Heavy-Traffic Limits for Queues with Many Exponential Servers

Shlomo Halfin; Ward Whitt

Operations Research, Vol. 29, No. 3 (May - Jun., 1981), 567-588.

Stable URL:
http://links.jstor.org/sici?sici=0030-364X%28198105%2F06%2929%3A3%3C567%3AHLFQWM%3E2.0.CO%3B2-1

Operations Research is currently published by INFORMS.

Heavy-Traffic Limits for Queues with Many Exponential Servers

SHLOMO HALFIN and WARD WHITT
Bell Laboratories, Holmdel, New Jersey
(Received November 1979; accepted November 1980)

Two different kinds of heavy-traffic limit theorems have been proved for s-server queues. The first kind involves a sequence of queueing systems having a fixed number of servers with an associated sequence of traffic intensities that converges to the critical value of one from below. The second kind, which is often not thought of as heavy traffic, involves a sequence of queueing systems in which the associated sequences of arrival rates and numbers of servers go to infinity while the service time distributions and the traffic intensities remain fixed, with the traffic intensities being less than the critical value of one. In each case the sequence of random variables depicting the steady-state number of customers waiting or being served diverges to infinity but converges to a nondegenerate limit after appropriate normalization. However, in an important respect neither procedure adequately represents a typical queueing system in practice because in the (heavy-traffic) limit an arriving customer is either almost certain to be delayed (first procedure) or almost certain not to be delayed (second procedure). Hence, we consider a sequence of ($GI/M/s$) systems in which the traffic intensities converge to one from below, the arrival rates and the numbers of servers go to infinity, but the steady-state probabilities that all servers are busy are held fixed. The limits in this case are hybrids of the limits in the other two cases. Numerical comparisons indicate that the resulting approximation is better than the earlier ones for many-server systems operating at typically encountered loads.

DIFFUSION APPROXIMATIONS for stochastic processes in queueing models are now quite common; see Borovkov (1976), Chandy and Sauer (1978), Halachmi and Franta (1978), Harrison (1978), Iglehart (1973a, b), Lemoine (1978), Newell (1973), and Whitt (1974) and the references in these sources. These diffusion approximations can be obtained by heuristic methods and limit theorems, with the limit theorems involving a sequence of queueing systems under "heavy-traffic" conditions. Regardless of the method used, the quality of the approximation can be judged by numerical comparisons. However, the limit theorems add some additional insight, especially about the regions where the approximation should work well.

The purpose of this paper is to prove a new heavy-traffic limit theorem for the standard $GI/M/s$ queue. Considering the extensive literature, it

Operations Research
Vol. 29, No. 3, May–June 1981
0030-364X/81/2903-0567 $01.25

is perhaps surprising that there is anything more to say. However, it is well known that a model with several parameters usually exhibits several different kinds of limiting behavior depending on how the parameters converge. This is illustrated nicely in Karlin and McGregor (1964) where several different diffusion approximations are displayed for genetics models. It is also illustrated here. We obtain our new limit theorem by letting the parameters converge in a different way, a way which we believe is particularly useful in applications.

To set the stage, we review the two existing kinds of heavy-traffic limit theorems for the $GI/G/s$ queue. In each case we consider a sequence of $GI/G/s$ systems indexed by n. The first and more familiar procedure is the one used by Kingman (1962, 1965), Prohorov (1963), Borovkov (1965), Iglehart and Whitt (1970), Köllerström (1974), and many others. With this procedure, the number of servers is held fixed and the sequence of traffic intensities $\{\rho_n\}$ converges to the critical value of one from below. Under these conditions, the sequence of normalized queue-length processes converges to Brownian motion with a negative drift and a reflecting barrier at the origin. (By "queue length" we mean the number of customers in the system, including the customers in service. In this case of heavy traffic, however, the number of servers is asymptotically negligible after normalization.) The associated sequence of normalized stationary queue-length distributions converges to an exponential distribution (the stationary distribution of the limiting diffusion process). As a consequence, an approximation for the stationary queue length $Q(\infty)$ is $s + X$ where X has an exponential distribution with mean

$$EX = (c_a^2 + \rho^2 c_s^2)/2(1 - \rho), \tag{0.1}$$

where ρ is the traffic intensity and c_a and c_s are the coefficients of variation (standard deviation divided by the mean) of the interarrival time and service time, respectively. We say "an" approximation rather than "the" approximation because the limit theorem does not completely determine an associated approximation. For example, another approximation from the same limit theorem can be obtained by replacing the term ρ^2 in the numerator of (0.1) by its heavy-traffic limit of 1; see Section 5 for further discussion. A significant feature of this limit theorem is that the approximating diffusion process and stationary distribution depend on the interarrival time and service time distributions only through their first two moments.

The second procedure is the one used by Iglehart (1965, 1973b) and Borovkov (1967); see Iglehart (1973a) and Whitt (1982) for further discussion. With this procedure, we hold the traffic intensity fixed and let the number of servers go to infinity. This can be achieved by holding the service-time distribution fixed and letting the arrival rate go to infinity

along with the number of servers. Since the traffic intensity is fixed at a value less than one, this situation is often not regarded as heavy traffic, but since the arrival rate and thus the steady-state mean number of customers in the system are going to infinity, we say the system is in heavy traffic. With this procedure, it turns out that s-server systems are asymptotically indistinguishable from infinite-server systems. If the service time distributions are exponential, i.e., for $GI/M/s$ systems, the sequence of appropriately normalized queue-length (number in system) processes converges to an Ornstein-Uhlenbeck diffusion process. However, if the service-time distributions are not exponential, i.e., for $GI/G/s$ systems, the sequence of normalized queue-length processes converges to a Gaussian process which is not a diffusion, i.e., it is not Markov. This suggests that a diffusion approximation may not be appropriate with a large number of servers and nonexponential service times. Upon reflection, this should not be surprising. With infinitely many servers, the residual service times for customers in service can have a significant impact on the future even under heavy loads. For further discussion, see Whitt (1982).

With this second procedure, the sequence of normalized stationary queue-length distributions converges to a simple limit for all $GI/G/s$ systems, namely, the normal distribution. Thus, an approximation for the distribution of the stationary queue length $Q(\infty)$ is a normal distribution with mean $s\rho = \lambda/\mu$ and variance $s\rho z$, where

$$z = 1 + (c_a{}^2 - 1)\mu \int_0^\infty [1 - G(x)]^2 dx, \tag{0.2}$$

λ is the arrival rate, μ^{-1} is the mean and $G(x)$ is the c.d.f. of the service-time distribution. Note that the parameter z in (0.2) is not determined by the first two moments of the service-time distribution, but rather by the mean and the parameter $\int_0^\infty [1 - G(x]^2 dx$.

Unfortunately, in an important respect neither of the regimes just described represents a typical queueing system in real life. The balance between service and economy usually dictates that the probability of delay be kept away from both zero and one, so that the number of customers present fluctuates between the regions above and below the number of servers. However, in the first procedure the steady-state probability that all servers are busy approaches one, and in the second procedure it approaches zero. This characterization is obvious for the first procedure, but somewhat less transparent for the second procedure. It follows because $Q(\infty)$ is approximately normally distributed with mean ρs and standard deviation of order $\sqrt{s}$.

To capture the essence of typical queueing systems, we consider a new limiting procedure. We let both the traffic intensity and the number of

servers increase while holding the probability of delay fixed. This turns out to be equivalent to letting $(1 - \rho_n)\sqrt{s_n}$ converge to a constant. (This itself is important. It means that to maintain a fixed probability of delay the number of servers s tends to be proportional to $1/(1 - \rho)^2$ when s is large or, equivalently, when ρ is near 1.) Under these conditions, the sequence of normalized queue-length processes associated with $GI/M/s$ systems converges to a diffusion process which is a hybrid of the limits in the first two cases, behaving like Brownian motion with negative drift above zero and the Ornstein-Uhlenbeck diffusion process below zero. (The boundary between the two processes can be thought of as the number of servers because this number is subtracted in the normalization.) Interestingly, the limit of the associated sequence of normalized stationary distributions is also a hybrid of the limits in the first two cases, having a continuous density with an exponential upper tail and a normal lower tail. As in the second case of heavy traffic, the limit of the sequence of normalized queue-length processes is not Markov if the service time distributions are not exponential. Partial results for this case appear in Section 4.

