EXPONENTIAL APPROXIMATIONS FOR TAIL PROBABILITIES IN QUEUES, I: WAITING TIMES

JOSEPH ABATE

Ridgewood, New Jersey

GAGAN L. CHOUDHURY

AT&T Bell Laboratories, Holmdel, New Jersey

WARD WHITT

AT&T Bell Laboratories, Murray Hill, New Jersey

(Received August 1992; revision received June 1993; accepted July 1993)

This paper focuses on simple exponential approximations for tail probabilities of the steady-state waiting time in infinite-capacity multiserver queues based on small-tail asymptotics. For the *GI/GI/s* model, we develop a heavy-traffic asymptotic expansion in powers of one minus the traffic intensity for the waiting-time asymptotic decay rate. We propose a two-term approximation for the asymptotic decay rate based on the first three moments of the interarrival-time and service-time distributions. We also suggest approximating the asymptotic constant by the product of the mean and the asymptotic decay rate. We evaluate the exponential approximations based on the exact asymptotic parameters and their approximations by making comparisons with exact results obtained numerically for the *BMAP/GI/1* queue, which has a batch Markovian arrival process, and the *GI/GI/s* queue. Numerical examples show that the exponential approximations are remarkably accurate, especially for higher percentiles, such as the 90th percentile and beyond.

In queueing applications we are often interested in tail probabilities of the steady-state waiting time (before beginning service). Our main purpose in this paper is to support and elaborate on the idea that these tail probabilities often have approximately an exponential form with parameters that can be determined. Abate, Choudhury and Whitt (1994a, b) discuss corresponding exponential approximations for the sojourn time (response time, i.e., waiting time plus service time) and the workload (virtual waiting time), and geometric approximations for the queue lengths (at arbitrary times, departure epochs, and arrival epochs).

Suppose that W is the steady-state waiting time. We are suggesting the simple exponential approximation

$$P(W > x) \approx \alpha e^{-\eta x} \qquad (1)$$

for suitably large x, where the *decay rate* η and the *constant* α are fixed positive real numbers independent of x. It is important to note that we do *not* suggest that (1) should hold for *all* x, but only for suitably large x.

If we are interested in the $(100p)$th percentile (or quantile, e.g., for $p = 0.95$), then we want w_p such that $P(W > w_p) = 1 - p$. Corresponding to (1), we suggest the approximation

$$w_p \approx \log\left(\frac{\alpha}{1-p}\right)\frac{1}{\eta}. \qquad (2)$$

For applications, it is important to note that the relative error in an approximation for a high percentile is typically substantially less than the relative error for the corresponding tail probability itself. (We elaborate on this point in Section 9.)

Moreover, the percentile (2) often does not depend greatly on the constant α, so that we can often use a relatively crude approximation for α; sometimes we can even set $\alpha = 1$ in (2). To see this, first note that $\log(\alpha/(1 - p)) = \log \alpha - \log(1 - p)$ and, second, note that $\log \alpha$ may be relatively small compared to $-\log(1 - p)$. This simplification in (2) obviously becomes more appropriate as p and α approach 1. Having the approximation not depend critically on α is helpful because an appropriate constant α is often much more difficult to determine than an appropriate decay rate η. (It is important to note, however, that an appropriate constant α in (1) is not always sufficiently close to 1; this difficulty occurs in models with superposition arrival processes; see Choudhury, Lucantoni and Whitt (1993a).)

The remarkable quality of exponential approximations for waiting-time tail probabilities in the *GI/GI/s* queue is pointed out with abundant numerical evidence in section 1.9 and Chapter 4 of Tijms (1986) and in Seelen, Tijms and van Hoorn (1985). Indeed, values of the parameters η and α in (1) are given in all the tables there (for queues

Subject classifications: Queues, approximations: exponential approximations for tail probabilities. Queues, limit theorems: small tail asymptotics and heavy traffic asymptotic expansions.
Area of review: STOCHASTIC PROCESSES AND THEIR APPLICATIONS

0030-364X/95/4305-0885 $01.25
© 1995 INFORMS

with infinite capacity). There the parameters η and α are defined precisely via the limit

$$\lim_{x\to\infty} e^{\eta x}P(W>x)=\alpha. \tag{3}$$

We also regard the small-tail asymptotics in (3) as the primary justification for (1) and (2). Given (3), we regard η as the *asymptotic decay rate* and α as the *asymptotic constant*.

There is now a substantial literature that establishes small-tail exponential asymptotics as in (3) for steady-state queueing distributions. An important early reference for the *GI/GI/*1 queue is Smith (1953). Based on early work on related risk models, the approximation based on (3) is sometimes called the *Cramér–Lundberg approximation*; see p. 131 of Seal (1969), section 12.5 of Feller (1971), section 12.5 of Asmussen (1987), section 6 of Asmussen (1989), Asmussen and Rolski (1991), and the references cited there. (Our work also has applications to risk theory.) For other contributions to small-tail asymptotics, see section 22 of Borovkov (1976), Neuts (1981, 1986), Neuts and Takahashi (1981), van Ommeren (1988, 1989), Elwalid, Mitra and Stern (1991), Asmussen and Perry (1992), Baiocchi (1992), Elwalid and Mitra (1993, 1995), Chang (1994), and Glynn and Whitt (1994). For nonexponential asymptotics with long tails (involving long-tail service times), see Abate, Choudhury and Whitt (1994b) and Willekens and Teugels (1992); for nonexponential asymptotics with subexponential tails (involving strong dependence between interarrival times and service times), see Jacquet (1992).

In this paper, we do not prove any limits like (3); here we are concerned with the applied relevance of (3). In Abate, Choudhury and Whitt (1994a) we contribute to the exponential asymptotic theory by establishing small-tail exponential asymptotics for steady-state distributions in models of the *M/G/*1 paradigm in Neuts (1989) and the *BMAP/GI/*1 queue, with batch Markovian arrival process, as in Lucantoni (1991).

Here we consider the standard *GI/GI/s* queue with unlimited waiting room, the first-come, first-served service discipline and i.i.d. (independent and identically distributed) service times that are independent of a renewal arrival process. (We also discuss models with nonrenewal arrival processes.) For the *GI/GI/s* model, let U/ρ be a generic interarrival time with mean $1/s\rho$ and let V be a generic service time with mean 1, each with general distributions. Thus, the traffic intensity is ρ. To have proper steady-state distributions, we assume that $\rho<1$. To have (3), the key condition is for there to be a root x of the transform equation

$$Ee^{xV/s}Ee^{-xU/\rho}=1. \tag{4}$$

When a root to (4) exists, it turns out to be unique and, under additional regularity conditions (e.g., see Theorem 2), (3) holds with the root being the asymptotic decay rate η. From (4), it follows that $\eta\equiv\eta(s)$ depends on the numbers of servers in a very simple way, in particular,

$$\eta(s)=s\eta(1). \tag{5}$$

For $s>1$, this conclusion has been established for phase-type service distributions by Takahashi (1981) and Neuts and Takahashi (1981); the case of general service-time distributions remains a conjecture. Equations similar to (4) hold when the independence among the interarrival times or among the service times is relaxed; e.g., see Neuts (1986), Asmussen (1989), Whitt (1993), Abate, Choudhury, Whitt (1994a), and Glynn and Whitt (1994). In many cases, expressions for the asymptotic constant α in (3) can be found, but α is substantially more complicated than η.

In this paper, we primarily do two things: First, we demonstrate through numerical examples that the exponential approximation (1) provided by the small-tail asymptotics in (3) is indeed remarkably good. Second, we show how to obtain convenient approximations for the asymptotic parameters η and α in (3) that still yield good exponential approximations.

For our numerical examples, we draw on *GI/GI/s* tables, such as Seelen, Tijms and van Hoorn, and we compute exact tail probabilities for the *BMAP/GI/*1 queue, exploiting algorithms in Choudhury, Lucantoni and Whitt (1995c), which are based on Lucantoni (1991) and Abate and Whitt (1992). We compute the exact asymptotic parameters η and α from transforms by applying the algorithm in Choudhury and Lucantoni (1995).

Our approximations for the asymptotic parameters η and α are based on two simple ideas. First, to approximate η we do a Taylor series expansion of the exponentials in (4) about $x=0$, which (after some analysis) produces an asymptotic expansion for η in powers of $1-\rho$, where ρ is the traffic intensity. Extensions of this idea to more general models are contained in Choudhury and Whitt (1994) and Abate and Whitt (1994). It is known that the second term in the heavy-traffic expansion for the steady-state mean EW in the *GI/GI/*1 model is relatively complicated; see Siegmund (1979), section 12.6 of Asmussen (1987), Whitt (1989), and Knessl (1990). It is significant that the asymptotic decay rate η in the *GI/GI/s* model admits a full asymptotic expansion directly in terms of the lower moments.

Having found the asymptotic decay rate η or a suitable approximation, we obtain an approximation for the asymptotic constant α by acting as if (1) were true and considering the mean. This yields the simple approximation

$$\alpha\approx\eta EW. \tag{6}$$

We also provide theoretical support for (6) by proving that

$$\alpha=\eta EW+O((1-\rho)^2)\quad\text{as } \rho\to 1 \tag{7}$$

in the *GI/GI/*1 queue.

Here is how the rest of this paper is organized. In Section 1 we review two familiar theoretical reference points supporting (1), namely, the $GI/M/s$ queue and heavy-traffic limit theorems. These support (1), but our intent is to go beyond what they suggest. In Section 2 we discuss numerical examples, which dramatically support the power of the small-tail asymptotics in (3) and the approximation for η and α.

In Section 3 we review the small-tail asymptotics for the steady-state waiting time in the $GI/GI/1$ queue. We try to explain *why* the waiting-time distribution should have an exponential tail and *why* its asymptotic decay rate takes the form it does. In Section 3 we also show that the *small-tail asymptotics in* (3) *does not always hold*, even in an $M/GI/1$ queue in which the service time has a finite moment generating function in a neighborhood of the origin. Moreover, when the pure-exponential asymptotics does not prevail, the actual asymptotics produces a remarkably poor approximation.

In Section 4 we develop a heavy-traffic expansion for the asymptotic decay rate η in (3) in the $GI/GI/s$ model as a function of $1 - \rho$. In Section 5 we develop simple two-moment approximations for the asymptotic decay rate, by using two-moment approximations for third moments, as in Whitt (1983) and Tijms (1986).

In Section 6 we develop the approximation (6) for the asymptotic constant α in (3). Our starting point is an exact analysis for the $M/G/1$ queue, where we find that (7) holds. We then show that (7) is also valid for all $GI/GI/1$ queues.

In Section 7 we show that for any $GI/GI/s$ queue there exists a threshold traffic intensity ρ^* below which the asymptotics (3) is not valid, and above which it is. (Usually $\rho^* = 0$, but not always.) In Section 8 we establish several stochastic comparisons that provide additional insight, and are useful for considering the sojourn time and workload in Part II. In Section 9 we discuss the relative errors associated with approximations (1) and (2). Finally, in Section 10 we draw some conclusions.

1. TWO THEORETICAL REFERENCE POINTS

In this section we briefly review two familiar reference points supporting (1). First, (1) is well known to be exact for the $GI/M/s$ queue with $\alpha = P(W > 0)$ and $\eta = s(1 - \sigma)$, where σ is the unique root in the interval (0, 1) of

$$Ee^{-s(1-\sigma)U/\rho} = \sigma, \tag{8}$$

and U/ρ is a generic interarrival time with mean $1/s\rho$, so that the traffic intensity is ρ; e.g., see Chapter 6 of Kleinrock (1975). Note that this is consistent with (5). For $s = 1$, $P(W > 0) = \sigma$; for $s > 1$, $P(W > 0)$ can be calculated numerically. As a consequence of this $GI/M/s$ result, we anticipate that the exponential approximations in (1) and (2) will perform well when a $G/GI/s$ model is suitably close to a $GI/M/s$ model.

