MANAGEMENT SCIENCE
Vol. 50, No. 10, October 2004, pp. 1449–1461
ISSN 0025-1909 | EISSN 1526-5501 | 04 | 5010 | 1449

informs

DOI 10.1287/mnsc.1040.0279
© 2004 INFORMS

Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, Mudd Building, 500 West 120th Street, New York, New York 10027-6699, ward.whitt@columbia.edu

To provide useful practical insight into the performance of service-oriented (non-revenue-generating) call centers, which often provide low-to-moderate quality of service, this paper investigates the efficiency-driven (ED), many-server heavy-traffic limiting regime for queues with abandonments. Attention is focused on the $M/M/s/r+M$ model, having a Poisson arrival process, exponential service times, s servers, r extra waiting spaces, exponential abandon times (the final $+M$), and the first-come–first-served service discipline. Both the number of servers and the arrival rate are allowed to increase, while the individual service and abandonment rates are held fixed. The key is how the two limits are related: In the now common quality-and-efficiency-driven (QED) or Halfin-Whitt limiting regime, the probability of initially being delayed approaches a limit strictly between 0 and 1, while the probability of eventually being served (not abandoning) approaches 1. In contrast, in the ED limiting regime, the probability of eventually being served approaches a limit strictly between 0 and 1, while the probability of initially being delayed approaches 1. To obtain the ED regime, it suffices to let the arrival rate and the number of servers increase with the traffic intensity ρ held fixed with $\rho > 1$ (so that the arrival rate exceeds the maximum possible service rate). The ED regime can be realistic because with the abandonments, the delays need not be extraordinarily large. When the ED appropriations are appropriate, they are appealing because they are remarkably simple.

Key words: call centers; contact centers; queues; multiserver queues; queues with customer abandonment; multiserver queues with customer abandonment; Erlang-A model; heavy-traffic limits; many-server heavy-traffic limits; efficiency-driven limiting regime
History: Accepted by Wallace J. Hopp, stochastic models and simulation; received February 6, 2004. This paper was with the author 9 days for 2 revisions.

1. Introduction

Recently, there has been great interest in multiserver queues with a large number of servers, motivated by applications to telephone call centers (see Mandelbaum 2001, Gans et al. 2003). For the basic Erlang-C model, i.e., for the $M/M/s/\infty$ model (having a Poisson arrival process, independent and identically distributed (IID) exponential service times, s servers, unlimited waiting space, and the first-come–first-served service discipline) and generalizations with more general arrival and service processes, useful insight can be gained by considering many-server heavy-traffic limits in which the number s of servers increases along with the arrival rate λ (with the individual service rate μ held fixed) so that the traffic intensity $\rho \equiv \lambda/s\mu$ increases too, with

$$(1-\rho)\sqrt{s} \to \beta \quad \text{as } s \to \infty, \tag{1.1}$$

where $0 < \beta < \infty$ (see Halfin and Whitt 1981; Puhalskii and Reiman 2000; Jelenkovic et al. 2004; Whitt 2005a, b). In this so-called *Halfin-Whitt limiting regime* or *Quality and Efficiency-Driven* (QED) *limiting regime*, the steady-state probability of delay approaches a limit strictly between 0 and 1. In contrast, if we only increase ρ, keeping s fixed, then the steady-state probability of delay approaches 1 as $\rho \uparrow 1$; if instead we only increase s, keeping ρ fixed with $\rho < 1$, then the steady-state probability of delay approaches 0.

Garnett et al. (2002) showed that the same QED limiting regime is also useful for multiserver queues with customer abandonment, specifically for the associated Erlang-A model, i.e., for the purely Markovian $M/M/s/\infty+M$ model with exponential abandonment times (the final $+M$) (also see Whitt 2005b for a generalization to $G/M/s/\infty+M$). Then, because the abandonment ensures stability, the limit β in (1.1) can be negative as well as positive. Again, in this QED limiting regime, the steady-state probability of initially being delayed approaches a limit strictly between 0 and 1. However, the probability of abandonment is asymptotically negligible; specifically, the steady-state probability of abandonment is asymptotically of order $1/\sqrt{s}$ as $s \to \infty$.

