise), with $pf(A_1) < pf(A_2)$ if $A_1 \neq A_2$; e.g., by Theorem 1.6 of Seneta [31, p. 23]. Hence, $\log p(e^s)$ is an increasing convex function of s on $(-\infty,\infty)$. Note that $h(s) \equiv \log p(e^s) = s$ if and only if $p(e^s) = e^s$. Since $A(1)$ is stochastic, $p(1) = 1$. The assumed positive recurrence of the Markov chain P implies that $\rho \equiv p'(1) < 1$; see [28, pp. 124-128, 481]. These relations in turn imply that $h(0) = 0$ and $h'(0) < 1$, so that, by the convexity, $h(s) = s$ for at most one other s, and such an s must be positive. Thus, $p(z) = z$ for at most one $z \neq 1$ and such a z must satisfy $z > 1$. ■

Example 5 in [1] involving the $M/G/1$ queue shows that the equation $p(z) = z$ may actually *not* have a root for $z > 1$, even when the service-time distribution has a finite moment generating function $\phi(s) \equiv Ee^{sV}$ for $s > 0$. A necessary and sufficient condition for a root $z > 1$ *not* to exist for some traffic intensity (arrival rate) is for ϕ to have a finite radius of convergence s^* with $\phi(s^*) < \infty$; see [1]. This can occur only when the rightmost singularity of the Laplace transform $\phi(-s)$ is a branch point singularity. When there is no root to $p(z) = z$ for $z > 1$, the steady-state waiting time W does not satisfy (1). Similarly, Q^d (which has the steady-state distribution of an $M/G/1$-type Markov chain) does not satisfy (2) or (3). In Example 5 of [1], $P(Q > k) \sim \bar{\beta} x^{-3/2} \sigma^k$ as $k \to \infty$, which is not a good approximation until k is very large. For further discussion, see [1], [3] and Borovkov [7, Section 22].

It is convenient to rewrite (5) as

$$X(z)(I - \bar{A}(z)) = x_0 B(z) - x_1 A_0 \tag{9}$$

where $\bar{A}(z) = A(z)/z$. Let $\bar{p}(z)$ be the Perron-Frobenius eigenvalue of

$\bar{A}(z)$ when z is real and positive. By Theorem 3, we are interested in the z such that $\bar{p}(z) = 1$ and $z > 1$. We next introduce a generalization of the fundamental matrix of Kemeny and Snell [21], which is a convenient form of a generalized inverse; see Hunter [20], Theorem 3.3.1 in Neuts [28, p. 144] and Baiocchi [6].

Consider an irreducible nonnegative matrix P with Perron-Frobenius eigenvalue $p \leq 1$ and let l and r be left and right eigenvectors associated with p normalized so that $le = lr = 1$, where e is a vector of 1's. Let the *fundamental matrix* associated with P be $Z = (I - P + prl)^{-1}$. Note that Z is the familiar fundamental matrix in Kemeny and Snell [21, pp. 75, 100] when P is stochastic. (Then l is the steady-state vector π and $r = e$, the vector of 1's.) Next, we relate the spectral radiuses of P and $P - prl$. The spectral radius of P is the Perron-Frobenius eigenvalue since P is nonnegative; that is not the case for $P - prl$.

Lemma 2. *If P is an irreducible aperiodic nonnegative matrix with positive Perron-Frobenius eigenvalue p and associated left and right eigenvectors l and r normalized so that $le = lr = 1$, then $P - prl$ is a matrix with spectral radius strictly less than p.*

Proof. By the orthogonality of eigenvectors, P and $P - prl$ have the same eigenvectors, with the eigenvalue of $P - prl$ associated with l and r being 0. The remaining eigenvalues of the matrices P and $P - prl$ coincide.

Since P is aperiodic, p is strictly greater than the modulus of any other eigenvalue of P. ■

Theorem 4. *If P is an irreducible nonnegative matrix with Perron-Frobenius eigenvalue $p \leq 1$ with associated left and right eigenvectors l and r normalized so that $le = lr = 1$, then $(I - P + prl)$ is nonsingular,*

$$Z \equiv (I - P + prl)^{-1} = I + \sum_{n=1}^{\infty} (P - prl)^n$$

$$= I + \sum_{n=1}^{\infty} (P^n - p^n rl) , \quad (10)$$

$Zr = r > 0$ and $lZ = l > 0$.

Proof. First, if $p < 1$, then $P^n \to 0$ and $p^n \to 0$, so that $(P^n - p^n rl) = (P - prl)^n \to 0$ as $n \to \infty$, so that (10) is valid; see Kemeny and Snell [21, pp. 22, 75] and Seneta [31, p. 252]. Henceforth, assume that $p = 1$. First suppose that P is aperiodic. By Lemma 2 $sp(P - prl) < 1$, so that $(P - prl)^n \to 0$ and we can reason as above. Next, suppose that P is periodic with period d. Then we can apply the argument in [21, Section 5.1]. Then P^d is a transition matrix associated with d separate ergodic sets, each of which is aperiodic. Treating each of these separately, we obtain limits for P^{nd+j} as $n \to \infty$ for each j. We then obtain a Cesàro limit for $(P - prl)^n \to 0$ as $n \to \infty$, which implies (10); see [21, p. 23]. Finally,

$$Zr = (I + \sum_{n=1}^{\infty} P^n - p^n rl)r = r + \sum_{n=1}^{\infty} (p^n r - p^n r) = r\,,$$

where r is strictly positive. A similar argument applies to lZ. ■

Theorem 5. *Consider an irreducible positive-recurrent Markov chain of M/G/1 type in which the matrix $A \equiv A(1)$ is irreducible. Let $r(z)$ and $l(z)$ be the eigenvectors of $\bar{A}(z)$ in (9) normalized so that $l(z)e = l(z)r(z) = 1$. Suppose that the equation $p(z) \equiv pf(A(z)) = z$ has a root σ^{-1} for $z > 1$, $p(\sigma^{-1}+\varepsilon) < \infty$ for some positive ε, $B(\sigma^{-1})$ is finite, $\sigma^{-i}x_{ij} \to y_j$ as $i \to \infty$ where $0 \le y_j \le \infty$ for each j, and $(x_0 B(\sigma^{-1}) - x_1 A_0) r(\sigma^{-1}) > 0$. Then $X(z)$ in (5) and (9) is finite for $1 < z < \sigma^{-1}$, the derivative $p'(\sigma^{-1})$ exists with $p'(\sigma^{-1}) > 1$ and*

$$\sigma^{-i}x_i \to \xi \equiv \frac{(x_0 B(\sigma^{-1}) - x_1 A_0) r(\sigma^{-1}) l(\sigma^{-1})}{p'(\sigma^{-1}) - 1} \quad \text{as } i \to \infty\,, \tag{11}$$

with all components of the limit in (11) being strictly positive and finite.

Proof. Since the matrix A is irreducible, so is $A(z)$ for all $z > 0$. Hence, $pf(A(z))$ is a simple eigenvalue. By Theorem 3, there is at most one root for $z > 1$. We now show that $X(z)$ in (9) is finite in a neighborhood of $z = \sigma^{-1}$ with $|z| < \sigma^{-1}$. First, since $B(\sigma^{-1})$ is finite by assumption, the right side of (9) is finite for $|z| < \sigma^{-1}$. For real z with $1 < z < \sigma^{-1}$, $1 < p(z) < z$ by the proof of Theorem 3. Hence, $sp\bar{A}(z) < 1$ and the inverse $(I - \bar{A}(z))^{-1}$ exists and is finite; see Seneta [31, p. 252].