Second, the exponential approximation is also strongly supported by heavy-traffic limit theorems. For the $G/GI/s$ model, which has i.i.d. service times independent of a general stationary arrival process, this involves a family of models indexed by ρ where $\rho \to 1$. We assume that we are given an arrival counting process $A \equiv \{A(t): t \geq 0\}$ with arrival rate 1; i.e., $A(t)/t \to 1$ as $t \to \infty$. We then introduce model ρ by scaling time in the arrival process by ρs; i.e., in model ρ the service times are unchanged and the arrival counting process is defined by $A_\rho(t) = A(\rho s t)$. (We scale by ρs to make the arrival rate ρs; since $EV = 1$, the overall service rate is s.)

To obtain a valid heavy-traffic limit theorem as $\rho \to 1$, we assume that $v_2 \equiv E[V^2] < \infty$, where V is a generic service time, so that the service times obey a functional central limit theorem (FCLT), and that the unscaled arrival process satisfies an FCLT, i.e.,

$$(nc_A^2)^{-1/2}[A(nt) - nt] \Rightarrow B(t) \quad \text{as } n \to \infty, \tag{9}$$

where $\Rightarrow$ denotes convergence in distribution (in the function space $D[0, \infty)$) and $B \equiv \{B(t): t \geq 0\}$ is standard (drift 0 and variance 1) Brownian motion. In the case of a renewal arrival process, the parameter c_A^2 in (9) is just the interarrival-time squared coefficient of variation (SCV, variance divided by the square of the mean), which we denote by c_a^2, but more generally, c_A^2 also reflects the dependence among the interarrival times. In great generality, c_A^2 is the *asymptotic variance constant*, i.e.,

$$c_A^2 = \lim_{t \to \infty} \frac{\text{Var}\, A(t)}{EA(t)}. \tag{10}$$

As a consequence, as $\rho \to 1$, the waiting times suitably normalized converge to reflected Brownian motion with a negative drift, which has an exponential steady-state distribution. We thus obtain (1) as a heavy-traffic approximation with $\alpha = 1$ and

$$\eta = \frac{2(1 - \rho)}{c_A^2 + c_s^2}, \tag{11}$$

where c_s^2 is the SCV of the service-time distribution; see Whitt (1989, 1992) and the references cited there. Fleming (1992) has observed that (1) with (11), sometimes with a heuristic refinement, performs remarkably well for percentiles in a class of $M/G/1$ queues. It turns out that (11) is the first term in the asymptotic expansion of η in powers of $(1 - \rho)$; see Section 4 and Choudhury and Whitt.

Both the $GI/M/s$ and heavy-traffic reference points are consistent with the idea that (1) might be good for *all* x. Then the obvious parameters are

$$\alpha = P(W > 0) \quad \text{and} \quad \eta^{-1} = E[W|W > 0]. \tag{12}$$

A nice treatment of full exponential approximations was carried out by Fredericks (1982). Simple approximations based on (1), (11), and (12) are also discussed in Whitt (1992). However, the approximations based on (1), (11), and (12) are relatively crude. By exploiting (3) and further asymptotic analysis, we obtain significant improvements over (1), (11), and (12).

2. NUMERICAL *G*/*GI*/1 EXAMPLES

In this section, we present examples to show that the simple exponential approximation provided by the small-tail asymptotics in (3) is often a remarkably good approximation, even when the service-time distribution is not phase-type and not nearly exponential, and even when the interarrival times are not independent. Moreover, the approximation based on (3) is significantly better than cruder "pure-exponential" approximations based on (11) or (12). Finally, simple approximations of the parameters α and η in (3), (6), and Section 4 perform well when the traffic intensity is not too small.

As indicated in the Introduction, we obtain the exact values of the tail probabilities from *BMAP*/*GI*/1 algorithms in Choudhury, Lucantoni and Whitt (1995c), which are based on Lucantoni and Abate and Whitt (1992). (An alternative numerical approach to *GI*/*G*/1 is in Abate, Choudhury and Whitt (1993).) We calculated the asymptotic parameters α and η in (3) using the moment-based numerical inversion algorithm of Choudhury and Lucantoni. In each case, we also tested for the exponential form (1) by looking for linearity in log $P(W > x)$. We estimated α and η at each x by performing a linear regression based on the numerical values of log $P(W > x)$ at points $x \pm j\delta$ for $0 \leqslant j \leqslant k$, for appropriate k and δ (e.g., $k = 5$ and $\delta = 2$). Convergence of the estimated parameters $\hat{\alpha}(x)$ and $\hat{\eta}(x)$ as x increases demonstrates (3).

The examples to be discussed are only a few of the examples considered. It is significant that the exact values are readily available with our algorithm. On the standard shared-computer-system environment in the AT&T Bell Laboratories Mathematical Sciences Research Center, the algorithm (when the number of MAP phases is 2) required less than 5 seconds to treat 9 values of the traffic intensity and 40 values of x. (For example, with $k = 5$, this means $9 \times 40 \times 5 = 7{,}200$ Laplace transform inversions.) The algorithm also can calculate related steady-state distributions, such as sojourn-time and workload distributions as well as waiting-time distributions. (We discuss these performance measures in Part II.)

Example 1. We first consider an $H_2^b/\Gamma_{1/2}/1$ queue, which has a hyperexponential interarrival-time distribution and a gamma service-time distribution with mean 1 and shape parameter 1/2. In particular, the $\Gamma_{1/2}$ service-time distribution has density

$$g(x) = (2\pi x)^{-1/2} e^{-x/2}, \quad x \geqslant 0,$$

and first three moments 1.0, 3.0, and 15.0. The Laplace transform of this density is

$$\hat{g}(s) \equiv \int_0^\infty e^{-sx} g(x)dx = 1/\sqrt{1 + 2s},$$

which is *not* rational. (Thus, the distribution is *not* phase-type.)

In Section 9.2 of Abate and Whitt (1992) it was shown that the small-tail asymptotic approximation (3) performs remarkably well for the $M/\Gamma_{1/2}/1$ queue. Now we consider what happens with a non-Poisson arrival process. The H_2^b arrival process used here is a renewal process with interarrival times having a two-phase hyperexponential distribution with balanced means; i.e., the interarrival-time density is of the form

$$f(x) = p_1 \lambda_1 e^{-\lambda_1 x} + p_2 \lambda_2 e^{-\lambda_2 x}, \quad x \geqslant 0,$$

where $p_1/\lambda_1 = p_2/\lambda_2$. As for the SCV, we used $c_a^2 = 2.0$. In particular, when the arrival rate (traffic intensity) ρ is 0.7, the overall mean is 10/7, and the parameters are $p_1 = 0.7885$, $\lambda_1 = 1.1039$, and $\lambda_2 = 0.2958$.

Table I compares the approximation for the tail probabilities based on (3) with exact values for the $H_2/\Gamma_{1/2}/1$ queue when $\rho = 0.7$. Also included in the table are the linear-regression estimates $\hat{\alpha}(x)$ and $\hat{\eta}(x)$ for each x, based on parameters $\delta = 2$ and $k = 2$. The rapid convergence of $\hat{\alpha}(x)$ and $\hat{\eta}(x)$ as x increases is evident. The quality of the approximation for $P(W > x)$ is also excellent for x suitably large. Indeed, there clearly is excellent accuracy by the 80th percentile, and there is spectacular accuracy when $P(W > x) \approx 10^{-4}$.

Table II compares approximations for the asymptotic parameters α and $1/\eta$ in (3). The approximations in (6) and Theorem 4 are denoted α_{ap} and $1/\eta_{ap}$. The approximations for $1/\eta$ based on (11) and (12) are $1/\eta_{HT}$ and $E[W|W > 0]$, respectively. The new approximations do dramatically better. For example, at $\rho = 0.7$ the percent relative errors of $1/\eta_{ap}$, $1/\eta_{HT}$, and $E[W|W > 0]$ as approximations of $1/\eta$ are, respectively, 0.3%, 4.2%, and

Table I

A Comparison of the Exponential Approximation for the Waiting-Time Tail Probabilities With Exact Values: The $H_2^b/\Gamma_{1/2}/1$ Queue With $\rho = 0.7$ and $c_a^2 = c_s^2 = 2.0$ in Example 1

Time	Exact	Approximation	$\hat{\alpha}(x)$	$1/\hat{\eta}(x)$
1.5	0.5855	0.5803	0.747	6.15
3.0	0.4607	0.4591	0.740	6.32
4.5	0.3638	0.3632	0.737	6.37
6.0	0.2876	0.2874	0.7354	6.390
7.5	0.22745	0.22738	0.7346	6.398
9.0	0.17993	0.17990	0.7341	6.401
10.5	0.14234	0.14233	0.73380	6.4024
12.0	0.11261	0.11261	0.73365	6.4031
13.5	0.08909	0.08909	0.73356	6.4035
15.0	0.07049	0.07049	0.73348	6.40386
18.0	0.04412	0.04412	0.733450	6.403925
36.0	0.00265	0.00265	0.73346147	6.40389005
48.0	0.000408	0.000408	0.73346152	6.40388999

Table II
A Comparison of Approximations for the Parameters α and $1/\eta$ in (3) With Estimated Exact Values for the $H_2^b/\Gamma_{1/2}/1$ Queue in Example 1

ρ	α	α_{ap}	$P(W > 0)$	$1/\eta$	$1/\eta_{ap}$	$1/\eta_{HT}$	$E[W\mid W > 0]$
0.9	0.9156	0.9112	0.925	19.743	19.751	20.00	19.55
0.8	0.8269	0.8200	0.846	9.739	9.755	10.00	9.54
0.7	0.7335	0.7262	0.762	6.404	6.425	6.67	6.20
0.6	0.6346	0.6300	0.673	4.738	4.762	5.00	4.52
0.5	0.5299	0.531	0.578	3.744	3.76	4.00	3.51
0.4	0.4193	0.429	0.477	3.089	3.10	3.33	2.84
0.3	0.3033	0.326	0.368	2.633	2.63	2.86	2.36
0.2	0.1844	0.220	0.253	2.312	2.27	2.50	1.99

3.1%. Typically, there is an order of magnitude improvement in going from $1/\eta_{HT}$ or $E[W|W > 0]$ to $1/\eta_{ap}$.

Example 2. It can be argued that the $\Gamma_{1/2}$ service-time density in Example 1 is reasonably close to an exponential service-time density. To make a more challenging comparison, we change the service-time distribution to a two-point distribution (and keep the H_2^b arrival process unchanged). In particular, this D_2 service-time distribution assumes the values 11.0 and 0.8 with the probabilities 0.019607843 and 0.980392157, respectively. This service-time distribution was chosen to be not phase-type and not nearly exponential. This distribution has mean 1.0 and SCV $c_S^2 = 2.0$, just as for $\Gamma_{1/2}$ in Example 1. This service-time distribution is an extremal distribution among all probability distributions on the interval [0, 11] with first two moments 1 and 3; see p. 120 of Whitt (1984a); e.g., it has the largest third moment among all these distributions.

Table III, the analog of Table I for Example 1, compares approximations for the tail probabilities based on (3) with exact values for the $H_2^b/D_2/1$ queue when ρ = 0.7. Since the arrival process and traffic intensity are the same as for Example 1, Tables I and III are directly comparable.

Table III shows that the accuracy of the exponential approximation is still spectacular for large x and excellent for moderately large x, but it does not become good quite as quickly as in Table I. Moreover, we see that the estimated parameters $\hat{\alpha}(x)$ and $1/\hat{\eta}(x)$ converge much more slowly as x increases for the $H_2^b/D_2/1$ queue than they do for the $H_2^b/\Gamma_{1/2}/1$ queue in Table I.

Paralleling Table II, Table IV compares approximations for the asymptotic parameters α and $1/\eta$ in (3). The quality of all the approximations is not as good in Table IV as it is in Table II, but the conclusions drawn about Example 1 are still valid for this D_2 service-time distribution. Indeed, the advantage of the new approximations α_{ap} and $1/\eta_{ap}$ over the others is even stronger here. Since the first two moments of the D_2 service-time distribution are the same as for the $\Gamma_{1/2}$ service-time distribution in Example 1, the approximation $1/\eta_{HT}$ is the same in Tables II and IV. These tables show that a refinement of the heavy-traffic approximation (11) can help significantly.

