

**EXPONENTIAL APPROXIMATIONS FOR TAIL PROBABILITIES
IN QUEUES: SOJOURN TIME AND WORKLOAD**

by

Joseph Abate,¹ Gagan L. Choudhury,² and Ward Whitt³

AT&T Bell Laboratories

November 12, 1992

¹AT&T Bell Laboratories Retired: 900 Hammond Road, Ridgewood, NJ 07450-2908

²AT&T Bell Laboratories, Room 3K-603, Holmdel, NJ 07733-3030

³AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636

Abstract

In this paper, we focus on simple exponential approximations for steady-state tail probabilities in G/GI/1 queues based on large-time asymptotics. We relate the large-time asymptotics for the steady-state waiting time, sojourn time and workload. We evaluate the exponential approximations based on the exact asymptotic parameters and their approximations by making comparisons with exact numerical results for BMAP/GI/1 queues. Numerical examples show that the exponential approximations are remarkably accurate at the 90th percentile and beyond.

Key words: queues; approximations; asymptotics; tail probabilities; sojourn time and workload.

1. Introduction and Summary

This paper is a sequel to [1], in which we studied exponential approximations for waiting-time tail probabilities in infinite-capacity queues based on large-time asymptotics. Let W be the steady-state waiting time (before beginning service) in an infinite-capacity queue with the first-come first-served queue discipline. In great generality,

$$P(W > x) \sim \alpha e^{-\eta x} \text{ as } x \rightarrow \infty, \quad (1)$$

i.e., $e^{\eta x} P(W > x) \rightarrow \alpha$ as $x \rightarrow \infty$, where η and α are positive constants (independent of x) called the *asymptotic decay rate* and the *asymptotic constant*, respectively. Moreover, the limiting exponential form is often a surprisingly good approximation when x is not too small. For example, the associated approximation for the $(100p)^{\text{th}}$ percentile $w_p \equiv \inf\{x : P(W > x) = 1 - p\}$,

$$w_p \approx \log \left[\frac{\alpha}{1-p} \right] \frac{1}{\eta}, \quad (2)$$

is usually remarkably good for $p \geq 0.90$. (It is easy to see that the relative error in an approximation for a high percentile is typically much lower than the relative error in an approximation for a tail probability itself.)

In [1] we presented numerical examples based on exact numerical solutions of the BMAP/GI/1 queue (with batch Markovian arrival process) and the GI/GI/s queue to show that the exponential approximation based on (1) is remarkably good, lending support to previous work in the same direction, notably by Tijms [26] and Asmussen [4]. Moreover we developed simple effective approximations for the asymptotic parameters α and η in (1).

The purpose of this paper is to relate the large-time asymptotic behavior of the steady-state waiting time W to the large-time asymptotic behavior of the *sojourn time* T (response time, i.e., waiting time plus service time) and the *workload* L (virtual waiting time). In particular, we show

that corresponding asymptotics for T and L are valid in any G/GI/1 queue (with general stationary arrival process) whenever (1) is valid. Moreover, we show that the resulting approximations are remarkably good by making comparisons with exact numerical values.

There is a substantial body of related literature. For additional work on asymptotics related to (1), see Abate, Choudhury and Whitt [2], Asmussen [5], [6], Asmussen and Perry [7], Baiocchi [8], Borovkov [9], Chang [10], Choudhury and Whitt [13], de Smit [14], Elwalid and Mitra [15], [16], Elwalid, Mitra and Stern [17], Fleming [18], Neuts [22], [23] and Takahashi [25]. In particular, in [2] we obtain asymptotic results for the steady-state waiting time, workload and queue lengths (at arrivals, at arbitrary times and at departures) in the BMAP/GI/1 queue. Our results here are different because we treat the sojourn time and the more general G/GI/1 model.

The remarkable quality of the exponential approximation (1) for M/GI/1 queues is discussed in Section 1.9 of Tijms [26] and Section 9 of Abate and Whitt [3]. However, Example 4 of [1] shows that (1) need *not* be valid for M/GI/1 queues, even when the service-time distribution has a finite moment generating function in some neighborhood of the origin. (It is evident that the limit (1) can fail when this condition does not hold.) In §22 of Borovkov [9] it is shown how to establish alternative asymptotic behavior (not pure-exponential) for GI/GI/1 queues when (1) does not hold. However, Example 4 of [1] shows that the quality of the approximation provided by the large-time asymptotics deteriorates dramatically when the pure-exponential form is lost. (This behavior also holds to T and L .) This phenomenon demonstrates that having a limit such as (1) does not by itself guarantee a good approximation. However, it turns out that, not only is the pure-exponential large-time asymptotics in (1) often valid, but it turns out to be a surprisingly good approximation.

It is now relatively well understood that an exponential approximation based on (1) is good for the waiting time. It may be surprising, though, that a similar exponential approximation is

also often good for the sojourn time (waiting time plus service time) without any special assumptions on the service-time distribution. This idea has been advanced by Fleming [18], who proposes simple heavy-traffic approximations for sojourn-time percentiles as well as waiting-time percentiles in a class of $M/GI/1$ queues. (He focuses on two-point service-time distributions, which are realistic for computer systems.) We provide additional support for this idea, as well as develop new approximations for more general models.

