Probability in the Engineering and Informational Sciences, 9, 1995, 581–602. Printed in the U.S.A.

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

JOSEPH ABATE* AND WARD WHITT†
AT&T Bell Laboratories
Room 2C-178
Murray Hill, New Jersey 07974-0636

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 c.d.f. 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 c.d.f. 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 the approximation based on 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.

1. INTRODUCTION

This paper is an extension of Abate and Whitt [3], in which we studied the M/M/1 busy-period distribution and proposed approximations for busy-period

*Retired: 900 Hammond Road, Ridgewood, New Jersey 07450-2908.
†Email: wow@research.att.com

 0269-9648/95 $11.00 + .10

distributions in more general single-server queues. Here we provide additional theoretical and empirical support for two approximations proposed in Abate and Whitt [3], the natural generalization of the *asymptotic normal approximation* in Eq. (4.3) there and the *inverse Gaussian approximation* in Eqs. (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 [26], 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 [27].

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 [13] and Section III.7.3 of Cohen [12] can be expressed differently, in terms of a scaled version of the heavy-traffic limit (Eq. (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 [6,7].

Here is how this paper is organized. We establish several M/G/1 results in Section 2, and we establish the heavy-traffic limit for other models in Section 3. In Section 4 we make the numerical comparisons with exact M/G/1 results. All proofs are presented in Section 5. Finally, we provide an example in Section 6 showing that a key equation for the asymptotics may have no root.

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 Cohen [12, p. 249], Cox and Smith [13, Sect. 5.6], or Abate and Whitt [8]). 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 c.d.f. $G(t)$ 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 when the system is empty again. Let $B(t)$ be the c.d.f. 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,

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

where $G_n(t)$ is the c.d.f. of the n-fold convolution of $G(t)$.

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

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

be the associated *stationary-excess c.d.f.* (or equilibrium residual lifetime c.d.f.). 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 c.d.f. $H_1(t)$ of regulated or reflecting Brownian motion (RBM) investigated in Abate and Whitt [2]. 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

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

where $\Phi(t)$ is the c.d.f. 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 $\frac{1}{2}$, i.e.,

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

and $\gamma_e(t)$ is the associated stationary-excess density. From Eq. (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 c.d.f. $\Phi(t)$ (e.g., Sect. 26.2.17 of Abramowitz and Stegan [10]).

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 0 starting with the exponential stationary distribution. In other words, it is an *exponential mixture of inverse Gaussian* densities; see Section 8 of Abate and Whitt [9]. The moment c.d.f. $H_1(t)$ is the only c.d.f. on $[0,\infty)$ with mean $\frac{1}{2}$ for which the 2-fold convolution coincides with stationary-excess c.d.f., i.e., for which the transforms satisfy

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

see Sections 1.2 and 1.3 of Abate and Whitt [2].

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 [26] the same limiting behavior holds in considerable generality. To obtain our heavy-traffic limit, we scale both inside (time) and outside the c.c.d.f. $B_\rho^c(t)$. 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 Eq. (1.32) of Ott [26] 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 Eq. (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 [3] for more discussion.

Understanding of Theorem 1 is enhanced by recognizing that the left side of Eq. (2.7) is a scaled version of the *density* of the busy-period stationary-excess c.d.f., 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.,

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

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 (page 111 of Feller [16]), plus inequality (5.1) below, we also obtain convergence of the associated scaled conditional equilibrium-time-to-emptiness c.d.f.'s from Theorem 1. The form of the limit comes from Corollary 1.1.1 and Eq. (4.3) of Abate and Whitt [2].

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 c.d.f. has a density $b(t)$. Under considerable generality (see Eq. (49) on page 156 of Cox and Smith [13] or Eqs. (11)–(13) of Abate, Choudhury, and Whitt [1]),

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

so that

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

where $\gamma(t)$ is the gamma density in Eq. (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, Eq. (2.12) need not have a solution, in which case Eq. (2.9) does not hold; we give an example in Section 6. The parameter α in Eqs. (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 Eq. (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 Eq. (2.4), that the derivative of $h_1(t)$ has the remarkably simple form

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

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

THEOREM 2: *If Eq.* (2.9) *holds, then*

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

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

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

COROLLARY: *If Eq.* (2.15) *holds, then*

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

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

We previously suggested approximation (2.15) for the M/M/1 queue in Eq. (4.3) of Abate and Whitt [3]; as before, we call it the *asymptotic normal approximation*, because it uses the normal density and c.d.f. in Eq. (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 [28, p. 109], but it does not seem to have been pursued.