(Moreover, since $sp(\bar{A}(z)) \le sp(\bar{A}(|z|))$, the inverse is also finite for complex z in the neighborhood of $z = \sigma^{-1}$ with $|z| < \sigma^{-1}$, although that is not required for the Tauberian theorem in [17].) Hence, let z be such that $1 < z < \sigma^{-1}$, so that $\bar{p}(z) < 1$. Let $l(z)$ and $r(z)$ be left and right eigenvectors associated with $\bar{p}(z)$. Multiplying by $r(z)$ on the right in (9), we see that

$$X(z)r(z)(1-\bar{p}(z)) = (x_0 B(z) - x_1 A_0)r(z) \equiv H(z)r(z) \ , \quad (12)$$

so that

$$X(z)(I - \bar{A}(z) + \bar{p}(z)r(z)l(z)) = H(z) + \frac{H(z)\bar{p}(z)r(z)l(z)}{1-\bar{p}(z)}$$

and, by Theorem 4,

$$X(z) = \left[H(z) + \frac{H(z)\bar{p}(z)r(z)l(z)}{1-\bar{p}(z)} \right] (I - \bar{A}(z) + \bar{p}(z)r(z)l(z))^{-1} \ .$$

Now we apply the Tauberian theorem in [17, p. 447]. (There we let the slowly varying function L be a constant and let $\rho = 1$.) Our assumption that $\sigma^{-i}x_i$ converges as $i \to \infty$ allows us to strengthen Cesàro convergence to ordinary convergence. We obtain

$$\begin{aligned} \lim_{i \to \infty} \sigma^{-i} x_i &= \lim_{z \to \sigma^{-1}} (1 - z\sigma) X(z) \\ &= \lim_{z \to \sigma^{-1}} (1 - z\sigma) \left[H(z) + \frac{H(z)\bar{p}(z)r(z)l(z)}{1-\bar{p}(z)} \right] \end{aligned}$$

$$\times (I - \bar{A}(\sigma^{-1}) + \bar{p}(\sigma^{-1}) r(\sigma^{-1}) l(\sigma^{-1}))^{-1}$$

$$= \frac{H(\sigma^{-1}) \bar{p}(\sigma^{-1}) r(\sigma^{-1}) l(\sigma^{-1})}{\sigma^{-1} \bar{p}'(\sigma^{-1})}$$

$$\times (I - \bar{A}(\sigma^{-1}) + \bar{p}(\sigma^{-1}) r(\sigma^{-1}) l((\sigma^{-1}))^{-1}$$

$$= \frac{H(\sigma^{-1}) r(\sigma^{-1}) l(\sigma^{-1})}{p'(\sigma^{-1}) - 1} (I - \bar{A}(\sigma^{-1}) + r(\sigma^{-1}) l(\sigma^{-1}))^{-1} ,$$

because

$$\lim_{z \to \sigma^{-1}} \frac{\bar{p}(z) - 1}{z\sigma - 1} = \frac{\bar{p}'(\sigma^{-1})}{\sigma} = p'(\sigma^{-1}) - 1 .$$

The derivative $p'(\sigma^{-1})$ exists because, by assumption, $p(\sigma^{-1} + \varepsilon) < \infty$; see the Appendix of [28]. Since $A(1)$ is positive recurrent, $p'(1) < 1$, as noted in the proof of Theorem 3. Since $h(s) \equiv \log p(e^s)$ is convex in s, we must have $h'(\sigma^{-1}) = p'(\sigma^{-1})\sigma^{-1}/p(\sigma^{-1}) = p'(\sigma^{-1}) > 1$. Next, by Theorem 4,

$$l(\sigma^{-1})(I - \bar{A}(\sigma^{-1}) + r(\sigma^{-1}) l(\sigma^{-1}))^{-1} = l(\sigma^{-1}) .$$

Finally, all components of the limit in (11) are strictly positive because $p'(\sigma^{-1}) > 1$, all components of $l(\sigma^{-1})$ are strictly positive and, by assumption, $(x_0 B(\sigma^{-1}) - x_1 A_0) r(\sigma^{-1})$ is positive. ■

Remark 1. The condition that $(x_0 B(\sigma^{-1}) - x_1 A_0) r(\sigma^{-1})$ be strictly positive needs to be checked in applications. In specific instances, this can often be easily done, as we show for the *BMAP/G/*1 queue in the next section. In general, we know that $r(1) = e$ since $A \equiv A(1)$ is stochastic and that $(x_0 B(1) - x_1 A_0)e = 0$; see [28, p. 145]. By (12), $(x_0 B(z) - x_1 A_0) r(z) > 0$ for all z, $1 < z < \sigma^{-1}$. For the *M/G/*1 queue, $x_1 = A_0^{-1}(1 - B_0)x_0$ by [28, p. 16], so that

$$x_0 B(\sigma^{-1}) - x_1 A_0 = x_0 (B(\sigma^{-1}) - (1 - B_0)) , \tag{13}$$

which is strictly positive for $\sigma^{-1} > 1$ because $B(z) > 0$ for $z > 0$. ∎

Corollary. *Under the conditions of Theorem 5,*

$$\sigma^{-i} x_i e \to \xi e = \frac{(x_0 B(\sigma^{-1}) - x_1 A_0) r(\sigma^{-1})}{p'(\sigma^{-1}) - 1} ,$$

$$\frac{x_{i+1,j}}{x_{ij}} \to \sigma , \ \frac{x_{i+1} e}{x_i e} \to \sigma \text{ and } \frac{x_{ij}}{x_i e} \to l(\sigma^{-1})$$

as $i \to \infty$.

3. The *BMAP/G/*1 Queue

The queue length at departure epochs in the *BMAP/G/*1 queue is a Markov chain of the *M/G/*1 type. Therefore, we can apply Theorem 5 to the *BMAP/G/*1 queue.

A *BMAP* can be defined by processes $N(t)$ and $J(t)$: $N(t)$ counts the number of arrivals in $[0,t]$ and $J(t)$ is an auxiliary state variable. The pair

$(N(t), J(t))$ is a continuous-time Markov chain with generator $\tilde{Q}$ in block-partitioned form, i.e.,

$$\frac{\tilde{Q}}{\rho} = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 \cdots \\ & D_0 & D_1 & D_2 \cdots \\ & & D_0 & D_1 \cdots \\ & & & D_0 \cdots \\ & & & \vdots \end{bmatrix} \tag{14}$$

where ρ is the overall arrival rate, D_k, $k \geq 0$, are $m \times m$ matrices, D_0 has negative diagonal elements and nonnegative off-diagonal elements, D_k is nonnegative for $k \geq 1$, and $D \equiv \sum_{k=0}^{\infty} D_k$ is an irreducible generator matrix for an m-state continuous-time Markov chain. (We choose ρ so that $\sum_{k=1}^{\infty} kD_k e = 1$.)

Let V denote a generic service time. We assume that $\phi(s) \equiv Ee^{sV} < \infty$ for some positive s. This is a necessary, but not sufficient, condition for the asymptotics in (1)–(3). If $\phi(s+\varepsilon) < \infty$ for some $\varepsilon > 0$, then $\phi'(s) = EVe^{sV} < \infty$ too. Key quantities are $\phi(\eta)$ and $\phi'(\eta)$ where η is the asymptotic decay rate in (1).