Here is how the rest of this paper is organized. In §2 we relate the asymptotic behavior of the waiting time, workload and sojourn time in the $G/GI/1$ model. In particular, we show that all three satisfy (1) with the *same* asymptotic decay rate η and asymptotic constants α_W , α_L and α_T that can be simply related. In §3 we briefly discuss light traffic. In particular, we note that the asymptotic decay rate η approaches the asymptotic decay rate of the service-time distribution, defined in (9) below, as $\rho \rightarrow 0$, where ρ is the traffic intensity. In §4 we apply the asymptotic exponential approximations to develop an approximation for the ratio EL/EW . In §5 we discuss numerical examples for the workload and sojourn time, drawing on Lucantoni [21], Abate and Whitt [3], Choudhury and Lucantoni [12] and Choudhury [11], just as in [1]. In §6 we relate the asymptotic decay rate η in (1) to the asymptotic decay rate for the steady-state queue length. Finally, we state our conclusions in §7.

2. Sojourn Time and Workload

In this paper we consider the $G/GI/1$ queueing model with one server, unlimited waiting space, the first-come first-served discipline and i.i.d. service times that are independent of a general stationary arrival process. We assume that the mean service time is 1 and that the arrival rate is $\rho < 1$. We assume that the various steady-state distributions discussed below exist as proper probability distributions. For the $GI/GI/1$ queue, it suffices for the interarrival-time distribution to be nonlattice; see Chapter 8 of Asmussen [4]. For the more general $G/GI/1$ model,

see Franken et al. [20].

Let V be a generic service time random variable. Theorem 11 and Example 4 of [1] show that in order for the large-time asymptotics (1) for W to be valid in the G/GI/1 queue, it is necessary, but not sufficient, to have $Ee^{sV} < \infty$ for some $s > 0$.

In this section we show that the steady-state sojourn time or response time T and the steady-state workload or virtual waiting time L tend to have the same asymptotic decay rate η as the steady-state waiting time W and asymptotic constants α_L and α_T that are easily related to the waiting-time constant $\alpha \equiv \alpha_W$.

For GI/PH/1 queues, the asymptotic behavior of W , L and T was described in detail by Neuts [22]. These relationships are also a consequence of interesting phase-type results in Asmussen [6]; see Corollary 2.2. In particular, Asmussen shows that if the service-time distribution is phase-type characterized by the pair (π, Q) where π is a d -dimensional vector and Q is a $d \times d$ generator matrix, then W , L and T have distributions, which except for a probability mass at the origin, are also phase-type with representations (π_W, \tilde{Q}) , (π_L, \tilde{Q}) and (π_T, \tilde{Q}) where \tilde{Q} is a common $d \times d$ generator matrix and π_W , π_L and π_T are in general different d -dimensional vectors. Since the asymptotic decay rate η is the Perron-Frobenius eigenvalue of \tilde{Q} , it is identical for all three random variables. The asymptotic constants involve the eigenvectors associated with the dominant eigenvalue and the vectors π_W , π_L and π_T .

Remark 1. We conjecture that this structural solidarity result extends to GI/PH/ s models with $s > 1$, but *without* having the number of phases in \tilde{Q} be equal to the number d of service-time phases. Indeed, we conjecture that the number of phases in \tilde{Q} is $\binom{d+s-1}{s} \equiv (d+s-1)!/(d-1)!s!$. This is based on the structural solidarity result for GI/H $_d$ / s queues established by de Smit [14]; the waiting-time distribution is again hyperexponential (plus a mass at the origin) with this larger number of exponential terms. ■

We extend Neuts [22] and Asmussen [6] for the sojourn time by replacing the GI and PH in GI/PH/1 by G and GI, respectively, but we only consider the asymptotic parameters. This next result extends easily to s servers.

Theorem 1. *In the G/GI/1 model, if $e^{\eta x}P(W > x) \rightarrow \alpha_W$ as $x \rightarrow \infty$, then $Ee^{\eta V} < \infty$ and*

$$e^{\eta x}P(T > x) \rightarrow \alpha_T \equiv \alpha_W E e^{\eta V} > \alpha_W \text{ as } x \rightarrow \infty .$$

Proof. By Theorem 11 of [1], $Ee^{\eta V} < \infty$. Since $T = W + V$ where W and V are independent,

$$\begin{aligned} e^{\eta x}P(T > x) &= \int_0^x e^{\eta(x-u)} P(W > x-u) e^{\eta u} dP(V \leq u) + e^{\eta x}P(V > x) \\ &= \int_0^\infty 1_{[0,x]} e^{\eta(x-u)} P(W > x-u) e^{\eta u} dP(V \leq u) + e^{\eta x}P(V > x) . \end{aligned}$$

Since $Ee^{\eta V} < \infty$, $e^{\eta x}P(V > x) \rightarrow 0$ as $x \rightarrow \infty$. Then the assumed convergence for W plus the bounded convergence theorem implies the desired conclusion. ■

Remark 2. From [22], [23] and [2], we know that the correction term $Ee^{\eta V}$ in Theorem 1 must be σ^{-1} , where σ is the queue-length asymptotic decay rate; see §6 here for further discussion.