Our purpose here is to point out that, in the presence of significant customer abandonment, the

QED limiting regime is not the only many-server heavy-traffic limiting regime worth considering to generate approximations that are useful in practice. Here we focus attention on the alternative *Efficiency-Driven* (ED) *limiting regime*, in which the probability of abandonment approaches a limit strictly between 0 and 1, while the probability of initially being delayed approaches 1. As shown by Garnett et al. (2002) for the $M/M/s+M$ model, such a limit occurs when ρ approaches a limit strictly greater than 1 as $\lambda \to \infty$ and $s \to \infty$. More simply, it suffices to keep ρ fixed with $\rho > 1$ as $\lambda \to \infty$ and $s \to \infty$.

In practice, of course, we are usually interested in a *single* queueing system with specified parameters, not a sequence of queueing systems. To apply one of these heavy-traffic limits, we think of our given system as one term (term s) in a sequence of systems indexed by s, where $s \to \infty$ (e.g., see pp. 58 and 158 of Whitt 2002 or Equations (2.21) and (3.4) here). Thus, if there are two different limits that might be considered, it is not evident a priori which of these limits should be more useful. Then, the resulting approximations can be judged by their effectiveness, i.e., by their accuracy and ease of use.

It is our experience that in typical scenarios with a large number of servers (e.g., $s = 100$) and only moderate abandonment that the QED approximations are usually more accurate than the ED approximations. However, when the quality of service is somewhat low, the ED approximations become appropriate. The ED approximations become appropriate when queue length and waiting times are relatively large, e.g., in environments such as the Internal Revenue Service help lines. The ED approximations may be relevant when absenteeism is a problem; then agent work scheduling may be aiming to be in the QED regime, but end up in the ED regime unintentionally. The great appeal of the ED approximations, when they are appropriate, is their simplicity; they often can be used for quick back-of-the-envelope calculations.

The ED approximations require having $\rho > 1$, so we may not expect them to be very useful. However, data from service-oriented call centers show that the arrival rate often exceeds the maximum possible service rate over measurement intervals, even when target performance levels occurring in service-level agreements (SLAs) are being met. In addition, computational results from computer simulations and numerical algorithms substantiate that performance targets often can be met when the arrival rate exceeds the maximum possible service rate. Specifically, in Whitt (2005c), we developed a numerical algorithm for calculating approximations for all the standard performance measures in the $M/GI/s/r+GI$ model with large s, in which the service times and abandon times come from independent sequences of IID random variables with general distributions. While studying how the approximations perform compared to simulations, we saw that it is often reasonable to have the arrival rate exceed the maximum possible service rate when there are many servers and significant abandonment.

An initial example has 100 agents (servers) handling calls with mean holding (service) time and mean abandon time both equal to five minutes. The SLA may stipulate that at most 5% of the customers should abandon and that 80% of the calls that eventually are served should be answered within 30 seconds. The $M/M/s+M$ model indicates that the 100 agents can handle an arrival rate of 20.4 calls per minute, yielding a traffic intensity of $\rho = 1.02 > 1$. With that arrival rate, the SLA is just met: 5% of the arrivals abandon and 80% of the answered calls are answered within 30 seconds. Some call centers provide even lower quality of service; then we may encounter even higher traffic intensities.

That initial example illustrates that, with substantial abandonments, we may well have $\rho > 1$ when the SLA is met. However, even though $\rho = 1.02 > 1$, the QED approximations perform well, being far superior to the ED approximations. For example, the QED approximation for the 0.05 abandonment probability is 0.051, whereas the ED approximation is 0.02. In this case, we will show that the ED approximations provide only a rough indication of performance. Nevertheless, the ED approximations may be useful.

