LIMITS AND APPROXIMATIONS FOR THE BUSY-PERIOD DISTRIBUTION IN SINGLE-SERVER QUEUES

by

Joseph Abate[1] *and Ward Whitt*[2]

AT&T Bell Laboratories

September 23, 1994

1 Retired: 900 Hammond Road, Ridgewood, NJ 07450-2908

2 AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636. Email: wow@research.att.com

Abstract

Limit theorems are established and relatively simple closed-form approximations are developed for the busy-period distribution in single-server queues. For the M/G/1 queue, the complementary busy-period cdf is shown to be asymptotically equivalent as $t \to \infty$ to a scaled version of the heavy-traffic limit (obtained as $\rho \to 1$), where the scaling parameters are based on the asymptotics as $t \to \infty$. We call this the asymptotic normal approximation, because it involves the standard normal cdf and density. The asymptotic normal approximation is asymptotically correct as $t \to \infty$ for each fixed ρ and as $\rho \to 1$ for each fixed t, and yields remarkably good approximations for times not too small, whereas the direct heavy-traffic ($\rho \to 1$) and asymptotic ($t \to \infty$) limits do not yield such good approximations. Indeed, even three terms of the standard asymptotic expansion does not perform well unless t is very large. As a basis for generating corresponding approximations for the busy-period distribution in more general models, we also establish a more general heavy-traffic limit theorem.

Key words: queues; busy period; M/G/1 queue; heavy traffic; diffusion approximations; Brownian motion; inverse Gaussian distribution; asymptotic expansions; relaxation time.

1. Introduction

This paper is an extension of Abate and Whitt (1988b), in which we studied the M/M/1 busy-period distribution and proposed approximations for busy-period distributions in more general single-server queues. Here we provide additional theoretical and empirical support for two approximations proposed in Abate and Whitt (1988b), the natural generalization of the *asymptotic normal approximation* in (4.3) there and the *inverse Gaussian approximation* in (6.6), (8.3) and (8.4) there. These approximations yield convenient closed-form expressions depending on only a few parameters, and they help reveal the general structure of the busy-period distribution. The busy-period distribution is known to be important for determining system behavior.

We first establish a heavy-traffic limit for the busy-period distribution in the M/G/1 queue, which involves letting $\rho \to 1$ from below, where ρ is the traffic intensity (Theorem 1). This M/G/1 result is contained in Theorem 4 of Ott (1977), but we provide a different representation and an interesting new proof. We also show that a variant of this heavy-traffic limit holds in much more general models (Theorem 6). Our heavy-traffic result for more general models complements early analysis by Rice (1962).

Next we show that asymptotics for the tail of the busy-period distribution as $t \to \infty$ in the M/G/1 queue in Section 5.6 of Cox and Smith (1961) and Section III.7.3 of Cohen (1982) can be expressed differently, in terms of a scaled version of the heavy traffic limit ((2.15) in Theorem 2). This representation is our asymptotic normal approximation. We show that it is asymptotically correct *both* as $\rho \to 1$ for each fixed t and as $t \to \infty$ for each fixed ρ less than 1. We show that it provides excellent approximations, much better than either limit separately, by making comparisons with exact numerical results for M/G/1 queues, using numerical transform inversion as in Abate and Whitt (1992a,b).

Here is how this paper is organized. We establish several M/G/1 results in Section 2. We establish the heavy-traffic limit for other models in Section 3. We make the numerical comparisons with exact M/G/1 results in Section 4. Finally, we present all proofs in Section 5.

2. M/G/1 Queue

We first consider the classical M/G/1 queue with one server, unlimited waiting space and some work-conserving discipline such as first-come first-served; see p. 249 of Cohen (1982) or Section 5.6 of Cox and Smith (1961). Customers arrive according to a Poisson process, whose rate we take to be ρ. The service times are independent and identically distributed, and independent of the arrival process. Let the service-time distribution have cdf (cumulative distribution function) $G(t)$ with mean 1 and finite second moment m_2. Thus the traffic intensity is ρ. Let $\hat{g}(s) = \int_0^\infty e^{-st}\, dG(t)$ be the Laplace-Stieltjes transform of G.

The busy period is the interval between the epoch of an arrival to an empty system and the next epoch that the system is empty again. Let $B(t)$ be the cdf of the busy period and $\hat{b}(s) = \int_0^\infty e^{-st} dB(t)$ its Laplace-Stieltjes transform. We assume that $\rho < 1$; then $B(t)$ is proper, i.e., $B(t) \to 1$ as $t \to \infty$, and it is characterized by the Kendall functional equation

$$\hat{b}(s) = \hat{g}(s + \rho - \rho\hat{b}(s)) \ . \tag{2.1}$$

Moreover,

$$B(t) = \sum_{n=1}^{\infty} \int_0^t \frac{e^{-\lambda u}\ (\lambda u)^{n-1}}{n!}\ dG_n(u)\ , \quad t \geq 0\ , \tag{2.2}$$

where $G_n(t)$ is the cdf of the n-fold convolution of $G(t)$.

For any cdf $F(t)$ with mean m, let $F^c(t) = 1 - F(t)$ be the *complementary cdf* (ccdf) and let

$$F_e(t) = m^{-1} \int_0^t F^c(u)\, du, \quad t \geq 0 , \qquad (2.3)$$

be the associated *stationary-excess cdf* (or equilibrium residual lifetime cdf). Note that $[1 - \hat{f}(s)]/sm$ is the Laplace-Stieltjes transform of F_e when $\hat{f}(s)$ is the Laplace-Stieltjes transform of F.

We characterize the heavy-traffic limit as the density $h_1(t)$ of the first-moment cdf $H_1(t)$ of regulated or reflecting Brownian motion (RBM) investigated in Abate and Whitt (1987). In particular, $H_1(t)$ is the time-dependent mean of RBM starting empty, normalized by dividing by the steady-state limit. Its density $h_1(t)$ can be expressed explicitly as

$$h_1(t) = 2t^{-1/2}\, \phi(t^{1/2}) - 2[1 - \Phi(t^{1/2})] = 2\gamma(t) - \gamma_e(t) , \quad t \geq 0 , \qquad (2.4)$$

where $\Phi(t)$ is the cdf and $\phi(t)$ is the density of a standard normal random variable with mean 0 and variance 1, $\gamma(t)$ is the gamma density with mean 1 and shape parameter 1/2, i.e.,

$$\gamma(t) = (2\pi t)^{-1/2} \exp(-t/2), \quad t \geq 0 , \qquad (2.5)$$

and $\gamma_e(t)$ is the associated stationary-excess density. From (2.4) we see that $h_1(t)$ is in convenient closed form; i.e., it is easy to evaluate directly, e.g., using rational approximations for the normal cdf $\Phi(t)$, e.g., 26.2.17 of Abramowitz and Stegan (1972).

The density $h_1(t)$ also has several other useful characterizations. It is the density of the equilibrium time to emptiness for RBM, i.e., the density of the first passage time to zero starting with the exponential stationary distribution. In other words, it is an *exponential mixture of inverse Gaussian* densities (an EMIG): see Section 8 of Abate and Whitt (1995). The moment cdf $H_1(t)$ is the only cdf on $[0, \infty)$ with mean 1/2 for which the two-fold convolution coincides with stationary-excess cdf, i.e., for which the transforms satisfy

$$\hat{h}_1(s)^2 = 2[1 - \hat{h}_1(s)] ; \qquad (2.6)$$

see Sections 1.2 and 1.3 of Abate and Whitt (1987).