Remark 3. It may seem surprising that the sojourn-time distribution should have the same asymptotic exponential form as the waiting time with the same asymptotic decay rate. However, Theorem 1 is especially easy to understand when the service time-distribution is deterministic; then $P(T > x) = P(W > x - 1) \sim \alpha e^{-\eta(x-1)}$ as $x \rightarrow \infty$, so that $\alpha_T = \alpha e^\eta$. The case of a service-time distribution with finite support is a minor modification. Theorem 1 is the natural generalization.

Remark 4. When ρ is not too small, so that η is sufficiently small, we can use the approximation

$$\begin{aligned} Ee^{\eta V} &\approx E \left[1 + \eta V + \frac{\eta^2 V^2}{2} + \frac{\eta^3 V^3}{6} \right] \\ &\approx 1 + \eta + \frac{\eta^2 (c_s^2 + 1)}{2} + \frac{\eta^3 v_3}{6} . \quad \blacksquare \end{aligned} \tag{3}$$

We now treat the workload in the G/GI/1 model. For this, we use a relation between a

distribution and its associated stationary-excess distribution. If X is a nonnegative random variable with cdf G and finite mean, then X_e is a random variable with the associated stationary-excess distribution, i.e.,

$$P(X_e > x) = \frac{1}{EX} \int_x^\infty P(X > y) dy, x \geq 0. \quad (4)$$

Lemma 1. If $Ee^{sX} < \infty$, then

$$Ee^{sX_e} = \frac{E(e^{sX} - 1)}{sEX}.$$

Proof. Apply integration by parts. ■

Theorem 2. In the $G/GI/1$ model, if $e^{\eta x} P(W > x) \rightarrow \alpha_W$ as $x \rightarrow \infty$, then

$$e^{\eta x} P(L > x) \rightarrow \alpha_L \equiv \frac{\alpha_W \rho}{\eta} (Ee^{\eta V} - 1).$$

Proof. By the generalized Takács formula, (4.5.9) on p. 129 of Franken et al. [20],

$$P(L > x) = \rho P(W + V_e > x) \quad (5)$$

for all x , where V_e is independent of W and has the stationary-excess distribution of the service-time distribution. The rest of the argument is as in Theorem 1. We use Lemma 1 (and the fact that $EV = 1$) to obtain

$$\alpha \rho E^{\eta V_e} = \frac{\alpha \rho}{\eta} (Ee^{\eta V} - 1). \quad \blacksquare$$

Remark 5. As with T in Theorem 1, we can express the correction term for L in Theorem 2 directly in terms of the asymptotic decay rates η and σ (see Remark 2); i.e., $\rho E(e^{\eta V} - 1)/\eta = \rho(1 - \sigma)/\eta\sigma$.

Remark 6. Notice that Theorem 2 is consistent with the well known property that L has the same distribution as W in the $M/G/1$ queue, because then $\rho(Ee^{\eta V} - 1)/\eta = 1$ by the defining property of η . Similarly, for $GI/M/1$,

$$\alpha_L = \alpha \cdot \rho \frac{(Ee^{\eta V} - 1)}{\eta} = \frac{\alpha \rho}{\sigma} = \rho$$

because $\alpha = \sigma = 1 - \eta$. Finally, for the GI/PH/1 queue, Theorems 1 and 2 agree with §2 of Neuts [22].

Remark 7. Paralleling Remark 4, we can use the approximation

$$\begin{aligned} \alpha_L &\approx \frac{\alpha_W \rho}{\eta} \left[\eta EV + \frac{\eta^2 EV^2}{2} + \frac{\eta^3 EV^3}{6} \right] \\ &\approx \alpha_W \rho \left[1 + \eta \frac{(c_s^2 + 1)}{2} + \frac{\eta^2 v_3}{6} \right]. \end{aligned} \quad (6)$$

Given formula (33) in [1] for GI/GI/1, we see that for GI/GI/1 as $\rho \rightarrow 1$

$$\begin{aligned} \alpha_L &\approx \alpha_W \rho (1 + (1-\rho) - \eta^* (1-\rho)^2 + \frac{4(1-\rho)^2 v_3}{6(c_a^2 + (c_s^2)^2)} + O((1-\rho)^3)) \\ &\approx \alpha_W (1 - \xi (1-\rho)^2 + O((1-\rho)^3)) \text{ as } \rho \rightarrow 1, \end{aligned} \quad (7)$$

where

$$\xi = \frac{2}{3} \frac{(u_3 - 6c_a^2)}{(c_a^2 + c_s^2)} - \frac{2(c_a^2 - 1)}{(c_a^2 + c_s^2)} + 1. \quad (8)$$

