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

JOSEPH ABATE

Ridgewood, New Jersey

GAGAN L. CHOUDHURY

AT&T Laboratories, Holmdel, New Jersey

WARD WHITT

AT&T Laboratories, Murray Hill, New Jersey

(Received December 1992; revision received October 1993; accepted April 1994)

We continue to focus on simple exponential approximations for steady-state tail probabilities in queues based on asymptotics. For the *G/GI/*1 model with i.i.d. service times that are independent of an arbitrary stationary arrival process, we relate the asymptotics for the steady-state waiting time, sojourn time, and workload. We show that the three asymptotic decay rates coincide and that the three asymptotic constants are simply related. We evaluate the exponential approximations based on the exact asymptotic parameters and their approximations by making comparisons with exact numerical results for *BMAP/G/*1 queues, which have batch Markovian arrival processes. Numerical examples show that the exponential approximations for the tail probabilities are remarkably accurate at the 90th percentile and beyond. Thus, these exponential approximations appear very promising for applications.

This paper is a sequel to Part I, Abate et al. (1995), in which we studied exponential approximations for steady-state waiting-time tail probabilities in infinite-capacity queues based on asymptotics. In Part I we presented numerical examples based on exact numerical solutions of the *BMAP/G/*1 queue (having i.i.d. service times independent of a batch Markovian arrival process) and the *GI/G/s* queue to show that the exponential approximation is remarkably good, lending support to previous work in the same direction, notably by Tijms (1986) and Asmussen (1987, 1989). Moreover we developed simple effective approximations for the asymptotic parameters.

The primary purpose of this paper is to relate the asymptotic behavior of the steady-state *waiting time W* to the asymptotic behavior of the steady-state *sojourn time T* (response time, i.e., waiting time plus service time) and the steady-state *workload L* (virtual waiting time). In particular, we show that corresponding asymptotics for *T* and *L* are valid in any *G/GI/*1 queue (having i.i.d. service times that are independent of a general stationary arrival process) whenever the exponential asymptotics for the waiting time are valid, and we show that the parameters are simply related. Moreover, we show by making comparisons with exact numerical values that the resulting approximations are remarkably good.

Even if we recognize that an exponential approximation is good for the waiting time, it may be surprising that a similar exponential approximation is also often good for the sojourn time without any special assumptions on the service-time distribution. This idea has been advanced by Fleming (1992), 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. Having a corresponding exponential approximation for sojourn times is very important for applications, because the sojourn time (response time) is often the critical variable.

Here is how the rest of this paper is organized. In Section 1 we relate the asymptotic behavior of the waiting time, workload, and sojourn time in the *G/GI/*1 model. In Section 2 we discuss numerical examples for the workload and sojourn time, drawing on Choudhury et al. (1996), which in turn draws upon Lucantoni (1991).

Additional related work (done after this paper) appears in Abate et al. (1994), Glynn and Whitt (1994), and Choudhury et al. (1996). Abate et al. (1994) directly establish exponential asymptotics for the steady-state variables in the BMAP/G/1 queue and give explicit expressions for the asymptotic parameters. Glynn and Whitt establish logarithmic limits for the steady-state waiting time and relate these limits to corresponding logarithmic limits for the steady-state workload and queue length (at an arbitrary time and at arrivals). These weaker logarithmic limits hold in greater generality than the limits considered in this paper. Finally, Choudhury et al. (1996) focus on the special

Subject classifications: Queues, approximations: exponential approximations for tail probabilites. Queues, limit theorems: asymptotics for delay and workload tail probabilites.
Area of review: Stochastic Processes and Their Applications.

0030-364X/96/4405-0758 $01.25

© 1996 INFORMS

~~Copyright © 2001 All Rights Reserved~~

case of superposition arrival processes, which is of special interest for communication networks, and show by numerical examples that the asymptotic approximations often do not perform as well in that setting. Due to developments in communication networks, this topic has recently received much attention. See the references (e.g., Part I) for additional references.

1. SOJOURN TIME AND WORKLOAD

We consider the $G/GI/1$ queueing model, which has 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.

Let W be the steady-state waiting time (before beginning service). As discussed in Part I, in great generality,

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

i.e., $e^{\eta x}P(W > x) \to \alpha_W$ as $x \to \infty$, where η and α_W are positive constants (independent of x) called the *asymptotic decay rate* and the *asymptotic constant*, respectively. Let V be a generic service time random variable. Theorem 10 and Example 5 of Part I show that in order for (1) 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 T and the steady-state workload L satisfy analogs of (1) when (1) holds with the *same* asymptotic decay rate η and asymptotic constants α_L and α_T that are easily related to α_W.

For $GI/PH/1$ queues, the asymptotic behavior of W, L, and T was described in detail by Neuts (1981). These relationships are also a consequence of interesting phase-type results in Asmussen (1992); 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 $(\pi_W, \tilde{Q})$, $(\pi_L, \tilde{Q})$ and $(\pi_T, \tilde{Q})$ 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.