The generating functions of Q^d and Q are given in (20) and (35) of Lucantoni [23], while the Laplace transform of L is given in (44) there. Let x_{ij}^d be the steady-state probability that the queue length is i and the auxiliary state is j just after a departure. Let $x_i^d = (x_{i1}^d, \ldots, x_{im}^d)$ and let $X^d(z) = \sum_{i=0}^{\infty} x_i^d z^i$. Let $D(z) = \sum_{k=0}^{\infty} D_k z^k$ and

$$A(z) = E[e^{\rho D(z) V}] = \int_0^{\infty} e^{\rho D(z) y} dP(V \le y) . \tag{15}$$

Then

$$X^d(z)[zI - A(z)] = (-x_0^d D_0^{-1}) D(z) A(z) , \tag{16}$$

where $(-x_0^d D_0^{-1})$ is a positive vector. The generating function of Q^d itself is $Q^d(z) = X^d(z) e$, where again e is a vector of 1's.

As we have indicated, (16) is a consequence of (5). We will be interested in z such that $z > 1$ and $pf(A(z)) = z$. The following is consistent with Neuts [27].

Theorem 6. *For the BMAP/G/1 queue, the equation $pf(A(z)) = z$ has at most one root with $z > 1$. Such a root exists if and only if there are solutions σ and η with $0 < \sigma < 1$ and $0 < \eta < \infty$ to the equations*

$$pf(D(1/\sigma)) = \frac{\eta}{\rho} \quad \text{and} \quad Ee^{\eta V} = 1/\sigma . \tag{17}$$

Proof. By Seneta [31, Section 2.3], the Perron-Frobenius theory applies to $D(z)$ for all $z > 0$, provided that $D(z)$ is finite. By (15), the matrices $D(z)$ and $A(z)$ have a common associated positive real right eigenvector $r(z)$:

$$\begin{aligned} A(z) r(z) &= \int_0^{\infty} e^{\rho D(z) y} r(z) \, dP(V \le y) \\ &= r(z) \int_0^{\infty} e^{\rho y pf(D(z))} \, dP(V \le y) , \end{aligned}$$

so that

$$pf(A(z)) = Ee^{\rho pf(D(z))V} . \tag{18}$$

From (18), we to see that (17) is equivalent to $pf(A(z)) = z$. ∎

The representation of the single equation $pf(A(z)) = z$ in terms of the two equations in (17) allows us to identify and separate the effects of the arrival process and the service-time distribution, see [13,14,19,27,33].

Next we apply Theorem 3 to deduce some properties of the Perron-Frobenius eigenvalue of $D(z)$.

Theorem 7. *The Perron-Frobenius eigenvalue $pf\ D(e^s)$ is a strictly increasing convex function of s with $pf(D(1)) = 0$.*

Proof. Since $D(z)$ is irreducible, $e^{D(z)}$ is a nonnegative matrix with $pf(e^{D(z)}) = e^{pf(D(z))}$; see (18) above and Theorem 2.7 of Seneta [31]. By Theorem 3, $pf(D(e^s)) = \log(e^{pf(D(e^s))})$ is increasing and convex function of s. Since $e^{D(1)}$ is stochastic, $pf(e^{D(1)}) = 1$, which implies that $pf(D(1)) = 0$. ∎

Remark 2. From Theorem 6 it follows that the asymptotic decay rates η and σ depend on the *BMAP* arrival process only via the function $p_D(z) \equiv pf(D(z))$. It is significant that $p_D(z)$ coincides with the limit of the time-average of the factorial cumulant generating function, i.e.,

$$p_D(z) = \lim_{t\to\infty} t^{-1} \log E z^{N(t)} ;$$

see Theorem 1 of Choudhury and Whitt [12]. Approximations for the asymptotic decay rates η and σ are developed in [12]. Logarithmic limits for more general models are established in [8,19]. ■

Theorem 8. *In the BMAP/G/1 queue, suppose that* $\rho < 1$, $D \equiv D(1)$ *is irreducible, the equations in (17) have solutions with* $0 < \sigma < 1$ *and* $0 < \eta < \infty$, $Ee^{(\varepsilon + \eta)V} < \infty$ *and* $pf(D(\sigma^{-1} + \varepsilon)) < \infty$ *for some positive* ε, *and* $\sigma^{-i} x_{ij}^d \to y_j$ *as* $i \to \infty$ *where* $0 \le y_j \le \infty$ *for each* j. *Then the generating function* $X^d(z)$ *in (16) is finite for* $1 < z < \sigma^{-1}$ *and*

$$\sigma^{-i} x_i^d \to \xi^d = \frac{\eta(-x_0^d D_0^{-1}) r(\sigma^{-1}) l(\sigma^{-1})}{\rho(p'(\sigma^{-1}) - 1)} \quad \text{as } i \to \infty , \qquad (19)$$

where $p(z) \equiv pf(A(z))$ *and all components of the limit* ξ^d *in (19) are positive and finite. In (19),* $p'(\sigma^{-1}) = \rho\phi'(\eta) p_D'(\sigma^{-1}) > 1$, *where* $\phi(\eta) = Ee^{\eta V}$ *and* $p_D(z) = pf(D(z))$.

Proof. The conditions here imply the conditions of Theorem 5, since (16) is a special case of (5). (Note, however, that the summation in $X(z)$ and $X^d(z)$ start at 1 and 0, respectively.) First, the Markov chain is irreducible and positive recurrent because D is irreducible and $\rho < 1$. Here $(-x_0^d D_0^{-1}) D(z) \bar{A}(z)$ plays the role of $x_0 B(z) - x_1 A_0$ in Theorem 5, and

$$(-x_0^d D_0^{-1}) D(z) \bar{A}(z) r(z) = (-x_0^d D_0^{-1}) pf(D(z)) pf(\bar{A}(z)) r(z) , \qquad (20)$$

which is strictly positive for all $z > 1$, because $pf(D(1)) = 0$, $pf(D(z))$ is strictly increasing in z, and all other components are strictly positive.

The right side of (20) simplifies at σ^{-1}; by (17) and (18), $pf(D(\sigma^{-1})) = \eta/\rho$ and $pf(\bar{A}(\sigma^{-1})) = 1$. From (18), we see that $p'(\sigma^{-1})$ is as given. ■

Remark 3. Let g be the steady-state probability vector associated with the stochastic matrix G in (23) of Lucantoni [23]. By (54) of [23], $(-x_0^d D_0^{-1})$ in (19) can be expressed as $-(x_0^d D_0^{-1}) = (1-\rho)g$. As noted in Section 1, (19) implies a limit $\sigma\xi^d/(1-\sigma)$ for the tail probabilities $\sigma^{-i} \sum_{k=i+1}^{\infty} x_i^d$, which agrees with the formula for the *MAP/G/*1 queue following (26) in Baiocchi [6]. ■

Expressions for the transforms of the other steady-state distributions in the *BMAP/G/*1 queue are also available. Let x_{ij}^t be the steady-state probability that the queue length is i and the auxiliary state is j at an arbitrary time. Let $x_i^t = (x_{i1}^t, \ldots, x_{im}^t)$ and $X^t(z) = \sum_{i=0}^{\infty} x_i^t z^i$. Then, by (35) of Lucantoni [23],

$$X^t(z)D(z) = (z-1)X^d(z) \ . \tag{21}$$

Relation (21) is established in Theorem 3.3.18 of Ramaswami [30] for the case $|z| < 1$, but it is valid more generally provided that everything is finite. It is convenient to use the following alternative expression for $X^t(z)$ from (3.3.20) of [30]:

$$X^t(z) - x_0^t = \rho x_0^d(D(z) - I) + \rho X^d(z) E[e^{\rho D(z) V_e}] \qquad (22)$$

where V_e is a random variable with the service-time stationary-excess distribution, i.e.,

$$P(V_e \le x) = (EV)^{-1} \int_0^x P(V > u)\, du \, . \qquad (23)$$

Lemma 3. *If $pf(A(z)) = z$ has a finite root $\sigma^{-1} > 1$, then $D(\sigma^{-1})$ and $Ee^{\rho D(\sigma^{-1}) V_e}$ are finite.*