The ED approximations become more appropriate with a further degradation of service. For example, suppose that the arrival rate increases by 8% from 20.4 calls per minute to 22.0 calls per minute, increasing the traffic intensity from $\rho = 1.02$ to $\rho = 1.10$. Now, as shown by Tables 1 and 2 in §5, the ED approximations are quite good. For example, the exact abandonment probability is 0.10, while the ED approximation is 0.09; the exact mean queue length is 10.9, while the ED approximation is 10.0. Moreover, further refined ED approximations have less than 1% error.

Without a finite waiting room or customer abandonment, steady-state distributions do not exist when $\rho > 1$. Then $\rho > 1$ is simply an overloaded regime, because the queue length explodes as time evolves. Then the overloaded regime is interesting primarily to describe transient behavior.

Without a finite waiting room or customer abandonment, it is possible to define an underloaded ED limiting regime, in which ρ is less than 1 but extremely close to 1. Indeed, that is the conventional heavy-traffic limiting regime with a single server. However, with a large number of servers, ρ has to be so close to 1 that the system becomes unstable in the sense that a very small increase in the traffic intensity pushes the system into the overloaded regime.

But the ED many-server heavy-traffic limiting regime becomes viable with customer abandonment. Even a small amount of abandonment can keep the system stable when the arrival rate exceeds the maximum possible service rate. Even though the steady-state probability of initially being delayed approaches 1 in the ED limit, this alternative ED heavy-traffic limit can be realistic because the delays experienced need not be extraordinarily large. We believe that the ED regime does describe the operation of many existing call centers remarkably well, especially when a great emphasis is placed on efficiency. An emphasis on efficiency is more common among call centers that are service oriented instead of revenue generating.

Even though in this paper we focus on a regime supporting low-to-moderate quality of service, we do not advocate providing low-to-moderate quality of service in call centers. Indeed, as suggested in Whitt (1999), it may be possible to provide spectacular quality of service without requiring a commitment of excessive resources with good planning and good execution. However, to improve the quality of service, it is important to understand the performance of existing call centers. We contend that the ED limiting regime can be helpful to understand the performance of existing call centers providing low-to-moderate quality of service.

In this paper, we establish limits in the ED heavy-traffic regime and develop approximations based on those limits. We only consider Markovian birth-and-death models. For the Markovian models considered here, the limits are not difficult to obtain; indeed, we establish them by the same argument used in the seminal heavy-traffic paper on multiserver queues by Iglehart (1965), drawing on Stone (1963). The stochastic-process limits here also can be viewed as consequences of more general results for state-dependent Markovian queues in Mandelbaum and Pats (1995; see Theorems 4.1 and 4.2 and §5.3 there), but the relatively complicated full framework of Mandelbaum and Pats (1995) is not needed to treat the special case considered here.

The main contribution here, we believe, is communicating that this ED many-server heavy-traffic limiting regime can indeed yield useful approximations. When this ED regime is appropriate, the heavy-traffic limit is very helpful because it generates remarkably simple approximations (e.g., see (3.1)–(3.5)). In particular, the ED approximations are simple even in comparison to the QED approximations. The ED approximations can be useful even if they are less accurate than the QED approximations. Indeed, because there already are effective exact numerical algorithms to calculate all desired performance measures in the $M/M/s/r+M$ model, as in Whitt (2005c), only truly simple approximations can provide significant value added.

The rest of this paper is organized as follows. We begin in §2 by establishing the stochastic-process limit for the number in system in the $M/M/s/r+M$ model in the ED regime. We also establish a deterministic fluid limit and a limit for the steady-state distributions in the ED regime. In §3, we develop heuristic approximations for steady-state performance measures based on the limits.