Our heavy-traffic limit is obtained by simply increasing the arrival rate ρ. It is possible to consider more general limits in which the service-time distributions also change with ρ, but as can be seen from Ott (1977) the same limiting behavior holds in considerable generality. To obtain our heavy-traffic limit, we scale both inside (time) and outside the complementary cdf $B_\rho^c(t)$. We introduce the subscript ρ to indicate the dependence upon ρ. All proofs appear in Section 5.

Theorem 1. For each $t > 0$,

$$\lim_{\rho \to 1} m_2 (1 - \rho)^{-1} B_\rho^c (t m_2 (1 - \rho)^{-2}) = h_1 (t) . \tag{2.7}$$

Theorem 1 can be obtained from (1.32) of Ott (1977) by letting his parameters be $\eta_n = (1 - \rho_n)^{-1}$, $\lambda_n = m_2 (1 - \rho_n)^{-1}$, $\mu_n = (1 - \rho_n) \rho_n / m_2$ and $a = \sigma = 1$, and by identifying his integral limit with $h_1 (t)$. However, we give a different proof.

The scaling in (2.7) is very important to establish the connection to RBM. Indeed, without the scaling, $B_\rho^c (t)$ is continuous in ρ for all $\rho > 0$ for each fixed t, so that the boundary for stability $\rho = 1$ plays no special role without scaling. Moreover, the behavior of $B_\rho^c (t)$ for small t obviously depends strongly on the form of the service-time distribution, but Theorem 1 shows that for suitably large t it does not. See Abate and Whitt (1988b) for more discussion.

Understanding of Theorem 1 is enhanced by recognizing that the left side of (2.7) is a scaled version of the *density* of the busy-period stationary-excess cdf, which in turn is a time-scaled version of the density of the equilibrium time to emptiness in the M/G/1 model conditional on the system not being empty; i.e.,

$$h_\rho (t) \equiv b_{\rho e} (t m_2 (1 - \rho)^{-2}) = m_2 (1 - \rho)^{-1} B_\rho^c (t m_2 (1 - \rho)^{-2}) , \quad t \geq 0 . \tag{2.8}$$

Theorem 1 thus can be regarded as a *local limit theorem* establishing convergence of the time-scaled M/G/1 conditional equilibrium-time-to-emptiness density $h_\rho(t)$ to the RBM equilibrium-time-to-emptiness density $h_1(t)$. (The M/G/1 conditioning event has probability ρ and thus converges to 1 as $\rho \to 1$.) As a consequence of the Lebesgue dominated convergence theorem, p. 111 of Feller (1971), plus inequality (2.1) below, we also obtain convergence of the associated scaled conditional equilibrium-time-to-emptiness cdf's from Theorem 1. The form of the limit comes from Corollary 1.1.1 and (4.3) of Abate and Whitt (1987).

Corollary. For each $t \geq 0$,

$$\lim_{\rho \to 1} H_\rho(t) = H_1(t) = 1 - 2(1+t)[1 - \Phi(t^{1/2})] + 2t^{1/2}\,\phi(t^{1/2}) \ .$$

We now turn to the asymptotic behavior as $t \to \infty$. Let $f(t) \sim g(t)$ as $t \to \infty$ mean that $f(t)/g(t) \to 1$ as $t \to \infty$. Assume that the busy-period cdf has a density $b(t)$. Under considerable generality, see (49) on p. 156 of Cox and Smith (1961) or (11)–(13) of Abate, Choudhury and Whitt (1994),

$$b(t) \sim \frac{\alpha e^{-t/2\beta}}{\sqrt{2\pi\beta t^3}} = \alpha\, t^{-1}\, \beta^{-1}\, \gamma\,(t/\beta) \quad as \quad t \to \infty \ , \tag{2.9}$$

so that

$$B^c(t) \sim 2\beta b(t) \sim 2\alpha t^{-1} \gamma(t/\beta) \text{ as } t \to \infty \ , \tag{2.10}$$

where $\gamma(t)$ is the gamma density in (2.5) and α and β are constants depending on ρ and $G(t)$. In particular, $\beta = \tau/2$ where τ is the *relaxation time*, with

$$\tau^{-1} = \rho + \zeta - \rho\hat{g}(-\zeta) \ , \tag{2.11}$$

where ζ is the unique real number u satisfying the equation

$$\hat{g}'(-u) = -\rho^{-1} \tag{2.12}$$

when it exists, which we assume is the case. In general, (2.12) need not have a solution, in which case (2.9) does not hold; we give an example in Section 6. The parameter α in (2.9) and (2.10) is

$$\alpha = [\rho^3 \beta^{-1} \hat{g}''(-\zeta)]^{-1/2} . \tag{2.13}$$

The following result is obtained by simply integrating both sides of (2.9) over the interval (t,∞). The key is to recognize that the right side is indeed integrable and then identify what that integral is. For this purpose, note that $\beta^{-1}\gamma(t/\beta)$ is a density function and, from (2.4), that the derivative of $h_1(t)$ has the remarkably simple form

$$h_1'(t) = -t^{-1}\gamma(t) , \quad t \geq 0 , \tag{2.14}$$

so that $h_1(t) \sim 2t^{-1}\gamma(t)$ as $t \to \infty$.

Theorem 2. If (2.9) holds, then

$$B_\rho^c(t) \sim \alpha\beta^{-1} \, h_1(t/\beta) \quad as \quad t \to \infty . \tag{2.15}$$

for $h_1(t)$ in (2.4) and α and β in (2.9)–(2.13).

Integrating over the interval (t,∞) once again, we obtain the following result from (2.15).

Corollary. If (2.15) holds, then

$$H_\rho^c(t) \sim \alpha \, H_1^c \, (t(1-\rho)^2 / m_2\beta) \quad as \quad t \to \infty \tag{2.16}$$

for $h_1(t)$ in (2.4) and α and β in (2.9)–(2.13).

We previously suggested the approximation (2.15) for the M/M/1 queue in (4.3) of Abate and Whitt (1988b); as before, we call it the *asymptotic normal approximation*, because it uses the normal density and cdf in (2.4). The general idea of using $h_1(t)$ for approximations seems to have been first proposed for the M/M/1-LIFO waiting-time distribution by Riordan (1962), p. 109, but it does not seem to have been pursued.

The difference between the two asymptotic expressions in (2.15) and (2.16) is due to the fact $H_\rho^c(t)$ has been time scaled while $B^c(t)$ has not. The expressions in Theorem 2 are to be contrasted with the standard asymptotic expansions, which are of the form

$$B^c(t) \sim e^{-t/2\beta} (\delta_1 t^{-3/2} + \delta_2 t^{-5/2} + \delta_3 t^{-7/2} + O(t^{-9/2})) \quad \text{as} \quad t \to \infty , \tag{2.17}$$

where δ_i are constants, with

$$\delta_1 = \alpha \sqrt{2\beta/\pi} \tag{2.18}$$

from (2.9); see Section 4 of Abate and Whitt (1988b). As shown for the M/M/1 queue in Abate and Whitt (1988b), (2.15) is a vastly superior approximation than the first few terms of (2.17). Interestingly, the direct asymptotics for the busy-period *density* yields a better approximation than the direct asymptotics for the busy-period *cdf* (e.g., see Table 1 of Abate, Choudhury and Whitt (1994)), and this good quality is inherited by the integral. The integral $h_1(t)$ in (2.15) has structure not inherited by its asymptotic form.