Proof. By Theorem 6, $pf(D(\sigma^{-1})) = \eta/\rho < \infty$, so that $D(\sigma^{-1})$ must be finite. Also, by Theorem 6, $Ee^{\eta V} = \sigma^{-1} < \infty$. Hence, using integration by parts, $Ee^{\eta V_e} = E(e^{\eta V} - 1)/\eta EV < \infty$. Finally, as in Theorem 6,

$$pf(Ee^{\rho D(\sigma^{-1}) V_e}) = Ee^{\rho pf(D(\sigma^{-1})) V_e} = Ee^{\eta V_e} < \infty \, . \; \blacksquare$$

Similarly, let x_{ij}^a and $X^a(z)$ be the corresponding probabilities and generating function seen by the first customer in a batch upon arrival. Let $D(j,k)$ be the $(j,k)^{\text{th}}$ element of D. Then, by the covariance formula in (8) of Melamed and Whitt [25] for instance,

$$X_j^a(z) = X_j^t(z) \sum_{k=1}^{m} (D - D_0)(j,k) \, , \qquad (24)$$

so that the generating function of

$$Q^a(z) = X^a(z) e = X^t(z)(D - D_0) e \, .$$

Remark 4. In the $MAP/G/1$ queue, where customers arrive and depart one at a time, Q^a has the same distribution as Q^d, but we need not have $X^a(z) = X^d(z)$. What we do have is $X^a(z)e = X^d(z)e$. To see that this is consistent with (21) and (24), note that in the *MAP/G/*1 case (21) becomes $X^t(z)(D_0 + D_1 z) = (z-1)X^d(z)$ and (24) yields $X^a(z)e = X^t(z)D_1 e$. Then note that $D_1 e = -D_0 e$, so that $X^t(z)D_1 e = X^d(z)e$. By (34) of Lucantoni [23], in the *BMAP/G/*1 queue

$$x_0^t = -x_0^d D_0^{-1} \text{ or } x_0^d = -x_0^t D_0 ,$$

whereas by the reasoning in (24) $x_0^a e = x_0^t(D - D_0)e$. ■

From (21), (22), (24) and Theorem 8, we have the following asymptotic results for x_i^t and x_i^a.

Theorem 9. *If, in addition to the conditions of Theorem 8,* $\sigma^{-i}x_{ij}^t \to y_j^t$ *and* $\sigma^{-i}x_{ij}^a \to y_j^a$ *as* $i \to \infty$ *where* $0 \le y_j^t, y_j^a \le \infty$ *for each j, then*

$$\sigma^{-i}x_i^t \to \xi^t \text{ and } \sigma^{-i}x_i^a \to \xi^a \text{ as } i \to \infty , \tag{25}$$

where

$$\xi^t = \frac{\rho(1-\sigma)}{\sigma\eta}\xi^d \tag{26}$$

and

$$\xi^a = \xi^t \sum_{k=1}^{m} (D-D_0)(j,k) \tag{27}$$

for ξ^d *in (19), so that* $\xi^a e = \xi^t(D-D_0)e$.

Proof. Again we apply the Tauberian theorem for generating functions in [17, p. 447]. Now we use (22) and (24). Note that

$$\xi^d \rho E(e^{\rho D(\sigma^{-1}) V_e}) = \xi^d \rho(1-\sigma)/\eta\sigma ,$$

because

$$l(\sigma^{-1}) E(e^{\rho D(\sigma^{-1}) V_e}) = l(\sigma^{-1}) E(e^{p_D(\sigma^{-1}) V_e})$$

$$= l(\sigma^{-1}) \frac{(e^{\eta V} - 1)}{\eta} = \frac{(1-\sigma) l(\sigma^{-1})}{\eta\sigma}$$

using Lemma 1 of [2]. Alternatively, from (19) and (21), since $l(\sigma^{-1}) D(\sigma^{-1}) = l(\sigma^{-1}) p_D(\sigma^{-1})$,

$$\xi^d = \frac{\sigma}{1-\sigma} \xi^t D(\sigma^{-1}) = \frac{\sigma}{1-\sigma} \xi^t p_D(\sigma^{-1}) = \frac{\sigma\eta\xi^t}{\rho(1-\sigma)} . \quad \blacksquare$$

Remark 5. In the *MAP/G/*1 queue, (25)–(27) yield $\xi^d e = \xi^a e = \xi^t D_1 e = \xi^t e \sigma\eta/\rho(1-\sigma)$. As a quick check on (25)–(27), note that in the *M/G/*1 queue $D - D_0 = D_1 = -D_0 = 1$, and $p_D(\sigma^{-1}) = \sigma^{-1} - 1 = \eta/\rho$, so that $\eta\sigma/\rho(1-\rho) = 1$. $\blacksquare$

We now consider the workload L and the waiting time of the first customer in a batch W^b. Let $F_j(x)$ be the joint probability that the workload is less than or equal to x and the auxiliary state variable is j at an arbitrary time in steady state. Let $F(x) \equiv (F_1(x), \ldots, F_m(x))$ and $\hat{F}(s) = \int_0^\infty e^{-sx} dF(x)$. Then, by (44) of Lucantoni [32],

$$\hat{F}(s)\left[I+\rho\frac{D(\hat{V}(s))}{s}\right] = x_0^t . \tag{28}$$

The Laplace transform of L is then $\hat{F}(s)\,e$.

Let $F^a(x) \equiv (F_1^a(x),\ldots,F_m^a(x))$ be the joint probability that the waiting time of the first customer in a batch is less than or equal to x and the auxiliary state variable upon arrival is j. Let $\hat{F}^a(s) = \int_0^\infty e^{-sx}dF^a(x)$. Then

$$\hat{F}_j^a(s) = \hat{F}(s)\sum_{k=1}^{m}(D-D_0)(j,k) \tag{29}$$

for $\hat{F}(s)$ in (28), by the same argument as for (24). The Laplace transform of W^b, where W^b is the waiting time of the first customer in a batch, is then $\hat{F}^a(s)e = \hat{F}(s)(D-D_0)e$.

To treat the workload and waiting time of the first customer in a batch in (28) and (29), let $\tilde{l}(s)$ and $\tilde{r}(s)$ be left and right eigenvectors of $\rho D(\hat{V}(s))/s$ associated with its Perron-Frobenius eigenvalue, which we denote by $f(s)$, normalized so that $\tilde{l}(s)e = \tilde{l}(s)\tilde{r}(s) = 1$. Note that $\tilde{l}(-\eta) = l(\sigma^{-1})$ and $\tilde{r}(-\eta) = r(\sigma^{-1})$. From (17), it is evident that the critical singularity is at $s = -\eta$. From (17), $f(-\eta) = -1$. By Theorem 7, $0 > f(s) \geq -1$ for $0 > s \geq -\eta$. We need the following analog of Theorem 4.

Lemma 4. *The matrix* $Y(\sigma^{-1}) \equiv I - \rho\eta^{-1}D(\sigma^{-1}) + r(\sigma^{-1})l(\sigma^{-1})$ *is nonsingular and* $l(\sigma^{-1})Y(\sigma^{-1}) = l(\sigma^{-1})$.