It appears that the limit theorems in this paper are new, but it turns out that the derived approximations are similar to ones that have been suggested without limit theorems; see Halachmi and Franta, and Newell. For example, the approximating stationary distribution with one exponential tail and one normal tail is discussed in Chapter 4 of Newell. It is significant that the approximation displayed in formula (15) of Halachmi and Franta is consistent with all three heavy-traffic limit theorems. Under each of the heavy-traffic conditions, the approximating distribution there converges to the appropriate heavy-traffic limiting distribution. However, there is no indication that the service-time distribution can affect the quality of the approximation when there are many servers.

The rest of this paper is organized as follows. We give some preliminary descriptions of a single $M/M/s$ queue in Section 1. Of particular interest is a recursive scheme for calculating the moments of the stationary queue length. As a consequence, we are able to obtain explicit expressions for the moments of the stationary distribution of the limiting diffusion process in the $GI/M/s$ case; see Corollary 1 and Theorems 3 and 4. In Section 2 we state our limit theorems for the $M/M/s$ queue. The $M/M/s$ results are extended to $GI/M/s$ queues in Section 3. The main limit theorem for $GI/M/s$ systems (Theorem 3) follows directly from simple criteria in Stroock and Varadhan (1979) for the convergence of Markov chains to diffusion processes. Properties of the limiting diffusion process are obtained from the $M/M/s$ case, which is even easier to treat. Perhaps the most interesting technical step here is the proof that the steady-state distributions of the $GI/M/s$ systems converge to the steady-

state distribution of the diffusion process (Theorem 4). For this we use stochastic comparison (stochastic order) properties.

We briefly discuss the more complicated picture of $GI/G/s$ queues in Section 4. We are able to generate approximations for $GI/G/s$ queues too, but our results in the case of nonexponential service times are much less satisfactory. First, we can only treat phase-type service-time distributions (mixtures of convolutions of exponential distributions), but this is not a serious limitation because any service-time distribution can be approximated arbitrarily well by a phase-type distribution. Second, we establish much weaker convergence, only for the infinitesimal means and covariances. Finally, the resulting limit process is complicated, so even the approximations are not very tractable with the added generality.

In Section 5 we discuss ways to obtain approximations from the limit theorems and we compare the three heavy-traffic approximations to the $M/M/s$ queue.

1. PRELIMINARY FACTS ABOUT THE $M/M/s$ QUEUE

In this section we describe a single $M/M/s$ queue with arrival rate λ, service rate μ and traffic intensity $\rho = \lambda/s\mu < 1$. We focus on the limiting distribution of the number $Q(t)$ of customers in the system (either waiting or being served) at time t. It is well known that $Q(t)$ converges in distribution as $t \to \infty$ to a random variable $Q(\infty)$ with a proper probability distribution. From standard $M/M/s$ theory, e.g., Cooper ([1972], p. 71)

$$p_k \equiv P(Q(\infty) = k) = \begin{cases} [(s\rho)^k/k!]\eta, & k \le s, \\ [s^s\rho^k/s!]\eta, & k \ge s, \end{cases} \tag{1.1}$$

and

$$\alpha \equiv P(Q(\infty) \ge s) = [(s\rho)^s/s!(1-\rho)]\eta, \tag{1.2}$$

where

$$\eta = [(s\rho)^s/(s!(1-\rho)) + \textstyle\sum_{k=0}^{s-1} (s\rho)^k/k!]^{-1}. \tag{1.3}$$

The quantity α in (1.2) is the Erlang delay formula or Erlang-C formula whose value we shall be fixing in our limit theorems.

We now present a recursive scheme for calculating the moments of $Q(\infty)$. It is convenient both for calculations here and for the limit theorems later to express the moments in terms of α. It is also convenient to break the moment sums into two parts. Let

$$\sigma_1^{(m)} = \textstyle\sum_{k=0}^{s-1} k^m p_k \quad \text{and} \quad \sigma_2^{(m)} = \sum_{k=s}^{\infty} k^m p_k, \qquad m = 0, 1, \cdots. \tag{1.4}$$

Clearly $\sigma_1^{(0)} = 1 - \alpha$, $\sigma_2^{(0)} = \alpha$ and $EQ(\infty)^m = \sigma_1^{(m)} + \sigma_2^{(m)}$. All higher-order terms can be calculated from the following recursive formulas.

LEMMA 1. $\sigma_1^{(m)} = \rho s \sum_{i=0}^{m-1} \binom{m-1}{i} \sigma_1^{(i)} - s^m \alpha(1-\rho)$

and $\sigma_2^{(m)} = \rho(1-\rho)^{-1} \sum_{i=0}^{m-1} \binom{m}{i} \sigma_2^{(i)} + s^m \alpha.$

Proof. From (1.1) we have

$$kp_k = \rho s p_{k-1}, \qquad k = 1, 2, \cdots, s-1, \tag{1.5}$$

and

$$p_k = \rho p_{k-1}, \qquad k = s, s+1, \cdots. \tag{1.6}$$

Consequently, $\rho(p_{s-1} + \alpha) = \alpha$ or $p_{s-1} = \alpha(1-\rho)\rho^{-1}$. With $\sigma_1^{(m)}$ defined as in (1.4) and with the aid of (1.5), we obtain

$$\begin{aligned}
\sigma_1^{(m)} &= \rho s \sum_{k=0}^{s-1} k^{m-1} \rho s p_{k-1} = \rho s \Big(\sum_{k=0}^{s-1} (k+1)^{m-1} p_k - s^{m-1} p_{s-1}\Big) \\
&= \rho s \sum_{i=0}^{m-1} \binom{m-1}{i} \sigma_1^{(i)} - \rho s^m p_{s-1} \\
&= \rho s \sum_{i=0}^{m-1} \binom{m-1}{i} \sigma_1^{(i)} - s^m \alpha(1-\rho).
\end{aligned}$$

Similarly, using (1.6), we obtain

$$\begin{aligned}
\sigma_2^{(m)} &= \rho \sum_{k=s}^{\infty} (k+1)^m p_k + \rho s^m p_{s-1} \\
&= \rho \sum_{i=0}^{m} \binom{m}{i} \sigma_2^{(i)} + \rho s^m p_{s-1},
\end{aligned}$$

so that

$$\begin{aligned}
\sigma_2^{(m)} &= \rho(1-\rho)^{-1} \sum_{i=0}^{m-1} \binom{m}{i} \sigma_2^{(i)} + \rho(1-\rho)^{-1} s^m p_{s-1} \\
&= \rho(1-\rho)^{-1} \sum_{i=0}^{m-1} \binom{m}{i} \sigma_2^{(i)} + s^m \alpha.
\end{aligned}$$

From Lemma 1 we obtain the following values:

$$\begin{aligned}
\sigma_1^{(1)} &= \rho s - s\alpha \\
\sigma_2^{(1)} &= \rho(1-\rho)^{-1}\alpha + s\alpha \\
\sigma_1^{(2)} &= (\rho s)^2 - s^2\alpha + \rho s(1-\alpha) \\
\sigma_2^{(2)} &= 2\rho^2(1-\rho)^{-2}\alpha + s^2\alpha + \rho(1-\rho)^{-1}\alpha(1+2s) \\
\sigma_1^{(3)} &= (\rho s)^3 - s^3\alpha + \rho s^2(2\rho - 2\alpha + \rho(1-\alpha)) \\
&\quad + \rho s(1-\alpha)
\end{aligned}$$

$$\begin{aligned}
\sigma_2^{(3)} &= 6\rho^3(1-\rho)^{-3}\alpha + s^3\alpha + 6\alpha\rho^2(1-\rho)^{-2}(s+1) \\
&\quad + \rho(1-\rho)^{-1}\alpha(3s^2+3s+1) \\
\sigma_1^{(4)} &= (\rho s)^4 - s^4\alpha + \rho s^3(6\rho^2 - \rho^2\alpha - 2\alpha\rho - 3\alpha) \\
&\quad + \rho s^2(7\rho - 4\rho\alpha - 3\alpha) + \rho s(1-\alpha) \\
\sigma_2^{(4)} &= 24\rho^4(1-\rho)^{-4}\alpha + s^4\alpha + 12\rho^3(1-\rho)^{-3}\alpha(2s+3) \\
&\quad + \rho^2(1-\rho)^{-2}\alpha(12s^2+24s+14) \\
&\quad + \rho(1-\rho)^{-1}\alpha(1+4s+6s^2+4s^3).
\end{aligned} \tag{1.7}$$

