HEAVY-TRAFFIC ASYMPTOTIC EXPANSIONS FOR THE ASYMPTOTIC DECAY RATES IN THE BMAP/G/1 QUEUE

Gagan L. CHOUDHURY

AT&T Bell Laboratories
Holmdel, NJ 07733-3030

Ward WHITT

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

Key Words: queues, $BMAP/G/1$ queue, batch Markovian arrival process, versatile Markovian point process, tail probabilities in queues, asymptotic decay rate, Perron-Frobenius eigenvalue, asymptotic expansion, caudal characteristic curve, heavy traffic

ABSTRACT

In great generality, the basic steady-state distributions in the $BMAP/G/1$ queue have asymptotically exponential tails. Here we develop asymptotic expansions for the asymptotic decay rates of these tail probabilities in powers of one minus the traffic intensity. The first term coincides with the decay rate of the exponential distribution arising in the standard heavy-traffic limit. The coefficients of these heavy-traffic expansions depend on the moments of the service-time distribution and the derivatives of the Perron-Frobenius eigenvalue $\delta(z)$ of the $BMAP$ matrix generating function $D(z)$ at $z = 1$. We give recursive formulas for the derivatives $\delta^{(k)}(1)$. The asymptotic expansions provide the basis for efficiently computing the asymptotic decay rates as functions of the traffic intensity, i.e., the caudal characteristic curves. The asymptotic expansions also reveal what features of the model the asymptotic decay rates primarily depend upon. In particular, $\delta(z)$ coincides with the limiting time-average of the factorial cumulant generating function (the logarithm of the generating function) of the arrival counting process, and the derivatives $\delta^{(k)}(1)$ coincide with the asymptotic factorial cumulants of the arrival counting process. This insight is important for admission control schemes in multi-service networks based in part on asymptotic decay rates. The interpretation helps identify appropriate statistics to compute from network traffic data in order to predict performance.

1. Introduction

In this paper we consider the $BMAP/G/1$ queue, which has a single server, unlimited waiting room, the first-in first-out service discipline and i.i.d. service times that are independent of a batch Markovian arrival process (*BMAP*). The *BMAP* is an alternative representation of the versatile Markovian point process of Neuts [28, 30] with an appealing simple notation, which was introduced by Lucantoni [25]. The $BMAP/G/1$ queue is equivalent to the N/G/1 queue considered by Ramaswami [31]. The *BMAP* generalizes the *MAP* by allowing batch arrivals; the *MAP* generalizes the Markov modulated Poisson processes (*MMPP*) by allowing an arrival and a change of environment state to occur simultaneously. *MMPP*s, *MAP*s and *BMAP*s are useful for studying superposition arrival processes, e.g., arising in models of statistical multiplexing, because superpositions of independent arrival processes of each type is again of the same type. Indeed, *MAP*s are sufficiently general that they can serve as approximations for any stationary point process (possibly at the expense of requiring large matrices); see Asmussen and Koole [6]. For an overview of the $BMAP/G/1$ queue, see Lucantoni [26].

In Abate, Choudhury and Whitt [2] we showed that in great generality the basic steady-state distributions in the $BMAP/G/1$ queue have asymptotically exponential tails. (For related work, see Asmussen [5], Asmussen and Perry [7], Baiocchi [8], Chang [9], Elwalid and Mitra [14,15], Glynn and Whitt [18] and van Ommeren [40].) Our purpose here is to obtain heavy-traffic asymptotic expansions (in powers of $1 - \rho$ where ρ is the traffic intensity) for the asymptotic decay rates. These asymptotic expansions provide a convenient way to compute the asymptotic decay rates. The asymptotic decay rates also can be computed by root finding, but the asymptotic expansions yield the asymptotic decay rates as functions of ρ (i.e., the caudal characteristic curves; see Neuts [29]), whereas the root finding must be repeated for each separate value of ρ. The asymptotic expansions also reveal what features of the model the asymptotic decay rates primarily depend upon. In particular, the asymptotic decay rates primarily depend on the *BMAP* through its lower asymptotic cumulants, the first three of which are the arrival rate, the asymptotic variance and the asymptotic central third moment. Although our proofs depend on the *BMAP* structure, this characterization does not; it applies to arbitrary stochastic point processes.

In Abate and Whitt [3] we showed that a heavy-traffic asymptotic expansion is possible for multi-channel queues in which the individual arrival and service channels are mutually independent renewal processes, and found the first two terms. Here we extend these results by treating *BMAP* arrival processes and finding more terms. Here we also provide an interpretation of the terms. For the *BMAP*, the key is to compute the derivatives of the Perron-Frobenius eigenvalue $\delta(z)$ of the *BMAP* matrix generating function $D(z)$ at $z = 1$, which we do here in §3. The analysis is similar to the analysis in the Appendix of Neuts [30]. We develop a recursive algorithm for computing any desired derivative of $\delta(z)$ at $z = 1$. The k^{th} derivative $\delta^{(k)}(1)$ is the k^{th} asymptotic factorial cumulant of the *BMAP*. We also develop an algorithm for computing the first seven coefficients in the heavy-traffic expansions for the asymptotic decay rates. (This can be extended if desired.)

One reason we are interested in the $BMAP/G/1$ queue is because it can serve as a model to aid in admission control in multi-service networks. (For that application, the service times can often be regarded as deterministic.) The idea is to construct simple admission criteria from the asymptotic decay rates. For this purpose, it is important to be able to quickly compute the asymptotic decay rates and understand what features of the model they primarily depend upon. Part of the efficiency stems from a separation of independent sources (see Theorem 3 in Section 2); the rest stems from the heavy-traffic expansion. For additional related work on admission control, see Chang [9], Elwalid and Mitra [14], Gibbens and Hunt [17], Guérin, Ahmadi and Naghshineh [19], Kelly [23], Sohraby [37,38], Whitt [43] and references therein. In this context Sohraby [37,38] also considers (one-term) heavy-traffic approximations for the decay rates.

It turns out that the first term of the heavy-traffic expansion for the asymptotic decay rates coincides with the rate of the steady-state exponential distribution of the diffusion process (reflected Brownian motion) in the familiar heavy-traffic limit theorem in which first $\rho \rightarrow 1$ and then $t \rightarrow \infty$. This might be anticipated, but it is not automatic because it involves an interchange of limits. (Here we first let $t \rightarrow \infty$ and then let $\rho \rightarrow 1$.) The contribution of the *BMAP* to this first term is via $1 + \delta^{(2)}(1)$, which corresponds to the asymptotic variance. Subsequent terms in the asymptotic expansion offer refinements to the basic heavy-traffic approximation. As we found for the $GI/G/1$ queue in [1], we find that a second term often provides a significant improvement, but that two terms is often a remarkably good approximation (for ρ not too small, e.g., $\rho \geq 0.6$). The algorithm here provides a means for investigating how many terms are needed as a function of ρ in any $BMAP/G/1$ queue. As one should anticipate, the number of required terms increases as ρ decreases;

see the numerical examples in Section 6.

The asymptotic expansions only yield approximations for the asymptotic decay rates. This applies directly to admission control based solely on asymptotic decay rates, e.g., on effective bandwidths [9,14,17,19,23,37,38,43], but for approximations of the tail probabilities themselves we also need the asymptotic constant. In [1] we suggested a simple approximation for the asymptotic constant, in particular, the product of the asymptotic decay rate and the mean (which becomes a relatively simple approximation upon applying approximations for the mean). We suggest that same approximation for the $BMAP/G/1$ queue as well. Fortunately, for the higher percentiles of the distributions, the asymptotic constant often has relatively little impact; often we can even approximate the asymptotic constant by 1 and get good approximations for the higher percentiles; see Example 6.2 below. However, we have found that the asymptotic constant can be very far from 1 when the arrival process is the superposition of a large number of independent sources [12]. In such circumstances, we evidently need more than the asymptotic decay rate to find good approximations for tail probabilities.

2. The Batch Markovian Arrival Process

In this section we review the basic properties of the *BMAP*. For more details, see Lucantoni [25,26]. The *BMAP* can be defined in terms of two processes $N(t)$ and $J(t)$: $N(t)$ counts the number of arrivals in the time interval $[0,t]$, while $J(t)$ indicates the auxiliary phase state at time t. The pair $(N(t), J(t))$ is a continuous-time Markov chain (*CTMC*) with *infinitesimal generator matrix* $\tilde{Q}$ in block partitioned form; i.e.,

$$\frac{\tilde{Q}}{\rho} = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 \dots \\ & D_0 & D_1 & D_2 \dots \\ & & D_0 & D_1 \dots \\ & & & D_0 \dots \\ & & & \vdots \end{bmatrix}, \tag{1}$$

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 all $k \geq 1$, and $D = \sum_{k=0}^{\infty} D_k$ is an irreducible infinitesimal generator matrix for an m-state *CTMC*.

Let π be the steady-state probability vector associated with D, i.e., determined by $\pi D = 0$ and $\pi e = 1$, where e is a vector of 1's and 0 is a vector of 0's (which should be clear from the context). A fundamental role is played by the *BMAP matrix generating function*

$$D(z) \equiv \sum_{k=0}^{\infty} D_k z^k \ . \qquad (2)$$

We assume that $D(z)$ has a radius of convergence z^* with $1 < z^* \leq \infty$. When $D_k = 0$ for all $k \geq k_0$, as is the case for the ordinary *MAP* (then $k_0 = 2$), $z^* = \infty$. Having $z^* > 1$ implies that $D(z)$ can be regarded as an analytic function of a complex variable z for $|z| < z^*$. The k^{th} derivative $D^{(k)}(z)$ is then finite and analytic for all k and $|z| < z^*$.