Interestingly, the approximation for the steady-state queue length in the ED regime (see (3.4) and (3.6)) depends on the number of servers, s, and the individual abandonment rate, α, only through the ratio s/α. It is thus natural to wonder if we can obtain a related limit as $\alpha \to 0$ and, indeed, such a limit was established by Ward and Glynn (2003) (assuming fixed s; see Part 4 of Theorem 1 and Remark 5 there). In §4, we show that the same ED limit holds more generally when ρ is held fixed with $\rho > 1$ and $s/\alpha \to \infty$, because either the individual abandonment rate α becomes small or the number s of servers becomes large or both. The main approximation developed by Ward and Glynn (2003) for fixed s stems from a double limit in which $\rho \to 1$ and $\alpha \to 0$, so that

$$(1-\rho)/\sqrt{\alpha} \to \beta, \tag{1.2}$$

in the spirit of (1.1), which defines the QED limiting regime. When $\alpha \to 0$, we also emphasize the value of the alternative ED regime in which ρ is fixed with $\rho > 1$.

In §5, we compare the ED approximations to exact numerical solutions for the $M/M/s/r+M$ model, which we obtain using the algorithm in Whitt (2005c). (That algorithm producing approximate performance measures for the $M/GI/s/r+GI$ model produces exact numerical results for the $M/M/s/r+M$ special case.) We show that the performance of the ED approximations ranges from poor to spectacular. We provide ways to judge when the ED approximations will be effective.

The paper ends in §6 with our conclusions. We conclude this introduction by mentioning two companion papers: In Whitt (2005a), we develop deterministic fluid approximations for the general $G/GI/s/r+GI$ model in the ED many-server heavy-traffic regime. Unlike here, the emphasis there is on trying to account for the impact on performance of a nonexponential service-time distribution and a nonexponential abandon-time distribution. In Whitt (2005e), we extend the ED many-server heavy-traffic limits here to $M(n)/M(n)/s/r+M(n)$ models with state-dependent rates. Our motivation there is to gain additional insight into the performance of the $M/M/s/r+M(n)$ approximation of the $M/GI/s/r+GI$ model developed in Whitt (2005c).

For additional discussion about customer abandonment in queues, see Brandt and Brandt (1999, 2002), Zohar et al. (2002), Ward and Glynn (2003), Mandelbaum and Zeltyn (2004), and the references therein.

2. Limits for the Erlang-A Model in the ED Regime

In this section, we establish ED many-server heavy-traffic limits for the $M/M/s+M$ model, also known as the Erlang-A model. We actually treat the $M/M/s/r+M$ model, allowing finite waiting room r as well as infinite waiting room ($r \le \infty$). When $r < \infty$, we require that r be sufficiently large that it is not a factor, e.g., see (2.2) below. Arrivals finding all servers busy and all waiting spaces full are blocked and lost, without affecting future arrivals. Entering customers are served in order of arrival by the first available server, but waiting customers may elect to abandon before they start service. Customers do not abandon after they start service.

We choose measuring units for time so that the individual mean service time is $1/\mu = 1$. The model is thus characterized by four parameters: (1) the arrival rate, λ, (2) the number of servers, s, (3) the number of extra waiting spaces, r, and (4) the individual abandonment rate, α. The assumption of exponential abandonment rates is equivalent to the customers having IID times to abandon before beginning service, with a common exponential distribution having mean $1/\alpha$.

We consider a sequence of these $M/M/s/r+M$ models indexed by s. Let λ_s, r_s, and α_s be the remaining parameters, as a function of s. We increase λ_s and r_s with s, but we leave the individual service rate $\mu = 1$ and the individual abandonment rate $\alpha_s = \alpha$ fixed, independent of s. We let the traffic intensity ρ remain fixed. (It suffices to let $\rho_s \to \rho > 1$.) In particular, we assume that

$$\lambda_s = \rho s \quad \text{where } \rho > 1 \tag{2.1}$$

and

$$r_s = \eta s \quad \text{where } \eta > q, \tag{2.2}$$

with

$$q \equiv \frac{\rho - 1}{\alpha} \tag{2.3}$$

for all s. Condition (2.1) determines the ED regime. Condition (2.2) ensures that, asymptotically, no customers are blocked, as we will show below.