From these, we can easily calculate the moments of $Q(\infty)$. Here are the mean and variance:

$$\begin{aligned}
&\mathrm{E}Q(\infty) = \rho s + (1-\rho)^{-1}\alpha\rho \\
\text{and} \qquad &\mathrm{Var}\, Q(\infty) = \rho s(1+\alpha) + (1-\rho)^{-2}(\alpha\rho + \alpha(1-\alpha)\rho^2).
\end{aligned} \tag{1.8}$$

Later we shall be interested in the moments about s. The first four are:

$$\begin{aligned}
E(Q(\infty)-s) &= -s(1-\rho) + \rho(1-\rho)^{-1}\alpha \\
E(Q(\infty)-s)^2 &= s^2(1-\rho)^2 + 2\rho^2(1-\rho)^{-2}\alpha \\
&\quad + \rho(1-\rho)^{-1}\alpha + \rho s(1-\alpha) \\
E(Q(\infty)-s)^3 &= -s^3(1-\rho)^3 - s^2\rho(1-\rho)(3-\alpha) \\
&\quad + 6\rho^3(1-\rho)^{-3}\alpha + \rho s(1-\alpha) \\
&\quad + 6\alpha\rho^2(1-\rho)^{-2} + \alpha\rho(1-\rho)^{-1} \\
E(Q(\infty)-s)^4 &= s^4(1-\rho)^4 + s^3(6\rho(1-\rho)^2 - \alpha\rho(1-\rho)^2) \\
&\quad + s^2\rho(7\rho - 4\rho\alpha + \alpha - 4) + 24\rho^4(1-\rho)^{-4}\alpha \\
&\quad + \rho s(1-\alpha) + 36\rho^3(1-\rho)^{-3}\alpha \\
&\quad + 14\rho^2(1-\rho)^{-2}\alpha + \rho(1-\rho)^{-1}\alpha.
\end{aligned} \tag{1.9}$$

These moments are important not only for describing the $M/M/s$ system but also for describing the diffusion approximations of the more general $GI/M/s$ systems.

2. LIMIT THEOREMS FOR THE *M/M/s* QUEUE

Now consider a sequence of $M/M/s$ queues indexed by n with $\lambda_n \to \infty$ as $n \to \infty$ and $\mu_n = \mu$, $s_n = n$ and $\rho_n = \lambda_n/n\mu < 1$ for all n. Let $Q(t)$, $Q(\infty)$, p_k, α and η also be subscripted by n when referring to the nth system. Our basic assumption is that $\alpha_n \equiv P(Q_n(\infty) \geq n) \to \alpha$, $0 < \alpha < 1$, as $n \to \infty$. It is evident that $\rho_n \to 1$ under this limiting constraint, but it is of

interest to see exactly at what rate. Let $\Phi(x)$ and $\phi(x)$ be the standard normal cumulative distribution function and density, respectively. We obtain the following from formula (1.2).

PROPOSITION 1. *The probability of delay has a nondegenerate limit, i.e.,*

$$\lim_{n\to\infty} P(Q_n(\infty) \geq n) = \alpha, \qquad 0 < \alpha < 1, \tag{2.1}$$

if and only if

$$\lim_{n\to\infty} (1 - \rho_n)\sqrt{n} = \beta, \qquad \beta > 0, \tag{2.2}$$

in which case

$$\alpha = [1 + \sqrt{2\pi}\beta\Phi(\beta)exp(\beta^2/2)]^{-1}. \tag{2.3}$$

Proof. After rearranging terms in (1.2), we see that $\alpha_n = [1 + A_n/B_n]^{-1}$ where $A_n = \sum_{k=0}^{n-1} (n\rho_n)^k/k!$ and $B_n = (n\rho_n)^n/n!(1 - \rho_n)$. After multiplying both A_n and B_n by $e^{n\rho_n}$, we have $\alpha_n = [1 + (\gamma_n/\xi_n)]^{-1}$, where

$$\gamma_n = \sum_{k=0}^{n-1} (1/k!)\ (n\rho_n)^k e^{-n\rho_n}!$$

and

$$\xi_n = (n\rho_n)^n e^{-n\rho_n}/n!(1 - \rho_n).$$

We recognize that $\gamma_n = P(X_n \leq n - 1)$ where X_n is a random variable with the Poisson distribution with parameter $n\rho_n$, and thus mean and variance both equal to $n\rho_n$; see Rhee (1977). Then

$$\gamma_n = P(X_n \leq n - 1) = P((n\rho_n)^{-1/2}[X_n - n\rho_n] \leq \nu_n), \tag{2.4}$$

where

$$\nu_n = (1 - \rho_n)n^{1/2}\rho_n^{-1/2} - (n\rho_n)^{-1/2}. \tag{2.5}$$

If $(1 - \rho_n)n^{1/2} \to \beta$, then $\nu_n \to \beta$. By the central limit theorem, as discussed by Feller (1968, pp. 190, 194, 244–245), we then have

$$\gamma_n \to \gamma \equiv P(N(0, 1) \leq \beta) = \Phi(\beta),$$

where $N(0, 1)$ is a standard normal random variable with cumulative distribution function Φ. (In the standard version of the central limit theorem ν_n does not depend on n, but it suffices for ν_n to converge. In general, if C_n and D_n are real-valued random variables such that $C_n \Rightarrow C$ and $D_n \xrightarrow{p} d$, then $D_nC_n \Rightarrow dC$; see Theorems 4.4 and 5.1 of Billingsley [1968].)

Turning to ξ_n, we first apply Stirling's formula to obtain $n! \sim (2\pi n)^{1/2}n^n e^{-n}$ (Feller, p. 52). Then

$$\xi_n \sim \exp(n[1 - \rho_n + \log \rho_n])/(\sqrt{2\pi n}(1 - \rho_n)), \tag{2.6}$$

where

$$\begin{aligned} \log \rho_n &= \log(1 - (1 - \rho_n)) \\ &= -(1 - \rho_n) - (1 - \rho_n)^2/2 + o(1 - \rho_n)^2. \end{aligned} \tag{2.7}$$

Hence,

$$\lim_{n\to\infty}\xi_n = \exp(\beta^2/2)(\beta\sqrt{2\pi})$$

if $(1-\rho_n)n^{1/2} \to \beta$, $0 < \beta < \infty$. Hence, if $(1-\rho_n)n^{1/2} \to \beta$, then $\alpha_n \to \alpha \equiv [1+\gamma/\xi]^{-1}$, $0 < \alpha < 1$, as claimed. Moreover, if $(1-\rho_n)n^{1/2} \to 0$, then $\nu_n \to 0$ so that $\gamma_n \to \Phi(0) = 2^{-1}$, and $\xi_n \to \infty$, which implies that $\alpha_n \to 1$. On the other hand, if $(1-\rho_n)n^{1/2} \to \infty$, then $\nu_n \to \infty$ so that $\gamma_n \to 1$, and $\xi_n \to 0$, which implies that $\alpha_n \to 0$. There is one more case to consider for the "only if" part of the proof. It is possible that the sequence $\{(1-\rho_n)n^{1/2}\}$ might fail to converge to any limit, finite or infinite. But then the sequence will have two subsequences which converge to different limits (one of which could be infinity). However, the reasoning above applies to each of these subsequences. Since $\alpha(\beta)$, the function in (2.3), is strictly decreasing in β, $\alpha(\beta_1) \neq \alpha(\beta_2)$ for $\beta_1 \neq \beta_2$. Hence, the sequence $\{\alpha_n\}$ will have two subsequences with different limits and thus not converge.