Specifying the overall arrival rate ρ separately means that

$$\pi\left(\sum_{k=1}^{\infty} kD_k\right)e = \pi D^{(1)}(1)e = 1 \ . \qquad (3)$$

As shown in [2, Section 3], $D(z)$ has a (simple real) *Perron-Frobenius eigenvalue* $\delta(z)$ for all real z with $0 \leq z < z^*$. In Section 3 below we show that $\delta(z)$ has derivatives of all orders for $0 < z < z^*$. Let $u(z)$ and $v(z)$ be the associated (positive real) left and right eigenvectors normalized so that $u(z)v(z) = u(z)e = 1$. By [2, Theorem 7], $\delta(e^s)$ is strictly increasing and convex function of s with $\delta(1) = 0$.

Let the marginal conditional distribution of $(N(t), J(t))$ be given by

$$P_{ij}(n,t) = P(N(t) = n \ , \ J(t) = j | N(0) = 0 \ , \ J(0) = i) \qquad (4)$$

and let

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n \tag{5}$$

be the associated *counting process matrix generating function*, which is given explicitly by

$$P^*(z,t) = e^{D(z)t} , \ t \geq 0 . \tag{6}$$

Given any initial vector $\tilde{\pi}$ on the phase space, the *counting process* $N(t)$ has probability distribution

$$P_{\tilde{\pi}}(N(t) = n) = \sum_{i=1}^{m} \sum_{j=1}^{m} \tilde{\pi}_i P_{ij}(n,t) = \tilde{\pi} P(n,t) e . \tag{7}$$

Combining (5) and (7), we see that $N(t)$ starting with $\tilde{\pi}$ has generating function $\tilde{\pi} P^*(z,t) e$. It is convenient to focus on the logarithm of this generating function, which is also known as the *factorial cumulant generating function*; see Johnson and Kotz [21, Section 1.5], Daley and Vere-Jones [13, Sections 5.2, 5.4, 7.4, 10.4], Chang [9] and Glynn and Whitt [18], i.e.,

$$c(z,t) \equiv \log E_{\tilde{\pi}} z^{N(t)} = \log(\tilde{\pi} P^*(z,t) e) . \tag{8}$$

We now show that $c(z,t)/t$ is bounded uniformly in t for $1 \leq z < z^*$. For this purpose, we introduce the *matrix norm* $|A| = \max_i \{ \sum_j |A_{ij}| \}$ for

$m \times n$ matrices, which has the property that $|A_1 A_2| \leq |A_1| \cdot |A_2|$ and $|e^A| \leq e^{|A|}$.

Lemma 1. *For real t and z with* $t > 0$ *and* $1 \leq z < z^*$,

$$0 = \frac{c(1,t)}{t} \leq \frac{c(z,t)}{t} \leq |D(z)| < \infty \ .$$

Proof. First, since $\tilde{\pi} e^{D(1)t}$ is a probability vector for each t, $c(1,t) = 0$ for all t. Second, $c(z,t)$ is increasing in z. Third, using the matrix norm,

$$c(z,t) \leq \log(|\tilde{\pi}| \cdot |e^{D(z)t}| \cdot |e|) = |D(z)| t \ ,$$

which is finite by the assumption on $D(z)$. ■

We now describe the asymptotic behavior of $c(z,t)/t$ as $t \to \infty$ in more detail. The following theorem is an immediate consequence of (6) and (8) above and Seneta [34, Theorem 2.7]. It gives an alternative interpretation of the Perron-Frobenius eigenvalue $\delta(z)$.

Theorem 1. *For any real z,* $0 < z < z^*$, *and any initial vector* $\tilde{\pi}$,

$$\tilde{\pi} P^* (z,t) e = e^{\delta(z) t} + O(e^{r(z)t}) \ \ as \ \ t \to \infty \ , \qquad (9)$$

where $r(z) < \delta(z)$, *so that*

$$c(z,t) \equiv \log(\tilde{\pi} P^* (z,t) e) = \delta(z) t + O(e^{-(\delta(z) - r(z))t}) \ \ as \ \ t \to \infty \quad (10)$$

and

$$t^{-1} c(z,t) \equiv t^{-1} \log(\tilde{\pi} P^*(z,t) e) \to \delta(z) \text{ as } t \to \infty \ . \tag{11}$$

Since $D(z)$ has been assumed to be an analytic function for $|z| < z^*$, $c(z,t)$ is analytic in z for $|z| < z^*$ and all t. (Apply Lemma 1 and Daley and Vere-Jones [13, pp. 113-114].) Hence, all derivatives with respect to z are analytic in the same region. Moreover, the k^{th} derivative of $c(z,t)$ with respect to z evaluated at $z = 1$, denoted by $c_k(t)$, is the *k^{th} factorial cumulant of* $N(t)$ (starting with $\tilde{\pi}$). It is helpful to work with the cumulants or factorial cumulants of $N(t)$ instead of the moments or factorial moments primarily for two reasons: First, the k^{th} cumulant of a sum of independent random variables is the sum of the k^{th} cumulants of the random variables being added, and similarly for the factorial cumulants. Second, as we show below, the cumulants and factorial cumulants are asymptotically linear as $t \to \infty$, whereas the k^{th} moment and factorial moment are $O(t^k)$ as $t \to \infty$.

The limiting time-average of the k^{th} factorial cumulant $c_k(t)$ as $t \to \infty$ is the k^{th} *asymptotic factorial cumulant*. We now show that the k^{th} asymptotic factorial cumulant is precisely $\delta^{(k)}(1)$, the k^{th} derivative of $\delta(z)$ evaluated at $z = 1$.

Theorem 2. *For each* $k \geq 1$,

$$\lim_{t \to \infty} \frac{c_k(t)}{t} = \delta^{(k)}(1) \ . \tag{12}$$

Proof. The limit (12) follows from the limit (11) because the terms in (12) are just the coefficients of the power series expansions of the terms in (11), i.e.,

$$\frac{c(z,t)}{t} = \sum_{k=1}^{\infty} \frac{c_k(t)}{t} \frac{(z-1)^k}{k!} \tag{13}$$

and

$$\delta(z) = \sum_{k=1}^{\infty} \frac{\delta^{(k)}(1)(z-1)^k}{k!} \ . \tag{14}$$

for complex z with $|z| < \bar{z}$ where $1 < \bar{z} < z^*$. To establish (12), we first show that (11) holds for all complex z uniformly in z for z in a neighborhood of $z = 1$. For general complex z, let $\delta(z)$ be the eigenvalue of $D(z)$ with maximum real part. Then $e^{\delta(z)t}$ is the maximum-modulus eigenvalue of $e^{D(z)t}$. Instead of (9), we have

$$\tilde{\pi} P^*(z,t) e = u(z) e^{\delta(z)t} v(z) + O(e^{r(z)t}) \ , \tag{15}$$

where $u(z)$ and $v(z)$ are the left and right eigenvector associated with $\delta(z)$ and $r(z)$ is real. Since $\delta(z)$ and $r(z)$ are continuous in z and since $\delta(z)$ is a simple eigenvalue for z real, in a neighborhood of $z = 1$ (15) holds with $r(z) < \mathrm{Re}(\delta(z))$. Moreover, since $\delta(1) > r(1)$ and the continuity holds, for all z in a neighborhood of $z = 1$,

$$\tilde{\pi} P^*(z,t) e = u(z) e^{\delta(z)t} v(z) + O(e^{rt}) \tag{16}$$

with $r < \mathrm{Re}(\delta(z))$ for a constant r. As a consequence of (16), (11) holds for complex z uniformly in z in a neighborhood of $z = 1$. Hence, we can apply the Cauchy integral formula for the k^{th} derivative, i.e.,

$$\frac{c_k(t)}{t} = \frac{d^k}{dz^k} \frac{c(z,t)}{t} \Big|_{z=1} = \frac{k!}{2t\pi i} \int_C \frac{c(z,t)}{(z-1)^{k+1}} \, dz \,, \tag{17}$$

where C is a simple closed contour about $z = 1$ (which can be put inside any neighborhood of $z = 1$). The uniform convergence of (11) in the neighborhood of $z = 1$ implies that the integrals in (17) converge, which establishes (12). ■

The role of cumulants in studying stochastic point processes has a long history, but the analysis seems somewhat obscure; we hope to pursue this further in a subsequent paper. An alternative approach to Theorem 2 is to apply corresponding results for stationary point processes, e.g., Daley and Vere-Jones [3, Exercise 10.4.7], together with a coupling argument to show that the initial nonstationarity is asymptotically negligible; see Lindvall

[24].