The good performance of (2.15) can be partly explained theoretically, because it is asymptotically exact as both $\rho \to 1$ for any fixed t (Theorem 1) and as $t \to \infty$ for any fixed ρ (Theorem 2). To see the connection to Theorem 1, we need to know how β and α behave as $\rho \to 1$. As shown in Abate, Choudhury and Whitt (1994),

$$\beta^{-1} = \frac{(1-\rho)^2}{m_2} \left[1 + (1-\rho)(1-\xi) + (1-\rho)^2 [1-\xi(2-(9/4)\xi)-\psi] + O((1-\rho)^3 \right] \tag{2.19}$$

and

$$\alpha = (1-\rho)^{-1} (1 + (1-\rho)(1-\xi) + O((1-\rho)^2) \tag{2.20}$$

as $\rho \to 1$, where $\xi = m_3/3m_2^2$, $\psi = m_4/12m_2^3$ and m_k is the k^{th} moment of the service time. Hence, we see that

$$\beta\alpha^{-1} \beta_\rho^c(\beta t) \sim m_2 (1-\rho)^{-1} \beta_\rho^c (t m_2 (1-\rho)^{-2}) \sim h_1(t) \tag{2.21}$$

as $\rho \to 1$ for each fixed positive t.

In Theorem 3.5 of Abate and Whitt (1988a) convergence was also established for the normalized M/M/1 busy-period density function as $\rho \to 1$. We also obtain such a result for M/G/1 under extra conditions. First, the busy-period cdf must have a density. A sufficient condition is for the service-time cdf $G(t)$ to be absolutely continuous. If the service-time cdf $G(t)$ is absolutely continuous with a density $g(t)$, then so are all n-fold convolutions G_n, p. 146 of Feller (1971). Thus, from (2.2) and Fubini, p. 111 of Feller (1971), $B(t)$ is absolutely continuous with density

$$b(t) = \sum_{n=1}^{\infty} \frac{e^{-\lambda t} (\lambda t)^{n-1}}{n!} g_n(t) , \quad t \geq 0 , \tag{2.22}$$

where $g_n(t)$ is the density of $G_n(t)$, from which we see that $b(0) = g(0)$ and $b'(0) \equiv g'(0)$.

Theorem 3. Suppose that the service-time cdf $G(t)$ is absolutely continuous with density $g(t)$, so that the busy-period cdf $B(t)$ is absolutely continuous with density $b(t)$, where $b(0) = g(0)$. If $b(0) < \infty$ and $b(t)$ is monotone, then

$$\lim_{\rho \to 1} m_2^2 (1 - \rho)^{-3} b(t m_2 (1 - \rho)^{-2}) = h_1'(t) = (2\pi t^3)^{-1/2} e^{-t/2} , \quad t > 0 . \tag{2.23}$$

Since $b'(0) = g'(0)$, where in general this is understood to be a one-sided derivative, a *necessary* condition in order for $b(t)$ to be monotone is $g'(0) < 0$. Keilson (1978) has shown that $b(t)$ is completely monotone (a mixture of exponentials) and thus monotone if $g(t)$ is completely monotone. Hence, a *sufficient* condition for Theorem 3 is the complete monotonicity of the service-time density. However, under this condition we can establish an even stronger result. For any function $f(t)$, let $f^{(k)}(t)$ be the k$^{\text{th}}$ derivative of f at t.

Theorem 4. If the service-time density $g(t)$ is completely monotone, then for all $k \geq 0$

$$\lim_{\rho \to 1} m_2^{2+k} (1 - \rho)^{-(3+2k)} b^{(k)}(t m_2 (1 - \rho)^{-2}) = h_1^{(k)}(t) , \quad t \geq 0 . \tag{2.24}$$

Theorem 4 describes a remarkable degree of local convergence. However, the good behavior is easy to understand via the complete monotonicity. By Theorem 2.1 of Keilson (1978), $b(t)$ in (2.22) and thus $B^c(t)$ and $h_\rho(t)$ in (2.8) are completely monotone when $g(t)$ is completely monotone; i.e., for each ρ, $0 < \rho < 1$,

$$h_\rho(t) = \int_0^\infty x^{-1} e^{-t/x} \, dW_\rho(x) \; , \quad t \geq 0 \; , \tag{2.25}$$

for some mixing cdf $W_\rho(x)$. Theorem 4 follows easily from a limit theorem for the mixing cdf's.

Theorem 5. If the service-time density $g(t)$ is completely monotone, so that $h_\rho(t)$ in (2.8) admits the spectral representation (2.25), then for each $x > 0$

$$\lim_{\rho \to 1} W_\rho(x) = W_1(x) = \int_0^x w_1(u) \, du \; , \tag{2.26}$$

where

$$w_1(x) = \begin{cases} 0 & x > 2 \; , \\ \dfrac{\sqrt{2-x}}{\pi\sqrt{x}} \; , & 0 \leq x \leq 2 \end{cases} \tag{2.27}$$

is the mixing density of $h_1(t)$ in (1.4).

We know the limit in Theorem 5 because we derived the spectral representation for $h_1(t)$ in Theorem 4.2 of Abate and Whitt (1988c). Explicit spectral representations are also given there for $b(t)$ and $h_\rho(t)$ in the M/M/1 model. However, in general we do not know the M/G/1 mixing cdf $W_\rho(x)$ in (2.25). From pp. 610-613 of Cohen (1982), we can identify the spectrum (the smallest interval containing the support) of $W_\rho(x)$ in any specific case.

3. More General Queues

For more general models, we propose again using the asymptotic normal approximation (2.15), where the parameters α and β are determined from analogs of the asymptotic expansion (2.10). Such asymptotic expansions can be obtained either analytically or numerically using transform inversion as in Choudhury and Lucantoni (1994). Rice (1962) in his (75) provides support for obtaining the asymptotic form (2.10) more generally at least for suitably large ρ. He expands the transforms in power series and obtains the form (2.10) from a square root transform expression.

We now establish a generalization of the heavy-traffic limit in Theorem 1 for a much larger class of single-server queues. As before we assume that some work-conserving discipline is used and that the mean service time is 1. For non-Poisson arrival processes, we assume that the system indexed by ρ is obtained by simply scaling a rate-one counting process $\{A(t) : t \geq 0\}$, i.e., $A_\rho(t) \equiv A(\rho t), t \geq 0$. We have two general conditions, one on the mean busy period and the other on the stationary workload process. Our conditions are in some sense not too appealing, because they are not directly for the elements of the model (e.g., for the interarrival-time and service-time distributions), but upon the descriptive quantities of interest. However, our conditions are intuitively appealing because they clearly reveal what needs to be verified in applications and they can indeed be verified in special cases, as we show.

Let B_ρ be the busy period in the model with traffic intensity ρ. The busy period is understood to mean the interval from when the server first becomes busy until the server is again idle. For models more general than GI/G/1, we can interpret this distribution as the long-run average of all such distributions over all busy periods.

Condition C1. For some constant b, $(1-\rho) EB_\rho \to b$ as $\rho \to 1$.

Let $\{W_\rho^*(t) : t \geq 0\}$ be the stationary workload process in the queue with traffic intensity ρ.

Here $W_\rho^*(t)$ should be interpreted as the time required for the system to become empty after time t if no new work were to arrive after time t. Let $\{R^*(t):t \geq 0\}$ be a stationary version of canonical RBM with drift coefficient -1 and diffusion coefficient 1. The stationary version is initialized by the exponential steady-state distribution with mean $1/2$. Let $\Rightarrow$ denote convergence in distribution or weak convergence. Let $D[0,\infty)$ be the function space of right-continuous real-valued functions with left limits, endowed with the usual Skorohod J_1 topology; e.g., see Ethier and Kurtz (1986).