Remarks. (1) Note that $\beta\alpha/(1-\alpha) = \phi(\beta)/\Phi(\beta)$, so that

$$1 - \Phi(\beta) \leq \alpha \leq [1-\Phi(\beta)]/[1-\beta^{-2}\Phi(\beta)]$$

for $\beta \geq 1$ by virtue of standard inequalities for the tail of the normal distribution; see (1.8) on p. 175 of Feller. Hence, if (2.2) holds, then $\alpha/[1-\Phi(\beta)] \to 1$ as $\beta \to \infty$. In contrast, using the limit theorem in (0.2) for the $M/M/s$ case (with $z = 1$), we would have $\alpha/[1-\Phi(\beta/\sqrt{\rho})] \to 1$ as $\beta \to \infty$.

(2) Proposition 1 provides a simple approximation for the probability that all servers are busy in an $M/M/s$ queue: just substitute $(1-\rho)s^{1/2}$ for β in (2.3). The proof suggests the following refinement:

$$\begin{aligned} &P(Q(\infty) \geq s) \approx [1 + \sqrt{2\pi s}(1-\rho) \\ &\cdot \exp(s\textstyle\sum_{k=2}^{m}[(1-\rho)^k/k])\Phi([(1-\rho)s-1]/\sqrt{\rho s})]^{-1}, \qquad m \geq 2, \end{aligned} \tag{2.8}$$

which we do not investigate further here. The term involving the normal c.d.f. in (2.8) comes from the central limit theorem for (2.4) using (2.5). The rest of (2.8) comes from (2.6) with m being the number of terms kept in the expansion of $\log \rho_n$ in (2.7). For large s and small m, (2.8) might be preferred to (1.2).

For the rest of this section we assume (2.1) or, equivalently, (2.2). Let $[x]$ be the greatest integer less than or equal to x. Elementary calculations using (1.1)–(1.3) yield

PROPOSITION 2. *Let δ be a positive constant.*

(i) *If $\{\delta_n\}$ is a sequence of constants such that $\delta_n \leq n$ for all n and $(n-\delta_n)n^{-1/2} \to \delta$ as $n \to \infty$, then*

$$\lim_{n\to\infty} P(Q_n(\infty) \leq \delta_n \mid Q_n(\infty) \leq n) = \Phi(\beta-\delta)/\Phi(\beta) \tag{2.9}$$

and

$$lim_{n\to\infty}\sqrt{n}P(Q_n(\infty) = [\delta_n] | Q_n(\infty) \le n) = \phi(\beta - \delta)/\Phi(\beta). \quad (2.10)$$

(ii) *If* $\{\delta_n\}$ *is a sequence of constants such that* $\delta_n \ge n$ *for all* n *and* $(\delta_n - n)n^{-1/2} \to \delta$ *as* $n \to \infty$, *then*

$$lim_{n\to\infty}P(Q_n(\infty) \ge \delta_n | Q_n(\infty) \ge n) = e^{-\delta\beta} \quad (2.11)$$

and

$$lim_{n\to\infty}\sqrt{n}P(Q_n(\infty) = [\delta_n] | Q_n(\infty) \ge n) = \beta e^{-\beta}. \quad (2.12)$$

Proof. Each of the four limits can easily be proved by applying Stirling's formula as in the proof of Proposition 1. We only give the details for (2.10):

$$\begin{aligned}
&\sqrt{n}P(Q_n(\infty) = [\delta_n] | Q_n(\infty) \le n) \\
&\quad = [\alpha/(1-\alpha)]\sqrt{n}(1-\rho_n)(n\rho_n)^{[\delta_n]-n}(n!/[\delta_n]!) \\
&\quad \sim [\alpha\beta/(1-\alpha)](n\rho_n)^{[\delta_n]-n}(n^{n+1/2}e^{-n}/[\delta_n]^{[\delta_n]+1/2}e^{-\delta_n}) \\
&\quad \sim [\alpha\beta/(1-\alpha)]\exp\{-(n-[\delta_n])\log(\rho_n)([\delta_n])\log(n/[\delta_n]) - n + [\delta_n]\} \\
&\quad \sim [\alpha\beta/(1-\alpha)]\exp\{\beta\delta - (\delta^2/2)\} = \phi(\beta-\delta)/\Phi(\beta).
\end{aligned}$$

We now focus on a normalization of the stationary distribution. In particular, let

$$X_n = (Q_n(\infty) - n)n^{1/2}, \qquad n \ge 1. \quad (2.13)$$

Proposition 2 immediately implies a weak convergence theorem for X_n. (Weak convergence is another name for convergence in law or distribution.) Let $\Rightarrow$ denote both weak convergence of probability measures and convergence in distribution of random variables; see Billingsley.

THEOREM 1. $X_n \Rightarrow X$, *where* $P(X \ge 0) = \alpha$, $P(X > x | X \ge 0) = e^{-x\beta}$, *and* $P(X \le x | X \le 0) = \Phi(\beta + x)/\Phi(\beta)$, $x \le 0$.

Proof. Apply Proposition 2 with $\delta_n = n + xn^{1/2}$.

Remarks. (1) Note that the limit X in Theorem 1 has a continuous density which is exponential for $x \ge 0$ and normal for $x \le 0$.

(2) It is interesting that we also have two other proofs of Theorem 1. We can show that all the moments converge to the limiting moments which uniquely determine a probability distribution; see p. 181 of Breiman (1968). For this proof, we apply Lemma 1 and an extension of Corollary 1 below. The second alternative proof is via the functional limit theorem, Theorem 2 below.

Using (1.9), we now investigate the convergence of moments.

COROLLARY 1. *As* $n \to \infty$,

$$\text{(i)}\ EX_n \to EX = -\beta + \alpha\beta^{-1},$$

$$\text{(ii)}\ EX_n^2 \to EX^2 = \beta^2 + 2\alpha\beta^{-2} + (1 - \alpha),$$

$$\text{(iii)}\ EX_n^3 \to EX^3 = -\beta^3 - (3 - \alpha)\beta + 6\alpha\beta^{-3}, \tag{2.14}$$

$$\text{(iv)}\ EX_n^4 \to EX^4 = \beta^4 + (6 - \alpha)\beta^2 + 3(1 - \alpha) + 24\alpha\beta^{-4}$$

and $\text{(v)}\ Var\ X_n \to Var\ X = (1 + \alpha) + (2\alpha - \alpha^2)\beta^{-2}.$

Proof. Since $X_n \Rightarrow X$ (Theorem 1) and $E|X_n|^5$ is uniformly bounded (Lemma 1), we have $EX_n^k \to EX^k$, $k = 1, 2, 3, 4$; see p. 32 of Billingsley. Finally, apply (1.9).

We now consider the entire stochastic process $\{Q_n(t), t \geq 0\}$. For each $n \geq 1$, form the normalized process

$$Y_n \equiv Y_n(t) = (Q_n(t) - n)/n^{1/2}, \qquad t \geq 0. \tag{2.15}$$

We shall show that $\{Y_n\}$ converges in distribution to a diffusion process on the real line, i.e., a Markov process with continuous paths whose evolution is characterized by its infinitesimal generator A, which is of the form

$$A = (\sigma^2(x)/2)(d^2/dx^2) + (m(x))(d/dx), \qquad -\infty < x < \infty, \tag{2.16}$$

where $\sigma^2(x) > 0$ and $\sigma^2(x)$ and $m(x)$ are continuous for all x. For this diffusion process, the boundary points $\pm\infty$ are inaccessible. For further discussion of diffusion processes, see Chapter 16 of Breiman (1968), and Stroock and Varadhan.