Example 3. To show that the quality of the exponential approximations does not depend critically on having a renewal arrival process, we change the arrival process in Example 2 to a nonrenewal Markov modulated Poisson process (*MMPP*). We let the *MMPP* have a two-state underlying continuous-time Markov chain. We let the mean holding time be 10.0 in each state. In one state the arrival rate is 1.1 and in the other it is 0.3. Thus, the overall arrival rate is ρ = 0.7, but the instantaneous arrival rate exceeds the long-run service rate 1 in one of the two states. To give a quick idea about this arrival process, the steady-state means EW in these $M/D_2/1$, $H_2^b/D_2/1$ and $MMPP_2/D_2/1$ models with ρ = 0.7 are, respectively, 3.50, 4.58, and 5.83.

The quality of the exponential approximation for this $MMPP_2/D_2/1$ queue is essentially the same as for

Table III
A Comparison of the Exponential Approximation for the Waiting Time Probabilities With Exact Values for The $H_2^b/D_2/1$ Queue With ρ = 0.7 and c_a^2 = c_s^2 = 2.0 in Example 2

Time	Exact	Approximation	$\hat{\alpha}(x)$	$1/\hat{\eta}(x)$
3.0	0.3582	0.4267	0.517	11.99
6.0	0.2740	0.2836	0.448	12.19
9.0	0.2045	0.1885	0.614	8.18
12.0	0.1223	0.1253	1.053	5.59
15.0	0.08034	0.0833	0.993	8.27
18.0	0.05625	0.05534	0.536	7.98
21.0	0.03731	0.03678	0.834	6.76
24.0	0.02410	0.02445	0.649	7.29
27.0	0.01623	0.01625	0.548	7.67
30.0	0.01088	0.01080	0.666	7.29
36.0	0.004758	0.004771	0.604	7.43
42.0	0.002110	0.002108	0.670	7.291
48.0	0.0009313	0.0009311	0.627	7.370
51.0	0.0006193	0.0006193	0.647	7.336
54.0	0.0004113	0.0004113	0.648	7.334
57.0	0.0002733	0.0002733	0.637	7.352
60.0	0.0001818	0.0001817	0.641	7.346
∞			0.6420	7.344

Table IV

A Comparison of Approximations for the Parameters α and $1/\eta$ in (3) With Estimated Exact Values for the $H_2^b/D_2/1$ Queue in Example 2

ρ	α	α_{ap}	$P(W > 0)$	$1/\eta$	$1/\eta_{ap}$	$1/\eta_{HT}$	$E[W\mid W > 0]$
0.9	0.8651	0.868	0.936	20.722	20.74	20.00	19.14
0.8	0.7450	0.743	0.863	10.709	10.77	10.00	9.16
0.7	0.6420	0.625	0.782	7.344	7.46	6.67	5.85
0.6	0.5548	0.514	0.692	5.627	5.84	5.00	4.23
0.5	0.4805	0.410	0.593	4.558	4.87	4.00	3.28
0.4	0.4158	0.314	0.487	3.801	4.14	3.33	2.66

$H_2^b/D_2/1$ queue in Example 2. Table V displays the exact values and the approximations. The approximation would be more than adequate for most engineering applications for the 90th percentile and beyond.

This paper does not consider approximations for η for nonrenewal arrival processes. (This is done in Choudhury and Whitt.) Thus, we present no analog of Tables II and IV for the $MMPP_2/D_2/1$ model. The performance of the approximations $1/\eta_{HT}$ and $E[W|W > 0]$ are about the same as for the $H_2^b/D_2/1$ model in Table IV.

Example 4. We now consider a $GI/GI/s$ queue with $s > 1$. From the tables in Seelen, Tijms and van Hoorn, it is easy to verify that, first, the exponential approximation based on (3) is remarkably accurate for $GI/GI/s$ queues for suitably large x and, second, that the approximations η_{ap}, α_{ap} and $\eta_{ap}EW_{ap}$ are good as well. The approximations for percentiles based on (2) are especially good; see Seelen and Tijms (1985) for related work. As s increases, $\alpha(s)$ often decreases, so that the approximation for α in the percentile approximation (2) often becomes more important as s increases.

Table V

A Comparison of the Exponential Approximation for the Waiting-Time Tail Probabilities With Exact Values for the $MMPP_2/D_2/1$ Queue With $\rho = 0.7$ in Example 3 (the asymptotic parameters are $\eta^{-1} = 8.9608$ and $\alpha = 0.6574$)

t	Exact	Approximation	Percent Error
3.0	0.4206	0.4704	11.9
6.0	0.3239	0.3365	3.9
9.0	0.2554	0.2408	−5.7
12.0	0.1737	0.1722	−0.9
15.0	0.1195	0.1232	3.1
18.0	0.08836	0.08819	−0.2
21.0	0.06414	0.06310	−1.8
24.0	0.04491	0.04515	0.5
27.0	0.03209	0.03230	0.7
30.0	0.02320	0.02311	−0.4
36.0	0.01180	0.01183	0.2
42.0	0.006065	0.006057	−0.1
48.0	0.003098	0.003101	0.1
54.0	0.001588	0.001587	0.0
60.0	0.000812	0.000813	0.0

To illustrate, Table VI displays approximations and exact values from Seelen, Tijms and van Hoorn of the 90th percentile of the conditional waiting-time distribution given that a customer is delayed, i.e.,

$$w_{90} = \inf\{x : P(W > x)/P(W > 0) < 0.10\},$$

for several values of s in the $H_2^b/E_2/s$ model with $c_a^2 = 2.0$ and $\rho = 0.8$. Since we focus on the conditional waiting time, as Seelen, Tijms and van Hoorn do, $\alpha(s)$ actually increases with s. To highlight the regularity, we display $s(w_{60}(s))$ in each case in Table VI; $s(w_p(s))$ is a nearly constant function of s.

For these examples, $\eta(s)/s = 0.17983$, consistent with (5). The one-term and two-term heavy-traffic approximations for $\eta(s)/s$ are $\eta_{HT}(s)/s = 0.1600$ and $\eta_{ap}(s)/s = 0.17493$. The relative errors for η_{HT} and η_{ap} are 11% and 2.7%. Approximation (2) using η and α agrees with the exact values in the precision given in the tables. (The relative error are less than 1%.) Using the approximation $\alpha \approx 1$ produces $sw_{90}(s) \approx 12.8$ for all s, which is not too bad. Approximation (2) with α_{ap} and η_{ap} has less than the 2.7% relative error of η_{ap} because the errors in α_{ap} and η_{ap} cancel each other to some extent.

In this case, we have used the exact conditional means $E(W|W > 0)$. Thus, it appears that our ability to estimate overall percentiles primarily depends on our ability to estimate the mean and the probability of delay, which have been studied, e.g., see Whitt (1992, 1993).

3. SMALL-TAIL ASYMPTOTICS FOR THE WAITING TIME IN THE *GI/GI/1* QUEUE

In this section, we review the small-tail asymptotics for the steady-state waiting time W in the $GI/GI/1$ queue. We assume that $\rho < 1$, so that the model is stable. For the waiting-time distribution to have the exponential tail in (3), it is necessary for the service-time moment generating function $Ee^{\gamma V}$ to be finite for some positive γ (see Proposition 1 and Theorem 8), and we assume that this is the case.

The first question is: *Why should we expect that the waiting-time distribution tail should be approximately exponential?* A partial explanation comes directly from the ladder variable representation of the random walk $\{S_n : n \geqslant 0\}$ with $S_0 = 0$ and steps distributed as

Table VI

A Comparison of Approximations for s Times the 90th Percentile of the Conditional Waiting Time Given That a Customer is Delayed With Exact Values in the $H_2^b/E_2/s$ Queue With $c_a^2 = 2.0$ and $\rho = 0.8$ for Several Values of s in Example 4 (the asymptotic decay rates are $\eta(s) = 0.17983\ s$; the exact values are from Seelen, Tijms and van Hoorn)

	Asymptotic Constant for Delayed Customers		$s \times$ 90th Percentile of Waiting Time (Delayed Customers)		
Servers s	$\dfrac{\alpha}{P(W > 0)}$	$\dfrac{\eta EW}{P(W > 0)}$	Exact	Approximation (2)	Approximation (2) with η_{ap} and α_{ap}
1	1.019	1.018	12.9	12.91	13.2
2	1.029	1.026	13.0	12.96	13.3
4	1.042	1.037	13.0	13.04	13.3
10	1.075	1.057	13.2	13.21	13.4
20	1.119	1.079	13.4	13.43	13.6
40	1.197	1.106	13.8	13.80	13.7

$X = V - \rho - 1 - U$, as discussed in Chapter 12 of Feller (1971) and Chapter 7 of Asmussen (1987). Let τ_+ and $S_{\tau+}$ be generic (strict) *ascending ladder epochs and heights* respectively, i.e.,

$$\tau_+ = \inf\{n \geqslant 1 : S_n > 0\}$$

with $S_{\tau+}$ defined on $\{\tau_+ < \infty\}$. Let $\{Z_n^+ : n \geqslant 0\}$ be an i.i.d. sequence distributed as the *conditional ladder height* given that the ladder epoch is finite, i.e.,

$$P(Z_n^+ \leqslant x) = P(S_{\tau+} \leqslant x | \tau_+ < \infty). \qquad (13)$$

Let $M = \max\{S_n : n \geqslant 0\}$. Then the *fundamental ladder variable representation for W* is

$$W \overset{d}{=} M = \sum_{k=1}^{N} Z_k^+, \qquad (14)$$

where $\overset{d}{=}$ denotes equality in distribution, $M = 0$ when $N = 0$, N is independent of $\{Z_k^+ : k \geqslant 1\}$ and N is geometrically distributed, i.e., $P(N = k) = (1 - p)p^k$, $k \geqslant 0$, where $p = P(\tau_+ < \infty)$. (It is easy to see that (13) is equivalent to (1.5) on p. 183 of Asmussen 1987.) From (13), we easily obtain an expression for the mean, namely,

$$EW = E(N)E(Z^+) = \frac{pEZ^+}{1 - p}$$
$$= \frac{P(\tau_+ < \infty)E[S_{\tau+} | \tau_+ < \infty]}{P(\tau_+ = \infty)} = \frac{E[S_{\tau+};\ \tau_+ < \infty]}{P(\tau_+ = \infty)}. \qquad (15)$$

The difficulty with (13) and (14) is that in general it is hard to say much about the distribution of $(\tau_+, S_{\tau+})$. However, (13) strongly supports (3) when N is large, because a large geometric sum of i.i.d. random variables is approximately exponentially distributed; see p. 1 of Feller and p. 132 of Keilson (1979).

Hence, we expect an exponential approximation for W to be good when p is suitably close to 1. However, in general p need not be especially close to 1. The small-tail asymptotics when p is not close to 1 involves large-deviation concepts. To partially understand the small-tail asymptotics, suppose that Z_k^+ is constant in (13), say z, then

$$P(W > x) = P(N > \lfloor x/z \rfloor) = p^{\lfloor x/z \rfloor} \approx e^{x(\log p)/z}.$$

However, a proper understanding seems to require the exponential-change-of-measure arguments; see pp. 406 and 411 of Feller and Chapter 12 of Asmussen (1987).

Next, suppose that we believe that the distribution of W has an asymptotically exponential tail as in (3). The second question is: *What should the asymptotic decay rate η be*? An explanation comes from a direct application of the *Lindley equation*

$$W \overset{d}{=} (W + X)^+, \qquad (16)$$

where $(x)^+ = \max\{x, 0\}$.

Proposition 1. *In the stable GI/GI/1 model, a necessary condition for* (3) *to hold for finite positive α and η is to have* (4) *hold with root η.*

Proof. Apply the Lindley equation (16) to obtain

$$\phi(x) \equiv e^{\eta X}P(W > x)$$
$$= \int_{-\infty}^{x} e^{\eta(x-u)}P(W > x - u)e^{\eta u}$$
$$\cdot\, dP(X \leqslant u) + e^{\eta x}P(X > x)$$
$$= \int_{-\infty}^{\infty} f_x(u)d\mu(u) + e^{\eta x}P(X > x), \qquad (17)$$

where

$$f_x(u) = 1_{(-\infty, x]}(u)e^{\eta(x-u)}P(W > x - u)$$

with 1_A being the indicator function of the set A and $d\mu(u) = e^{\eta u}dP(X \leqslant u)$. Given (3), the left side of (17) and $f_x(u)$ approach α as $x \rightarrow \infty$ and $f_x(u)$ is uniformly bounded in x. Hence, by Fatou's lemma,

$$\alpha \int_{-\infty}^{\infty} d\mu(u) \leqslant \lim_{x \to \infty} \int f_x(u)\ d\mu(u) \leqslant \alpha,$$

so that $\mu\ (R) < \infty$.