Note that the correction term ξ in (7) and (8) is $O((1-\rho)^2)$ instead of $O(1-\rho)$. Also note that ξ is independent of the third service-time moment. Finally, note that $\xi = 1$ in the case of M/G/1, as it must. ■

Theorems 1 and 2 show how to compute α_T and α_L given η and α_W or approximations for them. When it is not convenient to calculate $Ee^{\eta V}$, Remarks 4 and 7 show how to approximate α_T and α_L given η and α_W or approximations for them. Paralleling §6 of [1], we also suggest the approximations $\alpha_{Tap} = \eta ET$ and $\alpha_{Lap} = \eta EL$.

3. Light Traffic

It is possible to develop light-traffic and heavy-traffic interpolation formulas for tail probabilities in the spirit of Fleming and Simon [19], Whitt [27] and references therein, but we do not develop this idea here, because then we would lose the simple exponential form of the approximations considered here. (Recall that we have algorithms to compute the exact values.) However, it is useful to understand what happens in the light-traffic limit (as $\rho \rightarrow 0$). It is easy to show that the asymptotic decay rate η in (1) approaches the asymptotic decay rate $\bar{\eta}(V)$ of the service-time distribution, defined as

$$\bar{\eta}(V) = \sup\{\gamma > 0 : Ee^{\gamma V} < \infty\} . \quad (9)$$

Note that the definition of asymptotic decay rate in (9) is more general than (1), because we do not assume that the convergence in (1) necessarily holds. For example, we could have $P(V > x) \sim \alpha_V x^{-\beta_V} e^{-\eta_V x}$ as $x \rightarrow \infty$. The asymptotic decay rate $\bar{\eta}(V)$ in turn coincides with the asymptotic decay rate $\bar{\eta}(V_e)$ of the service-time stationary-excess distribution.

Indeed, from (5) it is easy to see that the steady-state workload distribution in the G/GI/1 model approaches the service-time stationary-excess distribution in the light-traffic limit; this is proved in Sigman [24]. This analysis applies directly only to the steady-state workload, but it applies to the other steady-state variables through the relationships we establish. Of course, Theorem 2 only goes from W to L . It is easy to go the other way in the context of (9), because (5) implies that $P(L > x | L > 0) = P(W + V_e > x)$ and

$$\frac{1 - \hat{L}(s)}{\rho} = 1 - \hat{W}(s) \hat{V}_e(s)$$

for G/GI/1 queues.

4. An Approximation for the Ratio EL/EW

In this section we discuss an approximation for the ratio EL/EW , which yields EL and EW if we have either one. The approximation is based on the exact form of α_L/α_W given in Theorem 2 and the approximations

$$\alpha_W \approx \eta EW \text{ and } \alpha_L \approx \eta EL . \quad (10)$$

In particular, we suggest

$$\frac{EL}{EW} \approx \frac{\alpha_L}{\alpha_W} = \frac{\rho(1-\sigma)}{\eta\sigma} \quad (11)$$

where $\sigma = Ee^{\eta V}$. By Remark 6, this approximation in (11) is exact for both $M/GI/1$ and $GI/M/1$.

For the $GI/GI/1$ queue we can combine (7), (8) and (11) to obtain the approximation

$$\frac{EL}{EW} \approx 1 - \zeta(1-\rho)^2 , \quad (12)$$

for ζ in (8).

This approximation in (11) can also be used for the ratio of mean queue lengths at arbitrary times and at arrivals. For theoretical support, see Theorem 11 of [2].

5. Numerical Examples for the Sojourn Time and the Workload

In Remark 3 we noted that it is easy to see that the asymptotic behavior of T and W are closely related when the service-time distribution is deterministic. We now consider what happens with service-time distributions that are substantially more variable than an exponential distribution.

As in [1], we obtain the exact tail probabilities from the algorithms in Lucantoni [21], with transform inversion from Abate and Whitt [3], as implemented by Choudhury [11]. We obtain the exact values of the asymptotic parameters from the moment-based generating-function-inversion algorithm in Choudhury and Lucantoni [12]. We also estimate the asymptotic

parameters by linear regression applied to the numerically calculated tail probabilities (after taking logarithms) as described in [1].

Example 1. Consider the $M/H_2^b/1$ queue with a hyperexponential service-time distribution with balanced means, as defined in Example 1 of [1]. Let the arrival rate be $\rho = 0.7$ and, as always, let the service-time distribution have mean 1. Consider the case of service-time squared coefficient of variation (variance divided by the square of the mean) $c_s^2 = 4.0$. Then the parameters of the density are $p_1 = 0.8872983$, $\lambda_1 = 1.7744966$ and $\lambda_2 = 0.2254034$.