As before, let $\Rightarrow$ denote convergence in distribution of random elements, but now the random elements take values in the space of sample paths (the space $D[0, \infty)$ endowed with the usual Skorokhod J_1 topology; see Chapter 3 of Billingsley, Lindvall [1973], and Section 2 of Whitt [1980]). The following result is obtained just like Iglehart's (1965) limit theorem for $M/M/s$ queues by applying Stone's (1961, 1963) simple criteria for the convergence of birth-and-death processes.

THEOREM 2. *If* $Y_n(0) \Rightarrow Y(0)$, *then* $Y_n \Rightarrow Y$ *in* $D[0, \infty)$, *where* Y *is a diffusion process with*

$$m(x) = \begin{cases} -\mu\beta, & x \geq 0 \\ -\mu(x + \beta), & x \leq 0 \end{cases}$$

and $\sigma^2(x) = 2\mu.$

Proof. It is easy to check that Stone's criteria as displayed in Theorem 3.2 of Iglehart (1965) are satisfied. As in Iglehart, because of the normalization, the state spaces associated with the birth-and-death processes Y_n

become dense in the real line as $n \to \infty$. Finally, the infinitesimal means $m_n(x)$ and infinitesimal variances $\sigma_n^2(x)$ of Y_n converge. To see this, note that the convergence $(1 - \rho_n)n^{1/2} = n^{1/2} - \lambda_n/\mu n^{1/2} \to \beta$ implies that, for $x > 0$,

$$m_n(x) = -n\mu(n^{-1/2}) + \lambda_n(n^{-1/2}) \to -\mu\beta$$

and

$$\sigma_n^2(x) = n\mu(n^{-1}) + \lambda_n(n^{-1}) \to 2\mu,$$

while, for $x < 0$,

$$m_n(x) = ([n^{1/2}x + n])\mu(n^{-1/2}) + \lambda_n(n^{-1/2}) \to -\mu(x + \beta)$$

and

$$\sigma_n^2(x) = ([n^{1/2}x + n])\mu(n^{-1}) + \lambda_n(n^{-1}) \to 2\mu,$$

where $[x]$ is the integer part of x.

Remark. An intuitively appealing statement involving real-valued random variables which is equivalent to Theorem 2 is: $f(Y_n) \Rightarrow f(Y)$ for all measurable real-valued functions f on $D[0, \infty)$ which are continuous at each continuous x in $D[0, \infty)$; see Section 5 of Billingsley.

We should expect that the limiting distribution of the diffusion process Y in Theorem 2 would coincide with the distribution of X in Theorem 1, but this is not automatic because an interchange of limits is involved. In general, we need to show that

$$P(Y(\infty) \le x) \equiv \lim_{t\to\infty} \lim_{n\to\infty} P(Y_n(t) \le x)$$

$$= \lim_{n\to\infty} \lim_{t\to\infty} P(Y_n(t) \le x) \equiv \lim_{n\to\infty} P(Y_n(\infty) \le x) \equiv P(X \le x).$$

While such an interchange of limits is often difficult, we can easily establish the desired result in this case.

COROLLARY 2. *$Y(\infty)$ has the same distribution as X.*

Proof. Let $Y_n(0)$ be distributed as X_n for each n. Then $\{Y_n(t), t \ge 0\}$ is a stationary process for each n and $Y_n(0) \Rightarrow X$ by Theorem 1. Since $Y_n(t)$ has the same distribution as $Y_n(0)$, $Y_n(t) \Rightarrow X$ as $n \to \infty$ too. But $Y_n(t) \Rightarrow Y(t)$ for each t as $n \to \infty$ by Theorem 2 and the continuous mapping theorem (see Billingsley, Theorem 5.1), using the projection mapping $\pi_t: D[0, \infty) \to R$, defined by $\pi_t(x) = x(t)$, which is continuous at all continuous x. Hence, $Y(t)$ is distributed as X for each t, so $Y(\infty)$ is distributed as X too. Since the distribution of $Y(\infty)$ is independent of the initial distribution, $Y(0)$, the proof is complete.

Theorem 2 can be applied to obtain limits for various traffic measurements. For example, let $1_B(x)$ be the indicator function of the

set B, which is equal to 1 for $x \in B$ and 0 for $x \notin B$. Some traffic measurements of considerable interest are obtained with the functional $\psi(\cdot, a, T)$:$D[0, \infty) \to R$ defined by

$$\psi(x, a, T) = T^{-1} \int_0^T 1_{\{x(t) \geq a\}} dt. \tag{2.17}$$

The following consequence of Theorem 2 is obtained by applying the continuous mapping theorem (Billingsley, Theorem 5.1).

COROLLARY 3. *Let $Q_n(0)$ have the distribution of $Q_n(\infty)$, i.e., the stationary distribution, for each n. For ψ in* (2.17) *and each a and T,*

$$\psi(Q_n, n + an^{1/2}, T) \Rightarrow \psi(Y, a, T) \text{ in } [0, 1].$$

where Y is strictly stationary, so that $E\psi(Y, a, T) = P(X \geq a)$.

Proof. We can apply the continuous mapping theorem (Billingsley, Theorem 5.1). On p. 231 of Billingsley, the functional ψ is shown to be measurable and continuous almost surely with respect to Brownian motion. The argument there using Fubini's theorem can be applied for the limit process Y too because $Y(t)$ has a density for each t.

Remark. Corollary 3 provides another proof of (2.1) under condition (2.2): just set $a = 0$.

3. *GI/M/s* QUEUES

We now show how the main convergence results—Theorems 1 and 2—extend to $GI/M/s$ queues. As before, let $s_n = n$, $\mu_n = \mu$ and $\rho_n = \lambda_n/n\mu < 1$ for all n. Let u_n be a generic interarrival time in the nth system. We need to make some assumptions about the way the distribution of u_n changes with n in addition to (2.2). We assume that

$$\text{(i) } Eu_n = \lambda_n^{-1},$$

$$\text{(ii) } \lim_{n\to\infty} \lambda_n^2 \operatorname{Var} u_n = c^2, \tag{3.1}$$

and

$$\text{(iii) } \sup_{n\geq 1} \lambda_n^3 Eu_n^3 < \infty.$$

For example, we could have $u_n = u/\lambda_n$ for all n where u is a fixed random variable with $Eu = 1$, $\operatorname{Var} u = c^2$ and $Eu^3 < \infty$.

We obtain the following extension of Theorem 2 for the processes Y_n in (2.15) by first considering the embedded Markov chains obtained by looking at the queue-length processes at arrival epochs. We apply the simple criteria for convergence of Markov chains in Theorems 10.2.2 and 11.2.3 of Stroock and Varadhan. We then use a random time change and the continuous mapping theorem to show that the original continuous-time processes obey a similar functional limit theorem.

To specify the initial conditions, let an arrival occur at time 0.

THEOREM 3. *Assume* (2.2) *and* (3.1). *If* $Y_n(0) \Rightarrow Y^*(0)$, *then* $Y_n \Rightarrow Y^*$ *in* $D[0, \infty)$, *where* Y^* *is a diffusion process with*

$$m(x) = \begin{cases} -\mu\beta, & x \geq 0 \\ -\mu(x + \beta), & x < 0 \end{cases}$$

and $$\sigma^2(x) = \mu(1 + c^2).$$

Proof. Let $Q_n^A(k)$ be the queue length in the nth system just prior to the kth arrival. It is well known that $\{Q_n^A(k), k \geq 0\}$ is an irreducible aperiodic positive-recurrent Markov chain for each n. Let

$$Z_n(t) = (Q_n^A([nt]) - n)/n^{1/2}, \qquad t \geq 0. \tag{3.2}$$

Under the conditions in (3.1), it is not difficult to show that $Z_n \Rightarrow Z$ in $D[0, \infty)$ for Z equal to Y^* in Theorem 3 with $\mu = 1$. By virtue of Theorems 10.2.2 and 11.2.3 of Stroock and Varadhan, it suffices to check the infinitesimal conditions as in the proof of Theorem 2 (see (2.4)-(2.6) on p. 268 of Stroock and Varadhan), which is a tedious but relatively straightforward task. For example, the infinitesimal mean is approximately $n^{1/2}\rho_n^{-1}(\rho_n - 1) \to -\beta$ for $x > 0$ and $n^{1/2}\rho_n(\rho_n - 1) - \rho_n^{-1}(n^{-1/2}[xn^{1/2}]) \to -\beta - x$ for $x < 0$. The most delicate point is showing that it suffices to assume that the number of busy servers does not change throughout an interarrival interval. Of course, the number of busy servers may change, but it is possible to show the adjustment is asymptotically negligible.