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

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

where δ_i are constants, with

$$\delta_1 = \alpha\sqrt{2\beta/\pi} \quad \textbf{(2.18)}$$

from Eq. (2.9); see Section 4 of Abate and Whitt [3]. As shown for the M/M/1 queue in Abate and Whitt [3], Eq. (2.15) is a vastly superior approximation than the first few terms of Eq. (2.17). Interestingly, the direct asymptotics for the busy-period *density* yields a better approximation than the direct asymptotics for the busy-period c.d.f. (e.g., see Table 1 of Abate et al. [1]), and this good quality is inherited by the integral. The integral $h_1(t)$ in Eq. (2.15) has structure not inherited by its asymptotic form.

The good performance of Eq. (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 et al. [1],

$$\begin{aligned}\beta^{-1} = \frac{(1-\rho)^2}{m_2}\,(1 &+ (1-\rho)(1-\xi) \\ &+ (1-\rho)^2[1 - \xi(2 - (9/4)\xi) - \psi] \\ &+ O((1-\rho)^3))\end{aligned} \quad \textbf{(2.19)}$$

and

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

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

$$\beta\alpha^{-1}B_\rho^c(\beta t) \sim m_2(1-\rho)^{-1}B_\rho^c(tm_2(1-\rho)^{-2}) \sim h_1(t) \quad \textbf{(2.21)}$$

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

In Theorem 3.5 of Abate and Whitt [5], 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 c.d.f. must have a density. A sufficient condition is for the service-time c.d.f. $G(t)$ to be absolutely continuous. If the service-time c.d.f. $G(t)$ is absolutely continuous with a density $g(t)$, then so are all n-fold convolutions G_n [16, p. 146]. Thus, from Eq. (2.2) and Fubini [16, p. 111], $B(t)$ is absolutely continuous with density

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

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 c.d.f. $G(t)$ is absolutely continuous with density $g(t)$, so that the busy-period c.d.f. $B(t)$ is absolutely continuous with density $b(t)$, where $b(0) = g(0)$. If $b(0) < \infty$ and $b(t)$ is monotone, then*

$$\lim_{\rho \to 1} m_2^2(1-\rho)^{-3} b(tm_2(1-\rho)^{-2}) = h_1'(t) = (2\pi t^3)^{-1/2} e^{-t/2}, \qquad t > 0. \tag{2.23}$$

Because $b'(0) = g'(0)$, 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 [20] 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 kth derivative of f at t.

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

$$\lim_{\rho \to 1} m_2^{2+k}(1-\rho)^{-(3+2k)} b^{(k)}(tm_2(1-\rho)^{-2}) = h_1^{(k)}(t), \qquad t \geq 0. \tag{2.24}$$

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 [20], $b(t)$ in Eq. (2.22) and thus $B^c(t)$ and $h_\rho(t)$ in Eq. (2.8) are completely monotone when $g(t)$ is completely monotone; i.e., for each ρ, $0 < \rho < 1$,

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

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

THEOREM 5: *If the service-time density $g(t)$ is completely monotone, so that $h_\rho(t)$ in Eq.* (2.8) *admits 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 Eq. (2.4).

where

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

V_i is the ith service time, and $N(0,1)$ is a standard (mean 0, variance 1) normal random variable.

Combining Eqs. (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 Eq. (5.205) of Cohen [12, p. 330]. 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 [12], Kingman [24] and Rice [27].

For practical purposes we suggest using the Kraemer and Langenbach-Belz [25] approximation, also given in Eq. (49) of Whitt [32]:

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

where

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

From Eqs. (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 [33].

We remark that Theorem 6 is consistent with Eq. (75) of Rice [27]. His approximate asymptotic formula for the busy-period density can be obtained from Eq. (3.1) by first taking the derivative and then letting $t \to \infty$. The corresponding formula for $B_\rho^c(t)$ is

$$B_\rho^c(t) \approx \frac{b(1-\rho)}{d} h_1(t(1-\rho)^2/d)$$
$$\approx (b/(1-\rho)^2)\sqrt{2d/\pi t^3}\exp(-t(1-\rho)^2/2d)$$
$$\approx EB\sqrt{D/\pi t^3}e^{-Dt} \tag{3.10}$$

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 Eq. (3.1) as our principal proposed approximation for $B^c(t)$. Our actual proposed approximation is Eq. (2.15) for α and β determined by Eq. (2.10), assuming that Eq. (2.10) holds for the more general model. We intend to discuss asymptotics of the form in Eq. (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 Eq. (77) of Takács [30]). More generally, we can obtain the desired parameters α and β in Eq. (2.10) numerically using the inversion algorithm in Choudhury and Lucantoni [11]. Rice's [27] formula (3.10) provides support for both Eq. (2.10) and the asymptotic behavior of the parameters α and β as in Eqs. (2.19) and (2.20).

4. NUMERICAL EXAMPLES

This section extends the numerical investigation of approximations for the busy-period c.c.d.f. $B^c(t)$ done for the M/M/1 queue in Abate and Whitt [3] to M/G/1 queues. Our previous investigation showed that even three terms of asymptotic expansion (2.17) yield 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 Eq. (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 Eq. (2.7) by moving the normalizing constants to the right-hand side.