Since the arrival process is Poisson (M), the distributions of W and L coincide. We apply the Pollaczek-Khintchine formula to obtain $EL = 5.833$ and $ET = 6.833$. The exact asymptotic parameters for L and T obtained from Choudhury and Lucantoni [12] and the linear regression are $\eta = 0.1000040$, $\alpha_L = \alpha_W = 0.5727238$ and $\alpha_T = 0.6545448$, so that $\sigma = \alpha_W/\alpha_T = 0.87500$.

The approximations from §4 and §6 of [1] are $\eta_{HT} = 0.1200$, $\eta_{ap} = 0.0984$, $\alpha_{Lap} = \eta EL = 0.5833$, $\alpha_{Tap} = \eta ET = 0.6833$, $\eta_{ap} EL = 0.5740$ and $\eta_{ap} ET = 0.6724$. As in [1], the approximations for the asymptotic parameters are quite good.

Tables 1 and 2 display exact values of the tail probabilities $P(L > x)$ and $P(T > x)$ and the associated exponential approximations. The regression estimates are displayed as well to show the (in this case, spectacular) rate of convergence to the exponential limit. In this case, the linear regression easily produces the exact asymptotic parameters.

Example 2. To see what happens with a non-renewal arrival process and a service-time distribution very unlike an exponential distribution, we now consider the $MMPP_2/D_2/1$ model of Example 3 in [1]. As before, $\rho = 0.7$ and $c_s^2 = 2.0$. First, the asymptotic decay rates calculated for W , T and L by the algorithm in Choudhury and Lucantoni [12] agreed to eight decimal places, yielding $\eta = 0.11159727$. For this model, it is easy to see that $\sigma^{-1} = Ee^{\eta V} = 1.13873$. The

successive approximations in (3) are: 1.0, 1.1115, 1.1301 and 1.1362. The relative error in the approximation for $Ee^{\eta V}$ is 0.8% and 0.2% using two and three moments.

The asymptotic constants are $\alpha_W = 0.65738$, $\alpha_T = 0.74867$ and $\alpha_L = 0.57261$. These provide empirical evidence supporting Theorems 1 and 2. For the asymptotic constant, $(\alpha_L/\alpha_W) = 0.87104$. The successive approximations in (6) are 0.7, 0.817, 0.8717. (Note that (7) does not apply because the arrival process is not renewal.)

Table 3 compares exponential approximations for the tail probabilities of the steady-state workload and sojourn time with exact values computed using the algorithm in [11]. Again the exponential approximations perform well. Our experience indicates that, consistent with intuition, the quality of the exponential approximations for the waiting time and workload is usually somewhat better than for the sojourn time. However, the difference is hardly perceptible in Table 3.

With regard to the approximation for EL/EW in §4, here EW = 5.831 and EL = 5.131, so that EL/EW = 0.880. Since $\alpha_L/\alpha_W = 0.871$, we see that the mean ratio approximation in (11) performs well in this example.

Example 3. We conclude with an MMPP/ $\Gamma_{1/2}/1$ example, which is used to evaluate heavy-traffic asymptotic expansions for the asymptotic decay rates of the waiting time in §7 of Choudhury and Whitt [13]. The service-time distribution is gamma with shape parameter 1/2, which is not rational and thus not PH. It is moderately highly variable, with first three moments 1, 3 and 15.

The arrival process is a two-phase MMPP, which has four parameters (the arrival rate and mean holding time in each phase), one of which we determine by letting the arrival rate be ρ . A second parameter is determined by assuming that the long-run arrival rate in each phase is $\rho/2$. A third parameter is determined by assuming that the expected number of arrivals during each visit

to each phase is 5. Finally, the last parameter is determined by making the ratio of the arrival rates in the two phases 4.

Tables 4 and 5 display approximations and exact values for higher percentiles of the steady-state workload and sojourn-time distributions, respectively. In each case, two values of ρ are considered: $\rho = 0.8$ and $\rho = 0.5$. Three approximations are considered. All approximations are exponential approximations $\alpha e^{-\eta x}$ with the exact η , converted to percentiles as in (2). The first approximation has the exact asymptotic constant, α_L and α_T , respectively; the second approximation approximates α_L by ηEL and α_T by ηET ; and the third approximation approximates α_L and α_T by 1.

From Tables 4 and 5, we see that the approximations for higher percentiles are very impressive. The accuracy improves as the percentile increases and as the traffic intensity increases. At $\rho = 0.8$, the relative error of the asymptotic approximation (with exact α) is less than 0.1% even at the 80th percentile. The approximation based on $\alpha \approx \eta^*$ mean performs remarkably well, substantially better than the approximation with $\alpha \approx 1.0$. However, for high percentiles such as 99.99, even $\alpha \approx 1.0$ yields a useful approximation.

Finally, with regard to the approximation for EL/EW in §4, the means EW and EL are and with $\rho = 0.5$ and and with $\rho = 0.8$, so that the ratio EL/EW is and with $\rho = 0.5$ and $\rho = 0.8$. On the other hand, the asymptotic constants α_W and α_L are and with $\rho = 0.5$ and and with $\rho = 0.8$, so that the ratio α_L/α_W is and with $\rho = 0.5$ and $\rho = 0.8$.