Having shown that Z_n converges, we can get the convergence of Y_n by performing a random time change; see Section 17 of Billingsley and Section 3 of Whitt (1980). Let $\{A_n(t), t \geq 0\}$ be the arrival process in the nth system and let $B_n(t) = A_n(t)/n$, $t \geq 0$. Then $B_n \xrightarrow{p} B$ in $D[0, \infty)$ where $B(t) = \mu t$, $t \geq 0$. Hence, $Z_n \circ B_n \Rightarrow Z \circ B = Y^*$ where $\circ$ is the composition map. The difference between $Z_n \circ B_n$ and Y_n is dominated by the jumps of Z_n, but the maximum jump in any bounded interval converges to 0 because Z_n has a limit with continuous paths. Hence, $d(Z_n \circ B_n, Y_n) \Rightarrow 0$ in $D[0, \infty)$, using the metric in Whitt say, and $Y_n \Rightarrow Y^*$ by Theorem 4.1 of Billingsley.

We now consider the associated heavy-traffic limit theorem for the steady-state distributions, i.e., for $Y_n(\infty)$. In order that $Y_n(t) \Rightarrow Y_n(\infty)$ as $t \to \infty$ for each n, we assume that u_n is nonlattice for each n. This is known to be necessary and sufficient; see p. 173 of Borovkov (1976) or Theorem 2.3 of Whitt (1972). It is important to note that convergence of $Y_n(\infty)$ as $n \to \infty$ does not follow immediately from Theorem 3. The gap is filled here by bounding the processes Y_n above and below by appropriate single-server systems for which heavy-traffic limit theorems for the steady-state distributions are known. In this way, we obtain

THEOREM 4. *If* (2.2) *and* (3.1) *hold, then*

(i) $Y_n(\infty) \Rightarrow Y^*(\infty)$, *where* $Y^*(\infty)$ *is distributed as* X *in Theorem* 1 *with* $\beta' = 2\beta/(1 + c^2)$ *instead of* β; *and*

(ii) $\lim_{n\to\infty} P(Q_n(\infty) \geq n) = P(Y^*(\infty) \geq 0) \equiv \alpha$, *where* α *is in* (2.3) *with* β' *instead of* β.

Remark. The detailed calculations for the $M/M/s$ queue in Sections 1 and 2 play a vital role in giving us explicit formulas for the limits in Theorem 4.

Proof. (i) First assume that the sequence $\{Y_n(\infty)\}$ is tight; see Section 6 of Billingsley. Then by Prohorov's Theorem (see Billingsley, Theorem 6.1) the sequence $\{Y_n(\infty)\}$ has a convergent subsequence $\{Y_{n'}(\infty)\}$. If we let $Y_{n'}(0)$ be distributed as $Y_{n'}(\infty)$, and let the time until the first arrival in the nth system have the stationary excess distribution associated with u_n (which is converging to 0 as $n \to \infty$ because $\lambda_n \to \infty$), then $\{Y_{n'}(t), t \geq 0\}$ is a strictly stationary stochastic process and, by a minor modification of Theorem 3 (to account for the new initial conditions), $Y_{n'} \Rightarrow \hat{Y}$ where $\hat{Y}$ is the limiting diffusion process with $\hat{Y}(0)$ having the distribution of the limit of $\{Y_{n'}(0)\}$. However, since $Y_{n'}$ is stationary for each n', so is $\hat{Y}$. Hence the limit of $\{Y_{n'}(\infty)\}$ must be the unique stationary distribution of $\hat{Y}$. (Uniqueness follows from Theorem 2 and Corollary 2.) Since every subsequence of $\{Y_n(\infty)\}$ that converges must converge to this same limit, the sequence $\{Y_n(\infty)\}$ itself must converge to this limit.

To complete the proof, it suffices to show that the sequence $\{Y_n(\infty)\}$ is tight. We shall do this by bounding $Y_n(\infty)$ above and below stochastically. In particular, we shall construct random variables $L_n(\infty)$ and $U_n(\infty)$ such that

$$P(L_n(\infty) \geq x) \leq P(Y_n(\infty) \geq x) \leq P(U_n(\infty) \geq x)$$

for all x and n, and

$$L_n(\infty) \Rightarrow L(\infty) \text{ and } U_n(\infty) \Rightarrow U(\infty) \text{ as } n \to \infty.$$

This convergence for $\{L_n(\infty)\}$ and $\{U_n(\infty)\}$ will be easy because $L_n(\infty)$ and $U_n(\infty)$ will correspond to normalizations of steady-state queue lengths in single-server queues, for which heavy-traffic limit theorems have already been established. Since $\{L_n(\infty)\}$ and $\{U_n(\infty)\}$ converge weakly, they are tight (Billingsley, Theorem 6.2). Hence, for any $\epsilon > 0$, there is an m such that

$$P(|L_n(\infty)| \geq m) < \epsilon/2 \text{ and } P(|U_n(\infty)| \geq m) \leq \epsilon/2$$

for all n. Consequently,

$$\begin{aligned} P(|Y_n(\infty)| \geq m) &\leq P(U_n(\infty) \geq m) + P(L_n \leq -m) \\ &\leq P(|U_n(\infty)| \geq m) + P(|L_n(\infty)| \geq m) \leq \epsilon, \qquad n \geq 1. \end{aligned}$$

We construct the stochastically bounding random variables $L_n(\infty)$ and $U_n(\infty)$ by constructing stochastically bounding stochastic processes $\{L_n(t), t \geq 0\}$ and $\{U_n(t), t \geq 0\}$ which converge weakly, i.e.,

$$P(L_n(t) \geq x) \leq P(Y_n(t) \geq x) \leq P(U_n(t) \geq x)$$

for all x, t and n, and

$$L_n(t) \Rightarrow L_n(\infty) \text{ and } U_n(t) \Rightarrow U_n(\infty) \text{ as } t \to \infty$$

for each n. This implies the desired bounding relation for the limiting distributions; see Proposition 3 of Kamae et al. (1977).

Hence it suffices to construct the two bounding processes $\{L_n(t), t \geq 0\}$ and $\{U_n(t), t \geq 0\}$. We shall actually use the stronger stochastic order discussed in Kamae et al. and Whitt (1981). We shall show that

$$P(f(L_n) \geq x) \leq P(f(Y_n) \geq x) \leq P(f(U_n) \geq x) \tag{3.3}$$

for all x, n and nondecreasing measurable real-valued functions f on the space $D[0, \infty)$ of sample paths. We construct the process U_n simply by introducing a lower impenetrable barrier at 0. (Since the barrier is for the normalized process, the barrier is at n in the unnormalized process.) Let $U_n(0) = \max\{0, Y_n(0)\}$. Clearly, $U_n(0) \Rightarrow \max\{0, Y(0)\}$ if $Y_n(0) \Rightarrow Y(0)$. We construct U_n by letting U_n and Y_n have identical arrival processes. Moreover, it is convenient to work with the embedded Markov chains, i.e., Z_n in (3.2) and the associated process U_n' obtained by introducing a lower impenetrable barrier at 0. By the known relations between the two stationary distributions (see Borovkov [1976], p. 182), we know both sequences $\{Y_n(\infty)\}$ and $\{Z_n(\infty)\}$ are tight if one is. It is easy to show that the criteria for two Markov chains to be stochastically ordered are satisfied in this case (Kamae et al.), so that Z_n is dominated by U_n' in the sense of (3.3) for each n. Moreover, it is known that $U_n'(\infty) \Rightarrow U'(\infty)$ because U_n' coincides with the normalization of the queue length process of a $GI/M/1$ queue having individual service rate $n\mu$, i.e., before normalization the sequence is of the form $\{\xi_k\}$ where $\xi_{k+1} = \max\{\xi_k + \zeta_k, 0\}$ with $\{\zeta_k\}$ being i.i.d. The heavy-traffic limit theorem for the stationary distributions was established for this system by Kingman (1962). Kingman actually studied the embedded sequence of waiting times at arrival epochs, but the embedded queue-length sequence in the $GI/M/1$ model has the same structure.