Condition C.2. For some constant $d, \{W_\rho^*(dt(1-\rho)^{-2}):t \geq 0\} \Rightarrow \{R^*(t):t \geq 0\}$ in $D[0,\infty)$ as $\rho \rightarrow 1$.

We prove the following generalization of Theorem 1 in Section 5.

Theorem 6. If conditions C1 and C2 hold, then

$$(d/b)(1-\rho)^{-1} B_\rho^c(dt(1-\rho)^{-2}) \rightarrow h_1(t) \;\; as \;\; \rho \rightarrow 1 \tag{3.1}$$

for each t.

For the M/G/1 queue, conditions C1 and C2 are known to hold with $b = 1$ and $d = m_2 = c_s^2 + 1$, where c_s^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of a service time. Hence, Theorem 6 actually contains Theorem 1 as a special case, but our proofs are somewhat different.

For GI/G/1 queues with mean service time 1, the mean busy period coincides with the reciprocal of the probability that an arrival finds an empty queue. For the M/G/1 queue, this probability is just $1-\rho$, but for other models it is more complicated. For the GI/M/1 queue, Halfin (1985) showed in his (4.6) that condition C1 holds with

$$b = \frac{c_a^2 + 1}{2} \, , \tag{3.2}$$

where c_a^2 is the SCV of the interarrival time. We now provide a general sufficient condition for condition C1 for GI/G/1 models by applying a result of Kella and Taksar (1994) about idle times. Let $\rho^{-1} U$ be an interarrival time in the GI/G/1 system with traffic intensity ρ.

Theorem 7. If $EU^{2+\varepsilon} < \infty$ *in the GI/G/1 model for some positive* ε*, then condition C1 holds.*

For the standard GI/G/1 queue, it is not difficult to establish condition C2 provided that the second moments of the interarrival-time and service-time distributions are finite. Heavy-traffic limits for the steady-state distributions were established by Kingman (1961,1962) and Szczotka (1990), while weak convergence theorems for workload processes given converging initial distributions were established by Iglehart and Whitt (1970) and Whitt (1971). In addition, here it is necessary to verify that the residual interarrival-time and service-time distributions associated with the stationary initial conditions have negligible effect on the heavy-traffic behavior. Since these residual distributions are stochastically bounded, they indeed have negligible effect; we omit the details. Then

$$d = c_a^2 + c_s^2 . \tag{3.3}$$

More generally, d appears as the variance constant in the process representing the total input of work; i.e., we obtain d from the limit

$$\frac{I(t) - t}{\sqrt{dt}} \Rightarrow N(0,1) \quad \text{as} \quad t \to \infty , \tag{3.4}$$

where

$$I(t) = \sum_{i=1}^{A(t)} V_i , \quad t \geq 0 , \tag{3.5}$$

V_i is the i^{th} service time and $N(0,1)$ is a standard (mean 0, variance 1) normal random variable.

Combining (3.2) and (3.3), we see that the heavy-traffic limit for GI/M/1 depends on the general interarrival-time distribution only through its first two moments. However, this nice property that holds for M/G/1 and GI/M/1 does *not* hold for general GI/G/1 queues. More generally, the busy-period mean is a relatively difficult quantity to obtain. For the $K_m/G/1$ queue, we can deduce that condition C2 holds and we can calculate b from (5.205) on p. 330 of Cohen (1982). From this expression, we see that b depends on the interarrival-time and service-time distributions beyond their first two moments. In particular, it depends on the m roots of the transform equation

$$\hat{\alpha}(-s)\hat{\beta}(s) = 1 \tag{3.6}$$

where $\hat{\alpha}(s)$ and $\hat{\beta}(s)$ are the Laplace-Stieltjes transforms of the interarrival-time and service-time distributions. For many other models the mean busy period can be calculated numerically. For more on the GI/G/1 busy period, see Cohen (1982), Kingman (1962) and Rice (1962).

For practical purposes, we suggest using the Kraemer and Langenbach-Belz (1976) approximation, also given in (49) of Whitt (1983),

$$E\beta_\rho = \frac{1}{P(W_\rho = 0)} \approx \frac{1}{(1-\rho)(1-\rho(c_a^2-1)h(\rho, c_a^2, c_s^2)} , \tag{3.7}$$

where

$$h(\rho, c_a^2, c_s^2) = \begin{cases} \dfrac{1+c_a^2+\rho c_s^2}{1+\rho(c_s^2-1)+\rho^2(4c_a^2+c_s^2)} , & c_a^2 \le 1 \\ \dfrac{4\rho}{c_a^2+\rho^2(4c_a^2+c_s^2)} , & c_a^2 \ge 1 . \end{cases} \tag{3.8}$$

From (3.7) and (3.8), we obtain

$$b \approx \frac{1}{1-(c_a^2-1)h(1, c_a^2, c_s^2)} . \tag{3.9}$$

For insights into the way the mean EB depends on the parameters c_a^2 and c_s^2, see Whitt (1984).

We remark that Theorem 6 is consistent with (75) of Rice (1962). His approximate asymptotic formula for the busy-period density can be obtained from (3.1) by first taking the derivative and then letting $t \to \infty$. The corresponding formula for $B^c_\rho(t)$ is

$$
\begin{aligned}
B^c_\rho(t) &\approx \frac{(b(1-\rho)}{d} h_1(t(1-\rho)^2/d) \\
&\approx (b/(1-\rho)^2)\sqrt{2d/\pi t^3} \exp(-t(1-\rho)^2/2d) \\
&\approx EB\sqrt{D/\pi t^3}\, e^{-Dt} \qquad (3.10)
\end{aligned}
$$

for $D \equiv (1-\rho)^2/2d$ and $EB = P(W = 0)^{-1} \approx b/(1-\rho)$. Formula (3.10) is intended for high ρ and large t.

Theorem 6 provides a pure heavy-traffic approximation for $B^c(t)$ in very general single-server queues. However, we do *not* regard (3.1) as our principal proposed approximation for $B^c(t)$. Our actual proposed approximation is (2.15) for α and β determined by (2.10), assuming that (2.10) holds for the more general model. We intend to discuss asymptotics of the form (2.10) for other GI/G/1 models in a future paper. For GI/M/1 the asymptotics can be obtained by exploiting the duality between M/G/1 and GI/M/1; see (77) of Takács (1967). More generally, we can obtain the desired parameters α and β in (2.10) numerically using the inversion algorithm in Choudhury and Lucantoni (1994). Rice's (1962) formula (3.10) provides support for both (2.10) and the asymptotic behavior of the parameters α and β as in (2.19) and (2.20).

4. Numerical Examples

This section extends the numerical investigation of approximations for the busy-period ccdf $B^c(t)$ done for the M/M/1 queue in Abate and Whitt (1988b) to M/G/1 queues. Our previous investigation showed that even three terms of the asymptotic expansion (2.17) yields a remarkably poor approximation; see Table 10 there. Hence, we do *not* consider the approximations for $B^c(t)$ in the M/G/1 queues based on the asymptotics as $t \to \infty$ in (2.10) or (2.17).

Here we consider three candidate approximations. First, we consider the *pure heavy-traffic approximation* obtained from Theorem 1, namely,

$$B^c(t) \approx (1-\rho) m_2^{-1} h_1((1-\rho)^2 t/m_2) \ . \tag{4.1}$$

Formula (4.1) is obtained from (2.7) by moving the normalizing constants to the righthand side.