Under appropriate regularity conditions, the statement in Theorem 2 can be improved to

$$c_k(t) = \delta^{(k)}(1)t + \gamma_k + r_k(t) \tag{18}$$

where $r_k(t) = o(1)$ as $t \to \infty$ and sometimes even $r_k(t) = O(e^{-s_k t})$ as $t \to \infty$ where s_k is a positive constant. In particular Smith [35,36] obtained such results for renewal processes and cumulative processes (associated with regenerative processes). Note that here $N(t)$ is indeed a cumulative process; as regeneration times we can take successive visits to any fixed phase state after leaving. Hence, Smith's [36] result of the form (18) applies here, except that he does not identify the asymptotic factorial cumulants with the derivatives $\delta^{(k)}(1)$. His expressions for the asymptotic factorial cumulants give alternative expressions for $\delta^{(k)}(1)$. His expressions for the second-order terms γ_k in (18) may be useful for developing refined approximations. For the first two cumulants $c_k(t)$, explicit expressions are also given in Narayana and Neuts [27] and Chapter 5 of Neuts [30].

The superposition of n independent *BMAP*s can be represented as another *BMAP* with an auxiliary phase state space equal to the product of the n individual auxiliary phase state spaces. Let the i^{th} component *BMAP* have arrival rate ρ_i and $m_i \times m_i$ matrices D_{ik}, $k \geq 0$, satisfying the earlier assumptions. We can characterize the superposition *BMAP* using basic properties of Kronecker products $\otimes$ and sums $\oplus$; See Neuts [29,30] for background. We assume that the arrival rates of all *BMAP*s are specified separately from their D_k matrices. When the arrival rates are included, there is important additivity in the matrix generating functions and the eigenvalue function.

Theorem 3. *Consider n independent BMAPs characterized by pairs $(N_i(t), J_i(t))$ with arrival rates ρ_i and $m_i \times m_i$ matrices $D_{ik}, k \geq 0$, $i \leq i \leq n$. Then the pair $(N_1(t) + \ldots + N_n(t), \ (J_1(t), \ldots, J_n(t)))$ determines another BMAP with arrival rate $\rho \equiv \rho_1 + \ldots + \rho_n$ and associated $m \times m$ matrices D_k, where $m = \prod_{i=1}^{n} m_i$, satisfying*

$$\rho D_k = \rho_1 D_{1k} \oplus \ldots \oplus \rho_n D_{nk}, \ k \geq 0 , \tag{19}$$

and matrix generating function

$$D(z) = \left[\frac{\rho_1}{\rho} \right] D_1(z) \oplus \ldots \oplus \left[\frac{\rho_n}{\rho} \right] D_n(z) , \tag{20}$$

which has PF eigenvalue function

$$\delta(z) = \left[\frac{\rho_1}{\rho}\right]\delta_1(z) + \ldots + \left[\frac{\rho_n}{\rho}\right]\delta_n(z) \tag{21}$$

and associated left and right eigenvectors

$$u(z) = u_1(z) \otimes \ldots \otimes u_n(z) \text{ and } v(z) = v_1(z) \otimes \ldots \otimes v_n(z) \tag{22}$$

satisfying $uv = ue = 1$.

Proof. By definition, $N_1(t) + \ldots + N_n(t)$ is the superposition arrival counting process. Since the component *BMAP*s are assumed to be independent, the pair $(N_1(t) + \ldots + N_n(t),\ (J_1(t), \ldots, J_n(t)))$ is a *CTMC* which determines a *BMAP*. The product structure for the auxiliary phase state space means that the D_k matrices should be defined by (19); recall that $A_1 \oplus A_2 = (A_1 \otimes I_2) + (I_1 \otimes A_2)$. Then (20)–(22) are elementary consequences of the Kronecker structure. ■

Formula (21) means that the *PF* eigenvalue function $\delta(z)$ of the superposition process can be determined by separately deriving the component *PF* eigenvalue functions $\delta_i(z)$ for $1 \leq i \leq n$ and then simply adding. This implies that the derivatives are simply additive too.

3. Derivatives of Eigenvalues and Eigenvectors

In this section we determine recursive formulas for the derivatives of the Perron-Frobenius (*PF*) eigenvalue $\delta(z)$ of $D(z)$ and the associated eigenvectors $u(z)$ and $v(z)$ at $z = 1$. The proof follows the Appendix in Neuts [30]. However, here we use a variant of the fundamental matrix of a *CTMC* instead of the fundamental matrix of a discrete-time Markov chain (*DTMC*). For additional discussion of fundamental matrices of *CTMC*s and more references, see Whitt [42]. Since $D \equiv D(1)$ is an infinitesimal generator of an irreducible *CTMC*, $\delta(1) = 0$, $u(1) = \pi$ and $v(1) = e$. Let

$$Y = (e\pi - D)^{-1} \text{ and } Z = Y - e\pi \ . \tag{23}$$

The matrix Z in (3.1) usually is called the *fundamental matrix* of the *CTMC* with generator D; see (13) and (55) of Whitt [42] and Neuts [30, Theorem 5.1.3]. A key fact is that the matrix $e\pi - D$ in (3.1) is nonsingular (when the dimension is two or more, which we assume is always the case).

Let $\delta^{(k)}$ denote the k^{th} derivative of $\delta(z)$ at $z = 1$, i.e., $\delta^{(k)} \equiv \delta^{(k)}(1)$, and similarly for other variables.

Theorem 4. *The derivatives of* $\delta(z)$, $u(z)$ *and* $v(z)$ *at* $z = 1$ *are given by*

$$\delta^{(1)} = \pi D^{(1)} e = 1 \ , \tag{24}$$

$$u^{(1)} = \pi(D^{(1)} - \delta^{(1)} I) Y = \pi D^{(1)} Z \ , \tag{25}$$

$$v^{(1)} = Y(D^{(1)} - \delta^{(1)} I) e = Z D^{(1)} e \ , \tag{26}$$

$$\delta^{(2)} = \pi D^{(2)} e + 2\pi D^{(1)} Z D^{(1)} e \ ,$$

$$
\begin{aligned}
u^{(2)} &= \pi D^{(2)} Z + 2\pi D^{(1)} Z D^{(1)} Z - 2\pi D^{(1)} Z^2 \\
&= \pi D^{(2)} Z + 2u^{(1)} (D^{(1)} - I) Z \, , \\
v^{(2)} &= Z D^{(2)} e + 2 Z D^{(1)} Z D^{(1)} e \\
&\quad -2 Z^2 D^{(1)} e - 2(\pi D^{(1)} Z^2 D^{(1)} e) e \\
&= Z D^{(2)} e + 2 Z (D^{(1)} - I) v^{(1)} - 2 u^{(1)} v^{(1)} \, ,
\end{aligned}
$$

$$
\begin{aligned}
\delta^{(3)} &= \pi D^{(3)} e + 3\pi D^{(1)} Z D^{(2)} e + 3\pi D^{(2)} Z D^{(1)} e \\
&\quad + 6\pi D^{(1)} Z D^{(1)} Z D^{(1)} e - 6\pi D^{(1)} Z^2 D^{(1)} e \\
&= \pi D^{(3)} e + 3 u^{(1)} D^{(2)} e + 3\pi D^{(2)} v^{(1)} \\
&\quad + 6 u^{(1)} (D^{(1)} - I) v^{(1)} \, ,
\end{aligned}
$$

and, for $n \geq 2$,

$$
\delta^{(n)} = \pi D^{(n)} e + \sum_{k=1}^{n-1} \begin{bmatrix} n \\ k \end{bmatrix} \pi (D^{(k)} - \delta^{(k)} I) v^{(n-k)} \, , \tag{27}
$$

$$
u^{(n)} = \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} u^{(n-k)} (D^{(k)} - \delta^{(k)} I) Y \, , \tag{28}
$$

$$
v^{(n)} = e \pi v^{(n)} + Y \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} (D^{(k)} - \delta^{(k)} I) v^{(n-k)} \, , \tag{29}
$$

where

$$
\pi v^{(n)} = - \sum_{k=1}^{n-1} \begin{bmatrix} n \\ k \end{bmatrix} u^{(k)} v^{(n-k)} \, . \tag{30}
$$

Proof. Follow the proof of Theorem A.2.1 of Neuts [30, p. 482]. Start by differentiating n times in the eigenvalue equation $D(z) v(z) = \delta(z) v(z)$ to obtain

$$\sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} (D^{(k)} - \delta^{(k)} I) v^{(n-k)} = 0 \ . \tag{31}$$

Premultiplying by π in (31), we obtain (24) and (27). From (31), we also obtain the singular system of linear equations

$$-Dv^{(n)} = \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} (D^{(k)} - \delta^{(k)} I) v^{(n-k)} \ . \tag{32}$$

Adding $e\pi v^{(n)}$ to both sides in (32), we obtain

$$(e\pi - D) v^{(n)} = e\pi v^{(n)} + \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} (D^{(k)} - \delta^{(k)} I) v^{(n-k)} \ . \tag{33}$$

Using (23) and $Ye = e$, we obtain (26) and (29) from (33). Differentiating n times in the relation $ue = 1$ and $uv = 1$, we obtain $u^{(n)} e = 0$ for all $n \geq 1$, $\pi v^{(1)} = -u^{(1)} e = 0$ and (30). To obtain the second formula in (26), note that

$$\begin{aligned} Y(D^{(1)} - \delta^{(1)} I) e &= Y(D^{(1)} - \pi D^{(1)} eI) e \\ &= (Y - e\pi) D^{(1)} e = ZD^{(1)} e \ . \end{aligned} \tag{34}$$

By similar reasoning, we obtain (25) and (28). The analog of $e\pi v^{(n)}$ is $u^{(n)} e\pi$. Since $u^{(n)} e = 0$ for $n \geq 1$, this term drops out. ■

We present one illustrative example.