Proof. As noted above, (17) implies that $pf(\rho\eta^{-1}D(\sigma^{-1})) = 1$. Consider a vector u such that $u(I - \rho\eta^{-1}D(\sigma^{-1}) + r(\sigma^{-1})l(\sigma^{-1})) = 0$. If we multiply on the right by $r(\sigma^{-1})$, we see that $u\, r(\sigma^{-1}) = 0$, but this implies that u is a left eigenvector of $D(\sigma^{-1})$. Since $ur(\sigma^{-1}) = 0$, we must have $u = 0$. ∎

Let S and S^a denote the auxiliary state at an arbitrary time and at an arrival epoch, respectively, in steady-state. As in (4) of [23], let π be the probability vector of S, i.e., satisfying $\pi D = 0$ and $\pi e = 1$. Let π^a be the probability vector of S^a, which by the argument for (24) satisfies $\pi_j^a = \pi_j \sum_{k=1}^{m} (D - D_0)(j,k)$. Then $F(\infty) = \pi$ and the Laplace transform for the tail probabilities $\pi - F(x)$ is $(\pi - \hat{F}(s))/s$. As before, let $\phi(\eta) = Ee^{\eta V}$ and $p_D(z) = pf(D(z))$.

We now apply the Tauberian theorem for Laplace transforms in Feller [17, p. 445]. As before, we assume that the quantities of interest, such as $e^{\eta x}P(L > x, S = j)$, converge to something as $x \to \infty$ in order to get ordinary convergence from the Cesàro convergence implied by the Tauberian theorem. For a large class of *MAP/G/*1 queues, this limit holds by virtue of Theorem 5.1 of Asmussen and Perry [8]. Our analysis supplements [5] by giving alternative expressions for the limits.

Theorem 10. *If, in addition to the conditions of Theorem 8,* $e^{\eta x}P(L > x, S = j) \to y_j$ *and* $e^{-\eta x}P(W^b > x, X^a = j) \to y_j^a$ *as* $x \to \infty$ *where* $0 \le y_j \le \infty$ *and* $0 \le y_j^a \le \infty$ *for each* j, *then the critical singularity of* $[\pi - \hat{F}(s)]/s$ *is at* $s = -\eta$ *and*

$$e^{\eta x}P(L > x, S = j) = e^{\eta x}(\pi_j - F_j(x)) \to \xi_j^L \text{ as } x \to \infty ,$$

where

$$\xi_j^L = \frac{x_0^\ell r(\sigma^{-1})l(\sigma^{-1})_j}{\rho p_D'(\sigma^{-1})\phi'(\eta) - 1} = \frac{\sigma\xi_j^\ell}{(1-\sigma)} , \tag{30}$$

$$e^{\eta x}P(W^b > x, S^a = j) = e^{\eta x}(\pi_j^a - F_j^a(x)) \to \xi^{W^b} \text{ as } x \to \infty ,$$

where

$$\xi^{W^b} = \xi_j^L \sum_{k=1}^{m} (D - D_0)(j,k) = \frac{\sigma\xi_j^a}{(1-\sigma)} , \tag{31}$$

$$e^{\eta x}P(L > x) \to \alpha_L \equiv \xi^L e \text{ as } x \to \infty \tag{32}$$

and

$$e^{\eta x}P(W^b > x) \to \alpha_{W^b} \equiv \xi^{W^b} e \text{ as } x \to \infty , \tag{33}$$

where ξ^l *and* ξ^{W^b} *are finite with all positive components.*

Proof. We apply the Tauberian theorem in [17, p. 445]. By (27),

$$\frac{(\pi - F(s))}{s}(I + \rho D(\hat{V}(s))/s)$$

$$= M(s) \equiv \frac{\pi(I + \rho D(\hat{V}(s))/s)}{s} - \frac{x_0^t}{s} . \tag{34}$$

Postmultiplying by $\tilde{r}(s)$ in (34), we see that, for $s < 0$,

$$\frac{(\pi - F(s))}{s} \tilde{r}(s)(1-f(s)) = \frac{\pi(1 - F(s))\tilde{r}(s) - x_0^t \tilde{r}(s)}{s} . \tag{35}$$

Therefore,

$$\begin{aligned}
&\lim_{s \to -\eta} (s + \eta) \frac{(\pi - F(s))}{s} \tilde{r}(s) \\
&= \lim_{s \to -\eta} \left[(s + \eta) \frac{\pi \tilde{r}(s)}{s} - \frac{(s + \eta) x_0^t \tilde{r}(s)}{(1 - f(s)) s} \right] \\
&= \frac{x_0^t \tilde{r}(-\eta)}{\eta f'(-\eta)} ,
\end{aligned} \tag{36}$$

where, with $p_D(z) \equiv pf(D(z))$,

$$f'(s) = \frac{\rho}{s} p_D'(\hat{V}(s)) \hat{V}'(s) - \frac{\rho p_D(\hat{V}(s))}{s^2} ,$$

so that

$$f'(\eta) = \frac{\rho p_D'(\sigma^{-1}) \phi'(\eta) - 1}{\eta} . \tag{37}$$

As in Theorem 8,

$$\begin{aligned}
p'(\sigma^{-1}) &\equiv \frac{d}{dz} pf(A(z))\Big|_{z = \sigma^{-1}} \\
&= \rho p_D'(\sigma^{-1}) E[Ve^{\eta V}] = \rho p'(\sigma^{-1}) \phi'(\eta)
\end{aligned} \tag{38}$$

and the denominator of (30) is strictly positive. Next, by (34) and (35),

$$(s+\eta)\frac{(\pi-F(s))}{s}\left[I+\rho\frac{D(\hat{V}(s))}{s}-f(s)\tilde{r}(s)\tilde{l}(s)\right]$$

$$=(s+\eta)\left[M(s)-\frac{(\pi(1-f(s))-x_0^t)\tilde{r}(s)\tilde{l}(s)f(s)}{s(1-f(s))}\right] \tag{39}$$

Taking the limit as $s \to -\eta$ in (39), we obtain

$$\lim_{s\to-\eta}(s+\eta)\left[\frac{(\pi-F(s))}{s}\right]\left[I-\rho\frac{D(\sigma^{-1})}{\eta}+\tilde{r}(-\eta)\tilde{l}(-\eta))\right]$$

$$=\frac{x_0^t\tilde{r}(-\eta)\tilde{l}(-\eta)}{\eta f'(-\eta)}\,, \tag{40}$$

where the matrix on the left is nonsingular by Lemma 4 and the right side is strictly positive and finite as in (36). We then note that $\tilde{l}(-\eta)=l(\sigma^{-1})$ and $\tilde{r}(-\eta)=r(\sigma^{-1})$. We take the inverse in (40) to obtain (30). By Lemma 4, $l(\sigma^{-1})Y(\sigma^{-1})=l(\sigma^{-1})$. Hence we can delete the $Y(\sigma^{-1})$. Our assumption that $e^{\eta x}P(L>x,\ S=j)$ converges enables to obtain ordinary convergence from the Cesàro convergence provided by the Tauberian theorem in [17]. We apply (29) and (30), and the Tauberian theorem again, to obtain (31)–(33). ■

We conclude by stating some relations among the asymptotic constants that follow from Theorems 8–10.