Our second approximation is the *asymptotic normal approximation* provided by Theorem 2, i.e., (2.15). The heavy-traffic approximation can be regarded as an approximation to the asymptotic normal approximation in which the asymptotic parameters β and α from (2.9)—(2.13) in (2.15) are replaced by the first terms in their heavy-traffic expansions in (2.19) and (2.20). Thus, we can see how these first two approximations differ by evaluating the quality of the one-term approximations in (2.19) and (2.20). As we would anticipate, these approximations get closer as ρ increases, but the pure heavy-traffic approximation has significantly bigger errors for lower values of ρ.

The third approximation considered here is the *inverse Gaussian (IG) approximation* in (6.6) and (8.3) of Abate and Whitt (1988b)

$$B^c(t) \approx IG^c((1-\rho)^2 t/(1+c_s^2);\ \nu, x) \ , \tag{4.2}$$

where

$$IG^c(t;\nu,x) = \Phi^c((t-x)/\sqrt{\nu t}) - e^{2x/\nu}\, \Phi^c((t+x)/\sqrt{\nu t}) \tag{4.3}$$

with $\Phi^c(x)$ the normal cdf and

$$x = \frac{1-\rho}{1 + c_s^2} \quad \text{and} \quad \nu = 1-x \ . \tag{4.4}$$

This scaling matches the first two moments.

Given that RBM is a natural heavy-traffic approximation for the workload process, the IG approximation is a natural approximation for the busy-period distribution, since it is a first-passage time distribution for RBM. This idea was the basis for an IG approximation proposed by Heyman (1974), but our IG approximation is a significant improvement, both because it is closed-form and because it yields better results, as shown for the M/M/1 queue before. For the M/M/1 queue, the asymptotic normal and IG approximations were the leading approximations among a fairly large set, with the asymptotic normal approximation performing better for large times and the IG approximation performing better for small times; see Tables 10 and 11 of Abate and Whitt (1988b).

Our numerical experience here for M/G/1 queues with other service-time distributions confirms our previous experience for M/M/1 queues. To illustrate, we display numerical results for $B^c(t)$ for two different service-time distributions and three traffic intensities. The service-time distributions are E_4, the four-stage Erlang with $c_s^2 = 0.25$, and $\Gamma_{1/2}$, the gamma density in (2.5) with shape parameter 1/2 and, thus, $c_s^2 = 2$. Clearly, E_4 is less variable than an exponential, while $\Gamma_{1/2}$ is more variable than an exponential. As before the mean service time is always 1. The three traffic intensities are: 0.5, 0.75 and 0.9.

The exact values of $B^c(t)$ and the three approximations are given for the six cases in Tables 1–6. The exact values are obtained by numerical transform inversion, using Abate and Whitt (1992a,b). Unlike for the M/M/1 queue, here we do not scale time within $B^c(t)$; the different tables would be more closely related if we did. As before, the asymptotic normal approximation in (2.15) performs remarkably well for times not too small, and all approximations improve as ρ increases.

The two service-time distributions we consider are both gamma distributions. In general, the gamma service-time transform is

$$\hat{g}(s;\omega) = (1 + s/\omega)^{-\omega} \tag{4.5}$$

for $\omega > 0$, where ω is the shape parameter and the mean is fixed at 1. The moments satisfy the recursions: $m_1 = 1$ and $m_{k+1} = (1+k/\omega) m_k$, $k \geq 1$. For the gamma transform in (4.5), the root of (2.12) is

$$\zeta = \omega(1-\rho^{1/(1+\omega)}) , \qquad (4.6)$$

so that the asymptotic parameters in (2.10) are

$$\beta^{-1} = 2(\rho + \omega-(1+\omega)\rho^{1/(1+\omega)}) \qquad (4.7)$$

and

$$\alpha = \rho^{-q} \sqrt{\beta/(1+\omega^{-1})} , \text{ where } q = (2\omega + 1)/2(\omega + 1) . \qquad (4.8)$$

We obtain deterministic (D) service by letting $\omega \to \infty$ in (4.5); i.e., then $\hat{g}(s;\omega) \to e^{-s}$, $\zeta = -\log \rho$,

$$\beta = 2(\log(\rho^{-1})-(1-\rho)) = (1-\rho)^2 \sum_{k=0}^{\infty} (1-\rho)^k/(1+k/2) \qquad (4.9)$$

and

$$\alpha = \sqrt{\beta/\rho} . \qquad (4.10)$$

Our two numerical examples involve the special cases $\omega = 4$ (E_4) and $\omega = 1/2$ ($\Gamma_{1/2}$). The first four moments for E_4 are 1, 5/4, 15/8 and 105/32 and for $\Gamma_{1/2}$ are 1, 3, 15, 105. The auxiliary parameters in (2.19) are $\xi = 2/5$ and $\gamma = 7/50$ for E_4 and $\xi = 5/9$ and $\gamma = 35/108$ for $\Gamma_{1/2}$.

We have noted that the heavy-traffic approximation is equivalent to the asymptotic normal approximation with the first terms of the heavy-traffic expansions for β^{-1} and α in (2.19) and (2.20). Refined heavy-traffic approximations can be obtained by using more terms in (2.19) and (2.20). Let β_k be the approximation of β based on k terms of the heavy-traffic asymptotic expansion for β in (1.19), and similarly for α. Table 7 shows the quality of these approximations for β for different values of ρ for the $\Gamma_{1/2}$ and D service. We use D because from (2.19) we see that the single term β_1 performs worst in that case because $1-\xi$ is largest for that case. We do not show any refined approximations in Tables 1–6. The approximations based on (α_2, β_2) and (α_2, β_3) are successive improvements over the basic heavy-traffic approximation based on (α_1, β_1). They fall between the heavy-traffic approximation based on (α_1, β_1) and the asymptotic normal approximation based on (α, β). The (α_2, β_3) refined approximation tends to be essentially the same as the asymptotic normal approximation at $\rho = 0.75$, but not at $\rho = 0.25$.

We might also evaluate the asymptotic normal approximation from a moment or integral-average point of view; i.e., we can ask about the quality of the approximation

$$\int_0^\infty B^c(t)\, dt \approx \alpha \int_0^\infty t \beta^{-1} h_1(t/\beta)\, dt \;, \tag{4.11}$$

from which we get

$$b_2/2 \approx \alpha\beta/2 \;, \tag{4.12}$$

where b_k is the k^{th} busy-period moment. However, $b_2 = \alpha_1 \beta_1$ and, by (2.19) and (2.20),

$$\alpha\beta = \alpha_1 \beta_1 (1 + O((1-\rho)^2) \quad \text{as} \quad \rho \to 1 \;, \tag{4.13}$$

so that the error in (4.11) is only $O((1-\rho)^2)$ as $\rho \to 1$.

5. Proofs

Proof of Theorem 1. By Chebychev's inequality using the first moment, p. 152 of Feller (1971), $B^c(t) \leq 1/t(1 - \rho)$ and

$$h_\rho(t) = m_2(1 - \rho)^{-1} B^c(t m_2 (1 - \rho)^{-2}) \leq 1/t \tag{5.1}$$

for all t and ρ. Since $B^c(t)$ is monotone, we can thus apply the Helly selection theorem, p. 267 of Feller (1967), to conclude that any sequence $\{h_{\rho_n}(t) : n \geq 1\}$ with $\rho_n \to 1$ has a subsequence that converges to a monotone function $f(t)$ (depending on the subsequence) with $0 \leq f(t) \leq 1/t$, where the convergence is pointwise at all continuity points of f. We establish convergence to h_1 by showing that h_1 is the only possible limit for a convergent subsequence. To do this we work with the transforms and the functional equation (2.1).