Example 3.1. Suppose that the *BMAP* is a two-phase Erlang (E_2) renewal process with rate 1, as in Lucantoni [25, pp. 32,36]. Then the phase-type (PH) and *BMAP* representations are (α, T) with $\alpha = (1,0)$ and

$$D_0 = T = \begin{bmatrix} -2 & 2 \\ 0 & -2 \end{bmatrix}, \quad D_1 = -Te\alpha = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} \tag{35}$$

and $D_k = 0$ for $k \geq 2$. Then $\pi = (1/2, 1/2)$, $\delta^{(1)} = 1$,

$$Y = \begin{bmatrix} 5/8 & 3/8 \\ 3/8 & 5/8 \end{bmatrix}, \quad Z = \begin{bmatrix} 1/8 & -1/8 \\ -1/8 & 1/8 \end{bmatrix} \tag{36}$$

$u^{(1)} = (1/8, -1/8)$, $v^{(1)} = (-1/4, 1/4)^t$, $\delta^{(2)} = -1/2$ and $\delta^{(3)} = 3/4$. The associated first three asymptotic cumulants of the *BMAP* are $c_1 = 1$, $c_2 = \delta^{(2)} + c_1 = 1/2$ and $c_3 = \delta^{(3)} + 3c_2 - 2c_1 = 1/4$; see [13, p. 114]. These agree with the known formulas for renewal arrival processes based on first three moments $m_1 = 1$, $m_2 = 3/2$ and $m_3 = 3$; see Riordan [32, p. 37] and Whitt [41, (2.7)]. This can be determined solely from relations between moments and cumulants, starting with the formula $m_k = k! m_1^k$ for the k^{th} moment of an exponential (M) random variable. An E_2 random variable with mean 1 is the sum of two independent M random variables. The first three moments of each M variable are $(1/2, 1/2, 3/4)$, so that the first three cumulants are $(1/2, 1/4, 1/4)$. The cumulants add, so that the cumulants of the E_2 variable are $(1, 1/2, 1/2)$.

Thus the moments of the E_2 variable are $(1, 3/2, 3)$. Then, by [41, (2.7)], the asymptotic cumulants are $(1, 1/2, 1/4)$.

4. Asymptotic Expansions for the Asymptotic Decay Rates

We now consider the $BMAP/G/1$ queue, which has i.i.d. service times that are independent of the batch Markovian arrival process (*BMAP*). Let V be a generic service time. We assume that $\phi(s) \equiv Ee^{sV} < \infty$ for some $s > 0$ and that $EV = 1$. We assume that $\rho < 1$, so that the model is stable. Let $\eta_V = \sup\{s : \phi(s) < \infty\}$; $\eta_V = \infty$ for some distributions; $\eta_V < \infty$ for other distributions.

Let Q and L be the steady-state queue length (number in system) and workload (virtual waiting time) at an arbitrary time, which we assume are well defined. See Ramaswami [31], Neuts [30], Lucantoni [25,26] and Abate, Choudhury and Whitt [2] for the transforms of the distributions of Q and L. We are interested in the asymptotic behavior of the tail probabilities. In great generality,

$$\sigma^{-k} P(Q > k) \to \beta \text{ as } k \to \infty \tag{37}$$

and

$$e^{\eta x} P(L > x) \to \alpha \text{ as } x \to \infty \ , \tag{38}$$

where σ, β, η and α are positive constants with $\alpha = \beta$; see Abate, Choudhury and Whitt [2] and references there.

Since here we are interested only in the asymptotic decay rates σ and η, we can also focus on the weaker limits

$$k^{-1} \log P(Q > k) \to \log \sigma \quad \text{as} \quad k \to \infty$$

and

$$x^{-1} \log P(L > x) \to -\eta \quad \text{as } x \to \infty ,$$

where log is the natural logarithm, see Chang [9] and Glynn and Whitt [18].

The pair of asymptotic decay rates (σ^{-1}, η) are characterized in [2] as the (necessarily unique) solution with $1 < \sigma^{-1} < z^*$ and $0 < \eta < \eta_V$ to the two equations

$$\delta(1/\sigma) = \frac{\eta}{\rho} \quad \text{and} \quad \phi(\eta) \equiv Ee^{\eta V} = \sigma^{-1} . \tag{39}$$

By [1, Theorem 11], $0 < \eta < \eta_V$. We remark that multiple solutions to (39) can hold above η_V when Ee^{sV} is not interpreted as $+\infty$ for $s > \eta_V$; i.e., there can be problems if we simply substitute $-s$ into an expression for the Laplace transform Ee^{-sV}.

We assume that there is indeed a solution (σ^{-1}, η) to (39), which we have noted is unique, and we develop algorithms to find σ and η as functions of ρ. First, a relatively simple approach is direct root finding. We can start with a candidate x for η with $0 < x < \eta_V$, and calculate $\rho\delta(\phi(x))$. If $\rho\delta(\phi(x)) > x$, then we decrease x; if $\rho\delta(\phi(x)) < x$, then we increase. In particular, if $\eta_V < \infty$, then it is a possible initial x. Otherwise, start with 1. Go from x to $2x$ if x is the largest value considered so far and we should increase; otherwise use bisection. Given an x and a y such that $\rho\delta(\phi(x)) \leq x$ and $\rho\delta(\phi(y)) \geq y$, we know that $x \leq \eta \leq y$.

Second, we express η and $(\sigma^{-1} - 1)$ as asymptotic expansions in powers of $(1 - \rho)$. When we use many terms, this is a numerical algorithm for finding the exact solution; when we use relatively few terms, this can be regarded as a simple approximation. A combination of the two approaches yields an efficient algorithm for computing σ and η as functions of ρ. We use the asymptotic expansion for ρ in the interval $[\rho_0, 1]$ and root finding below ρ_0. We use the root finding to identify ρ_0 and confirm the accuracy above ρ_0.

As in [3], we develop approximations for σ^{-1} and η by expanding $\delta(z)$ and $\phi(s) \equiv Ee^{sV}$ in Taylor series about $z = 1$ and $s = 0$, respectively. We have

$$\delta(z) = \delta(1) + \delta^{(1)}(1)(z-1) + \ldots + \delta^{(k)}(1)\frac{(z-1)^k}{k!} + \ldots$$

$$= (z-1) + \delta^{(2)}\frac{(z-1)^2}{2} + \ldots + \delta^{(k)}\frac{(z-1)^k}{k!} + \ldots \qquad (40)$$

and

$$\phi(s) = \phi(0) + \phi^{(1)}(0)s + \phi^{(2)}(0)\frac{s^2}{2} + \ldots + \phi^{(k)}(0)\frac{s^k}{k!} + \ldots$$

$$= 1 + s + m_2\frac{s^2}{2} + \ldots + m_k\frac{s^k}{k!} + \ldots , \qquad (41)$$

where $\delta^{(k)} \equiv \delta^{(k)}(1)$ and $m_k = EV^k = \phi^k(0)$. We obtain $\delta^{(k)}$ from the algorithm in Section 3. Given an explicit expression for the transform $\phi(s) \equiv Ee^{sV}$, we can calculate any desired number of moments m_k via the algorithm in Choudhury and Lucantoni [10].

Given either σ or η (or approximations) and the derivatives, we can obtain an approximation for the other from (39)–(41), i.e.,

$$\frac{\eta}{\rho} = \delta(\sigma^{-1}) \approx (\sigma^{-1}-1) + \ldots + \delta^{(k)}\frac{(\sigma^{-1}-1)^k}{k!} + \ldots \qquad (42)$$

and

$$\sigma^{-1} = \phi(\eta) \approx 1 + \eta + \frac{m_2\eta^2}{2} + \ldots + \frac{m_k\eta^k}{k!} + \ldots \qquad (43)$$

We first develop an asymptotic expansion in powers of $1-\rho$ for η by considering the composite function $\delta(\phi(\eta))$. We then use the (43) to develop an asymptotic expansion for σ^{-1} in powers of $1-\rho$.

By (39), $\delta(\phi(\eta)) = \eta/\rho$. Consequently,

$$\psi(\eta) \equiv \frac{\delta(\phi(\eta))}{\eta} - 1 = \frac{1-\rho}{\rho} \equiv \varepsilon \ . \tag{44}$$

We apply (40) and (41) to determine the coefficients in a Taylor series expansion of (44), i.e.

$$\psi(\eta) = a_1\eta + a_2\frac{\eta^2}{2} + \ldots + \frac{a_k\eta^k}{k!} + \ldots \tag{45}$$

Then we use reversion of series, see 3.6.25 of Abramowitz and Stegun [4, p. 16] and Riordan [32, Section 2.8], to obtain

$$\eta = f(\varepsilon) \equiv b_1\varepsilon + b_2\frac{\varepsilon^2}{2} + \ldots + b_k\frac{\varepsilon^k}{k!} + \ldots \tag{46}$$

from (45). The idea is that $\eta = f(\psi(\eta))$, so that the coefficients b_k in (46) can be obtained from the coefficients a_k in (45) using the formula for the derivatives of a composite function. (In this step, and others to follow, we could also use a symbolic language such as MAPLE.)