Theorem 11. *(a) In the BMAP/G/*1 *queue, if the limits exist, then*

$$\beta = \alpha_L \ , \ \beta^a = \alpha_W \tag{41}$$

and

$$\frac{\beta^a}{\beta} = \frac{\xi^t(D-D_0)e}{\xi^t e} = \frac{l^L(D-D_0)e}{\xi^L e}$$

$$= \frac{\xi^{W^b} e}{\xi^L e} = l(\sigma^{-1})(D-D_0)e \ . \tag{42}$$

*(b) In the MAP/G/*1 *queue,*

$$\frac{\beta^a}{\beta} = \frac{\beta^d}{\beta} = \frac{\xi^t D_1 e}{\xi^t e}$$

$$= \frac{\xi^L D_1 e}{\xi^L e} = \frac{\xi^W e}{\xi^L e} = \frac{\alpha_W}{\alpha_L} = l(\sigma^{-1}) D_1 e = \frac{\eta\sigma}{\rho(1-\sigma)} \ . \tag{43}$$

Proof. (a). Apply Theorems 8–10, recalling that $\beta = \sigma\xi^t e/(1-\sigma)$ and $\beta^a = \sigma\xi^a e/(1-\sigma)$. (b) In the *MAP/G/*1 queue, customers arrive one at a time, so that $\beta^a = \beta^d$ and $\xi^W e = \alpha_W$. Then $D - D_0 = D_1$. Finally, by (26), $\beta^d/\beta = \eta\sigma/\rho(1-\sigma)$. Theorem 2 of [2] also shows that $\alpha_W/\alpha_L = \eta\sigma/\rho(1-\sigma)$ in any *G/GI/*1 queue (with i.i.d. service times that are independent of the arrival process). ■

Corollary. *(a) In the MAP/M/1 queue,* $\sigma = 1 - \eta$ *so that*

$$\frac{\beta^a}{\beta} = \frac{\alpha_W}{\alpha_L} = \frac{\sigma}{\rho} = \frac{1-\eta}{\rho} \ . \tag{44}$$

(b) In the MAP/D/1 queue, $\sigma = e^{-\eta}$*, so that*

$$\frac{\beta^a}{\beta} = \frac{\alpha_W}{\alpha_L} = \frac{\eta e^{-\eta}}{\rho(1-e^{-\eta})} = \frac{-\sigma \log \sigma}{\rho(1-\sigma)} . \tag{45}$$

Remark 6. In applications we are often interested in the steady-state sojourn time or response time T, i.e., the waiting time W plus the service time V. Theorem 1 of [2] shows that in any $G/GI/1$ queue if $e^{\eta x} P(W > x) \to \alpha_W$ as $x \to \infty$, then $e^{\eta x} P(T > x) \to \alpha_W / \sigma$. Moreover, the exponential approximation for the sojourn-time distribution is also remarkably good, even when the service-time distribution is not nearly exponential. ■

4. The *MAP/MSP/1* Queue

In this section we consider a related class of Markov models with dependence among the service times. We let the number of service completions during the first t units of time that the server is busy be a *MAP* that is independent of the arrival process. We call this service process a *Markovian service process* (*MSP*).

For the *MAP/MSP/1* queue, the queue length at an arbitrary time, together with the phases of the arrival and service processes is a *quasi-birth-and-death* (*QBD*) process. A *QBD* process has a generator of the form

$$\tilde{Q} = \begin{bmatrix} \tilde{B}_0 & \tilde{A}_0 & & & \\ \tilde{B}_1 & \tilde{A}_1 & \tilde{A}_0 & & \\ & \tilde{A}_2 & \tilde{A}_1 & \tilde{A}_0 & \\ & & \tilde{A}_2 & \tilde{A}_1 & \tilde{A}_0 \dots \\ & & & \vdots & \end{bmatrix} \tag{46}$$

Let $D_0^{\uparrow}$ and $D_1^{\uparrow}$ be the coefficient matrices of the *MAP* and $D_0^{\downarrow}$ and $D_1^{\downarrow}$ the coefficient matrices of the *MSP*, as in (14), normalized so that they each have overall rate 1, i.e., $D_1^{\uparrow} e = D_1^{\uparrow} e = 1$. Let the overall service rate and arrival rate in the queue be 1 and ρ, respectively. Then the matrices in (46) can be expressed using the Kronecker product ⊗ and sum ⊕ operations as

$$\tilde{A}_0 = I_1 \otimes \rho D_1^{\uparrow}\ ,\ \tilde{A}_1 = D_0^{\downarrow} \oplus \rho D_0^{\uparrow} \text{ and } \tilde{A}_2 = D_1^{\downarrow} \otimes I_2\ , \tag{47}$$

where I_1 and I_2 are identity matrices. By our assumption that the *MSP* operates only when the server is busy,

$$\tilde{B}_0 = I_1 \oplus \rho D_0^{\uparrow} \quad \text{and} \quad \tilde{B}_1 = \tilde{A}_2\ . \tag{48}$$

We can obtain alternative models by changing the definition of $\tilde{B}_0$. For example, we could let the phase process run without generating any real service completions. Then we would have $\tilde{B}_0 = D^{\downarrow} \oplus \rho D_0^{\uparrow}$.

Since the *MAP/MSP/*1 queue produces a Markov chain of *GI/M/*1 type, it has the asymptotic behavior in (4), where in this case the rate matrix R is the minimal nonnegative solution to the equation

$$R^2 \tilde{A}_2 + R\tilde{A}_1 + \tilde{A}_0 = 0 . \tag{49}$$

The asymptotic decay rate σ is then the Perron-Frobenius eigenvalue of R, $pf(R)$, which satisfies the equation

$$pf(\tilde{A}(\sigma)) = 0 , \tag{50}$$

where $\tilde{A}(z) = z^2 \tilde{A}_2 + z\tilde{A}_1 + \tilde{A}_0$. As with $D(z)$ in §3, the Perron-Frobenius theory applies to $\tilde{A}(z)$ since $\tilde{A}(z)$ has nonnegative off-diagonal elements. By the same argument as for Theorem 7, we have the following result.

Theorem 12. *In a QBD process if $\tilde{A}(1)$ is irreducible, then the Perron-Frobenius eigenvalue $pf(\tilde{A}(e^s))$ is a strictly increasing convex function of s with $pf(\tilde{A}(1)) = 0$, so that the equation $pf(\tilde{A}(z)) = 0$ has at most one root σ with $0 < \sigma < 1$.*

We apply Theorem 12 to establish asymptotics for the *MAP/MSP/*1 queue.

Theorem 13. *In the MAP/MSP/*1 *queue, if $\tilde{Q}$ is irreducible and positive recurrent, and $D^{\uparrow}(1)$ and $D^{\downarrow}(1)$ are irreducible, then*

$$x_i = x_0 R^i = \sigma^i (x_0) \hat{r}(\sigma)) \hat{l}(\sigma) + o(\sigma^i) \ \text{as} \ i \to \infty , \tag{51}$$

where the asymptotic decay rate σ is the unique root with $0 < \sigma < 1$ of the equation $pf(\tilde{A}(z)) = 0$ or, equivalently, the equation

$$pf(D^{\downarrow}(z)) = -pf(\rho D^{\uparrow}(1/z)) \ , \tag{52}$$

and $\hat{l}(\sigma)$ *and* $\hat{r}(\sigma)$ *are left and right eigenvectors of* $\tilde{A}(\sigma)$ *associated with* $pf(\tilde{A}(\sigma))$ *such that* $\hat{l}(\sigma)r(\sigma) = \hat{l}(\sigma)e = 1$.