We begin by expressing the busy-period functional equation (2.1) in terms of the busy-period stationary-excess transform $\hat{b}_e(s)$. First,

$$\hat{b}(s) = \hat{g}(s + s\rho(1-\rho)^{-1}\hat{b}_e(s)) \tag{5.2}$$

and then

$$\hat{b}_e(s) = \frac{(1-\rho)[1-\hat{b}(s)]}{s} = \frac{(1-\rho)}{s}(1-\hat{g}(s + s\rho(1-\rho)^{-1}\hat{b}_e(s)) \ . \tag{5.3}$$

We then change the time scale to obtain form (2.8) and (5.3)

$$\hat{h}_\rho(s) = \hat{b}_e((1-\rho)^2 m_2^{-1} s) = \frac{m_2}{(1-\rho)s}\left[1 - \hat{g}\left[\frac{(1-\rho)^2 s}{m_2}\left[1 + \frac{\rho}{1-\rho}\hat{h}_\rho(s)\right]\right]\right] \ . \tag{5.4}$$

Now we assume $\hat{h}_\rho(s) \to \hat{f}(s)$ as $\rho \to 1$ for some subsequence, and show that we must have $\hat{f}(s) = \hat{h}_1(s)$. Note that the service-time distribution does not change with ρ, but the argument of $\hat{g}$ in (5.4) is getting small as $\rho \to 1$. Since the service-time distribution has a finite second moment,

$$\hat{g}(s) = 1 - s + \frac{m_2 s^2}{2} + o(s) \quad \text{as} \quad s \to 0 \ . \tag{5.5}$$

Expanding $\hat{g}$ in (5.4), we obtain

$$\hat{h}_\rho(s) = 1 - \rho + \rho\hat{h}_\rho(s) - \frac{\rho^2(1-\rho)s\hat{h}_\rho(s)^2}{2} + o(1-\rho) \tag{5.6}$$

so that, after subtracting $\rho \hat{h}_\rho(s)$ from both sides,

$$\hat{h}_\rho(s) = 1 - \frac{\rho^2 s \hat{h}_\rho(s)^2}{2} + o(1) \quad \text{as} \quad 1-\rho \rightarrow 0 \tag{5.7}$$

and any limit $\hat{f}(s)$ must satisfy (2.6), which implies that $\hat{f}(s) = \hat{h}_1(s)$. (By Corollary 1.5.2 of Abate and Whitt (1987), $\hat{h}_2(s)$ is the unique solution to (1.6).)

Proof of Theorem 3. Since $b(t)$ is monotone with $b(0) = g(0) < \infty$, $g(0)^{-1} b(t)$ can be regarded as a complementary cdf, say $C^c(t)$, where $C_e(t) = B(t)$. Hence, $C(t)$ has mean $g(0)^{-1}$ and second moment $2/(1 - \rho)\, g(0)$. As in the proof of Theorem 1, apply Chebychev's inequality, but now with the second moment, to conclude that $C^c(t) \leq 2/t^2 (1 - \rho)\, g(0)$ and

$$h'_\rho(t) = m_2^2 (1 - \rho)^{-3} b(t m_2 (1 - \rho)^{-1}) \leq 2/t^2 g(0) \tag{5.8}$$

for all t an ρ. Hence, any sequence $\{h'_{\rho_n}(t) : n \geq 1\}$ with $\rho_n \rightarrow 1$ has a subsequence that converges to a monotone function $f(t)$ with $0 \leq f(t) \leq g/t^2 g(0)$ for all t. Given some convergent subsequence, apply the Lebesgue dominated convergence theorem, p. 111 of Feller (1971), on $[\varepsilon, \infty)$ for any $\varepsilon > 0$ to get the integrals to converge. Hence, the limit $f(t)$ must satisfy $\int_{t_1}^{t_2} f(u)\, du = h_1(t_2) - h_1(t_1) = \int_{t_1}^{t_2} h'_1(u)\, du$ for all $t_2 > t_1 > \varepsilon$, so that $t(t) = h'_1(t)$; i.e., f of any convergent subsequence must be $h'_1(t)$. Hence, the proof is complete.

Proof of Theorem 4. The proof of Theorem 3 can be extended by induction: Since $b(t)$ is completely monotone with $b^{(k)}(0) = g^{(k)}(0) < \infty$ for all k, $g^{(k)}(0) b^{(k)}(t)$ can be regarded as a complementary cdf, cay $C_k^c(t)$ for each k, where $C_{ke}(t) = C_{k-1}(t)$. Hence, if m_{kj} denotes the j^{th} moment of C_k, then $m_{k,j} = m_{k-1,j+1}/(j + 1) m_{k-1,1}$. As in the proofs of Theorems 1 and 3, apply Chebychev's inequality, but now with the $(k + 1)^{\text{st}}$ moment to conclude that $C_k^c(t) \leq K_k / t^{2+k} (1 - \rho)$ for a constant K_k independent of ρ. The rest of the proof follows the proof of Theorem 3.

Proof of Theorem 5. Since the service-time cdf $G(t)$ has mean 1 and finite second moment m_2, the busy-period cdf has mean $1/(1 - \rho)$ and second moment $m_2/(1 - \rho)^3$. Thus $h_\rho(t)$ in (2.8) is a density with mean $1/2$ for all ρ. Consequently,

$$1/2 = \int_0^\infty t h_\rho(t)\, dt = \int_0^\infty dt \int_0^\infty t x^{-1} e^{-t/x}\, dW(x) = \int_0^\infty x dW_\rho(x)\ . \tag{5.9}$$

Hence, the family of cdf's $\{W_\rho(x) : 0 < \rho < 1\}$ is tight or stochastically bounded, p. 254 of Feller (1971), so that every subsequence has a convergent subsequence with a proper limit, p. 267 of Feller (1971). Since $h_\rho(t) \to h_1(t)$ as $\rho \to 1$ for each $t > 0$, the associated cdf's converge, i.e., $H_\rho(t) \to H_1(t)$ as $\rho \to 1$ for each $t > 0$, as noted in the Corollary to Theorem 1, but the cdf's can be regarded as Laplace-Stieltjes transforms of the mixing cdf's. Hence, if $W_{\rho_n} \to W$ as $\rho_n \to 1$, then we must have $W = W_1$. Thus every convergent subsequence with $\rho_n \to 1$ from $\{W_\rho(x) : 0 < \rho < 1\}$ must converge to W_1 in (2.26). Hence, the mixing cdf's must converge as stated in (2.26).

Proof of Theorem 4 from Theorem 5. In terms of the mixing cdf $W_\rho(x)$ in (2.25), the k^{th} derivative of $h_\rho(t)$ is

$$h_\rho^{(k)}(t) = \int_0^\infty x^{-(k+1)} e^{-t/x}\, dW_\rho(x)\ , \quad t \geq 0\ . \tag{5.10}$$

Since the function $x^{-(k+1)} e^{-t/x}$ is a modification of a gamma density, it is continuous and bounded. Since $W_\rho(x) \to W_1(x)$ as $\rho \to 1$ for each x, $h_\rho^{(k)}(t) \to h_1^{(k)}(t)$ as $\rho \to 1$ for each k and t; p. 249 of Feller (1971).