First, Table 1 gives the first seven coefficients a_k in (45) in terms of the derivatives $\delta^{(k)} \equiv \delta^{(k)}(1)$ and $m_k = EV^k = \phi^{(k)}(0)$. These are obtained from the composite function $\delta(\phi(\eta))$ plus (44). In Section 2.8 and Problem 32 on p. 47 of Riordan [29] and Riordan [33], a recursive algorithm is given for the derivatives. This can be used to calculate a_k for arbitrary k. Second, Table 2 gives the first seven coefficients b_k in (46) in terms of the coefficients a_k in (45). Again the algorithm in Riordan [32] enables us to calculate b_k for arbitrary k.

Note that these formulas simplify in special cases. For example, if the service-time distribution is deterministic (D), then $m_k = 1$ for all $k \geq 1$. If the service-time distribution is exponential (M), then $m_k = k!$ If the arrival process is Poisson, then $\delta(z) = D(z) = z - 1$, so that $\delta^{(k)} = 0$ for all $k \geq 2$. For a Poisson arrival process, $(n + 1) a_n = m_{n+1}$ for all $n \geq 1$.

Next, let $g(x) = x/(1-x)$ and note that $g(1-\rho) = (1-\rho)/\rho = \varepsilon$. Hence, we can express the asymptotic decay rate η in an asymptotic expansion in powers of $(1-\rho)$, i.e.,

$$\eta = h(1-\rho) \equiv f(g(1-\rho)) = c_1 (1-\rho) + \ldots + c_k \frac{(1-\rho)^k}{k!} + \ldots \quad (47)$$

Once again, we use the formula for the derivatives of a composite function to obtain the coefficients c_k in (47) from the coefficients b_k in (46).

Since

$$g(x) = \frac{x}{1-x} = \sum_{k=1}^{\infty} x^k \ , \quad (48)$$

we have $g^{(k)}(0) = k!$ Table 3 below gives the first seven coefficients c_k in (47) in terms of the coefficients b_k in (46). Once again, the algorithm in Riordan [32, Section 2.8] can be used to calculate c_k for arbitrary k.

Finally, we can combine the previous results to express the coefficients c_k in (47) first in terms of the coefficients a_k in (45) and then in terms of the derivatives $\delta^{(k)}$ and m_k. The first four coefficients c_k are displayed in Table 4. For the first two, we give expressions in terms of $\delta^{(k)}$ and m_k.

From the analysis so far, we see that the coefficient c_k in (47) depends on the first $(k+1)$ derivatives $\delta^{(j)}$ and m_j. (The first derivatives are fixed; by convention $\delta^{(1)} = m_1 = 1$.) Note that c_2 is decreasing in m_2 and $\delta^{(2)}$. More generally, c_k is decreasing in m_{k+1} and $\delta^{(k+1)}$. (Smaller η means more congestion.)

It is well known that in the M/M/1 queue, $\eta = 1-\rho$ and $\sigma = \rho$. hence, for the M/M/1 queue, $c_1 = 1$ and $c_k = 0$ for $k \geq 2$. It is easy to see that Table 4 is consistent with this, because then $\delta(z) = z-1$ and $m_k = k!$, so that $a_k = k!$

The simple heavy-traffic approximation is the first term, i.e.,

$$\eta \approx c_1(1-\rho) = \frac{2(1-\rho)}{m_2 + \delta^{(2)}} \; . \tag{49}$$

Note that $c_A^2 \equiv 1 + \delta^{(2)}$ is the asymptotic variance of the arrival process, while $c_s^2 \equiv m_2 - 1$ is the squared coefficient of variation (variance divided by the square of the mean) of the service time. Hence, (49) agrees with the familiar heavy-traffic formula; e.g., see [16].

Paralleling [3], the two-term refined approximation here is

$$\eta \approx c_1(1-\rho) + c_2 \frac{(1-\rho)^2}{2} = \frac{2(1-\rho)}{m_2+\delta^{(2)}} + \frac{2(3m_2^2+3(\delta^{(2)})^2-2m_3-2\delta^{(3)})}{3(m_2+\delta^{(2)})^3}(1-\rho)^2 \tag{50}$$

For the $M/G/1$ queue, (50) here agrees with (14) of [3]; then $\delta^{(k)} = 0$ for $k \geq 2$. Higher-order refined approximations follow from Tables 1-4 and the associated recursive algorithm for the derivatives of composite functions. For the $M/G/1$ queue, $a_k = m_{k+1}/(k+1)$ so that

$$\eta \approx c_1(1-\rho)+c_2(1-\rho)^2+c_3(1-\rho)^3+c_4(1-\rho)^4 , \tag{51}$$

where

$$c_1 = \frac{2}{m_2}\ , \quad c_2 = \frac{6m_2^2 - 4m_3}{3m_2^3}$$

$$c_3 = \frac{36m_2^4 - 48m_3 m_2^2 + 32m_3^2 - 12m_2 m_4}{3m_2^5}$$

$$c_4 = [2400m_2 m_3 m_4 - 288m_2^2 m_5 - 3200m_3^3 + 5760m_3^2 m_2^2 \\ - 2160m_2^3 m_4 - 4320m_3 m_2^4 + 2160m_2^6]/45m_2^7\ . \tag{52}$$

In the M/M/1 case, $m_k = k!$ and $\eta = 1 - \rho$. Formula (52) is consistent with this result. It is interesting that the individual terms cancel in this case. This suggests that we might rewrite (52) as

$$\begin{aligned} c_2 &= \frac{2(3m_2^3 - 2m_3)}{3m_2^3} \\ c_3 &= \frac{12m_2(3m_2^3 - m_4) - 16m_3(3m_2^2 - 2m_3)}{3m_2^5} \\ c_4 &= [240(3m_2^2 - 2m_3)(4m_3^2 - 3m_2 m_4)\ -\ 320m_3(4m_3^2 - 3m_2 m_4) + \\ &\quad 144m_2^2(15m_2^4 - 2m_5) -\ 1440m_2^2 m_3(3m_2^2 - 2m_3)]/45m_2^7\ . \end{aligned} \tag{53}$$

In other words, we evaluate m_3 in relation to m_2 via the term $(3m_2^2 - 2m_3)$ and we evaluate m_4 relative to m_2 via the term $(3m_2^3 - m_4)$. In the fourth term we also evaluate m_4 relative to m_2 and m_3 via $(4m_3^2 - 3m_2 m_4)$ and m_5 relative to m_2 via $(15m_2^4 - 2m_5)$. These individual terms are all 0 for the exponential distribution.

Now we consider the other asymptotic decay rate σ. We can combine (43), (47) and Riordan [32, p. 49] to obtain an asymptotic expansion in powers of $(1-\rho)$ for σ^{-1}, i.e.,

$$\sigma^{-1} - 1 = \phi(h(1-\rho)) - \phi(0) = \sum_{k=1}^{\infty} d_k \frac{(1-\rho)^k}{k!} . \qquad (54)$$

The first seven coefficients d_k in (54) are expressed in terms of the coefficients c_k in (47) in Table 5.

As a consequence, the simple heavy-traffic approximation is the first term, i.e.,

$$\sigma^{-1} - 1 \approx c_1(1-\rho) = \frac{2(1-\rho)}{m_2+\delta^{(2)}} , \qquad (55)$$

so that

$$1-\sigma = \frac{2(1-\rho)}{m_2+\delta^{(2)}} + O((1-\rho)^2) . \qquad (56)$$

The two-term refined approximation is

$$\sigma^{-1} - 1 \approx c_1(1-\rho) + (c_2 + m_2 c_1^2)\frac{(1-\rho)^2}{2} = \frac{2(1-\rho)}{m_2\delta^{(2)}}$$

$$+ \left[\frac{2(3m_2^2 + 3(\delta^{(2)})^2 - 2m_3 - 2\delta^{(3)})}{3(m_2+\delta^{(2)})^3} + \frac{2m_2}{(m_2+\delta^{(2)})^2} \right] (1-\rho)^2 . \qquad (57)$$

5. Non-BMAP Arrival Processes

We have seen that the asymptotic decay rates in a *BMAP/G/*1 queue depend on the *BMAP* through the Perron-Frobenius eigenvalue $\delta(z)$ of the *BMAP* matrix generating function $D(z)$ in (2). Moreover, for z near 1, $\delta(z)$ is primarily determined by its derivatives $\delta^{(k)} \equiv \delta^{(k)}(1)$ at $z = 1$. Furthermore, we have characterized $\delta(z)$ and $\delta^{(k)}$ in terms of the limiting time-average of the factorial cumulant generating function in (8) and (6), and its derivatives, the factorial cumulants $c_k(t)$. It is significant that these last characterizations extend beyond *BMAP*s.

When the arrival counting process $N(t)$ is a general stochastic point process, we *define* $\delta(z)$ by

$$\delta(z) = \lim_{t \to \infty} \frac{\log E z^{N(t)}}{t} , \tag{58}$$

assuming that the limit exists; see Chang [9], Glynn and Whitt [18] and Whitt [43]. Moreover, we *define* $\delta^{(k)}$ as

$$\delta^{(k)} = \lim_{t \to \infty} \frac{c_k(t)}{t} , \tag{59}$$

again assuming that the limit exists.