6. The Queue Length

In this section we indicate how the asymptotic decay rate η for the steady-state waiting time, sojourn time and workload is related to the asymptotic decay rate σ for the steady-state queue length in a large class of G/GI/1 models. To establish a connection, we assume that the service-

time distribution is phase-type (PH). What we present here extends results in Neuts [22] for GI/PH/1 queues and Abate, Choudhury and Whitt [2] for BMAP/GI/1 queues. It still remains to give a proof for the general G/GI/1 model.

We assume that the steady-state distributions are well defined. Let Q be the steady-state queue length at an arbitrary time. Paralleling (1), typically we have

$$P(Q > k) \sim \beta \sigma^k \text{ as } k \rightarrow \infty, \quad (13)$$

but we define the asymptotic decay rate σ more generally by setting

$$\bar{\sigma}(Q)^{-1} = \sup \{z \geq 1 : Ez^Q < \infty\} \quad (14)$$

as in (9). We define $\bar{\eta}(L)$ as in (9) too. We discuss only the steady-state distributions at arbitrary times, but a corresponding result holds for the steady-state distributions at arrival epochs.

To establish our result, we need to know about the asymptotic decay rates of phase-type distributions.

Lemma 2. *Let V have a k -state phase-type distribution characterized by the pair (α, T) where α is the initial distribution, T is the infinitesimal generator of the absorbing continuous-time Markov chain and every phase can occur. Then*

$$\bar{\eta}(V) = 1/\min \{ |T_{ii}| : 1 \leq i \leq k \} \geq \bar{\eta}(R_i)$$

where R_i represents the residual service time starting in phase i .

Proof. Since α_i is the probability of starting in phase i ,

$$Ee^{sV} = \sum_{i=1}^k \alpha_i Ee^{sR_i},$$

so that $\bar{\eta}(V) = \max \{ \bar{\eta}(R_i) : 1 \leq i \leq k, \alpha_i > 0 \}$. Let R_i be an exponential random variable with mean $-1/T_{ii}$, let $P_{ij} = T_{ij}/|T_{ii}|$ if $|T_{ii}| > 0$ with $P_{ij} = 0$ otherwise, and let

$$q_i = 1 - \sum_{j=1}^k P_{ij}. \text{ Then, for any } i,$$

$$Ee^{sR_i} = Ee^{sT_i} \left(q_i + \sum_{j=1}^k Ee^{sR_j} P_{ij} \right),$$

so that

$$\bar{\eta}(R_i) = \max \{ \bar{\eta}(T_i), \bar{\eta}(R_j) : 1 \leq j \leq k, P_{ij} > 0 \}.$$

Since all phases can be reached,

$$\bar{\eta}(V) = \max_{1 \leq i \leq k} \bar{\eta}(R_i) = \max_{1 \leq i \leq k} \bar{\eta}(T_i) = \frac{1}{\min_{1 \leq i \leq k} |T_{ii}|}. \quad \blacksquare$$

We now state our queue length theorem.

Theorem 3. *In the G/PH/1 queue,*

$$\bar{\sigma}(Q)^{-1} = Ee^{\bar{\eta}(L)V}.$$

Proof. Let the service-time distribution have k phases. Let L_i and Q_i be the steady-state workload and queue length at an arbitrary time conditional on the server being busy in service phase i . Let R_i be a residual service time starting in phase i . Let π_i be the steady-state probability of being in service phase i conditional on the server being busy. It is well known that $P(L > 0) = P(Q > 0) = \rho$, so that

$$Ee^{sL} = \rho \sum_{i=1}^k \pi_i Ee^{sL_i} \quad \text{and} \quad Ez^Q = \rho \sum_{i=1}^k \pi_i Ez^{Q_i} \quad (15)$$

and

$$\bar{\eta}(L) = \max_{1 \leq i \leq k} \bar{\eta}(L_i) \quad \text{and} \quad \bar{\sigma}(Q)^{-1} = \max_{1 \leq i \leq k} \bar{\sigma}(Q_i)^{-1}. \quad (16)$$

The fundamental relation connecting L and Q is

$$Ee^{sL_i} = Ee^{s(R_i + \sum_{j=1}^{Q_i} V_j)} = Ee^{sR_i} E(Ee^{sV})^{Q_i}. \quad (17)$$

By Theorem 11 of [1] and Theorem 2 here, $\bar{\eta}(V) < \bar{\eta}(L)$. By Lemma 2, $\bar{\eta}(R_i) \leq \bar{\eta}(V)$, so that

we can apply (15)–(17) to obtain the desired conclusion. ■

Remark 8. In [2] we showed for the BMAP/GI/1 queue that the asymptotic constants α_L for L and β in (13) coincide. It does not seem easy to deduce this conclusion from the proof of Theorem 3.