Proof of Theorem 6. Just as in the proof of Theorem 1, we can apply Chebychev's inequality and condition C1 to deduce that $B_\rho^c(t) \leq K/t(1-\rho)$ for some $K > b$ and ρ suitably larger. Hence

$$(d/b)(1-\rho)^{-1} B_\rho^c(dt(1-\rho)^{-2}) \leq K/bt \tag{5.11}$$

for all t and for all ρ suitably close to 1. As in the proof of Theorem 1, we have established relative compactness, i.e., every sequence $\{ (d/b)(1-\rho_n)^{-1} B^c_{\rho_n} (dt(1-\rho_n)^{-2}) : n \geq 1 \}$ obtained by taking $\rho_n \rightarrow 1$ has a further subsequence converging to a monotone limit. It thus remains to show that any such limit must be $h_1(t)$.

For this purpose, consider a subsequence converging to some $f(t)$. We then obtain convergence of the associated integrals

$$(d/b)(1-\rho)^{-1} \int_0^t B^c_\rho (du(1-\rho)^{-2})\, du \rightarrow \int_0^t f_1(u)\, du \equiv F(t) \ , \qquad (5.12)$$

but the left side of (5.12) can be reexpressed after a change of variables as

$$b^{-1}(1-\rho) \int_0^{td(1-\rho)^{-2}} B^c_\rho (v)\, dv = b^{-1}(1-\rho)(EB_\rho) B_{e\rho}(td(1-\rho)^{-2}) \ , \qquad (5.13)$$

where $B_{e\rho}(t)$ is the stationary-excess cdf associated with B_ρ, defined in (2.3). By condition C1, $b^{-1}(1-\rho) EB_\rho \rightarrow 1$ as $\rho \rightarrow 1$. Moreover, by condition C2 and the continuous mapping theorem with the first passage time map

$$B_{e\rho}(td(1-\rho)^{-2}) \rightarrow H_1(t) \quad \text{as} \quad \rho \rightarrow 1 \ , \qquad (5.14)$$

because, as noted before, $H_1(t)$ is also the cdf of the equilibrium time to emptiness of *B*. Also $(1-\rho) + \rho B_{e\rho}(t)$ is the equilibrium time to emptiness for $W^*_\rho(t)$. (With probability ρ, the server is busy. Conditional on the server being busy, the remaining busy period has cdf $B_{e\rho}(t)$.) Hence, we must have $F(t) = H_1(t)$ and thus $f(t) = h_1(t)$ for all t. (The limit $h_1(t)$ is continuous.)

Proof of Theorem 7. In the GI/G/1 model (and more generally) with traffic intensity ρ,

$$EI_\rho / EB_\rho = (1-\rho)/\rho \ , \qquad (5.15)$$

where I_ρ is the idle time; e.g., see p. 286 of Cohen (1982). To understand (5.15), recall that B_ρ is a random number of service times, while $B_\rho + I_\rho$ is the same random sum of interarrival times. Given (5.15), we see that to establish $C1$ it suffices to show that $EI_\rho \rightarrow EI_1$ as $\rho \rightarrow 1$, but the sufficiency of the moment condition is established in Lemma 3.1 of Kella and Taksar (1994).

6. An Example With No Root

The asymptotics in (2.9) and (2.10) is based on the equation (2.12) having a root. We have noted that in general (2.12) need not have a root. Here we give an example, using the generalized inverse Gaussian service-time density

$$g(t) = \sqrt{3/4\pi t^5} \exp(-(t-3)^2/12t) , \quad t > 0 , \tag{6.1}$$

with Laplace transform

$$\hat{g}(s) = (1 + (2/3) r(s)) e^{-r(s)} , \tag{6.2}$$

where

$$r(s) = (\sqrt{1+12s} - 1)/2 , \tag{6.3}$$

as in (38) of Abate, Choudhury and Whitt (1994). In this case the service-time distribution has a finite moment generating function in the neighborhood of the origin. In this case, equation (12) becomes

$$1 - \sqrt{1-12u} = 2 \log \rho^{-1} . \tag{6.4}$$

Therefore, (12) has a root only when $\rho > e^{-1/2} \approx 0.607$. For $\rho > e^{-1/2}$,

$$\xi = [1-(1-2\log\rho^{-1})^2]/12 , \tag{6.5}$$

and

$$\tau^{-1} = \log\rho^{-1} - (1-\rho) - (\log\rho^{-1})^2/3 . \tag{6.6}$$

When $\rho < e^{-1/2}$, the asymptotics evidently has the form of the service-time distribution itself with a new constant, as in the long-tail case treated by De Meyer and Teugels (1980). In particular, by that argument we should have

$$b(t) \sim (1-\rho) g((1-\rho)t) \quad \text{as} \quad t \to \infty ; \tag{6.7}$$

see Section 3 of Abate, Choudhury and Whitt (1994).

Acknowledgment. We thank Offer Kella for helping us establish Theorem 7.

REFERENCES

[1] Abate, J., Choudhury, G. L. and Whitt, W. (1994) Calculating the M/G/1 busy-period density and the LIFO waiting-time distribution by direct numerical transform inversion. submitted for publication.

[2] Abate, J. and Whitt, W. (1987) Transient behavior of regulated Brownian Motion, I and II. *Adv. Appl. Prob.* 19, 560-631.

[3] Abate, J. and Whitt, W. (1988a) Transient behavior of the M/M/1 queue via Laplace transforms. *Adv. Appl. Prob.* 20, 145-178.

[4] Abate, J. and Whitt, W. (1988b) Approximations for the M/M/1 busy-period distribution. *Queueing Theory and its Applications, Liber Amicorum Professor J. W. Cohen*, O. J. Boxma and R. Syski, eds. North-Holland, Amsterdam, 149-191.

[5] Abate, J. and Whitt, W. (1988c) Simple spectral representations for the M/M/1 queue, *Queueing systems*, 3, 321-346.

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

[7] Abate, J. and Whitt, W. (1992b) Solving probability transform functional equations for numerical inversion. *Oper. Res. Letters* 12, 245-251.

[8] Abate, J. and Whitt, W. (1994) Transient behavior of the M/G/1 workload process. *Opns. Res.* 42, 750-764.

[9] Abate, J. and Whitt, W. (1995) An operational calculus for probability distributions via Laplace transforms. *Adv. Appl. Prob.*, to appear.

[10] Abramowitz, M. and Stegun, I. A. (eds.) (1972) *Handbook of Mathematical Functions*,

Dover, New York.

[11] Choudhury, G. L. and Lucantoni, D. M. (1994) Numerical computation of moments with application to asymptotic analysis. *Oper. Res.,* to appear.

[12] Cohen, J. W. (1982) *The Single Server Queue*, 2nd edn., North-Holland, Amsterdam.

[13] Cox, D. R. and Smith, W. L. (1961) Queues, Methuen, London.

[14] De Meyer, A. and Teugels, J. L. (1980) On the asymptotic behavior of the distributions of the busy period and service time in M/G/1. *Adv. Appl. Prob.* 17, 802-813.

[15] Ethier, S. N. and Kurtz, T. G. (1986) *Markov Processes, Characterization and Convergence,* Wiley, New York.

[16] Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd edn., Wiley, New York.

[17] Halfin, S. (1985) Delays in queues, properties and approximations. *Teletraffic Issues in an Advanced Information Society, Proceedings of ITC-11,* M. Akiyama, ed., Elsevier, Amsterdam, 47-52.