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 (1983); the waiting-time distribution is again hyperexponential (plus a mass at the origin) with this larger number of exponential terms.

We extend Neuts (1981) and Asmussen (1992) for the sojourn time by replacing the GI and PH in $GI/PH/1$ by G and GI, respectively, but we consider only the asymptotic parameters. This next result extends easily to s servers.

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

$$e^{\eta x}P(T > x) \to \alpha_T \equiv \alpha_W Ee^{\eta V} > \alpha_W \ \textit{as}\ x \to \infty\,.$$

Proof. By Theorem 10 of Part I, $Ee^{\eta V} < \infty$. Since $T = W + V$ where W and V are independent,

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

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

From Proposition 9 of Glynn and Whitt (1994), we know that the correction term $Ee^{\eta V}$ in Theorem 1 is σ^{-1}, where σ is the queue-length asymptotic decay rate. Glynn and Whitt relate the logarithmic asymptotics for the waiting time and queue length. It remains to relate (1) to analogs of (1) for the queue length. For special cases, connections are established in Neuts (1981, 1986) and Abate et al. (1994). The $BMAP/G/1$ model in the last paper is very close to the $G/GI/1$ model considered here.

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 $x \to \infty$, 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.

We now consider 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 finite mean, then X_e is a random variable with the associated stationary-excess distribution, i.e.,

$$P(X_e > x) = \frac{1}{EX}\int_x^\infty P(X > y)dy,\ x \geq 0\,. \tag{2}$$

Applying integration by parts, we see that if $Ee^{sX} < \infty$, then

$$Ee^{sX_e} = \frac{E(e^{sX} - 1)}{sEX}\,. \tag{3}$$

Copyright © 2001 All Rights Reserved

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

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

Proof. By the generalized Takács formula in Franken et al. (1981 (4.5.9) on p. 129),

$$P(L > x) = \rho P(W + V_e > x) , \tag{4}$$

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 (3) (and the fact that $EV = 1$) to obtain

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

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

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 Section 2 of Neuts (1981).

When η is sufficiently small, we can use the approximation

$$\begin{aligned} Ee^{\eta V} &\approx E\left(1 + \eta V + \frac{\eta^2 V^2}{2} + \frac{\eta^3 V^3}{6}\right) \\ &\approx 1 + \eta + \frac{\eta^2(c_s^2 + 1)}{2} + \frac{\eta^3 \nu_3}{6} . \end{aligned} \tag{5}$$

Inserting (5) into the formula for α_L, we get

$$\begin{aligned} \alpha_L &\approx \frac{\alpha_W \rho}{\eta}\left(\eta EV + \frac{\eta^2 EV^2}{2} + \frac{\eta^3 EV^3}{6}\right) \\ &\approx \alpha_W \rho\left(1 + \eta\frac{(c_s^2 + 1)}{2} + \frac{\eta^2 \nu_3}{6}\right) . \end{aligned} \tag{6}$$

Given formula (33) in Part I for *GI/GI/*1, we see that for *GI/GI/*1 as $\rho \to 1$

$$\begin{aligned} \alpha_L \approx \alpha_W\Bigg(1 &- \left(\frac{c_a^2 - 1}{c_a^2 + c_s^2}\right)(1 - \rho) - \left(\left(\frac{1 + c_s^2}{c_a^2 + c_s^2}\right)(\eta^* + 1)\right. \\ &\left. - \frac{2\nu_3}{3(c_a^2 + c_s^2)^2}\right)(1 - \rho)^2 + O((1 - \rho)^3) \quad \text{as } \rho \to 1 , \end{aligned} \tag{7}$$

where

$$\eta^* = \frac{(2\nu_3 - 3c_s^2(c_s^2 + 2)) - (2u_3 - 3c_a^2(c_a^2 + 2))}{3(c_a^2 + c_s^2)^2} . \tag{8}$$

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 $Ee^{\eta V}$, (5)–(8) show how to approximate α_T and α_L given η and α_W or approximations for them. Paralleling Section 6 of Part I, we also suggest the approximations $\alpha \approx \eta ET$ and $\alpha \approx \eta EL$.

2. NUMERICAL EXAMPLES FOR THE SOJOURN TIME AND THE WORKLOAD

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 Part I, we obtain the exact tail probabilities from the algorithm described in Choudhury et al. (1996), which draws upon Lucantoni. We obtain the exact values of the asymptotic parameters from the moment-based generating-function-inversion algorithm in Choudhury and Lucantoni (1996). We also estimate the asymptotic parameters by linear regression applied to the numerically calculated tail probabilities (after taking logarithms) as described in Part I.