Since $\int_{-\infty}^{\infty} d\mu(u) = Ee^{\eta X} = 1$, $E[e^{\eta X}; X \geqslant x] \rightarrow 0$ as $x \rightarrow \infty$, but $E[e^{\eta X}; X > x] \geqslant e^{\eta x}P(X > x)$, so that $e^{\eta x}P(X > x) \rightarrow 0$ as $x \rightarrow \infty$. Finally, by the bounded convergence theorem,

$$\int_{-\infty}^{\infty} f_x(u)d\mu(u) \rightarrow \alpha \int_{-\infty}^{\infty} d\mu(u) \quad \text{as } x \rightarrow \infty.$$

Hence, $\mu(R) = Ee^{\eta x} = 1$.

Remark 1. Note that (17) is similar to the renewal equation, but the integral is over $(-\infty, x]$ instead of $[0, x]$. When we work with the ladder height and use the exponential change of measure, (17) is indeed replaced with the renewal equation. From (4), we obtain $pEe^{\eta Z^+} = 1$ too.

Here is the basic $GI/GI/1$ small-tail asymptotic result; see p. 269 of Asmussen (1987) and p. 123 of Borovkov. We include an inequality showing that the asymptotic formula (3) is conservative for all x. Related upper and lower bounds appear in Ross (1974). However, it is important to note that the bound (18) below need not hold for more general models; see Choudhury, Lucantoni and Whitt (1995a).

Theorem 1. *If* (4) *holds in the stable* $GI/GI/1$ *queue, then*

$$P(W > x) \leqslant e^{-\eta x} \quad \textit{for all } x. \tag{18}$$

If, in addition, the distribution of X *is nonlattice and* $Ee^{(\eta+\epsilon)X} < \infty$ *for some* $\epsilon > 0$ *(or just* $EXe^{\eta X} < \infty$*), then* (3) *holds, where*

$$\alpha = \frac{p(\tau_+ = \infty)}{\eta E(Z^+ e^{\eta Z^+})}. \tag{19}$$

Since, in general, we do not know much about Z^+, from (19) we see that the asymptotic constant α is not readily available. Fortunately, the asymptotic decay rate η is not difficult to obtain from (4), because the function $\phi(x) = \log Ee^{xX}$ is convex with $\phi(0) = 0$ and $\phi'(0) < 0$. (This guarantees that there is at most one solution.) Moreover, we typically expect such a root to exist. For example, it is easy to see that a root always exists when the service-time distribution has a rational Laplace transform. (The limit (3) in this case is due to Smith.) However, *even with the* $M/G/1$ *model, for general service-time distributions with finite moment generating functions,* (4) *need not have a solution, so that* (3) *need not be valid.*

Borovkov described the asymptotic behavior of $P(W > x)$ as $x \rightarrow \infty$ when (4) does not hold. By Theorem 12, p. 132, of Borovkov, under considerable generality in this exceptional case

$$P(W > x) \sim \alpha' P(V_e > x) \quad \text{as } x \rightarrow \infty \tag{20}$$

for some constant α', where V_e has the service-time, *stationary-excess* (or equilibrium residual lifetime) *distribution*, i.e.,

$$P(V_e > x) = \frac{1}{EV}\int_x^{\infty} P(V > y)\,dy. \tag{21}$$

Note that the service-time, stationary-excess distribution inherits the asymptotic behavior of the service-time distribution, i.e., if

$$P(V > x) \sim \psi(x) \quad \text{as } x \rightarrow \infty, \tag{22}$$

then

$$(EV)P(V_e > x) \sim \psi_e(x) \equiv \int_x^{\infty} \psi(y)\,dy \quad \text{as } x \rightarrow \infty; \tag{23}$$

see p. 17 of Erdélyi (1956). Hence, if $\psi(x) = \alpha x^p e^{-\eta x}$ for some α, p, and η then $\psi_e(x) = (\alpha/\eta)x^p e^{-\eta x}$ for the same α, p, and η; see (6.5.3) and (6.5.32) of Abramowitz and Stegun (1972).

Example 5. We now exhibit a service-time distribution with a finite moment generating function for which (4) does *not* always have a solution in the $M/G/1$ model. The service-time density is

$$g(t) = \left(\frac{a(a+1)}{2\pi t^3}\right)^{1/2} e^{-at/2(a+1)}\left(1 - e^{-(1+2a)t/2a(a+1)}\right), \tag{24}$$

and its cdf is

$$G(t) \equiv \int_0^t g(x)\,dx = 2(a+1)\Phi(\sqrt{(a+1)t/a})$$
$$- 2a\Phi(\sqrt{at/(a+1)}) - 1 - 2tg(t),$$

where Φ is the cdf of the standard (mean 0, variance 1) normal distribution. It has first three moments $m_1 = 1$,

$$m_2 = \frac{(a+1)^2}{a} - \frac{a^2}{a+1} \quad \text{and}$$

$$m_3 = 3\left(\frac{a^3}{(a+1)^2} - \frac{(a+1)^3}{a^2}\right) \tag{25}$$

and Laplace transform

$$\hat{g}(s) = \sqrt{(a+1)(a+1+2as)} - \sqrt{a(a+2(a+1)s)}. \tag{26}$$

This distribution is obtained from transform pair 29.3.36 in Abramowitz and Stegun. From (26), we see that $\hat{g}(-s)$ is real and finite (a finite moment generating function) for all $s \leqslant s^* = a/2(a+1)$. Note that s^* is the radius of convergence of the moment generating function $\hat{g}(-s)$ and that $-s^*$ is the right-most singularity of the Laplace transform $\hat{g}(s)$.

In the case of an $M/G/1$ queue with arrival rate (traffic intensity) ρ, (4) becomes

Table VII
A Comparison of Asymptotic Approximations With Exact Solutions for Example 5 where for $\rho = 0.7$, the Exponential Asymptotics (3) Applies With $\alpha = 0.4865$ and $\eta^{-1} = 13.1199$, and for $\rho = 0.3$, the Nonexponential Asymptotics (30) Applies

	$\rho = 0.3$		$\rho = 0.7$	
t	Exact	Nonexponential Asymptotics (30)	Exact	Exponential Asymptotics
1.0	0.2027	39.9	0.5894	0.451
2.0	0.1559	12.9	0.5218	0.418
4.0	0.1024	3.7	0.42468	0.359
8.0	0.05146	0.875	0.29601	0.2644
12.0	0.02808	0.319	0.21157	0.1949
16.0	0.01597	0.1400	0.152903	0.1437
20.0	0.00932	0.0667	0.111183	0.1059
40.0	0.000777	0.00319	0.0234668	0.02307
60.0	0.0000745	0.000235	0.0050598	0.005023
80.0	0.00000777	0.0000206	0.0010975	0.001094

$$\hat{g}(-s)\,\frac{\rho}{\rho + s} = 1. \tag{27}$$

Putting $s = s^*$ in (27), we obtain the critical value for ρ to be

$$\rho^* = \frac{1 + \sqrt{1 + 2a}}{4(a + 1)}. \tag{28}$$

For $\rho \geqslant \rho^*$, (27) has a root, but for $\rho < \rho^*$, (27) has no root. From (28) we see that $\rho^*(a)$ decreases from 1/2 to 0 as a increases. When no root exists to (27), we can apply Theorem 12, p. 132, of Borovkov to obtain the asymptotic behavior. For the $M/G/1$ special case considered here, it is perhaps easier to apply Lemma 2 on p. 133 of Borovkov, noting that there $U(x)$ is the complementary cdf of V_e in (21), $p = \rho$ and $b = 1/\rho^*$. (To obtain this expression for b, note that (27) can be written as $\hat{g}_e(-s) = 1/\rho$, where $\hat{g}_e(s)$ is the Laplace–Stieltjes transform of V_e, so that $\hat{g}_e(-s^*) = 1/\rho^*$.) Consequently, when $\rho < \rho^*(a)$,

$$\begin{aligned} P(W > t) &\sim \frac{\rho(1-\rho)}{(1-(\rho/\rho^*))^2}\,P(V_e > t) \quad \text{as } t \to \infty \\ &\sim \frac{\rho(1-\rho)}{(1-(\rho/\rho^*))^2}\left(\frac{a(a+1)}{2\pi t^3}\right)^{1/2} \\ &\quad \cdot \left(\frac{2(a+1)}{a}\right)^2 e^{-at/2(a+1)} \quad \text{as } t \to \infty. \end{aligned} \tag{29}$$

Note that the exponential decay rate in (29) is s^*, which is independent of ρ. The fact that it is independent of ρ suggests that the quality of the approximation might not be so good. Also note that the asymptotic constant in (29) explodes as $\rho \uparrow \rho^*$. On the other hand, as $\rho \downarrow \rho^*$, the asymptotic constant goes to 0, as can be seen from (42) and (26). Thus, the asymptotic approximations tend to be useless for this example in the neighborhood of ρ^*.

To consider a specific case let, $a = 0.25$. For $a = 0.25$, $\rho^* = 0.445$. In this case, the first three service-time moments are 1, 6.20, and 93.72. We consider two arrival rates: $\rho = 0.70$ and $\rho = 0.30$. When $\rho = 0.70$, (4) and (3) hold; when $\rho = 0.30$, (29) applies and we have

$$P(W > x) \sim \frac{44.10}{x^{3/2}}\,e^{-0.1x} \quad \text{as } x \to \infty. \tag{30}$$

Table VII compares the asymptotic approximations in (30) for $\rho = 0.3$ and in (3) with $\alpha = 0.4865$ and $\eta^{-1} = 13.1199$ for $\rho = 0.7$ with exact values computed from Choudhury, Lucantoni and Whitt (1995). For $\rho = 0.3$ the nonexponential asymptotics is revealed through the steady change in the estimates $\hat{\eta}^{-1}(x)$; these estimates are 3.35, 5.23, 7.57, 8.35, and 8.96 at $x = 1$, 4, 20, 40, and 80, respectively. These estimates are *gradually* heading to the limit 10.0 in (30).

Consistent with previous experience, e.g., Table III in Abate and Whitt (1987) and Section 4 of Abate and Whitt (1988), we see that the quality of the approximations based on the exact asymptotics is remarkably poor when $\rho = 0.30$, when the asymptotic form is *not* a pure exponential. Methods for obtaining much better approximations for distributions that do not have pure-exponential asymptotics were developed in Abate and Whitt (1987, 1988).

The poor quality of asymptotic approximation (30) can be explained, at least in part, because the next term in the asymptotic expansion is $\hat{\alpha}x^{-5/2}e^{-0.1x}$, by virtue of Heaviside's theorem, p. 254 of Doetsch (1974), whereas the next term when (3) holds is typically $\hat{\alpha}e^{-\hat{\eta}x}$ for $\hat{\eta} > \eta$. In Abate and Whitt (1988) we found that even the first three terms of an asymptotic expansion related to (30) was not a very good approximation.

When $\rho = 0.7$, the quality of the approximations is not exceptional (e.g., compared with Table I), but it is quite good. Evidently, the quality of the approximation for times of interest is much better when the pure exponential limit (3) is valid than when it is not. The approximate asymptotic parameters when $\rho = 0.7$ are $\eta_{ap}^{-1} = 12.72$ and $\alpha_{ap} = \eta EW = 0.5687$, while the exact asymptotic parameters computed via Choudhury and Lucantoni are $\eta^{-1} = 13.1199$ and $\alpha = 0.4865$.

Remark 2. The service-time density (24) with the asymptotic behavior of $\alpha x^{-3/2}e^{-\eta x}$ as $x \to \infty$ may seem a curiosity, but it routinely arises in priority queues; (see paper in preparation). The waiting times of lower priority customers are related to ordinary $M/G/1$ waiting times with the busy periods of higher priority customers playing the role of the service-time, stationary-excess variable V_e, and these busy-period distributions often have the $\alpha x^{-3/2}e^{-\eta x}$ asymptotic form. As a curiosity, we point out that the convolution of the service-time distribution in Example 5, which has transform $\hat{g}(s)^2$, coincides with the distribution of the busy period in an $M/M/1$ queue with $\lambda = 1/8a(a + 1)$ and $\mu = \lambda + 1/2$, see p. 215 of

Kleinrock. As a partial check, note that $\hat{g}(s)^2$ has mean 2, while the busy period has mean $1/(\mu - \lambda) = 2$.