Proof. $\tilde{A}(1)$ is irreducible if $D^{\uparrow}(1)$ and $D^{\downarrow}(1)$ are. By Theorem 12, the equation $pf(\tilde{A}(z)) = 0$ has at most one root with $0 < z < 1$. By Theorem 3.1.1 of [26], this root σ exists and satisfies $0 < \sigma < 1$, and (51) is valid. By basic properties of the Kronecker operations,

$$pf(\tilde{A}(z)) = pf\left[z(I_1 \otimes \rho D_1^{\uparrow}) + (I_1 \otimes \rho D_0^{\uparrow}) + (D_0^{\downarrow} \otimes I_2) + \frac{(D_1^{\downarrow} \otimes I_2)}{z}\right]$$

$$= pf((I_1 \otimes (\rho z D_1^{\uparrow} + \rho D_0^{\uparrow})) + ((D_0^{\downarrow} + \frac{D_1^{\downarrow}}{z}) \otimes I_2))$$

$$= pf(z\rho D_1^{\uparrow} + \rho D_0^{\uparrow}) + pf(D_0^{\downarrow} + \frac{D_1^{\downarrow}}{z}))$$

$$= pf(D^{\uparrow}(z)) + pf(D^{\downarrow}(1/z)) \ . \tag{53}$$

By Theorem 7, $pf(D^{\uparrow}(z))$ and $pf(D^{\downarrow}(z))$ are increasing convex functions of z. Hence $pf(D^{\downarrow}(1/z))$ is a convex function of z as well. ■

We can obtain related results for the steady-state distributions just before arrivals and just after departures by applying the covariance formula (8) in Melamed and Whitt [25]. To this end, let x_{ij}^{t} denote the steady-state probability that the queue length is i and the (joint arrival and service) phase is j at an arbitrary time, and similarly for x_{ij}^a and x_{ij}^d. Then, paralleling (24), we obtain

$$x_{ij}^a = x_{ij}^t \sum_k \tilde{A}_0(j,k) \text{ and } x_{ij}^d = x_{ij}^t \sum_k \tilde{A}_2(j,k) \, , \tag{54}$$

from [25], using time reversal in §4 there for x_{ij}^d. Hence,

$$x_i^a e = x_i^t \tilde{A}_0 e \text{ and } x_i^d e = x_i^t \tilde{A}_2 e \, , \tag{55}$$

from which the asymptotics for x_i^a and x_i^d follow. Since the long-run flow rate from level i up (down) is $x_i^t \tilde{A}_0 e$ ($x_i^t \tilde{A}_2 e$), we have $x_i^a e = x_i^d e$, as we should.

Remark 7. The asymptotic behavior in (4) also holds for the multi-server *MAP/MSP/m* and *GI/MSP/m* models, even with heterogeneous servers, because these models can be analyzed via Markov chains of *GI/M/*1 type; see [26,27,29,32]. The complicated behavior occurring when not all m servers are busy is captured by the boundary states; see of Neuts [26, p. 207]. In the non-boundary states, the multiple servers are represented by Kronecker product and sum operations. Multichannel (superposition) arrival processes can be represented, just as in Neuts [27, pp. 243-248].

5. A Numerical Example

To illustrate the *BMAP/G/*1 results in Section 3, we present a numerical example. We consider a simple *MAP*, a two-state Markov modulated

Poisson process ($MMPP_2$). An $MMPP$ is a $BMAP$ with $D_0 = M - \Lambda$, $D_1 = \Lambda$ and $D_j = 0$ for $j \geq 2$, where M is the infinitesimal-generator matrix of the Markovian environment process and Λ is the associated Poisson rate matrix. With two phases,

$$M = \begin{bmatrix} -m_0 & m_0 \\ m_1 & -m_1 \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 \end{bmatrix}, \tag{56}$$

where m_0, m_1, λ_0 and λ_1 are positive constants and $m_0 \lambda_1 + m_1 \lambda_0 = m_0 + m_1$, so that the arrival rate is 1. The overall arrival rate is then ρ where ρ is specified separately. Then

$$D(z) = \begin{bmatrix} -m_0 + \lambda_0(z-1) & m_0 \\ m_1 & -m_1 + \lambda_1(z-1) \end{bmatrix}. \tag{57}$$

The Perron-Frobenius eigenvalue $pf(D(z)) \equiv p_D(z)$ can be found by solving the characteristic equation. Let $x = z - 1$, $m = m_0 + m_1$, $s = \lambda_0 + \lambda_1$ and $d = \lambda_0 - \lambda_1$. Then the characteristic equation is

$$\det(D(z) - \gamma I) = \gamma^2 - \gamma(sx - m) + \lambda_0 \lambda_1 x^2 - mx = 0. \tag{58}$$

Therefore,

$$p_D(z) = \gamma(x) = \frac{sx - m + \sqrt{d^2 x^2 + 2mx(2-s) + m^2}}{2}. \tag{59}$$

Here the key equations in (17) are

$$\gamma(\sigma^{-1} - 1) = \frac{\eta}{\rho} \text{ and } Ee^{\eta V} = \sigma^{-1} \tag{60}$$

for $\gamma(x)$ in (59).

For our example, consider the two-point service-time distribution:

$$P(V = 11) = 0.01960784 = 1 - P(V = 0.8) .$$

Clearly, $EV = 1$ and $EV^2 = 3$. Let $\rho = 0.7$, $\rho\lambda_0 = 1.1$, $\rho\lambda_1 = 0.3$ and $\rho m_0 = \rho m_1 = 0.1$. The overall arrival rate is 0.7, but the arrival rate in phase 0 is 1.1, which exceeds the overall service rate of 1. Then $m = 0.285714$, $s = 2.0$ and $d = 1.142857$, so that (59) becomes

$$\gamma(x) = x - 0.142857 + 0.5\sqrt{1.306122x^2 + 0.0816326} . \quad (61)$$

The solution to (60) is $\eta = 0.1115972$ and $\sigma = 0.878066$. Thus $p_D(\sigma^{-1}) = 0.159424$, $l(\sigma^{-1}) = (0.629544, 0.370456)$,

$$D(\sigma^{-1}) = \begin{bmatrix} 0.075362 & 0.142857 \\ 0.142857 & -0.083342 \end{bmatrix} \text{ and } r(\sigma^{-1}) = \begin{bmatrix} 1.179891 \\ 0.694306 \end{bmatrix} \quad (62)$$

with $l(\sigma^{-1})$ and $r(\sigma^{-1})$ being chosen so that $l(\sigma^{-1})r(\sigma^{-1}) = l(\sigma^{-1})e = 1$.

To calculate ξ^d in (19), we also need $\phi'(\eta)$, $p'_D(\sigma^{-1})$ and g. (Recall that $(-x_0 D_0^{-1}) = (1-\rho)g$). First,

$$\phi'(\eta) = EVe^{\eta V} = 1.593677 . \quad (63)$$

From (59),

$$p'_D(z) = \gamma'(x)$$

$$= \frac{s}{2} + \frac{(d^2x^2 + 2mx(2-s)+m^2)^{-1/2}}{4}(2d^2x + 2m(2-s)) ,$$

so that here

$$p_D'(\sigma^{-1}) = \gamma'(0.138866) = 1.277475 . \tag{65}$$

From [11], we obtain $g = (0.241185, 0.758815)$, so that $g\, r(\sigma^{-1}) = 0.811421$. Thus, by (19),

$$\xi^d = \frac{\eta(1-\rho)g\, r(\sigma^{-1})l(\sigma^{-1})}{\rho(\rho p_D'(\sigma^{-1})\phi'(\eta) - 1)} = 0.091288 l(\sigma^{-1}) \tag{66}$$

and $\xi^a e = \xi^d e = 0.091287$. From (26), we obtain

$$\xi^t = \frac{\rho(1-\sigma)}{\sigma\eta}\xi^d = 0.0795156 l(\sigma^{-1}) \tag{67}$$

and $\xi^t e = 0.0795156$. We also see that (66) and (67) are consistent with (27), i.e., $\xi^a e = \xi^t D_1 e = \xi^d e$ as it should.