Given (58) and (59), we can treat single-server queues with non-*BMAP* arrival processes if we can obtain convenient expressions for the generating function $Ez^{N(t)}$ and/or the factorial cumulants. For example, we can treat renewal processes and superpositions of independent renewal processes by applying expressions for the asymptotic cumulants obtained by Smith [35]. To illustrate, Table 6 gives expressions for the first four asymptotic factorial cumulants of a rate-one renewal process in terms of the first four moments of the inter-renewal times. (Since the renewal process has rate 1, the first moment of the inter-renewal time is 1.) A specific numerical example is given below in Example 6.3.

Smith [35] gives formulas for the first eight asymptotic cumulants, the first four of which are given in [41, (2.7)]. The asymptotic factorial cumulants are related to the asymptotic cumulants the same way that factorial moments are related to moments. Smith [36] also gives expressions for asymptotic cumulants of a cumulative process associated with a regenerative process. Since the *BMAP* is a special case, this gives alternative expressions for the asymptotic factorial cumulants for a *BMAP*. It also gives formulas for a large class of non-*BMAP* arrival processes.

By invoking the additivity discussed in Section 2, we can treat arrival processes that are independent superpositions of *BMAP*s, non-*BMAP* renewal processes and other non-*BMAP* non-renewal cumulative processes.

6. Numerical Examples

In this section we look at some numerical examples. Our first two examples are $MMPP/\Gamma_\gamma/1$ queues, with a gamma service-time distribution having shape parameter γ and a Markov modulated Poisson process ($MMPP$) as an arrival process, which is a MAP. As before, we assume that the mean service time is 1, so that γ is the sole service parameter. The parameter γ allows us to consider a range of variability; i.e. the SCV is $c_s^2 = 1/\gamma$. In particular, $\gamma = 1$ is exponential (M) , $\gamma = k$ is Erlang (E_k) which approaches D as $k \to \infty$, and $\gamma < 1$ corresponds to distributions that are more variable than exponential.

Example 6.1. We start by considering a two-phase $MMPP$ ($MMPP_2$). The $MMPP_2$ has four parameters (the arrival rate and mean holding time in each phase), one of which we determine by letting the overall arrival rate be ρ. We fix one parameter by assuming that the long-run arrival rate in each phase is $\rho/2$; we fix another parameter by assuming that the expected number of arrivals during each visit to each phase is 5. Our final parameter is the ratio r of the arrival rates in the two phases. We see what happens as we vary the three parameters γ, r and ρ.

First, to obtain a concrete model with ρ the only free parameter, we let $r = 4$ and $\gamma = 0.5$. This gamma service-time distribution does not have a rational Laplace transform so it is not *PH*. Since it has $c_s^2 = 2.0$, it is moderately highly variable. This service-time distribution was used previously in Example 1 of [1].

For this case, we now give the approximating caudal characteristic curve (η as a function of ρ) based on the seven-term heavy-traffic expansion. Let $\eta_k(\rho)$ be the k-term approximation for η as a function of ρ. Then

$$\begin{aligned}\eta \equiv \eta(\rho) \approx \eta_7(\rho) &= 0.41667(1-\rho) - 0.07668(1-\rho)^2 \\ &+ 0.00116(1-\rho)^3 + 0.05428(1-\rho)^4 0.07884(1-\rho)^5 \\ &+ 0.07243(1-\rho)^6 + 0.04636(1-\rho)^7 \ . \end{aligned} \tag{60}$$

From (60), it is evident that when ρ is close to 1 $\eta_k(\rho)$ should be a good approximation for small k and improve dramatically as k increases. However, we cannot expect the quality of the approximation to be good for very small ρ. (There we should rely on the root finding.)

The exact values of the asymptotic decay rates η and σ were found for 12 values of ρ by root finding and are displayed in Table 7. Also displayed in Table 7 are the values η_k of the k-term expansions for η for each k, $1 \leq k \leq 7$, (i.e., η_7 is given in (60)) and the approximation $\sigma_7 = 1/\phi(\eta_7)$ for σ. (We do not display the alternative approximation for σ based on its asymptotic expansion in (54).)

From Table 7, we see that the approximations η_7 and σ_7 essentially coincide with the exact values provided that ρ is not too small, say for $\rho \geq 0.3$. No value of σ_7 is given for $\rho = 0.05$ because $\eta_7 > \eta_V = 0.5$ in that case.

Note that the heavy-traffic approximation η_1 is remarkably good for $\rho \geq 0.2$. Indeed, η_1 is actually better than η_2 for $\rho \leq 0.4$. However, we suggest considering η_k truly a good approximation only if η_j remains good for all $j \geq k$. (It seems evident that nothing strange will happen with η_k for $k > 7$, but we have not verified it.) In this sense, η_1 and η_2 might be judged very good only for $\rho \geq 0.9$ and $\rho \geq 0.6$, respectively. This is based on a criterion of two percent relative error: The percent relative error of η_1 and η_2 at $\rho = 0.95$, 0.90, 0.80, 0.70, 0.60 and 0.50 are, respectively, 0.9%, 1.9%, 3.7%, 5.2%, 6.1%, 5.9% and 0.0%, 0.0%, 0.1%, 0.6%, 1.7% and 3.9%. As in [1], we conclude that for engineering purposes η_2 does remarkably well for ρ not too small. However, for small ρ, e.g. $\rho \leq 0.5$, the convergence of the asymptotic expansion is quite slow (which should be no surprise). For example, at $\rho = 0.2$, η_k is better than η_1 for the first time at $k = 6$.

We next consider what happens as we vary each of the parameters r and γ. Table 8 displays η and η_k, $1 \leq k \leq 7$, for 4 values of r with $\gamma = 0.5$

and $\rho = 0.5$ and 0.8, while Table 9 displays η and η_k, $1 \leq k \leq 7$, for 6 values of γ with $r = 4.0$ and $\rho = 0.5$ and 0.8. The case $\gamma = \infty$ corresponds to a deterministic service-time distribution.

As anticipated, fewer terms suffice with higher ρ in all these cases. At $\rho = 0.8$, we might consider the heavy-traffic approximation η_1 good enough and we would certainly consider η_2 good enough. At $\rho = 0.5$, neither η_1 nor η_2 is consistently good (especially by our criterion that η_j remain good for all $j \geq k$).

Table 9 shows that for $\rho = 0.5$ the quality of the approximations degrade as γ increases. The deterministic service-time distribution is the most difficult case we considered. Indeed, for $\rho = 0.5$, η_7 still has 1.1% and 1.6% error for $\alpha = 4.0$ and $\alpha = \infty$. The relatively big change from η_6 to η_7 in that case shows that convergence has not yet occurred.

Example 6.2. To begin learning about the impact of multiple sources, we now compare one *MMPP* source with the superposition of two independent *MMPP* sources. We consider the $MMPP/\Gamma_\gamma/1$ model in Example 6.1 with $\gamma = 0.5$, $r = 4.0$ and two values of ρ: $\rho = 0.5$ and $\rho = 0.8$. The component *MMPP*s in the superposition are the same as the single-source *MMPP* except time has been scaled by a factor of 2; i.e., the component process parameters are obtained by taking the single-source *MMPP* and changing its free parameters from (ρ, r) to $(\rho/2, r)$. (We have chosen the parameters so that time scaling corresponds simply to changing ρ).

We compare approximations with exact values for the percentiles of the steady-state waiting time when there are one and two component streams. The exact percentile values are computed using the program in [11]. (Bisection search is used to find the percentiles with the algorithm for the tail probabilities.) For the approximations, we always use the exact value of the asymptotic decay rate η (which does not change as the number of component streams changes). We consider three exponential approximations. The first is the asymptotic exponential approximation $\alpha e^{-\eta x}$ with the exact asymptotic constant α (computed using [10]); the second has α approximated by ηEW, as suggested in [1]; and the third has α approximated crudely by 1.0.

Tables 10 and 11 display the results for $\rho = 0.5$ and $\rho = 0.8$, respectively. Consistent with the theoretical results in [2], the exact asymptotic formula becomes a very good approximation as the percentile increases. Indeed, for very high percentiles all the approximations are pretty good, which helps justify our having focused on the asymptotic decay rate η in this paper. From Tables 10 and 11, we see that the quality of the approximations degrades as (1) the number of streams increases, (2) the percentile required decreases, and (3) the traffic intensity decreases.

Focusing on the number of streams, we see from Tables 10 and 11 that for any given ρ and any given required percentile, the quality of the asymptotic exponential approximation degrades as we increase the number of streams from 1 to 2. In [12] we investigate this issue further. There we show that the percentile where the asymptotic exponential approximation is judged good typically increases as the number of streams increases. This phenomenon occurs because the asymptotic constants α and β in (37) and (38) themselves are exponential in the number of sources when the sources are scaled to keep the total arrival rate fixed.

Hence, when the arrival process is a superposition of a large number of independent processes, the asymptotic decay rate alone often does not yield good approximations for tail probabilities. As in previous work on steady-state means [16,20,39], more intricate approximations are evidently needed.

Example 6.3 To illustrate how the results extend beyond *BMAP* arrival processes, we consider the $\Gamma_{0.5}/\Gamma_2/1$ queue, which has $\Gamma_2 \equiv E_2$ service times independent of a $\Gamma_{1/2}$ renewal arrival process. Since the $\Gamma_{1/2}$ distribution does not have a rational Laplace transform, it is not phase-type and not a *BMAP*.