Remark 3. From (27), we see that for the $M/G/1$ queue (4) will fail to have a root for some ρ if and only if $\hat{g}(s)$ has a singularity on the negative real line and $\hat{g}(-s^*) < \infty$, where $-s^*$ is the right-most singularity. This can only occur when the right-most singularity of $\hat{g}$ is a branch-point singularity, but it does not happen for all branch-point singularities. The right-most singularity $-s^*$ of $\hat{g}$ in Example 1 is a branch-point singularity, but $\hat{g}(-s^*) = \infty$, so that (4) has a root for all ρ in that case. (In Example 5, $\hat{g}(-s^*) = \sqrt{1 + 2a}$.) If the right-most singularity $-s^*$ is a pole (possibly with multiplicity) or an essential singularity, then $\hat{g}(-s^*) = \infty$, so that (4) always has a root for all ρ. It is possible for the right-most singularity of the service-time transform to be an essential singularity. For example, this occurs with the density

$$g(t) = 2e^{-(2t+1)}I_0(2\sqrt{2t}), \ t \geq 0,$$

where I_0 is the modified Bessel function, which has transform

$$\hat{g}(s) = \left(\frac{2}{2+s}\right)e^{-s/(2+s)}$$

and tail behavior

$$g(t) \sim \frac{e^{-(1+\sqrt{2t})^2}}{\sqrt{\pi}(2t)^{1/4}} \quad \text{as } t \to \infty;$$

see p. 374, 29.2.14, and 29.3.81 of Abramowitz and Stegun. The right-most singularity of $\hat{g}$ is -2. Since $\hat{g}(-s^*) = \infty$, as would be the case for any essential singularity, (4) has a root for all ρ and the right-most singularity of the associated waiting-time transform in this case is a simple pole for all ρ.

4. HEAVY-TRAFFIC ASYMPTOTIC EXPANSION FOR THE DECAY RATE

In this section, we develop a heavy-traffic asymptotic expansion for the asymptotic decay rate η in (3) and (4). In fact, this idea was already proposed by Smith, p. 461, but he considers only one term, as on p. 233 of Neuts (1986) and in Proposition 6.1 of Asmussen (1987). (Interestingly, Smith's argument does achieve much of the later heavy-traffic result due to Kingman 1962.) Our asymptotic expansion is also in the spirit of heavy-traffic refinements in Siegmund (1979, 1985) and Section 12.6 of Asmussen (1987), but they start with the exact value of η, which indeed is easy to compute, and develop refinements for the asymptotic constant α as $\eta \to 0$. Instead, we focus on η itself. Our approximation here is η_{ap} in Section 2.

Our asymptotic expansion is in the spirit of asymptotic expansions for the mean EW, as discussed in Section 2 of Whitt (1989) and Knessl (1990). The first terms agree, i.e., the limits of $(1 - \rho)EW(\rho)$ and $(1 - \rho)/\eta(\rho)$ as $\rho \to 1$ both coincide with the standard heavy-traffic limit $(c_a^2 + c_s^2)/2$, but the next term in the expansion of EW is known to be complicated, essentially requiring the solution of a difficult Wiener–Hopf problem. We will show that the decay rate is better behaved.

Our result is a representation of η as an asymptotic expansion in powers of $1 - \rho$. For this purpose, it is convenient to consider a family of models indexed by the traffic intensity ρ. Let V be a generic service time and U/ρ be a generic interarrival time in the model with traffic intensity ρ. Thus, U and V are both fixed random variables with mean 1. Let u_k and v_k denote the kth moment of U and V, respectively. Thus, $u_1 = v_1 = 1$, $c_a^2 = u_2 - 1$ and $c_s^2 = v_2 - 1$. (Note that the SCVs of U and U/ρ are the same. Note that u_3 is the third moment of U, not U/ρ.)

We assume that the transforms Ee^{sV} and $Ee^{-sU/\rho}$ admit expansions in powers of s, i.e.,

$$\begin{aligned} Ee^{sV} &= 1 + sv_1 + \frac{s^2v_2}{2} + \frac{s^3v_3}{6} + \ldots \\ &= 1 + s + \frac{s^2(c_s^2 + 1)}{2} + \frac{s^3v_3}{6} + O(s^4) \quad \text{as } s \to 0 \end{aligned} \tag{31}$$

and

$$\begin{aligned} Ee^{-sU/\rho} &= 1 - \frac{s}{\rho} + \frac{s^2(c_a^2 + 1)}{2\rho^2} - \frac{s^3u_3}{6\rho^3} \\ &\quad + O(s^4) \quad \text{as } s \to 0. \end{aligned} \tag{32}$$

The expansion (31) holds provided that $Ee^{sV} < \infty$ for $s > 0$, while the expansion (32) holds provided that relevant moments of U are finite. (By Taylor's theorem, $u_3 < \infty$ yields $o(s^3)$, while $u_4 < \infty$ yields $O(s^4)$.) We will only keep terms up to the third moment, but it is easy to go further; see Choudhury and Whitt. (Note that our convention of letting the interarrival time be U/ρ prevents $1 - \rho$ terms from being hidden in the interarrival-time moments.)

Theorem 2. *If $Ee^{sV} < \infty$ for $s > 0$ and Ee^{-sU} admits the four-term expansion in* (32), *then*

$$\begin{aligned} \eta = \frac{2(1-\rho)}{c_a^2 + c_s^2} & (1 - (1 - \rho)\eta^* \\ & + O((1-\rho)^2)) \quad \text{as } \rho \to 1, \end{aligned} \tag{33}$$

where

$$\eta^* = \frac{(2v_3 - 3c_s^2(c_s^2 + 2)) - (2u_3 - 3c_a^2(c_a^2 + 2))}{3(c_a^2 + c_s^2)^2}. \tag{34}$$

Proof. Using (31) and (32), we can express (4) as

$$\left(1 + \eta + \eta^2 \frac{(c_s^2 + 1)}{2} + \frac{\eta^3v_3}{6} + O(\eta^4)\right)$$

$$\cdot \left(1 - \frac{\eta}{\rho} + \eta^2 \frac{(c_a^2 + 1)}{2\rho^2} - \frac{\eta^3u_3}{6\rho^3} + O(\eta^4)\right) = 1$$

or

$$1 + \eta\left(1 - \frac{1}{\rho}\right) + \eta^2\left(\frac{c_s^2 + 1}{2} - \frac{1}{\rho} + \frac{c_a^2 + 1}{2\rho^2}\right)$$
$$+ \eta^3\left(\frac{v_3}{6} - \frac{(c_s^2 + 1)}{2\rho} + \frac{(c_a^2 + 1)}{2\rho^2} - \frac{u_3}{6\rho^3}\right) + O(\eta^4) = 1$$

or, after subtracting 1 from both sides and dividing by η/ρ,

$$\eta\left(\frac{\rho(c_s^2 + 1)}{2} - 1 + \frac{(c_a^2 + 1)}{2\rho}\right)$$
$$+ \eta^2\left(\frac{\rho v_3}{6} - \frac{(c_s^2 + 1)}{2} + \frac{(c_a^2 + 1)}{2\rho} - \frac{u_3}{6\rho^2}\right)$$
$$+ O(\eta^3) = 1 - \rho. \tag{35}$$

Expanding the ρ terms on the left in (35) in powers of $(1 - \rho)$, we get

$$\eta\left(\frac{c_s^2 + c_a^2}{2}\right) + \eta\left(\frac{c_a^2 - c_s^2}{2}\right)(1 - \rho)$$
$$+ \eta^2\left(\frac{v_3}{6} + \frac{c_a^2 - c_s^2}{2} - \frac{u_3}{6} + O(1 - \rho)\right)$$
$$+ O(\eta^3) = 1 - \rho. \tag{36}$$

Substituting the entire left side of (36) in for $(1 - \rho)$ wherever it appears, we get

$$\eta\left(\frac{c_a^2 + c_s^2}{2}\right) + \eta^2\left(\frac{v_3}{6} - \frac{u_3}{6} + \left(\frac{c_a^2 - c_s^2}{2}\right)\left(1 + \frac{c_a^2 + c_s^2}{2}\right)\right)$$
$$+ O(\eta^3) = 1 - \rho. \tag{37}$$

Note that (37) is in the form to use the inverse function theorem as with reversion of power series to express η as a partial power series in $(1 - \rho)$; see 3.6.25 of Abramowitz and Stegun. (This argument was already used by Halfin (1985) to analyze $GI/M/1$.) This amounts to matching the leading coefficients in $\eta = \sum_{k=1}^{\infty} b_k(1 - \rho)^k$ and $(1 - \rho) = \sum_{k=1}^{\infty} a_k\eta^k$, which yields $b_1 = 1/a_1$, $b_2 = -a_2/a_1^3$, and $b_3 = (2a_2^2 - a_1a_3)/a_1^5$. We apply this with the first two terms in (37) to obtain (33) and (34).

We call η^* in (34) the *(asymptotic decay-rate) asymptotic correction factor*. We summarize some of its properties next. These translate immediately into corresponding properties for η. Let $\eta^*(U, V)$ denote η^* in (34) as a function of the interarrival time U and service time V (actually their distributions).

Proposition 2. *The asymptotic correction constant η^* in* (34) *has the following properties:*

a. $\eta^*(V, U) = -\eta^*(U, V)$;
b. $\eta^*(U, V) = 0$ *when $U = V$ or just when $u_2 = v_2$ and $u_3 = v_3$;*
c. η^* *is linearly increasing in v_3 and linearly decreasing in u_3 (with the first two moments held fixed).*

Remark 4. Property b is consistent with considerable experience that simple approximations for $GI/GI/1$ systems tend to perform better when the interarrival and service times have similar distributions. This seems to be related to the necessary and sufficient conditions for quasireversibility of a Brownian node in Harrison and Williams (1992).

5. TWO-MOMENT APPROXIMATIONS FOR THE DECAY RATE

In many cases the third moments of interarrival times and service times are not readily available. Thus, as in Section 5.1 of Whitt (1983) and Chapter 4 of Tijms (1986), it is natural to use two-moment approximations for third moments. First, if the distribution is *gamma* (Γ, which includes Erlang (E_k) with $c^2 = 1/k$ as a special case, but covers all values of c^2), then

$$v_3 = (2c_s^2 + 1)(c_s^2 + 1) \tag{38}$$

(assuming that $v_1 = 1$). Second, if the distribution is *hyperexponential with balanced means* (H_2^b), then

$$v_3 = 3c_s^2(c_s^2 + 1). \tag{39}$$

(In the terminology of (16) of Whitt (1984b), a gamma distribution with $c_\Gamma^2 > 1$ corresponds approximately to (has the same first three moments as) an H_2 distribution with weighting factor $r \approx 0.2$ instead of 0.5, which means a higher third moment ($r = (p_1/\lambda_1)/((p_1/\lambda_1) + (p_2/\lambda_2))$ for $\lambda_1 < \lambda_2$). In fact, the corresponding r depends on c_Γ^2, but this rough correspondence applies for a broad range.)