7. Conclusions

For the G/GI/1 queue, we have shown that the steady-state waiting time W , sojourn time T and workload L have the same large-time asymptotic behavior with the same asymptotic decay rate η and the asymptotic constants α_W , α_T and α_L that are simply related. As shown in §6, the relations among the asymptotic constants α_W , α_L and α_T is intimately connected to the asymptotic decay rate for the steady-state queue length. We have proposed approximations for α_T , α_L and EL/EW to go with previous approximations for $\eta_1 \alpha_W$ and EW. We have presented numerical examples showing that these exponential approximations based on large-time asymptotics perform remarkably well.

References

- [1] J. Abate, G. L. Choudhury and W. Whitt, "Exponential approximations for tail probabilities in queues, I: waiting times," 1992, submitted for publication.
- [2] J. Abate, G. L. Choudhury and W. Whitt, "Asymptotics for steady-state tail probabilities in structured Markov chains," 1992, submitted for publication.
- [3] J. Abate and W. Whitt, "The Fourier-series method for inverting transforms of probability distributions," *Queueing Systems* 10 (1992) 5-88.
- [4] S. Asmussen, *Applied Probability and Queues*, John Wiley, New York, 1987.
- [5] S. Asmussen, "Risk theory in a Markovian environment," *Scand. Act. J.* (1989) 69-100.
- [6] S. Asmussen, "Phase-type representations in random walk and queueing problems," *Ann. Probab.* 20 (1992) 772-789.
- [7] S. Asmussen and D. Perry, "On cycle maxima, first passage problems and extreme value theory of queues," *Stochastic Models* 8 (1992), to appear.
- [8] A. Baiocchi, "Asymptotic behavior of the loss probability of the MAP/G/1/K queue, Part I: Theory," INFOCOM Dept., University of Rome "La Sapienza," 1992.
- [9] A. A. Borovkov, *Stochastic Processes in Queueing Theory*, Springer-Verlag, New York, 1976.
- [10] C. S. Chang, "Stability, queue length and delay, part II: stochastic queueing networks," IBM T. J. Watson Research Center, Yorktown Heights, NY, 1992.
- [11] G. L. Choudhury, "An algorithm for a large class of G/G/1 queues," in preparation.
- [12] G. L. Choudhury and D. M. Lucantoni, "Numerical computation of the moments of a probability distribution from its transforms," in preparation.

- [13] G. L. Choudhury and W. Whitt, "Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/GI/1 queue," 1992, submitted for publication.
- [14] J. H. A. de Smit, "The queue GI/M/s with customers of different types or the queue GI/H_m/s. *Adv. Appl. Prob.* 15 (1983) 392-419.
- [15] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," (1992a) to appear.
- [16] A. I. Elwalid and D. Mitra, "Markovian arrival and service communication systems: spectral expansions, separability and Kronecker-product forms," (1992b) in preparation.
- [17] A. I. Elwalid, D. Mitra and T. E. Stern, "Statistical multiplexing of Markov modulated sources: theory and computational algorithms." pp. 495-500 in *Teletraffic and Data Traffic in a Period of Change, ITC-13*, A. Jensen and B. Iversen (eds.), Elsevier, Amsterdam, 1991.
- [18] P. J. Fleming, "Simple accurate formulas for approximating percentiles of delay through a single device." Motorola, Inc., Arlington Heights, IL, 1992.
- [19] P. J. Fleming and B. Simon, "Interpolation approximations of sojourn time distributions," *Oper. Res.* 39 (1991) 251-260.
- [20] P. Franken, D. König, U. Arndt and V. Schmidt, *Queues and Point Processes*, Akademie-Verlag, Berlin, 1981.
- [21] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process," *Stochastic Models* 7 (1991) 1-46.
- [22] M. F. Neuts, "Stationary waiting-time distributions in the GI/PH/1 queue," *J. Appl. Prob.* 18 (1981) 901-912.

- [23] M. F. Neuts, "The caudal characteristic curve of queues," *Adv. Appl. Prob.* 18 (1986) 221-254.
- [24] K. Sigman, "Light traffic for workload in queues," *Queueing Systems*, to appear.
- [25] Y. Takahashi, "Asymptotic exponentiality of the tail of the waiting-time distribution in a PH/PH/c queue," *Adv. Appl. Prob.* 13 (1981) 619-630.
- [26] H. C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach*, John Wiley, New York, 1986.
- [27] W. Whitt, "An interpolation approximation for the mean workload in a GI/G/1 queue," *Oper. Res.* 37 (1989) 936-952.

x	exact	$\alpha_L e^{-\eta x}$	$\hat{\alpha}_L(x)$	$\hat{\eta}(x)$
3.0	0.4278	0.4243	0.5931	9.178
6.0	0.31441	0.31431	0.5740	9.966
9.0	0.232846	0.232844	0.57279	9.9984
12.0	0.17249290	0.17249283	0.5727272	9.999554
18.0	0.094663802	0.094663802	0.572723848	9.99960019
24.0	0.051951350	0.051951350	0.572723841	9.99960026