[18] Heyman, D. P. (1974) An approximation for the busy period of the M/G/1 queue using a diffusion model. *J. Appl. Prob.* 11, 159-169.

[19] Iglehart, D. L. and Whitt, W. (1970) Multiple channel queues in heavy traffic, I and II. *Adv. Appl. Prob.* 2, 150-177 and 355-369.

[20] Keilson, J. (1978) Exponential spectra as a tool for the study of server-systems with several classes of customers. *J. Appl. Prob.* 15, 162-170.

[21] Kella, O. and Taksar, M. I. (1994) A heavy traffic limit for the cycle counting process in G/G/1, optional interruptions and elastic screen Brownian motion. *Meth. Oper. Res.* 19,

132-151.

[22] Kingman, J. F. C. (1961) The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* 57, 902-904.

[23] Kingman, J. F. C. (1962a) On queues in heavy traffic. *J. Roy. Stat. Soc. Ser. B* 24, 383-392.

[24] Kingman, J. F. C. (1962b) The use of Spitzer's identity in the investigation of the busy period and other quantities in the queue GI/G/1. *J. Aust. Math. Soc.* 2, 345-356.

[25] Kraemer, W. and Langenbach-Belz, M. (1976) Approximate formulae for the delay in the queueing system GI/G/1. *Proceedings Eighth Int. Teletraffic Congress,* Melbourne, Australia, 235-1/8.

[26] Ott, T. J. (1977) The stable M/G/1 queue in heavy traffic and its covariance function. *Adv. Appl. Prob.* 9, 169-186.

[27] Rice, S. O. (1962) Single server systems – II. Busy periods. *Bell System Tech. J.* 41, 279-310.

[28] Riordan, J. (1962) *Stochastic Service Systems,* Wiley, New York.

[29] Szczotka, W. (1990) Exponential approximation of waiting time and queue size for queues in heavy traffic. *Adv. Appl. Prob.* 22, 230-240.

[30] Whitt, W. (1971) Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Prob.* 8, 74-94.

[31] Whitt, W. (1983) The queueing network analyzer. *Bell System Tech. J.* 62, 2779-2815.

[32] Whitt, W. (1984) Minimizing delays in the GI/G/1 queue. *Oper. Res.* 32, 41-51.

time	exact by transform inversion	inverse Gaussian apoproximation in (4.2)–(4.4)	asymptotic normal approximation in (2.15)	heavy-traffic approximation in (4.1)
0.5	.881	.811	1.10	.66
1	.582	.560	.617	.384
2	.289	.306	.300	.202
3	.175	.188	.180	.130
5	.0800	.0852	.0815	.0667
9	.0239	.0240	.0242	.0248
12	.0110	.0106	.0112	.0135
15	.00551	.00494	.00555	.00781
20	.00186	.00152	.00188	.00340
32	.000175	.000113	.000176	.000578

Table 1. A comparison of approximations with exact values for the busy-period ccdf $B^c(t)$ in the $M/E_4/1$ queue with $\rho = 0.5$.

time	exact by transform inversion	inverse Gaussian apoproximation in (4.2)–(4.4)	asymptotic normal approximation in (2.15)	heavy-traffic approximation in (4.1)
0.5	.891	.800	1.05	.82
1	.640	.599	.670	.531
2	.399	.400	.410	.329
6	.159	.166	.161	.134
10	.0924	.0964	.0930	.0780
15	.0554	.0575	.0556	.0493
30	.0179	.0182	.0180	.0174
40	.00982	.00978	.00986	.01005
80	.00136	.00126	.00136	.00170
120	.000252	.000218	.000252	.000382

Table 2. A comparison of approximations with exact values of the busy-period ccdf $B^c(t)$ in the $M/E_4/1$ queue with $\rho = 0.75$.

time	exact by transform inversion	inverse Gaussian apoproximation in (4.2)–(4.4)	asymptotic normal approximation in (2.15)	heavy-traffic approximation in (4.1)
0.5	.897	.796	1.02	.93
1	.671	.618	.697	.637
5	.265	.266	.268	.246
15	.124	.127	.125	.115
30	.0703	.0717	.0704	.0656
60	.0353	.0360	.0354	.0334
120	.0147	.0148	.0147	.0141
200	.00626	.00628	.00626	.00620
400	.00125	.00123	.00125	.00131
600	.000330	.000318	.000330	.000366

Table 3. A comparison of approximations with exact values of the busy-period ccdf $B^c(t)$ in the $M/E_4/1$ queue with $\rho = 0.9$.

time	exact by transform inversion	inverse Gaussian apoproximation in (4.2)–(4.4)	asymptotic normal approximation in (2.15)	heavy-traffic approximation in (4.1)
0.1	.755	.94	2.0	1.3
1	.369	.368	.467	.313
2	.237	.223	.269	.186
5	.103	.095	.109	.081
8	.0585	.0546	.0606	.0477
15	.0218	.0211	.0222	.0197
20	.0123	.0122	.0125	.0120
30	.00452	.00476	.00457	.00512
40	.00186	.00207	.00188	.00244
60	.000378	.000466	.000380	.000657

Table 4. A comparison of approximations with exact values of the busy-period ccdf $B^c(t)$ in the $M/\Gamma_{1/2}/1$ queue with $\rho = 0.5$.

time	exact by transform inversion	inverse Gaussian apoproximation in (4.2)–(4.4)	asymptotic normal approximation in (2.15)	heavy-traffic approximation in (4.1)
0.1	.757	.94	1.7	.1.4
1	.392	.404	.457	.382
5	.152	.147	.157	.133
8	.106	.102	.108	.093
15	.0607	.0587	.0615	.0537
30	.0284	.0276	.0286	.0258
60	.0103	.0102	.0104	.0099
80	.00607	.00605	.00609	.00599
120	.00244	.00248	.00244	.00256
250	.000222	.000239	.000222	.000282

Table 5. A comparison of approximations with exact values of the busy-period ccdf $B^c(t)$ in the $M/\Gamma_{1/2}/1$ queue with $\rho = 0.75$.

time	exact by transform inversion	inverse Gaussian apoproximation in (4.2)–(4.4)	asymptotic normal approximation in (2.15)	heavy-traffic approximation in (4.1)
0.1	.758	.93	1.5	.1.4
1	.406	.424	.458	.428
5	.183	.181	.186	.174
10	.121	.119	.122	.115
20	.0772	.0762	.0777	.0731
50	.0392	.0387	.0393	.0372
100	.0212	.0209	.0212	.0202
250	.00738	.00732	.00738	.00716
500	.00241	.00241	.00241	.00240
1000	.000470	.000475	.000470	.000488

Table 6. A comparison of approximations with exact values of the busy-period ccdf $B^c(t)$ in the $M/\Gamma_{1/2}/1$ queue with $\rho = 0.9$.

$\Gamma_{1/2}$ service-time distribution				
ρ	exact β	3 terms β_3/β	2 terms β_2/β	1 terms β_1/β
0.25	3.232	1.12	1.24	1.65
0.50	9.084	1.03	1.08	1.32
0.75	42.43	1.003	1.02	1.13
0.90	286.5	1.000	1.002	1.05
D service-time distribution				
ρ	exact β	3 terms β_3/β	2 terms β_2/β	1 term β_1/β
0.25	0.786	1.27	1.51	2.26
0.50	2.59	1.06	1.16	1.55
0.75	13.26	1.007	1.03	1.21
0.90	93.33	1.000	1.005	1.07

Table 7. A comparison of heavy-traffic expansion approximations for the M/G/1 asymptotic parameter β.