Table VIII displays the values of the asymptotic correction factor η^* in (34) in the four cases in which the interarrival and service times are Γ and H_2^b. When both distributions are the same type, the sign of η^* is the

Table VIII
Two-Moment Approximations for the $GI/GI/1$ Decay-Rate Asymptotic Correction Factor η^* in (34) Based on Gamma (Γ) and Hyperexponential (H_2^b) Distributions

Interarrival-Time Distribution	Service-Time Distribution	η^*
Γ	Γ	$\dfrac{c_a^2 - c_s^2}{3(c_a^2 + c_s^2)}$
Γ	$H_2^b(c_s^2 \geq 1)$	$\dfrac{3c_s^2 - c_a^2 - 2}{3(c_a^2 + c_s^2)}$
$H_2^b(c_a^2 \geq 1)$	Γ	$\dfrac{c_s^2 + 2 - 3c_a^2}{3(c_a^2 + c_s^2)}$
$H_2^b(c_a^2 \geq 1)$	$H_2^b(c_s^2 \geq 1)$	$\dfrac{c_a^2 - c_s^2}{(c_a^2 + c_s^2)}$

sign of $c_a^2 - c_s^2$ and $0 \leq \eta^* \leq 1$. When $c_\Gamma^2 \leq 1$ and $0 \leq \eta^* \leq 1$,

$$\eta^*(\Gamma, H_2) < 0 < \eta^*(H_2, \Gamma). \tag{40}$$

To illustrate, the asymptotic expansion for η in Theorem 2 with the two-moment approximation to the asymptotic correction factor η^* in (34), as given in Table II, provides a quick qualitative explanation of the numerical results in Tables 4.1 and 4.8 on pages 274 and 305 of Tijms. When we add our approximation for the asymptotic constant α in (3), our approximations provide good quantitative estimates as well.

6. APPROXIMATIONS FOR THE ASYMPTOTIC CONSTANT

In this section, we investigate approximations for the asymptotic constant α in (3). We assume that the conditions of Theorem 1 hold. We start with the exact solution for the $M/G/1$ queue. For any nonnegative random variable X, let $\hat{X}(s) = Ee^{-sX}$. Then for the $M/G/1$ queue

$$\hat{W}^c(s) \equiv \int_0^\infty P(W > t)e^{-st}\,dt = \frac{1 - \hat{W}(s)}{s}$$
$$= \frac{\rho}{s}\frac{(1 - \hat{V}_e(s))}{(1 - \rho\hat{V}_e(s))} = \frac{\rho}{s}\frac{(s - 1 + \hat{V}(s))}{(s - \rho + \rho\hat{V}(s))},$$

so that $-\eta$ in (3) is the negative root of $s - \rho + \rho\hat{V}(s) = 0$ closest to the origin. Then we can apply the final-value theorem to get α; i.e.,

$$\alpha = \lim_{s \to -\eta} (s + \eta)\hat{W}^c(s) = \frac{-(1 - \rho)}{1 + \rho\hat{V}'(-\eta)}; \tag{42}$$

e.g., see p. 346 of Kleinrock or p. 254 of Doetsch. To provide a quick check on (42), note that $\hat{V}(s) = (1 + s)^{-1}$ for the $M/M/1$ queue; then $\hat{V}'(-\eta) = -(1 - \eta)^{-2}$ and $\eta = 1 - \rho$, so that $\alpha = \rho$, as it should.

We now provide a case for approximating the asymptotic constant α by ηEW; this is α_{ap} in Section 2. If $P(W > x) = \alpha e^{-\eta x}$ for all x (as in $GI/M/1$), then this is exact. We first show that in the $M/G/1$ queue the error is $O((1 - \rho)^2)$ as $\rho \to 1$. At the same time, we develop a more refined explicit approximation for α for $M/GI/1$ in terms of the first four moments of the service-time distribution.

Theorem 3. *In the $M/G/1$ queue, if $Ee^{sV} < \infty$ for some $s > 0$, then*

$$\frac{\eta EW}{\rho} = 1 - (1 - \rho)(\xi - 1) + (1 - \rho)^2$$
$$\cdot (1 + 2\xi(\xi - 1) - \zeta) + O((1 - \rho)^3 \quad \text{as } \rho \to 1 \tag{43}$$

and

$$\frac{\alpha}{\rho} = \frac{1}{1 + (1 - \rho)(\xi - 1) + (1 - \rho)^2 2(\zeta - \xi^2) + O((1 - \rho)^3)}, \tag{44}$$

where

$$\xi = \frac{2v_3}{3v_2^2} \quad \text{and} \quad \zeta = \frac{v_4}{3v_2^3}, \tag{45}$$

so that (7) *holds.*

Proof. Given that $Ee^{sV} < \infty$, (31) is valid and

$$\hat{V}'(s) = -1 + v_2 s - \frac{v_3 s^2}{2} + \frac{v_4 s^3}{6} + O(s^4) \quad \text{as } s \to 0, \tag{46}$$

The rest of the proof is just like that for Theorem 2, exploiting (42).

We write $\eta EW/\rho$ and α/ρ in (43) and (44) because we are thinking of $\eta^{-1} \approx E(W|W > 0)$ and $\alpha \approx P(W > 0)$ as in (12). Note that $\eta EW/\rho$ is also consistent with (33). Note that for the $M/M/1$ queue $\xi = \zeta = 1$ in (45), so that both the one-term and two-term approximations for ηEW and α are exact.

We now show that (7) also holds in $GI/GI/1$ queues. We apply Siegmund's (1979) corrected heavy-traffic approximations, as given in Asmussen (1987).

Theorem 4. *In the $GI/GI/1$ queue, if the conditions of Theorem* 1, (31) *and* (32) *hold, then* (7) *holds.*

Proof. By Theorem 7.7 of Asmussen (1987),

$$\eta EW = 1 - \beta\eta + O(\eta^2) \quad \text{as } \eta \to 0,$$

and thus as $\rho \to 1$ for the designated constant β. By Theorem 7.2 of Asmussen (1987), for the same β,

$$\alpha = e^{-\eta\beta} + o(\eta^2) \qquad \text{as } \eta \to 0$$
$$= 1 - \eta\beta + O(\eta^2) \quad \text{as } \eta \to 0,$$

and thus as $\rho \to 1$.

Based on Theorems 3 and 4, we suggest ηEW as an approximation for α more generally. This, in turn, leads to the associated approximation $\eta_{ap}EW_{ap}$, where η_{ap} is the approximation developed for η in (33) and EW_{ap} is an approximation for EW. Many approximations for EW already have been developed; e.g., see Tijms (1986) and Whitt (1993). As a specific new simple approximation for the $GI/GI/1$ queue (with service time having mean 1), we propose

$$EW_{ap} = \frac{\rho}{1 - \rho}\left(\frac{c_a^2 + c_s^2}{2} - (1 - \rho)g(c_a^2, c_s^2, u_3)\right), \tag{47}$$

where

$$g(c_a^2, c_s^2, u_3) = \frac{2u_3/3 + (c_a^2 + 1)(c_s^2 - 1)}{2(c_a^2 + c_s^2)} - c_a^2. \tag{48}$$

The heuristic correction (48) is based on an exact analysis of the $K_2/GI/1$ queue, p. 329 of Cohen (1982). For the $GI/M/1$ case, (48) agrees with the heavy-traffic expansion for the mean in the $GI/M/1$ queue, as worked out by Halfin; see (38) of Whitt (1989). For $M/G/1$, $g = 0$ as well.

In the spirit of Section 5, we give simplified two-moment approximations for (48). For H_2^b interarrival-time distributions, we have

$$g(c_a^2, c_s^2, b) \equiv g(c_a^2, c_s^2, 3c_a^2(c_a^2 + 1)) = \frac{(c_a^2 - 1)(1 - c_s^2)}{2(c_a^2 + c_s^2)}. \tag{49}$$

For gamma interarrival-time distributions, we have

$$g(c_a^2, c_s^2, \gamma) \equiv g(c_a^2, c_s^2, (c_a^2 + 1)(2c_a^2 + 1)) = \frac{(1 - c_a^2)}{3} + g(c_a^2, c_s^2, b). \tag{50}$$

These approximations in (48)–(50) have reasonably good accuracy; a few comparisons with exact values are given in Table IX. They are especially appealing for providing insight into the impact of the parameters (e.g., in contrast to the more complicated approximations by Kraemer and Langenbach-Belz used in Whitt 1983.)

We conclude this section by briefly discussing a rather complicated approximation for the asymptotic constant α given in (4.110) on p. 304 of Tijms. This approximation, which we denote by α_t, can be approximately represented as

$$\alpha_t \approx \left(\frac{\eta}{\hat{U}(\eta)^{-1} - 1}\right)\left(\frac{1}{\rho}\right)\eta(M/GI/1)$$
$$\approx \rho\left(1 - \left[\xi - 1 - \frac{c_a^2 - 1}{c_a^2 + c_s^2}\right](1 - \rho)\right). \tag{51}$$

For the $H_2^b/H_2^b/1$ queue, (52) coincides with $\eta_{ap}EW_{ap}$ when either $c_a^2 = 1$ or $c_s^2 = 1$. However, in general α_t seems not to be as good as $\alpha_{ap} \equiv \eta EW$ or $\eta_{ap}EW_{ap}$. To illustrate, Table X compares the approximations for the $E_2/D/1$ queue.

Table IX
A Comparison of the Approximation for the Mean EW in (47) With Exact Values in $GI/GI/1$ Queues

ρ	Method	$H_2^b/D/1$ $c_a^2 = 3$ $c_s^2 = 0$	$E_2/H_2^b/1$ $c_a^2 = 0.5$ $c_s^2 = 2.5$	$H_2^b/M/1$ $c_a^2 = 2.0$ $c_s^2 = 1.0$	$H_2^b/H_2^b/1$ $c_a^2 = 1.5$ $c_s^2 = 1.5$
0.5	Exact	1.08	1.35	1.45	1.51
	Approximate	1.33	1.40	1.50	1.52
0.8	Exact	5.61	5.83	5.97	6.01
	Approximate	5.73	5.84	6.00	6.03

Table X
A Comparison of Approximations for the Asymptotic Constant α in the $E_2/D/1$ Queue

ρ	α Exact	$\alpha_{ap} = \eta EW$	$\eta_{ap}EW_{ap}$	α_t
0.9	0.89	0.89	0.87	0.84
0.7	0.69	0.70	0.63	0.56
0.5	0.51	0.45	0.41	0.33

7. A THRESHOLD FOR SMALL-TAIL ASYMPTOTICS

We have seen in Example 5 that the small-tail asymptotics in (3) need not always be valid, even in an $M/GI/1$ queue where the service-time distribution has a finite moment generating function in a neighborhood of the origin. However, in Example 5 there exists a threshold traffic intensity ρ^* (as a function of the parameter a) such that (3) is valid for all $\rho > \rho^*$, and not for $\rho < \rho^*$. In Remark 3 we noted that this threshold phenomenon holds for all $M/GI/1$ queues. We now show that this threshold phenomenon holds in all $GI/GI/s$ queues. Similar results will hold for more general models based on analogs of (4). Recall that, by Proposition 1 and Theorem 1, (4) is necessary and sufficient for (3) for the $GI/GI/1$ queue when X has a nonlattice distribution.

Theorem 5. *In the $GI/GI/s$ queue, if $Ee^{sV} < \infty$ for some $s > 0$, then there exists ρ^* with $0 \leqslant \rho^* < 1$ such that, for all $\rho > \rho^*$ there exists a root $\eta(\rho)$ to* (4) *and, for all $\rho < \rho^*$ there does not exist a root $\eta(\rho)$ to* (4). For $\rho^* > 0$, there exists a root $\eta(\rho^*)$.

Proof. For simplicity, let there be one server. Since $EV = EU = 1$ and $Ee^{sV} < \infty$, $\psi(x) \equiv Ee^{x(V-U)}$ exists for $0 < x < x^*$. Hence we can apply Taylor's theorem to conclude that

$$Ee^{x(V-U)} = 1 + x^2\frac{E(V-U)^2}{2} + o(x^2) \quad \text{as } x \to 0,$$

so that $Ee^{x(V-U)} > 1$ for all suitably small x. Choose a strictly positive x^* such that $Ee^{xV} < \infty$ and $Ee^{x(V-U)} > 1$ for all x with $0 < x \leqslant x^*$. Since $Ee^{x(V-U/\rho)}$ is strictly increasing in ρ, approaching 0 as $\rho \to 0$, there is a ρ and an $\eta(\rho)$ such that $\eta(\rho) = x$ and $Ee^{x(V-U/\rho)} = 1$ for any x with $0 < x \leqslant x^*$. Letting $\eta(\rho)$ be defined by $Ee^{\eta(\rho)(V-U/\rho)} = 1$, we see that $\eta(\rho)$ is strictly decreasing in ρ. Finally, suppose that $Ee^{\eta(\rho_i)(V-U/\rho_i)} = 1$ for $\rho_1 < \rho < \rho_2$. Then there must exist $\eta(\rho)$ with $\eta(\rho_1) > \eta(\rho) > \eta(\rho_2)$ such that $Ee^{\eta(\rho)(V-U/\rho)} = 1$ because $Ee^{x(V-U/\rho)}$ is continuous in x where it is finite and $Ee^{\eta(\rho_1)(V-U/\rho)} > Ee^{\eta(\rho_1)(V-U/\rho_1)} = 1 = Ee^{\eta(\rho_2)(V-U/\rho_2)} > Ee^{\eta(\rho_2)(V-U/\rho)}$.