The construction of the lower bound is similar, but slightly more complicated. Again we let L_n have the same arrival process as Y_n and focus on the embedded chains. We construct the lower bounding process by replacing the asymptotic positive state-dependent drift of $-(x + \beta)$ by the constant drift of $+\beta$ for $x \leq -2\beta$ and by introducing an impenetrable upper barrier at $x = -2\beta$. We achieve this in the embedded Markov

chains by introducing corresponding barriers and increasing the rates of the exponential departures during the interarrival times. Again it is easy to see that the conditions for the discrete-time Markov chains to be stochastically ordered are satisfied. We let $L_n(0) = \min\{-2\beta, Y_n(0)\}$, so the initial conditions are appropriate too. We obtain the desired convergence by noticing that $-2\beta - L_n(t)$ has the same structure as the normalized queue-length process in the $M/G/1$ queue, i.e., before normalization the sequence has the form $\{-\xi_k\}$, where $\xi_{k+1} = \max\{\xi_k, 1\} + \zeta_k$ with $\{\zeta_k\}$ being i.i.d. After subtracting from -2β, the embedded process of interest corresponds to the departure points in the $M/G/1$ queue. The heavy-traffic limit theorem for the stationary distributions of this discrete-time process is again known; see Gnedenko and Kovalenko ([1968], p. 147).

4. *GI/G/s* QUEUES

The most important fact about $GI/G/s$ queues with nonexponential service times under condition (2.2) is that the properly normalized queue-length process is not asymptotically Markov. Because the number of servers is very large, the residual service times are not asymptotically negligible in heavy traffic. Moreover, it does not seem possible to obtain heavy-traffic limit theorems for $GI/G/s$ queues by the same elementary proofs, and the prospective limit process is not well understood. One approach is to represent the service time as a finite random sum or mixture of exponential phases, as in Whitt (1982). It is known that the class of distributions of this form is dense in the family of all service time distributions, so this procedure covers essentially all $GI/G/s$ systems. This procedure makes the particular $GI/G/s$ system equivalent to an acyclic network of queues with a single external arrival process and a constraint on the total number of customers that can be in the network. If the total number of customers in the network reaches s (the number of servers), then external arrivals must wait before entering the network. To convert this into the usual open network, we add a node for the waiting customers not yet in service. The vector-valued discrete-time process indicating the number of customers waiting and in each service phase at arrival epochs is again a Markov chain but the simple criteria in Stroock and Varadhan to obtain convergence to a diffusion process are not satisfied.

We illustrate this approach by stating (without displaying the calculations) a heavy-traffic limit theorem for $GI/H_2/s$ systems, where the service time distribution is hyperexponential, i.e., a mixture of two exponential distributions, with density

$$g(x) = p_1\mu_1 e^{-\mu_1 x} + p_2\mu_2 e^{-\mu_2 x}, \qquad x \geq 0, \tag{4.1}$$

where $p_2 = 1 - p_1$. The mean service time is thus $\mu^{-1} = p_1\mu_1^{-1} + p_2\mu_2^{-1}$. However, here our heavy-traffic limit theorem involves only the convergence of infinitesimal means and variances. Thus this result falls far short of Theorems 2 and 3, but it can serve as the basis for approximating because this weaker result also identifies a limit process. Hence, the limit process can be used to generate approximations. However, the limit process is much less tractable than in the exponential case. Nevertheless, we believe the multivariate diffusion process arising in this $GI/H_2/s$ case has great potential for approximating the behavior of $GI/G/s$ queues. What we would propose is first approximating a general service-time distribution (with coefficient of variation greater than one) by an H_2 distribution. Then use the $GI/H_2/s$ diffusion approximation or a related random walk.

Let $Q_n^0(k)$ be the number of customers waiting and $Q_n^i(k)$ the number of customers in phase i ($i = 1, 2$) of service in the nth system at the epoch of the kth arrival. Let

$$\mathbf{Z}_n(t) = [Z_n^0(t), Z_n^1(t), Z_n^2(t)]$$
$$= n^{-1/2}[Q_n^0([nt]), Q_n^1([nt]) - n\alpha_1, Q_n^2([nt]) - n\alpha_2], \qquad t \geq 0, \tag{4.2}$$

$$\alpha_i = p_i\mu/\mu_i, \qquad i = 1, 2. \tag{4.3}$$

PROPOSITION 3. *If* (2.2), (3.1) *and* (4.1) *hold, then the sequence of processes* $\{\mathbf{Z}_n\}$ *converges to a diffusion process in the sense that infinitesimal mean vectors and covariance matrices of* $\mathbf{Z}_n$ *converge to an infinitesimal mean vector* $[m_0(\mathbf{x}), m_1(\mathbf{x}), m_2(\mathbf{x})]$ *and an infinitesimal covariance matrix* $(\sigma_{ij}(\mathbf{x}))$ *of the form*:

(i) $m_0(\mathbf{x}) = 0$ *and* $m_i(\mathbf{x}) = -x_i\mu_i/\mu$, $x_0 = 0$ *and* $x_1 + x_2 < 0$;
$m_0(\mathbf{x}) = -\beta - (x_1\mu_1 + x_2\mu_2)/\mu$ *and*
$m_1(\mathbf{x}) = (p_1x_2\mu_2 - (1 - p_1)x_1\mu_1)/\mu = -m_2(\mathbf{x})$, $x_0 > 0$ *and* $x_1 + x_2 = 0$;

(ii) $\sigma_{0i}^2(\mathbf{x}) = 0$, $\sigma_{ii}^2(x) = 2p_i + (c^2 - 1)p_i^2$,
$\sigma_{12}^2(\mathbf{x}) = p_1p_2(c^2 - 1)$, *for* $x_0 = 0$ *and* $x_1 + x_2 < 0$;
$\sigma_{00}^2(\mathbf{x}) = 1 + c^2$, $\sigma_{0i}^2(\mathbf{x}) = 0$,
$\sigma_{ii}^2(\mathbf{x}) = 2p_1p_2$, $\sigma_{12}^2(\mathbf{x}) = -2p_1p_2$, *for* $x_0 > 0$ *and* $x_1 + x_2 = 0$.

5. APPROXIMATIONS AND NUMERICAL COMPARISONS

The limit theorems yield approximations for a given queueing system if we regard the given system as the nth system in a converging sequence and replace the limit by an approximate equality. The limit theorems for stochastic processes yield approximations for the time-dependent behavior as well as for the stationary distributions, but here we focus on the

stationary distributions. For example, Theorem 4 yields the following approximation for the stationary queue length in a $GI/M/s$ system:

$$Q(\infty) \approx s + \sqrt{s}\, X(\beta) \tag{5.1}$$

with

$$\beta = 2(1 - \rho)\sqrt{s}/(1 + c^2), \tag{5.2}$$

where $X(\beta)$ is the limit in Theorem 1 and c is the coefficient of variation (standard deviation divided by the mean) of the interarrival time distribution.