Next,

$$\xi^L = \frac{\sigma}{1-\sigma}\xi^t = 0.572608 l(\sigma^{-1}) \tag{68}$$

and

$$\xi^W e = \frac{\sigma \xi^a e}{1-\sigma} = 0.657379 , \tag{69}$$

so that $\alpha_L = \beta$, $\alpha_W = \beta^a$ and

Table 1. A comparison of approximations with exact values for the steady-state queue-length tail probabilities.

k	$P(Q^a > k)$ exact	$P(Q^a > k)$ approx.	$P(Q > k)$ exact	$P(Q > k)$ approx.
2	0.5039	0.5068	0.4366	0.4415
4	0.3660	0.3908	0.3223	0.3404
8	0.2267	0.2323	0.1989	0.2023
10	0.1790	0.1791	0.1563	0.1560
12	0.1395	0.1381	0.1214	0.1203
14	0.1076	0.1065	0.09348	0.09273
16	0.08249	0.08208	0.07174	0.07150
20	0.04869	0.04879	0.04242	0.04250
24	0.02898	0.02901	0.02525	0.02526
28	0.017250	0.017242	0.015024	0.015019
36	0.0060921	0.0060927	0.0053066	0.0053070
48	0.0012798	0.0012798	0.0011148	0.0011148

$$\frac{\alpha_L}{\alpha_W} = \frac{\beta}{\beta^a} = \frac{\xi^t e}{\xi^d e} = \frac{\rho(1-\sigma)}{\eta\sigma} = 0.871049 \qquad (70)$$

as in Theorem 11.

Finally, in Table 1 we compare the exponential approximations $\beta\sigma^k$ and $\beta^a\sigma^k$ to the exact values $P(Q > k)$ and $P(Q^a > k)$ for this model obtained from [11]. From Table 1 it is clear that the exponential approximations are excellent at the 80th percentile and beyond. As in [1], we find that the exponential approximations often perform remarkably well, but of course, in general, the point where the exponential approximations become good depends on the model; e.g. see [10].

Acknowledgments. We are grateful to Søren Asmussen, David Lucantoni, Marcel Neuts and the referees for helpful comments.

REFERENCES

[1] Abate, J., G. L. Choudhury and W. Whitt. Exponential approximations for tail probabilities in queues, I: waiting times. *Opns. Res.*, forthcoming.

[2] Abate, J., G. L. Choudhury and W. Whitt. Exponential approximations for tail probabilities in queues, II: sojourn time and workload. *Opns. Res.*, forthcoming.

[3] Abate, J., G. L. Choudhury and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. submitted, 1993.

[4] Abate, J. and W. Whitt. the Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10 (1992) 5-88.

[5] Asmussen, S. and D. Perry. On cycle maxima, first passage problems and extreme value theory for queues. *Stochastic Models* 8 (1992) 421-458.

[6] Baiocchi, A. Asymptotic behavior of the loss probability of the *MAP/G/1/K* queue, part I: theory. *Stochastic Models*, forthcoming.

[7] Borovkov, A. *Stochastic Processes in Queueing Theory*, New York: Springer-Verlag, 1976.

[8] Chang, C. S. Stability, queue length and delay of deterministic and stochastic queueing networks, manuscript, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1993.

[9] Choudhury, G. L. and D. M. Lucantoni. Numerical computation of the moments of a probability distribution from its transforms. *Opns. Res.*, forthcoming.

[10] Choudhury, G. L., D. M. Lucantoni and W. Whitt. Squeezing the most out of ATM. submitted, 1993.

[11] Choudhury, G. L., D. M. Lucantoni and W. Whitt. An algorithm for a large class of G/G/1 queues, in preparation, 1993.

[12] Choudhury, G. L. and W. Whitt. Heavy-traffic asymptotic expansions for the asymptotic decay rates in the *BMAP/G/*1 queue. submitted, 1992.

[13] Elwalid, A. I. and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE Trans. Networks,* forthcoming.

[14] Elwalid, A. I. and D. Mitra. Markovian arrival and service communication systems: spectral expansions, separability and Kronecker-product forms. manuscript, AT&T Bell Laboratories, Murray Hill, NJ, 1993.

[15] Falkenberg, E. Über Markov-ketten vom M/G-typ. Ph.D. dissertation, Technische Hochschule Darmstadt, 1989.

[16] Falkenberg E. On the asymptotic behavior of the stationary distribution of Markov chains of M/G/1-type. *Stochastic Models,* forthcoming.

[17] Feller, W. *An Introduction to Probability Theory and Its Applications,* Vol. II, second ed. New York: Wiley, 1971.

[18] Gail, H. R., S. L. Hantler and B. A. Taylor. Spectral analysis of *M/G/*1 type Markov chains, manuscript, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1992.

[19] Glynn, P. W. and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* 31 (1994), forthcoming.

[20] Hunter, J. J. Generalized inverses and their applications to applied probability problems. *Lin. Alg. Appl.* 45 (1982) 157-198.

[21] Kemeny, J. G. and J. L. Snell. *Finite Markov Chains.* Princeton, NJ: Van Nostrand, 1960.

[22] Kingman, J. F. C. A convexity property of positive matrices. *Quart. J. Math.* 12 (1961) 283-284.

[23] Lucantoni, D. M. New results on the single server queue with a batch Markovian arrival process. *Stochastic Models* 7 (1991) 1-46.

[24] Lucantoni, D. M. The *BMAP/G/*1 queue: a tutorial. in *Models and Techniques for Performance Evaluation of Computer and Communications Systems,* L. Donatiello and R. Nelson (eds.), Springer Verlag, New York, 1993, forthcoming.

[25] Melamed, B. and W. Whitt. On arrivals that see time averages. *Oper. Res.* 38 (1990) 156-179.

[26] Neuts M. F., *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Baltimore: The Johns Hopkins Press, 1981.

[27] Neuts, M. F. The caudal characteristic curve of queues. *Adv. Appl. Prob.* 18 (1986) 221-254.

[28] Neuts, M. F. *Structured Stochastic Matrices of M/G/*1 *Type and Their Applications,* New York: Marcel Dekker, 1989.

[29] Neuts, M. F and Y. Takahashi. Asymptotic behavior of the stationary distributions in the *GI/PH/c* queue with heterogeneous servers. *Z. Wahrscheinlichkeitsth.* 57 (1981) 441-452.

[30] Ramaswami, V., The *N/G/*1 queue and its detailed analysis. *Adv. Appl. Prob.* 12 (1980) 229-261.

[31] Seneta, E. *Nonnegative Matrices and Markov Chains.* second ed., New York: Springer-Verlag, 1981.

[32] Takahashi, Y. Asymptotic exponentiality of the tail of the waiting-time distribution in a *PH/PH/c* queue. *Adv. Appl. Prob.* 13 (1981) 619-630.

[33] Whitt, W. Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues. *Telecommunication Systems,* forthcoming.

[34] Wilf, H. S. *Generatingfunctionology.* New York: Academic Press, 1990.

Received: 9/12/1992
Revised: 4/1/1993
Accepted: 8/22/1993

Recommended by Marcel F. Neuts, Editor