Example 1. Consider the $M/H_2^b/1$ queue with a hyperexponential service-time distribution with balanced means. Let the arrival rate be $\rho = 0.7$ and, as always, let the service-time distribution have mean 1. The H_2^b distribution is defined in Example 1 of Part I. 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 (1994) 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 Section 4 and Section 6 of Part I are $\eta_{HT} = 0.1200$, $\eta_{ap} = 0.0984$, $\alpha_{Lap} \equiv \eta EL = 0.5833$, $\alpha_{Tap} \equiv \eta ET = 0.6833$, $\eta_{ap}EL = 0.5740$, and $\eta_{ap}ET = 0.6724$. As in Part I, the approximations for the asymptotic parameters are quite good.

Tables I and II 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 nonrenewal arrival process and a service-time distribution very unlike an exponential distribution, we now consider the $\text{MMPP}_2/\text{D}_2/1$ model of Example 3 in Part I. 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 agreed to eight decimal places, yielding $\eta = 0.11159727$. For this model, it is easy to see that $\sigma^{-1} = Ee^{\eta V} =$

Copyright © 2001 All Rights Reserved

Table I

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	$\alpha_L e^{-\eta x}$	$\hat{\alpha}_L(x)$	$\hat{\eta}(x)$
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

1.13873. The successive approximations in (5) 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 III compares exponential approximations for the tail probabilities of the steady-state workload and sojourn time with exact values. 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 not perceptible in Table III.

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 Section 7 of Choudhury and Whitt (1994). The service-time distribution is gamma with shape parameter 1/2, so that the transform is not rational and thus the distribution is not PH. It is moderately highly variable, with first three moments 1, 3, and 15.

Table II

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	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 III

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

	Workload			Sojourn		
			Percent			Percent
x	Exact	Approx.	error	Exact	Approx.	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	0.0	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	0.1	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

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 IV and V 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) of Part I. 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 IV and V, 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.

However, as shown in Choudhury et al. (1996), the quality of these asymptotic approximations can deteriorate when the arrival process is the superposition of many independent component arrival processes. Nevertheless, since

Copyright © 2001 All Rights Reserved

Table IV
A comparison of approximations with exact values of high percentiles of the steady-state workload in the $MMPPP/\Gamma_{1/2}/1$ queue in Example 3

		$\rho = 0.8$, $\eta = 0.08039$		
		Percentile value		
Percentile		Approx., exact α	Approx., $\alpha \approx \eta$EL	Approx.
required	Exact	$\alpha_L = 0.73234$	$\alpha_L \approx 0.7439$	$\alpha_L \approx 1.0$
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 value		
Percentile		Approx., exact α	Approx., $\alpha \approx \eta$EL	Approx.
required	Exact	$\alpha_L = 0.36257$	$\alpha_L \approx 0.4135$	$\alpha_L \approx 1.0$
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 V
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

		$\rho = 0.8$, $\eta = 0.08039$		
		Percentile value		
Percentile		Approx., exact α	Approx., $\alpha \approx \eta$ET	Approx.
required	Exact	$\alpha_T = 0.87702$	$\alpha_T \approx 0.8899$	$\alpha_T \approx 1.0$
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 value		
Percentile		Approx., exact α	Approx., $\alpha \approx \eta$ET	Approx.
required	Exact	$\alpha_T = 0.64493$	$\alpha_L \approx 0.7288$	$\alpha_L \approx 1.0$
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

the desired probabilities are often extremely small, even in that demanding setting the asymptotic approximation is often good enough provided that we know (or have a good approximation for) the asymptotic constant as well as the asymptotic decay rate.

ACKNOWLEDGMENT

Again we thank Dave Lucantoni for assistance and Phil Fleming for showing us his unpublished work.

REFERENCES

Abate, J., G. L. Choudhury, and W. Whitt. 1994. Asymptotics for Steady-State Tail Probabilities in Structured Markov Queueing Models. *Stochastic Models*, **10**, 99–143.

Abate, J., G. L. Choudhury, and W. Whitt. 1995. Exponential Approximations for Tail Probabilities in Queues. I: Waiting Times. *Opns. Res.*, **43**, 885–901.

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. 1992. Phase-Type Representations in Random Walk and Queueing Problems. *Ann. Probab.* **20,** 772–789.

Choudhury, G. L. and D. M. Lucantoni. 1996. Numerical Computation of the Moments of a Probability Distribution From Its Transform. *Opns. Res.*, **44**, 368–381.

Choudhury, G. L., D. M. Lucantoni, and W. Whitt. 1996. Squeezing the Most Out of ATM. *IEEE Trans. Commun.* **44,** 203–217.

Copyright © 2001 All Rights Reserved

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

DE SMIT, J. H. A. 1983. The Queue *GI/M/s* with Customers of Different Types or the Queue *GI/H$_m$/s*. *Adv. Appl. Prob.* **15,** 392–419.

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

FRANKEN, P., D. KÖNIG, U. ARNDT, AND V. SCHMIDT. 1981. *Queues and Point Processes*. Akademie-Verlag, Berlin.

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

LUCANTONI, D. M. 1991. New Results on the Single Server Queue with a Batch Markovian Arrival Process. *Stochastic 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.

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

Copyright © 2001 All Rights Reserved