The moments of a Γ_γ random variable with mean γ are $\Gamma(\gamma + r)/\Gamma(\gamma)$, where $\Gamma(x)$ is the gamma function; see (9) on p. 168 of Johnson and Kotz [22]. Hence, the k^{th} moment of a $\Gamma_{1/2}$ distribution with mean 1 is $(2k-1)!/2^{(k-1)}(k-1)!$. The first four moments of the mean-1 $\Gamma_{1/2}$ distribution are 1, 3, 15 and 105. By Table 6, The first four asymptotic factorial cumulants are 1, 1, 0 and 0. Interestingly, the first four coefficients c_k in the expansion for η in (47) are all positive. (When this is true for all k, the approximations improve monotonically with k for all ρ.) Table 12 displays the first four approximations η_k and the exact values for four values of ρ. As with the $BMAP/G/1$ queues, there is good accuracy.

Acknowledgment. We thank Joseph Abate, David Lucantoni and the referees for assistance.

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. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models* 10 (1994).

[3] Abate, J. and W. Whitt. A heavy-traffic expansion for asymptotic decay rates in multi-channel queues. *Opns. Res. Letters*, forthcoming.

[4] Abramowitz, M. and I. A. Stegun. *Handbook of Mathematical Functions*, Washington, D.C.: National Bureau of Standards, U.S.

Government Printing Office, 1972.

[5] Asmussen, S. Risk theory in a Markovian environment. *Scand. Act. J.* (1989) 69-100.

[6] Asmussen, S. and G. Koole. Marked point Processes as limits of Markovian arrival streams. *J. Appl. Prob.* 30 (1993) 365-372.

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

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

[9] Chang, C.-S. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Aut. Cont.,* forthcoming.

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

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

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

[13] Daley, D. J. and D. Vere-Jones. *An Introduction to the Theory of Point Processes.* New York: Springer-Verlag, 1988.

[14] Elwalid, A. I and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking* 1 (1993) 329-343.

[15] 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.

[16] Fendick, K. W. and W. Whitt. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proc. IEEE* 77 (1989) 171-194.

[17] Gibbens, R. J. and P. J. Hunt. Effective bandwidths for the multi-type UAS channel. *Queueing Systems* 9 (1991) 17-28.

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

[19] Guerin, R., H. Ahmadi, and N. Naghshineh. Equivalent capacity and its application to bandwidth application in high-speed networks, *IEEE J. Sel. Areas Commun.* 9 (1991) 968-981.

[20] Heffes, H. and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. Sel. Areas Commun.* SAC-4 (1986) 856-868.

[21] Johnson, N. L and S. Kotz. *Discrete Distributions.* New York: Wiley, 1969.

[22] Johnson, N. L. and S. Kotz. *Distributions in Statistics, Continuous Univariate Distributions-1.* New York: Wiley, 1970.

[23] Kelly, F. P. Effective bandwidths at multi-class queues. *Queueing Systems* 9 (1991) 5-16.

[24] Lindvall, T. *Lectures on the Coupling Method.* New York: Wiley, 1992.

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

[26] Lucantoni, D. M. The $BMAP/G/1$ queue: a tutorial. In *Models and Techniques for Performance Evaluation of Computer and Communication Systems,* L. Danatiello and R. Nelson (eds.), New York: Springer-Verlag, 1993, 330-358.

[27] Narayana, S. and M. F. Neuts. The first two moment matrices of the counts for the Markovian arrival process. *Stochastic Models* 8 (1992) 459-477.

[28] Neuts, M. F. A versatile Markovian point process. *J. Appl. Prob.* 16 (1979) 764-779.

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

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

[31] Ramaswami, V. The $N/G/1$ queue and its detailed analysis. *Adv. Appl. Prob.* 12 (1980) 222-261.

[32] Riordan, J. An *Introduction to Combinatorial Analysis.* New York: Wiley, 1958.

[33] Riordan, J. *Combinatorial Identities.* New York: Wiley, 1968.

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

[35] Smith, W. L. On the cumulants of renewal processes. *Biometrika* 43 (1959) 1-29.

[36] Smith W. L. On the cumulants of cumulative processes. Department of Statistics, University of North Carolina, 1977.

[37] Sohraby, K. On the asymptotic behavior of heterogeneous statistical multiplexer with applications. *Infocom '92,* Florence, Italy, 1992.

[38] Sohraby, K. On the theory of general on-off sources with applications in high-speed networks. *Infocom '93,* San Francisco, CA, 1993.

[39] Sriram, K. and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Sel. Areas Commun.,* SAC-4 (1986) 833-846.

[40] van Ommeren, J. C. W. Exponential expansion for the tail of the waiting-time probability in the single-server queue with batch arrivals. *Adv. Appl. Prob.* 20 (1988) 880-895.

[41] Whitt, W. Approximating a point process by a renewal process, I: two basic methods. *Opns. Res.* 30 (1992) 125-147.

[42] Whitt, W. Asymptotic formulas for Markov processes with applications to simulation. *Opns. Res.* 40 (1992) 279-291.

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

k	$(k+1)\,a_k$
1	$2a_1 = m_2 + \delta^{(2)}$
2	$3a_2 = m_3 + \delta^{(2)}(3m_2) + \delta^3$
3	$4a_3 = m_4 + \delta^{(2)}(4m_3 + 3m_2^2) + \delta^{(3)}(6m_2) + \delta^{(4)}$
4	$5a_4 = m_5 + \delta^{(2)}(5m_4 + 10m_3 m_2) + \delta^{(3)}(10m_3 + 15m_2^2) + \delta^{(4)}(10m_2) + \delta^{(5)}$
5	$6a_5 = m_6 + \delta^{(2)}(6m_5 + 15m_4 m_2 + 10m_3^2) + \delta^{(3)}(15m_4 + 60m_3 m_2 + 15m_2^3)$ $+\delta^{(4)}(20m_3 + 45m_2^2) + \delta^{(5)}(15m_2) + \delta^{(6)}$
6	$7a_6 = m_7 + \delta^{(2)}(7m_6 + 21m_5 m_2 + 35m_4 m_3)$ $+\delta^{(3)}(21m_5 + 105m_4 m_2 + 70m_3^2 + 105m_3 m_2^2)$ $+\delta^{(4)}(35m_4 + 210m_3 m_2 + 105m_2^3)$ $+\delta^{(5)}(35m_3 + 105m_2^2) + \delta^{(6)}(21m_2) + \delta^{(7)}$
7	$8a_7 = m_8 + \delta^{(2)}(8m_7 + 28m_6 m_2 + 56m_5 m_3 + 35m_4^2)$ $+\delta^{(3)}(28m_6 + 168m_5 m_2 + 280m_4 m_3 + 210m_4 m_2^2 + 280m_3^2 m_2)$ $+\delta^{(4)}(56m_5 + 420m_4 m_2 + 280m_3^2 + 840m_3 m_2^2 + 105m_2^4)$ $+\delta^{(5)}(70m_4 + 560m_3 m_2 + 420m_2^3) + \delta^{(6)}(56m_3 + 210m_2^2)$ $+\delta^{(7)}(28m_2) + \delta^{(8)}$

Table 1. The first seven coefficients a_k of $\psi(\eta) = \sum_{k=1}^{\infty} a_k \eta^k / k!$ in (44) in terms of the derivatives $\delta^{(k)} \equiv \delta^{(k)}(1)$ and the moments $m_k = EV^k = \phi^{(k)}(0)$, where $\delta^{(1)} = m_1 = 1$, obtained from the Bell polynomials in Table 3 of Riordan [32, p. 49].

k	$a_1^{2k-1} b_k$
1	$a_1 b_1 = 1$
2	$a_1^3 b_2 = -a_2$
3	$a_1^5 b_3 = 3a_2^2 - a_1 a_3$
4	$a_1^7 b_4 = 10 a_1 a_2 a_3 - a_1^2 a_4 - 15 a_2^3$
5	$a_1^9 b_5 = 15 a_1^2 a_2 a_4 + 10 a_1^2 a_3^2 + 105 a_2^4 - a_1^3 a_5 - 105 a_1 a_2^2 a_3$
6	$a_1^{11} b_6 = 21 a_1^3 a_2 a_5 + 35 a_1^3 a_3 a_4 + 1260 a_1 a_2^3 a_3$ $- a_1^4 a_6 - 280 a_1^2 a_2 a_3^2 - 945 a_2^5 - 210 a_1^2 a_2^2 a_4$
7	$a_1^{13} b_7 = 28 a_1^4 a_2 a_6 + 56 a_1^4 a_3 a_5 + 35 a_1^4 a_4^2 + 3150 a_1^2 a_2^3 a_4$ $+ 6300 a_1^2 a_2^2 a_3^2 + 10395 a_2^6 - a_1^5 a_7 - 378 a_1^3 a_2^2 a_5$ $- 1260 a_1^3 a_2 a_3 a_4 - 280 a_1^3 a_3^3 - 17325 a_1 a_2^4 a_3$

Table 2. The first seven coefficients b_k of $f(\varepsilon) = \sum_{k=1}^{\infty} b_k \varepsilon^k / k!$ in (46) in terms of the coefficients a_k of $\varepsilon = \psi(\eta) = \sum_{k=1}^{\infty} a_k \eta^k / k!$ obtained by reversion of series; see 3.6.25 of Abramowitz and Stegun [4, p. 16] or Riordan [32, Section 2.8].