TABLE I

A COMPARISON OF THE THREE HEAVY-TRAFFIC APPROXIMATIONS: CASE OF $s = 100$ SERVERS

Method	Characteristic	Traffic Intensity	
		$\rho = 0.949$	$\rho = 0.834$
$M/M/s$	$EQ(\infty)$	105.3	84.6
True system	$P(Q(\infty) \geq s)$	0.50	0.05
No. 1[a]	$EQ(\infty)$	118.6	105.1
Exponential	$P(Q(\infty) \geq s)$	1.00	1.00
No. 2[b]	$EQ(\infty)$	94.9	83.4
Normal	$P(Q(\infty) \geq s)$	0.30	0.035
No. 3[c]	$EQ(\infty)$	104.9	83.7
Hybrid	$P(Q(\infty) \geq s)$	0.50	0.060

[a] The traffic intensity ρ was determined by fixing s and $P(Q \geq s)$ for the $M/M/s$ system.

[b] Nos. 1 and 2 come from (0.1) and (0.2), respectively.

[c] For No. 3, we let $\beta = (1 - \rho)\sqrt{s}$.

However, the limit theorem does not uniquely specify the approximation. For example,

$$Q(\infty) \approx s + \sqrt{s}\, X(\beta) + s^{1/4} W \tag{5.3}$$

is also consistent with Theorem 4 for any random variable W. Of course, extraneous terms such as $s^{1/4}W$ in (5.3) do not usually arise, but this phenomenon indicates that some care should be taken to select an appropriate approximation. A real difficulty of this kind arises with the first heavy-traffic limit theorem; see the approximation in (0.1). Since the number of servers is fixed, the heavy-traffic limit including the customers in service is the same as the heavy-traffic limit excluding the customers in service. However, for a given system with a relatively large number of

servers the resulting approximations are quite different. Experience indicates that it is much better to regard the exponential approximation as applying only to the customers waiting, excluding the customers in service. In fact, for moderately high loads such as $\rho = 0.9$ and large numbers of servers such as 100 the exponential approximation is fairly reasonable if we exclude the customers in service, but ridiculous if we do not. Another possible approximation for $Q(\infty)$ suggested by D. P. Heyman is the exponential random variable X with mean in (0.1) plus ρs, the mean number of busy servers. The approximate mean in (0.1) was obtained by using the mean for the waiting time in the limit theorem of Köllerström, and the relation $L = \lambda W$.

We have compared the three heavy-traffic approximations with the actual stationary distribution for the $M/M/s$ queue. As should be expected, each approximation has regions where it tends to perform well. When the number of servers is relatively large (such as $s = 100$) and the load is fairly heavy (such as $\rho = 0.85$), the new hybrid approximation tends to be best. In particular, the conditional distribution of the number of customers waiting given that all servers are busy is better described by the exponential distribution associated with the hybrid approximation than by the normal distribution. A specific numerical comparison is given in Table I.

ACKNOWLEDGMENTS

We are grateful to the referees for their helpful comments. We would like to thank David Burman for showing us the paper by Rhee.

REFERENCES

BILLINGSLEY, P. 1968. *Convergence of Probability Measures.* John Wiley & Sons, New York.

BOROVKOV, A. A. 1965. Some Limit Theorems in the Theory of Mass Service, I. *Theor. Prob. Appl.* **10,** 375–400.

BOROVKOV, A. A. 1967. On Limit Laws for Service Processes in Multi-channel Systems. *Siberian Math. J.* **8,** 746–763.

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

BREIMAN, L. 1968. *Probability.* Addison-Wesley, Reading, Mass.

CHANDY, K. M., AND C. H. SAUER. 1978. Approximate Methods for Analyzing Queueing Network Models for Computer Systems. *Comput. Surv.* **10,** 281–317.

COOPER, R. B. 1972. *Introduction to Queueing Theory.* Macmillan, New York.

FELLER, W. 1968. *An Introduction to Probability Theory and Its Applications, Vol. I,* Ed. 3. John Wiley & Sons, New York.

GNEDENKO, B. V., AND I. N. KOVALENKO. 1968. *Introduction to Queueing Theory.* Israel Program for Scientific Translation, Jerusalem.

HALACHMI, B., AND W. R. FRANTA. 1978. A Diffusion Approximation to the Multi-Server Queue. *Mgmt. Sci.* **24,** 522–529.

HARRISON, J. M. 1978. The Diffusion Approximation for Tandem Queues in Heavy Traffic. *Adv. Appl. Prob.* **10,** 886–905.

IGLEHART, D. L. 1965. Limit Diffusion Approximations for the Many-Server Queue and the Repairman Problem. *J. Appl. Prob.* **2,** 429–441.

IGLEHART, D. L. 1973a. Weak Convergence in Queueing Theory. *Adv. Appl. Prob.* **5,** 570–594.

IGLEHART, D. L. 1973b. Weak Convergence of Compound Stochastic Processes, I. *Stoch. Proc. Appl.* **1,** 11–31.

IGLEHART, D. L., AND W. WHITT. 1970. Multiple Channel Queues in Heavy Traffic; II. Sequences, Networks, and Batches. *Adv. Appl. Prob.* **2,** 355–369.

KAMAE, T., U. KRENGEL AND G. O'BRIEN. 1977. Stochastic Inequalities on Partially Ordered Spaces. *Ann. Prob.* **5,** 899–912.

KARLIN, S., AND J. MCGREGOR. 1964. On Some Stochastic Models in Genetics. In *Stochastic Models in Medicine and Biology*, pp. 245–279, J. Gurland (ed.). University of Wisconsin Press, Madison.

KINGMAN, J. F. C. 1962. On Queues in Heavy Traffic. *J. Roy. Stat. Soc.* Ser. B, **24,** 383–392.

KINGMAN, J. F. C. 1965. The Heavy Traffic Approximation in the Theory of Queues. In *Proceedings of Symposium on Congestion Theory*, pp. 137–159, W. Smith and W. Wilkinson (eds.). University of North Carolina Press, Chapel Hill, N.C.

KÖLLERSTRÖM, J. 1974. Heavy Traffic Theory for Queues with Several Servers, I. *J. Appl. Prob.* **11,** 544–552.

LEMOINE, A. J. 1978. Networks of Queues—A Survey of Weak Convergence Results. *Mgmt. Sci.* **24,** 1175–1193.

LINDVALL, T. 1973. Weak Convergence of Probability Measures and Random Functions in the Function Space $D[0, \infty)$. *J. Appl. Prob.* **10,** 109–121.

NEWELL, G. F. 1973. *Approximate Stochastic Behavior of n-Server Service Systems with Large n* (Lecture Notes in Economics and Mathematical Systems, No. 87). Springer-Verlag, New York.

PROHOROV, YU. 1963. Transient Phenomena in Processes of Mass Service (in Russian). *Litovsk. Mat. Sb.* **3,** 199–205.

RHEE, J. J. 1977. The Expectation Value and the Variance for the Poisson Distribution. *Am. J. Phys.* **46,** 769–770.

STONE, C. 1961. *Limit Theorems for Birth and Death Processes and Diffusion Processes*, Ph.D. thesis, Department of Mathematics, Stanford University.

STONE, C. 1963. Limit Theorems for Random Walks, Birth and Death Processes, and Diffusion Processes. *Ill. J. Math.* **4,** 638–660.

STROOCK, D., AND S. R. S. VARADHAN. 1979. *Multidimensional Diffusion Processes.* Springer-Verlag, New York.

WHITT, W. 1972. Embedded Renewal Processes in the $GI/G/s$ Queue. *J. Appl. Prob.* **9,** 650–658.

WHITT, W. 1974. Heavy Traffic Limit Theorems for Queues: A Survey. In *Mathematical Methods in Queueing Theory* (Lecture Notes in Economics and Mathematical Systems, No. 98), pp. 307–350. Springer-Verlag, Berlin.

WHITT, W. 1980. Some Useful Functions for Functional Limit Theorems. *Math. Opns. Res.* **5,** 67–85.

WHITT, W. 1981. Comparing Counting Processes and Queues. *Adv. Appl. Prob.* (to appear).

WHITT, W. 1982. On the Heavy-Traffic Limit Theorem for $GI/G/\infty$ Queues. *Adv. Appl. Prob.* (to appear).