Table 1. A comparison of exponential approximations with exact values of the workload tail probabilities, $P(L > x)$, in the $M/H_2^b/1$ queue with $\rho = 0.7$ and $c_s^2 = 4.0$ in Example 1. Also included are the local linear regression estimates of the asymptotic parameters.

x	exact	$\alpha_T e^{-\eta x}$	$\hat{\alpha}_T(x)$	$\hat{\eta}(x)$
3.0	0.4943	0.4849	0.7107	8.263
6.0	0.35947	0.35921	0.6581	9.921
9.0	0.266115	0.266108	0.6547	9.99667
12.0	0.1971358	0.1971356	0.6545	9.99949
18.0	0.108187743	0.108187743	0.654544812	9.9996010
24.0	0.059373268	0.059373268	0.654544803	9.99960026

Table 2. A comparison of exponential approximations with exact values of the sojourn-time tail probabilities, $P(T > x)$, in the $M/H_2^b/1$ queue with $\rho = 0.7$ and $c_s^2 = 4.0$ in Example 1. Also included are the local linear regression estimates of the asymptotic parameters.

x	workload			sojourn		
	exact	approx.	percent error	exact	approx.	percent error
3.0	0.3765	0.4097	8.8	0.4801	0.5356	11.6
6.0	0.2900	0.2931	1.0	0.3564	0.3833	7.6
9.0	0.2230	0.2097	-6.0	0.2884	0.2742	-4.9
12.0	0.1506	0.1501	-0.3	0.2033	0.1962	3.5
15.0	0.1049	0.1074	2.4	0.1355	0.1403	3.5
18.0	0.0771	0.0768	-0.4	0.0997	0.1004	0.7
21.0	0.0557	0.0550	-1.3	0.0733	0.0719	-1.9
24.0	0.03913	0.03932	0.5	0.05137	0.05142	0.1
27.0	0.02800	0.02814	0.5	0.03644	0.03678	0.9
30.0	0.02020	0.02013	-0.3	0.02638	0.02632	-0.2
36.0	0.010304	0.01030	0.2	0.01344	0.01347	0.2
42.0	0.005282	0.005275	-0.1	0.006908	0.006898	-0.1
48.0	0.002699	0.002701	0.7	0.003528	0.003521	0.1
54.0	0.001383	0.001383	0.0	0.001808	0.001808	0.0
60.0	0.000708	0.000708	0.0	0.000925	0.000925	0.0

Table 3. A comparison of exponential approximations for the steady-state workload and sojourn-time tail probabilities with exact values in the $MMPP_2/D_2/1$ queue in Example 2.

$\rho = 0.8, \eta = 0.08039$				
percentile required	percentile value			
	exact	approx., exact α $\alpha_L =$	approx., $\alpha \approx \eta * \text{mean}$ $\alpha_L \approx$	approx. $\alpha_L \approx 1.0$
80	16.1555	16.1489	16.34	20.0
90	24.7714	24.7709	24.97	28.6
99	53.4126	53.4126	53.61	87.3
99.9	82.0542	82.0542	82.25	85.9
99.99	110.6959	110.6959	110.89	114.6
$\rho = 0.5, \eta = 0.19677$				
percentile required	percentile value			
	exact	$\alpha_L = 0.50219$	approx., exact α $\alpha_L \approx 0.53255$	approx., $\alpha \approx \eta * \text{mean}$ $\alpha_L \times 1.0$
80	3.6059	3.0228	3.70	8.2
90	6.8173	6.5455	7.22	11.7
99	18.2667	18.2476	18.92	23.4
99.9	29.9509	29.9496	30.62	35.1
99.99	41.6518	41.6517	42.32	46.8

Table 4. A comparison of approximations with exact values of high percentiles of the steady-state workload in the MMPPP/T_{1/2}/1 queue in Example 3.

$\rho = 0.8, \eta = 0.08039$				
percentile required	percentile value			
	exact	approx., exact α $\alpha_T =$	approx., $\alpha \approx \eta * \text{mean}$ $\alpha_T \approx$	approx. $\alpha_T \approx 1.0$
80	18.3940	18.3914	18.57	20.0
90	27.0136	27.0134	27.19	28.6
99	55.6551	55.6551	55.83	57.3
99.9	84.2968	84.2968	84.48	85.9
99.99	112.9384	112.9384	113.12	114.6
$\rho = 0.5, \eta = 0.19677$				
percentile required	percentile value			
	exact	$\alpha_T =$	approx., exact α $\alpha_L \approx$	approx., $\alpha \approx \eta * \text{mean}$ $\alpha_L \times 1.0$
80	6.2030	5.9497	6.58	8.2
90	9.5853	9.4724	10.10	11.7
99	21.1821	21.1745	21.80	23.4
99.9	32.8770	32.8765	33.50	35.1
99.99	44.5786	44.5786	45.20	46.8

Table 5. A comparison of approximations with exact values of high percentiles of the steady-state sojourn time in the MMPP/ $\Gamma_{1/2}/1$ queue in Example 3.