Our second approximation is the *asymptotic normal approximation* provided by Theorem 2, i.e., Eq. (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 Eqs. (2.9)–(2.13) in Eq. (2.15) are replaced by the first terms in their heavy-traffic expansions in Eqs. (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 Eqs. (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 Eqs. (6.6) and (8.3) of Abate and Whitt [3],

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

where

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

with $\Phi^c(x)$ the normal c.d.f. 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, because it is a first-passage time distribution for RBM. This idea was the basis for an IG approximation proposed by Heyman [18], 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 [3].

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 Eq. (2.5) with shape parameter 1/2 and, thus, $c_s^2 = 2$. Clearly, E_4 is less variable than an exponential, whereas $\Gamma_{1/2}$ 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 [6,7]. 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 Eq. (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

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

TABLE 1. A Comparison of Approximations with Exact Values of the Busy-Period c.c.d.f. $B^c(t)$ in the M/E_4/1 Queue with $\rho = 0.5$

Time	Exact by Transform Inversion	IG Approximation in Eqs. (4.2)–(4.4)	Asymptotic Normal Approximation in Eq. (2.15)	Heavy-Traffic Approximation in Eq. (4.1)
0.5	0.881	0.811	1.10	0.66
1	0.582	0.560	0.617	0.384
2	0.289	0.306	0.300	0.202
3	0.175	0.188	0.180	0.130
5	0.0800	0.0852	0.0815	0.0667
9	0.0239	0.0240	0.0242	0.0248
12	0.0110	0.0106	0.0112	0.0135
15	0.00551	0.00494	0.00555	0.00781
20	0.00186	0.00152	0.00188	0.00340
32	0.000175	0.000113	0.000176	0.000578

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 Eq. (4.5), the root of Eq. (2.12) is

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

TABLE 2. A Comparison of Approximations with Exact Values of the Busy-Period c.c.d.f. $B^c(t)$ in the M/E_4/1 Queue with $\rho = 0.75$

Time	Exact by Transform Inversion	IG Approximation in Eqs. (4.2)–(4.4)	Asymptotic Normal Approximation in Eq. (2.15)	Heavy-Traffic Approximation in Eq. (4.1)
0.5	0.891	0.800	1.05	0.82
1	0.640	0.599	0.670	0.531
2	0.399	0.400	0.410	0.329
6	0.159	0.166	0.161	0.134
10	0.0924	0.0964	0.0930	0.0780
15	0.0554	0.0575	0.0556	0.0493
30	0.0179	0.0182	0.0180	0.0174
40	0.00982	0.00978	0.00986	0.01005
80	0.00136	0.00126	0.00136	0.00170
120	0.000252	0.000218	0.000252	0.000382

Table 3. A Comparison of Approximations with Exact Values of the Busy-Period c.c.d.f. $B^c(t)$ in the $M/E_4/1$ Queue with $\rho = 0.9$

Time	Exact by Transform Inversion	IG Approximation in Eqs. (4.2)–(4.4)	Asymptotic Normal Approximation in Eq. (2.15)	Heavy-Traffic Approximation in Eq. (4.1)
0.5	0.897	0.796	1.02	0.93
1	0.671	0.618	0.697	0.637
5	0.265	0.266	0.268	0.246
15	0.124	0.127	0.125	0.115
30	0.0703	0.0717	0.0704	0.0656
60	0.0353	0.0360	0.0354	0.0334
120	0.0147	0.0148	0.0147	0.0141
200	0.00626	0.00628	0.00626	0.00620
400	0.00125	0.00123	0.00125	0.00131
600	0.000330	0.000318	0.000330	0.000366

so that the asymptotic parameters in Eq. (2.10) are

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

and

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

Table 4. A Comparison of Approximations with Exact Values of the Busy-Period c.c.d.f. $B^c(t)$ in the $M/\Gamma_{1/2}/1$ Queue with $\rho = 0.5$

Time	Exact by Transform Inversion	IG Approximation in Eqs. (4.2)–(4.4)	Asymptotic Normal Approximation in Eq. (2.15)	Heavy-Traffic Approximation in Eq. (4.1)
0.1	0.755	0.94	2.0	1.3
1	0.369	0.368	0.467	0.313
2	0.237	0.223	0.269	0.186
5	0.103	0.095	0.109	0.081
8	0.0585	0.0546	0.0606	0.0477
15	0.0218	0.0211	0.0222	0.0197
20	0.0123	0.0122	0.0125	0.0120
30	0.00452	0.00476	0.00457	0.00512
40	0.00186	0.00207	0.00188	0.00244
60	0.000378	0.000466	0.000380	0.000657