8. STOCHASTIC COMPARISONS

In this section, we show that the waiting-time asymptotic decay rate η decreases when the interarrival times and service times become more variable. The notion of "more variable" is made precise by convex stochastic order; see Stoyan (1983).

We say that one random element X_1 is *less variable in the convex stochastic order* than another random element X_2, and we write $X_1 \leqslant_c X_2$, if $Ef(X_1) \leqslant Ef(X_2)$ for all convex real-valued functions f for which the expectations are well defined. We apply this notion for random elements of $\mathbb{R}$ and $\mathbb{R}^\infty$, i.e., when X_i is a real-valued random variable and when X_i is a sequence of real-valued random variables. We use the $\mathbb{R}^\infty$ setting when the interarrival times or service times are not independent. It is important to note that $EX_1 = EX_2$ if $X_1 \leqslant_c X_2$ for real-valued random variables (because $f(x) = x$ and $f(x) = -x$ are both convex).

Our first result is for the *GI/GI/s* queue. To appreciate this positive result for the decay rate, note that the steady-state waiting-time distributions are *not* necessarily stochastically ordered under these conditions, i.e., we do not necessarily have $F_1^c(x) \leqslant F_2^c(x)$ for all x where $F_i^c \equiv 1 - F_i$ is the complementary cdf of the steady-state waiting time in model i; see Whitt (1984c).

Theorem 6. *Let U_i and V_i be generic interarrival-time and service-time random variables in two GI/GI/s queueing models indexed by i. Suppose that there are positive constants η_i such that $Ee^{\eta_i V_i/s}Ee^{-\eta_i U_i/\rho} = 1$ for each i. If $U_1 \leqslant_c U_2$ and $V_1 \leqslant_c V_2$, then $\eta_1 \geqslant \eta_2$.*

Proof. The convex stochastic order implies that $Ee^{xV_1} \leqslant Ee^{xV_2}$ and $Ee^{-xU_1} \leqslant Ee^{-xU_2}$ for all x, so that $\phi_1(x) \equiv Ee^{x[(V_1/s)-(U_1/\rho)]} \leqslant Ee^{x[(V_2/s)-(U_2/\rho)]} \equiv \phi_2(x)$ for all x. Since $\log \phi_i$ is convex with $\phi_i(0) = 1$ and $\phi_i'(0) < 0$, $\phi_1(\eta_2) \leqslant \phi_2(\eta_2) = 1 = \phi_1(\eta_1) \leqslant \phi_2(\eta_1)$ and $\eta_1 \geqslant \eta_2$.

We now obtain a corresponding result for general *G/G/1* models without the independence assumptions.

Theorem 7. *Let $\{(U_n^i, V_n^i): -\infty < n < \infty\}$ be a stationary sequence of interarrival times and service times in two G/G/1 queueing models indexed by i. If there exist positive constants α_i and η_i such that*

$$\lim_{x\to\infty} e^{\eta_i x}F_i^c(x) = \alpha_i \tag{52}$$

for each i and $\{(U_n^1, V_n^1): -\infty < n < \infty\} \leqslant_c \{(U_n^2, V_n^2): -\infty < n < \infty\}$, then $\eta_1 \geqslant \eta_2$ and if $\eta_1 = \eta_2$, then $\alpha_1 \leqslant \alpha_2$.

Proof. It is well known that we can express the steady-state waiting time W_i of model i as

$$W_i = \sup\left\{0, \sum_{j=0}^{k} (V_{-j}^i - U_{-j}^i): k \geqslant 0\right\}, \tag{53}$$

which implies that W is a convex function of $\{(U_j^i, V_j^i): -\infty < j < \infty\}$; e.g., see Chapter 1 of Borovkov. Hence, by the assumed convex stochastic order, $Ee^{xW_1} \leqslant Ee^{xW_2}$ for all x. By (52), $Ee^{xW_i} < \infty$ for all $x < \eta_i$ and $Ee^{xW_i} = \infty$ for all $x > \eta_i$. Hence $\eta_1 \geqslant \eta_2$. By the final-value theorem for Laplace transforms, e.g., p. 346 of Kleinrock or p. 254 of Doetsch,

$$\alpha_i = \lim_{x\to\infty} e^{\eta_i x}F_i^c(x) = \lim_{x\to-\eta} (x + \eta_i)\,\frac{(1 - Ee^{-xW_i})}{x},$$

so that, if $\eta_1 = \eta_2$, then $\alpha_1 \leqslant \alpha_2$.

Corollary. *Among G/G/1 models satisfying the conditions of Theorem 7, the decay rate η of the steady-state waiting time is maximized (a) among stationary interarrival-time sequences with given stationary service-time sequence, and (b) among stationary service-time sequences with given stationary interarrival-time sequence by the deterministic sequence, i.e.,*

$$\eta_{D/G/1} \geqslant \eta_{G/G/1} \quad \text{and} \quad \eta_{G/D/1} \geqslant \eta_{G/G/1}.$$

We now apply the Corollary to Theorem 7 to show that the decay rate of the steady-state waiting time in any *G/GI/1* model is strictly less than the decay rate of the service-time distribution. Since the service-time distribution need not have a pure exponential tail, we first need to define its decay rate. For an arbitrary random variable X, let its *decay rate* be defined as

$$\bar{\eta}(X) = \sup\{\gamma \leqslant \infty : Ee^{\gamma X} < \infty\}. \tag{54}$$

If $P(V \leqslant y) = 1$ for some y, then $\bar{\eta}(V) = \infty$, but in many cases (e.g., when the service-time distribution is phase-type) $\bar{\eta}(V) < \infty$.

Theorem 8. *Consider a stable G/GI/1 model satisfying (52) for all $\rho > \rho^*$ for some ρ^* with steady-state waiting-time decay rate $\eta(\rho)$. Assume that the associated D/GI/1 model with the same service times and arrival rate also satisfies (52) for all $\rho > \rho^*$ with a steady-state waiting-time decay rate $\eta_D(\rho)$. Then $\eta(\rho) \leqslant \eta_D(\rho) < \bar{\eta}(V)$ for all $\rho > \rho^*$, where $\bar{\eta}(V)$ is the service-time decay rate in (54).*

Proof. By Theorem 7, $\eta(\rho) \leqslant \eta_D(\rho)$ for all $\rho \geqslant \rho^*$. By Proposition 1, we must have $Ee^{\eta_D(\rho)V}Ee^{-\eta_D(\rho)d} = 1$, where d is the constant interarrival time in the *D/GI/1* model, which implies that $Ee^{\eta_D(\rho)V} = e^{d\eta_D(\rho)} < \infty$ and $\bar{\eta}(V) \geqslant \eta_D(\rho)$ for all $\rho > \rho^*$. Since $\eta_D(\rho)$ is strictly decreasing in ρ, $\bar{\eta}(V) > \eta_D(\rho)$ for all $\rho > \rho^*$.

9. RELATIVE ERRORS

We now examine the relative errors in the exponential approximations (1) and (2). We first explain why the relative error of the exponential approximation for high percentiles in (2) is usually substantially less than the relative error for the exponential approximation for the tail probability itself in (1). Suppose that the relative error of the exponential approximation in (1) is ϵ for some suitably small ϵ and high p, e.g., $p = 0.90$ or $p = 0.99$. This means that

$$RE(P(W > w_p)) \equiv \frac{P(W > w_p) - \alpha e^{-\eta w_p}}{P(W > w_p)} = \epsilon.$$

Consequently, we have

$$\begin{aligned} P(W > w_p) &= \alpha e^{-\eta w_p} + P(W > w_p)\epsilon \\ &= \alpha e^{-\eta w_p}(1 - \epsilon)^{-1} \\ &= \alpha e^{-\eta w_p}(1 + \epsilon + O(\epsilon^2)) \quad \text{as } \epsilon \to 0. \end{aligned}$$

and

$$\begin{aligned} \log P(W > w_p) &\approx \log \alpha - \eta w_p + \log(1 + \epsilon) \\ &= \log(1 - p). \end{aligned}$$

Hence, we obtain the following approximations for the relative error of the percentile w_p:

$$RE(w_p) \equiv \frac{w_p - \log(\alpha/(1-p))/\eta}{w_p} \approx \frac{\log(1+\epsilon)}{\eta w_p}$$

$$\approx \frac{\epsilon}{\eta w_p} \approx \frac{\epsilon}{\log(\alpha/(1-p))}. \qquad (55)$$

From (55), note that $\log(\alpha/(1-p)) \to \infty$, as $p \uparrow 1$, so that $RE(w_p)/RE(P(W > w_p)) \to 0$ as $p \uparrow 1$. For a rough indication of the advantage in typical cases, note that $\log(\alpha/(1-p)) = 2.3k$ when $\alpha = 1$ and $1 - p = 10^{-k}$.

We now turn to the relative error for the tail probabilities themselves. In typical cases,

$$P(W > x) \sim \alpha e^{-\eta x} + \hat{\alpha} e^{-\hat{\eta} x} \quad \text{as } x \to \infty, \qquad (56)$$

where $\hat{\eta} > \eta$; see Theorem 11 (C_1) on p. 129 of Borovkov. Then

$$RE(P(W > w_p)) \equiv \frac{P(W > w_p) - \alpha e^{-\eta w_p}}{P(W > w_p)}$$

$$\approx \frac{\hat{\alpha} e^{-\hat{\eta} w_p}}{\alpha e^{-\eta w_p} + \hat{\alpha} e^{-\hat{\eta} w_p}}$$

$$= \frac{1}{1 + (\hat{\alpha}/\alpha) e^{+(\hat{\eta} - \eta) w_p}}$$

$$\approx \frac{1}{1 + \left(\frac{\hat{\alpha}}{\alpha}\right)\left(\frac{\alpha}{1-\rho}\right)^{((\hat{\eta}/\eta)-1)}}. \qquad (57)$$

From (57), we see how $RE(P(W > w_p))$ depends on the three parameters $(\hat{\alpha}/\alpha)$, $(\alpha/(1-p))$ and $(\hat{\eta}/\eta) - 1$.

10. CONCLUSIONS

In this paper, we have developed simple exponential approximations for the steady-state waiting distribution of the form (1) and (2) based on the small-tail asymptotics in (3) and approximations for the asymptotic parameters η and α. We are most impressed by the empirical evidence that the exponential form (1) is appropriate in regions of practical interest (the 80th or 90th percentile and beyond). Even though steady-state waiting-time distributions often have relatively complicated mathematical expressions, *these distributions usually appear to be essentially exponential in the region of primary interest* (not too near the origin). For example, as a consequence, we would start looking at data on a steady-state queueing distribution by plotting the logarithm and looking for linearity.

The simple exponential approximations obviously are helpful when we cannot calculate the exact values, but we are also interested in approximations when the model can be solved numerically (which is more and more the case). First, it often occurs that the $G/GI/s$ model is only partially specified; e.g., a distribution may only be partly specified by its first two moments. Then we want approximations based on this partial information. For example, the parametric-decomposition approximation method for non-Markovian open queueing networks in Whitt (1983) and the references cited therein treats the queues as independent $GI/GI/s$ queues with interarrival-time distributions partially specified by the first two moments. Since the arrival process to each queue typically is not actually a renewal process, there typically is no full interarrival-time distribution to discover. Similar reasoning applies with BMAP arrival processes partially characterized by a few parameters.

Even if the model is fully specified and the exact solution is available, an approximation can provide a convenient simple representation of the exact result. In subsequent applications, it is often much more convenient to work with the exponential form (1) based on two parameters than an algorithm for computing $P(W > x)$ for any x. However, in some cases we may need something better than the simple exponential approximation considered here, but still much simpler than the full solution; for this purpose, refined approximations such as the three-exponential approximations in Choudhury, Lucantoni and Whitt (1995) are useful.