k	$k!$	c_k
1	1	b_1
2	2	$2b_1 + b_2$
3	6	$6b_1 + 6b_2 + b_3$
4	24	$24b_1 + 36b_2 + 12b_3 + b_4$
5	120	$120b_1 + 240b_2 + 120b_3 + 20b_4 + b_5$
6	720	$720b_1 + 1800b_2 + 1200b_3 + 300b_4 + 30b_5 + b_6$
7	5040	$5040b_1 + 15120b_2 + 12600b_3 + 4200b_4 + 630b_5 + 42b_6 + b_7$

Table 3. The first seven derivatives $c_k \equiv h^{(k)}(0)$ of $h(1-\rho)$ in (47) in terms of derivatives $b_k \equiv f^{(k)}(0)$ where $h(1-\rho) = f(g(1-\rho))$ and $\varepsilon \equiv g(1-\rho) = (1-\rho)/\rho$, based on Riordan [32, Section 2.8].

k	c_k
1	$\dfrac{1}{a_1} = \dfrac{2}{m_2 + \delta^{(2)}}$
2	$\dfrac{2a_1^2 - a_2}{a_1^3} = \dfrac{4(3m_2^2 + 3(\delta^{(2)})^2 - 2m_3 - 2\delta^{(3)})}{3(m_2 + \delta^{(2)})^3}$
3	$\dfrac{6a_1^4 - 6a_2 a_1^2 + 3a_2^2 - a_1 a_3}{a_1^5}$
4	$\dfrac{10a_1 a_2 a_3 - a_1^2 a_4 - 15a_2^3 + 36a_2^2 a_1^2 - 12a_1^3 a_3 - 36a_2 a_1^4 + 24a_1^6}{a_1^7}$

Table 4. The first four coefficients c_k in the heavy-traffic expansion (47) expressed in terms of the coefficients a_k and, for the first two, the basic derivatives $\delta^{(k)}$ and m_k.

k	d_k
1	c_1
2	$c_2 + m_2 c_1^2$
3	$c_3 + m_2(3c_2c_1) + m_3c_1^3$
4	$c_4 + m_2(4c_3c_1 + 3c_2^2) + m_3(10c_2c_1^2) + m_4c_1^4$
5	$c_5 + m_2(5c_4c_1 + 10c_3c_2) + m_3(10c_3c_1^2 + 15c_2^2c_1) + m_4(10c_2c_1^3) + m_1c_1^5$
6	$c_6 + m_2(6c_5c_1 + 15c_4c_2 + 10c_3^2) + m_3(15c_4c_1^2 + 60c_3c_2c_1 + 15c_2^3)$ $+ m_4(20c_3c_1^3 + 45c_2^2c_1^2) + m_5(15c_2c_1^4) + m_6c_1^6$
7	$c_7 + m_2(7c_6c_1 + 21c_5c_2 + 35c_4c_3)$ $+ m_3(21c_5c_1^2 + 105c_4c_2c_1 + 70c_3^2c_1 + 105c_3c_2^2)$ $+ m_4(35c_4c_1^3 + 210c_3c_2c_1^2 + 105c_2^3c_1)$ $+ m_5(35c_3c_1^4 + 105c_2^2c_1^3) + m_6(21c_2c_1^5) + m_7c_1^7$

Table 5. The first seven coefficients d_k in the asymptotic expansion for $\sigma^{-1} - 1$ in (54) in terms of the coefficients c_k of the asymptotic expansion for η.

k	asymptotic factorial cumulant $\delta^{(k)}$
1	1
2	$\nu_2 - 2$
3	$-\nu_3 + 3\nu_2^2 - 6\nu_2 + 6$
4	$\nu_4 - 10\nu_3\nu_2 + 15\nu_2^3 + 12\nu_3$ $- 36\nu_2^2 + 36\nu_2 - 24$

Table 6. The first four asymptotic factorial cumulants $\delta^{(k)}$ of a rate-1 renewal process in terms of the moments ν_k of the inter-renewal time ($\nu_1 = 1$).

r	ρ	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η
1	0.5 0.8	.33333 .13333	.31481 .13037	.31070 .13011	.30956 .13008	.30920 .13007	.30908 .13007	.30904 .13007	.30902 .13007
2	0.5 0.8	.28125 .11250	.26669 .11017	.26550 .11009	.26646 .11019	.26734 .11013	.26779 .11013	.26794 .11013	.26786 .11013
4	0.5 0.8	.20833 .08333	.18916 .08027	.18931 .08028	.19270 .08036	.19516 .08039	.19630 .08039	.19666 .08039	.19677 .08039
8	0.5 0.8	.16598 .06639	.14278 .06268	.14186 .06262	.14447 .06269	.14611 .06271	.14670 .06271	.14687 .06271	.14708 .06271

Table 8. The heavy-traffic asymptotic expansion for the asymptotic decay rate η as a function of the *MMPP* parameters r and ρ in Example 6.1. The other parameter is fixed at $\gamma = 0.5$.

γ	ρ	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η
∞	0.5 0.8	.35714 .14286	.29094 .13226	.32075 .13417	.35609 .13508	.36482 .13517	.36772 .13518	.37367 .13519	.37961 .13519
4	0.5 0.8	.32787 .13115	.27628 .12289	.29414 .12404	.32031 .12471	.32983 .12480	.33188 .12481	.33451 .12482	.33838 .12482
2	0.5 0.8	.30303 .12121	.26143 .11456	.27243 .11526	.29159 .11575	.30030 .11584	.30237 .11585	.30358 .11585	.30594 .11585
1	0.5 0.8	.26315 .10526	.23363 .10054	.23793 .10081	.24832 .10108	.25436 .10114	.25636 .10115	.25688 .10115	.25765 .10115
0.5	0.5 0.8	.20833 .08333	.18916 .08027	.18931 .08028	.19270 .08036	.19516 .08039	.19630 .08039	.19666 .08039	.19677 .08039
0.25	0.5 0.8	.14706 .05882	.13396 .05673	.13259 .05664	.13296 .05665	.13339 .05665	.13365 .05665	.13377 .05665	.13381 .05665

Table 9. The heavy-traffic asymptotic expansion for the asymptotic decay rate η as a function of the *MMPP* parameters γ and ρ in Example 6.1. The other parameter is fixed at $r = 4$.

one stream, $\rho = 0.5$, $\eta = 0.19677$				
percentile required	percentile value			
	exact	approx., exact α $\alpha = 0.50219$	approx., $\alpha \approx$ mean $*\eta$ $\alpha \approx 0.53255$	approx. $\alpha \approx 1.0$
80	4.8558	4.6789	4.9773	8.1794
90	8.2781	8.2016	8.4999	11.7021
99	19.9087	19.9036	20.2020	23.4041
99.9	31.6060	31.6057	31.9041	35.1062
99.99	43.3078	43.3078	43.6061	46.8082
two streams				
percentile required	percentile value			
	exact	approx., exact α $\alpha = 0.37188$	approx., $\alpha \approx$ mean $*\eta$ $\alpha \approx 0.47361$	approx. $\alpha \approx 1.0$
80	4.2672	3.1523	4.3811	8.1794
90	7.4446	6.6749	7.9038	11.7021
99	18.3770	18.3770	19.6059	23.4041
99.9	30.0791	30.0791	31.3079	35.1062
99.99	41.8015	41.7811	43.0100	46.8082

Table 10. A comparison of approximations with exact values of percentiles of the steady-state waiting time in the $MMPP/\Gamma_{1/2}/1$ queue with $\rho = 0.5$ and one or two arrival streams in Example 6.2.

one stream, $\rho = 0.8$, $\eta = 0.08039$				
	percentile value			
percentile required	exact	approx., exact α $\alpha = 0.80368$	approx., $\alpha \approx$ mean $*\eta$ $\alpha \approx 0.80957$	approx. $\alpha \approx 1.0$
80	17.3032	17.3012	17.3920	20.0197
90	25.9233	25.9232	26.0140	28.6417
99	54.5649	54.5649	54.6557	57.2834
99.9	83.2065	83.2065	83.2973	85.9250
99.99	111.8482	111.8482	111.9391	114.5668
two streams				
	percentile value			
percentile required	exact	approx., exact α $\alpha = 0.76992$	approx., $\alpha \approx$ mean $*\eta$ $\alpha \approx 0.78538$	approx. $\alpha \approx 1.0$
80	16.8030	16.7673	17.0146	20.0197
90	25.3978	25.3893	25.6366	28.6417
99	54.0311	54.0310	54.2783	57.2834
99.9	82.6727	82.6727	82.9200	85.9251
99.99	111.3144	111.3144	111.5617	114.5667

Table 11. A comparison of approximations with exact values of percentiles of the steady-state waiting time in the $MMPP/\Gamma_{1/2}/1$ queue with $\rho = 0.8$ and one or two arrival streams in Example 6.2.

traffic intensity	η_1	η_2	η_3	η_4	exact η
0.6	.32000	.34560	.35174	.35342	.35414
0.7	.24000	.25440	.25699	.25752	.25767
0.8	.16000	.16650	.16717	.16727	.16729
0.9	.08000	.08160	.08170	.08170	.08170

Table 12. A comparison with approximations for the asymptotic decay rate η with exact values for the $\Gamma_{1/2}/\Gamma_2/1$ queue in Example 6.3.