TABLE 5. A Comparison of Approximations with Exact Values of the Busy-Period c.c.d.f. $B^c(t)$ in the $M/\Gamma_{1/2}/1$ Queue with $\rho = 0.75$

Time	Exact by Transform Inversion	IG Approximation in Eqs. (4.2)–(4.4)	Asymptotic Normal Approximation in Eq. (2.15)	Heavy-Traffic Approximation in Eq. (4.1)
0.1	0.757	0.94	1.7	1.4
1	0.392	0.404	0.457	0.382
5	0.152	0.147	0.157	0.133
8	0.106	0.102	0.108	0.093
15	0.0607	0.0587	0.0615	0.0537
30	0.0284	0.0276	0.0286	0.0258
60	0.0103	0.0102	0.0104	0.0099
80	0.00607	0.00605	0.00609	0.00599
120	0.00244	0.00248	0.00244	0.00256
250	0.000222	0.000239	0.000222	0.000282

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

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

and

$$\alpha = \sqrt{\beta}/\rho. \tag{4.10}$$

TABLE 6. A Comparison of Approximations with Exact Values of the Busy-Period c.c.d.f. $B^c(t)$ in the $M/\Gamma_{1/2}/1$ Queue with $\rho = 0.9$

Time	Exact by Transform Inversion	IG Approximation in Eqs. (4.2)–(4.4)	Asymptotic Normal Approximation in Eq. (2.15)	Heavy-Traffic Approximation in Eq. (4.1)
0.1	0.758	0.93	1.5	1.4
1	0.406	0.424	0.458	0.428
5	0.183	0.181	0.186	0.174
10	0.121	0.119	0.122	0.115
20	0.0772	0.0762	0.0777	0.0731
50	0.0392	0.0387	0.0393	0.0372
100	0.0212	0.0209	0.0212	0.0202
250	0.00738	0.00732	0.00738	0.00716
500	0.00241	0.00241	0.00241	0.00240
1000	0.000470	0.000475	0.000470	0.000488

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, and 105. The auxiliary parameters in Eq. (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 Eqs. (2.19) and (2.20). Refined heavy-traffic approximations can be obtained by using more terms in Eqs. (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 Eq. (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 Eq. (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

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

Table 7. A Comparison of Heavy-Traffic Expansion Approximations for the M/G/1 Asymptotic Parameter β

ρ	Exact β	3 Terms β_3/β	2 Terms β_2/β	1 Term β_1/β
	$\Gamma_{1/2}$ Service-Time Distribution			
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			
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

from which we get

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

where b_k is the kth busy-period moment. However, $b_2 = \alpha_1\beta_1$ and, by Eqs. (2.19) and (2.20),

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

so that the error in Eq. (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, page 152 of Feller [16], $B^c(t) \le 1/t(1-\rho)$ and

$$h_\rho(t) = m_2(1-\rho)^{-1}B^c(tm_2(1-\rho)^{-2}) \le 1/t \tag{5.1}$$

for all t and ρ. Because $B^c(t)$ is monotone, we can thus apply the Helly selection theorem, page 267 of Feller [16], to conclude that any sequence $\{h_{\rho_n}(t): n \ge 1\}$ with $\rho_n \to 1$ has a subsequence that converges to a monotone function $f(t)$ (depending on the subsequence) with $0 \le f(t) \le 1/t$, 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 functional Eq. (2.1).

We begin by expressing busy-period functional Eq. (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 forms (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 Eq. (5.4) is getting small as $\rho \to 1$. Because the service-time distribution has a finite second moment,

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

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

23. Kingman, J.F.C. (1962). On queues in heavy traffic. *Journal of the Royal Statistical Society Series B* 24: 383–392.
24. Kingman, J.F.C. (1962). The use of Spitzer's identity in the investigation of the busy period and other quantities in the queue GI/G/1. *Journal of the Australian Mathematical Society* 2: 345–356.
25. Kraemer, W. & Langenbach-Belz, M. (1976). Approximate formulae for the delay in the queueing system GI/G/1. *Proceedings of the Eighth International 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. *Advances in Applied Probability* 9: 169–186.
27. Rice, S.O. (1962). Single server systems—II. Busy periods. *Bell System Technical Journal* 41: 279–310.
28. Riordan, J. (1962). *Stochastic service systems*. New York: Wiley.
29. Szczotka, W. (1990). Exponential approximation of waiting time and queue size for queues in heavy traffic. *Advances in Applied Probability* 22: 230–240.
30. Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes*. New York: Wiley.
31. Whitt, W. (1971). Weak convergence theorems for priority queues: Preemptive-resume discipline. *Journal of Applied Probability* 8: 74–94.
32. Whitt, W. (1983). The queueing network analyzer. *Bell System Technical Journal* 62: 2779–2815.
33. Whitt, W. (1984). Minimizing delays in the GI/G/1 queue. *Operations Research* 32: 41–51.