As in Whitt (1992), we contend that the most important reason for simple approximations is to develop insight. Armed with simple approximations, we are better able to think about how the system behaves. We illustrated this feature in Section 5 when we compare waiting-time tail probabilities in $G_1/G_2/1$ and $G_2/G_1/1$ models (e.g., $M/E_k/1$ versus $E_k/M/1$) with a common traffic intensity ρ. The approximation formulas quickly reveal the dominant effect, as is borne out in numerical examples.

The simple exponential approximation (1) is even important when it is *not* a good approximation, because it provides a useful frame of reference to help us interpret numerical results. Anticipating that (1) usually holds, we are prepared to notice and appreciate departures from (1). We illustrated this feature in Section 3 when we presented an $M/GI/1$ example (where the service-time distribution *has* a finite moment generating function) for which (3) is *not* valid, so (1) is not good. Moreover, in this case the exact asymptotic formula is a remarkably poor approximation until very large times. Two other cases in which (1) is not good are the waiting times of low-priority customers in $M/GI/1$ priority queues with sufficiently small traffic intensities, and the waiting times in queues having an arrival process consisting of a superposition of a large number of independent non-Poisson processes, see Choudhury, Lucantoni and Whitt (1995). These difficult problems require different methods.

ACKNOWLEDGMENT

We thank David Lucantoni for helpful discussions, Phil Fleming for showing us his related work, and Charles Goldie for helpful comments.

REFERENCES

ABATE, J., AND W. WHITT. 1987. Transient Behavior of Regulated Brownian Motion, I: Starting at the Origin. *Adv. Appl. Prob.* **19,** 560–598.

ABATE, J., AND W. WHITT. 1988. Approximations for the $M/M/1$ Busy-Period Distribution. In *Queueing Theory and its Applications, Liber Amicorum for J. W. Cohen*, O. J. Boxma and R. Syski (eds.) North-Holland, Amsterdam, 149–191.

ABATE, J., AND W. WHITT. 1992. The Fourier-Series Method for Inverting Transforms of Probability Distributions. *Queue. Syst.* **10,** 5–88.

ABATE, J., AND W. WHITT. 1994. A Heavy-Traffic Expansion for the Asymptotic Decay Rates of Tail Probabilities in Multi-Channel Queues. *O.R. Letts.* **15,** 223–230.

ABATE, J., G. L. CHOUDHURY AND W. WHITT. 1993. Calculation of the $GI/G/1$ Waiting-Time Distribution and Its Cumulants From Pollaczek's Formulas. *AEU* **47,** 299–314.

ABATE, J., G. L. CHOUDHURY AND W. WHITT. 1994a. Asymptotics for Steady-State Tail Probabilities in Structured Markov Queueing Models. *Stoch. Models* **10,** 99–143.

ABATE, J., G. L. CHOUDHURY AND W. WHITT. 1994b. Waiting-Time Tail Probabilities in Queues With Long-Tail Service-Time Distributions. *Queue. Syst.* **16,** 311–338.

ABATE, J., G. L. CHOUDHURY AND W. WHITT. 1995. Exponential Approximations for Tail Probabilities in Queues, II: Sojourn Time and Workload. *Opns. Res.* (to appear).

ABRAMOWITZ, M., AND I. A. STEGUN. 1972. *Handbook of Mathematical Functions*, 10th printing. National Bureau of Standards, U.S. Government Printing Office, Washington, D.C.

ASMUSSEN, S. 1987. *Applied Probability and Queues.* John Wiley, New York.

ASMUSSEN, S. 1989. Risk Theory in a Markovian Environment. *Scand. Act. J.* 69–100.

ASMUSSEN, S., AND D. PERRY. 1992. On Cycle Maxima, First Passage Problems and Extreme Value Theory of Queues. *Stoch. Models* **8,** 421–458.

ASMUSSEN, S., AND T. ROLSKI. 1991. Computational Methods in Risk Theory: A Matrix-Algorithmic Approach. *Insurance: Math. and Econ.* **10,** 259–274.

BAIOCCHI, A. 1992. Asymptotic Behavior of the Loss Probability of the $MAP/G/1/K$ Queue, Part I: Theory. INFOCOM Department, University of Rome "La Sapienza."

BOROVKOV, A. A. 1976. *Stochastic Processes in Priority Queueing Theory.* Springer-Verlag, New York.

CHANG, C. S. 1994. Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks. *IEEE Trans. Auto. Control* **39,** 913–931.

CHOUDHURY, G. L., AND D. M. LUCANTONI. 1995. Numerical Computation of the Moments of a Probability Distribution From Its Transforms. *Opns. Res.* (to appear).

CHOUDHURY, G. L., AND W. WHITT. 1994. Heavy-Traffic Expansions for the Asymptotic Decay Rates in the $BMAP/G/1$ Queue Stoch. Models **10,** 453–498.

CHOUDHURY, G. L., D. M. LUCANTONI AND W. WHITT. 1995. Squeezing the Most Out of ATM IEEE Trans. Commun. (to appear).

COHEN, J. W. 1982. *The Single Server Queue*, 2nd ed. North-Holland, Amsterdam.

DOETSCH, G. 1974. *Introduction to the Theory and Application of the Laplace Transformation.* Springer-Verlag, New York.

ELWALID, A. I., AND D. MITRA. 1993. Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks. *IEEE AEM Trans. Networks* **1,** 329–343.

ELWALID, A. I., AND D. MITRA. 1994. Markovian Arrival and Service Communication Systems: Spectral Expansions, Separability and Kronecher-Product Forms In *Computations with Markov Chains*, W. J. Stewart (ed.). Kluwen, Boston, 507–546.

ELWALID, A. I., D. MITRA AND T. E. STERN. 1991. Statistical Multiplexing of Markov Modulated Sources: Theory and Computational Algorithms. In *Teletraffic and Data Traffic in a Period of Change, ITC-13*, A. Jensen and B. Iversen (eds.). Elsevier, Amsterdam, 495–500.

ERDELYI, A. 1956. *Asymptotic Expansions.* Dover, New York.

FELLER, W. 1971. *An Introduction to Probability Theory and Its Applications.* Vol. II, 2nd ed. John Wiley, New York.

FLEMING, P. J. 1992. Simple, Accurate Formulas for Approximating Percentiles of Delay Through a Single Device. Motorola, Inc., Arlington Heights, Ill.

FRANKEN, P., D. KONIG, U. ARNDT AND V. SCHMIDT. 1981. *Queues and Point Processes.* Akademic Verlag, Berlin.

FREDERICKS, A. A. 1982. A Class of Approximations for the Waiting Time Distribution in a $GI/G/1$ Queueing System. *Bell Syst. Tech. J.* **61,** 295–325.

GLYNN, P. W., AND W. WHITT. 1994. Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single-Server Queue. *J. Appl. Prob. Studies in Applied Probability, Papers in Honor of Lajos Takáos*, J. Galambos and J. Gani (eds.). Applied Prob. Trust, Sheffield, England, 131–156.

HALFIN, S. 1985. Delays in Queues, Properties and Approximations. In *Teletraffic Issues in an Advanced Information Society, ITC-11*, M. Akiyama (ed.). Elsevier, Amsterdam, 47–52.

HARRISON, J. M., AND R. WILLIAMS. 1992. Brownian Models of Feedforward Queueing Networks. *Anns. Appl. Prob.* **2,** 263–293.

JACQUET, P. 1992. Subexponential Tail Distribution in La Palice Queues. *Perf. Eval. Rev.* **20,** 60–69.

KEILSON, J. 1979. *Markov Chain Models—Rarity and Exponentiality.* Springer-Verlag, New York.

KINGMAN, J. F. C. 1962. On Queues in Heavy Traffic. *J. Roy. Statist. Soc.* **B24,** 383–392.

KLEINROCK, L. 1975. *Queueing Systems, Vol. 1: Theory.* John Wiley, New York.

KNESSL, C. 1990. Refinements to Heavy Traffic Limit Theorems in Queueing Theory. *Opns. Res.* **38,** 826–837.

LUCANTONI, D. M. 1991. New Results on the Single Server Queue with a Batch Markovian Arrival Process. *Stoch. Models* **7,** 1–46.

NEUTS, M. F. 1981. Stationary Waiting-Time Distributions in the $GI/PH/1$ Queue. *J. Appl. Prob.* **18,** 901–912.

NEUTS, M. F. 1986. The Caudal Characteristic Curve of Queues. *Adv. Appl. Prob.* **18,** 221–254.

NEUTS, M. F. 1989. *Structured Stochastic Matrices of M/G/1 Type and Their Applications.* Marcel Dekker, New York.

NEUTS, M. F., AND Y. TAKAHASHI. 1981. Asymptotic Behavior of the Stationary Distributions in the *GI/PH/C* Queue With Heterogeneous Servers. *Z. Wahrscheinlichkeitsth.* **57,** 441–452.

ROSS, S. M. 1974. Bounds on the Delay Distribution in *GI/G/*1 queues. *J. Appl. Prob.* **11,** 417–421.

SEAL, H. 1969. *Stochastic Theory of a Risk Business.* John Wiley, New York.

SEELEN, L. P., AND H. C. TIJMS. 1985. Approximations to the Waiting Time Percentiles in the *M/G/c* Queue. In *Teletraffic Issues in an Advanced Information Society, ITC-11,* M. Akiyama (ed.). Elsevier, Amsterdam, 53–57.

SEELEN, L. P., H. C. TIJMS AND M. H. VAN HOORN. 1985. *Tables for Multi-Server Queues.* North-Holland, Amsterdam.

SIEGMUND, D. 1979. Corrected Diffusion Approximation in Certain Random Walk Problems. *Adv. Appl. Prob.* **11,** 701–719.

SIEGMUND, D. 1985. *Sequential Analysis.* Springer-Verlag, New York.

SMITH, W. L. 1953. On the Distribution of Queueing Times. *Proc. Camb. Phil. Soc.* **49,** 449–461.

STOYAN, D. 1983. *Comparison Methods for Queues and Other Stochastic Models.* John Wiley, Chichester, U.K.

TAKAHASHI, Y. 1981. Asymptotic Exponentiality of the Tail of the Waiting-Time Distribution in a *PH/PH/c* Queue. *Adv. Appl. Prob.* **13,** 619–630.

TIJMS, H. C. 1986. *Stochastic Modeling and Analysis: A Computational Approach.* John Wiley, New York.

VAN OMMEREN, J. C. W. 1988. Exponential Expansion for the Tail of the Waiting Time Probability in the Single Server Queue with Batch Arrivals. *Adv. Appl. Prob.* **20,** 880–895.

VAN OMMEREN, J. C. W. 1989. *Asymptotic Analysis of Queueing Systems.* Ph.D. Dissertation, Free University, Amsterdam.

WHITT, W. 1983. The Queueing Network Analyzer. *Bell Syst. Tech. J.* **62,** 2779–2815.

WHITT, W. 1984a. On Approximations for Queues, I: Extremal Distributions. *AT&T Bell Lab. Tech J.* **63,** 115–138.

WHITT, W. 1984b. On Approximations for Queues, III: Mixtures of Exponential Distributions. *AT&T Bell Lab. Tech. J.* **63,** 163–175.

WHITT, W. 1984c. Minimizing Delays in the GI/G/1 Queue. *Opns. Res.* **32,** 41–51.

WHITT, W. 1989. An Interpolation Approximation for the Mean Workload in a *GI/G/*1 Queue. *Opns. Res.* **37,** 936–952.

WHITT, W. 1992. Understanding the Efficiency of Multi-Server Service Systems. *Mgmt. Sci.* **38,** 708–723.

WHITT, W. 1993a. Tail Probabilities with Statistical Multiplexing and Effective Bandwidths for Multi-Class Queues. *Telecomm. Syst.* **2,** 71–107.

WHITT, W. 1993b. Approximations for the *GI/G/m* Queue. *Prod. and Opns. Mgmt.* **2,** 141–161.

WILLEKENS, E., AND J. L. TEUGELS. 1992. Asymptotic Expansions for Waiting Time Probabilities in an M/G/1 Queue With Long-Tailed Service Time. *Queue. Syst.* **10,** 295–312.