Let $N_s(t)$ be the number of customers in the system at time t when there are s servers. The ED regime is relatively tractable because, in the ED regime, $N_s(t)$ tends to concentrate about a fixed value; i.e., for large s, we will show that

$$N_s(t) \approx (1+q)s \tag{2.4}$$

for q defined in (2.3). Heuristically, we obtain q in (2.3) by finding the point, say xs, where the input rate equals the output rate. Clearly, that can occur only with all servers busy ($x > 1$). Before scaling by dividing by s, we have the equation

$$\lambda_s = s + \alpha(xs - s); \tag{2.5}$$

after dividing by s and letting $s \to \infty$, we obtain the equation

$$\rho = 1 + \alpha x. \tag{2.6}$$

The solution to these equations is $x = q$ for q in (2.3).

The diffusion approximation is a refinement of the deterministic approximation in (2.4) and (2.3). To establish convergence to a diffusion process, we form the normalized stochastic process

$$\mathbf{N}_s(t) \equiv \frac{N_s(t) - s(1+q)}{\sqrt{s}}, \quad t \ge 0 \tag{2.7}$$

for q in (2.3). (Throughout this paper, we use boldface to denote normalized processes and their limits.) Let the initial state $N_s(0)$ be specified independently, so that the stochastic process $\{N_s(t): t \ge 0\}$ is Markov. To establish a stochastic-process limit for the processes $\mathbf{N}_s$, let $D \equiv D([0,\infty), \mathbb{R})$ denote the space of all right-continuous real-valued functions on the positive half line $[0,\infty)$ with left limits everywhere in $(0,\infty)$, endowed with the usual Skorohod J_1 topology (see Billingsley 1999 or Whitt 2002). Let $\Rightarrow$ denote convergence in distribution (weak convergence), both for sequences of stochastic processes in D or for sequences of random variables in $\mathbb{R}$. Let $\mathrm{Nor}(m, \sigma^2)$ denote a random variable that is normally distributed with mean m and variance σ^2.

Theorem 2.1 (Stochastic-Process Limit for the Erlang-A Model in the ED Regime). *Consider the sequence of $M/M/s/r+M$ models specified above, satisfying* (2.1)–(2.3). *If $\mathbf{N}_s(0) \Rightarrow \mathbf{N}(0)$ as $s \to \infty$, then*

$$\mathbf{N}_s \Rightarrow \mathbf{N} \quad \text{in } D \quad \text{as } s \to \infty, \tag{2.8}$$

where $\mathbf{N}_s$ is the scaled process in (2.7) *and $\mathbf{N}$ is an Ornstein-Uhlenbeck (OU) diffusion process with infinitesimal mean (state-dependent drift)*

$$m(x) = -\alpha x \tag{2.9}$$

and infinitesimal variance

$$\sigma^2(x) = 2\rho, \tag{2.10}$$

which has steady-state distribution

$$\mathbf{N}(\infty) \stackrel{\mathrm{d}}{=} \mathrm{Nor}(0, \rho/\alpha). \tag{2.11}$$

PROOF. Because N_s is a birth-and-death process and the limiting OU diffusion process has no boundaries, we can apply the weak convergence theory in Stone (1963), just as Iglehart (1965) did in his seminal paper. Given Stone (1963), with the scaling in (2.7), it suffices to show that the infinitesimal mean and variances converge to the infinitesimal mean and variance of the limit process.

Because $N_s(t)$ is nonnegative integer valued, the possible values of $\mathbf{N}_s(t)$ are $[k - s(1+q)]/\sqrt{s}$ for $k \geq 0$. Hence, for arbitrary real number x, we consider a sequence $\{x_s\colon s \geq 1\}$, where x_s is an allowed value of $\mathbf{N}_s(t)$ for each s and $x_s \to x$ as $s \to \infty$. For example, for all sufficiently large s, we can construct an allowed value by letting

$$x_s \equiv \frac{\lfloor s(1+q) + x\sqrt{s} \rfloor - s(1+q)}{\sqrt{s}},$$

where $\lfloor t \rfloor$ is the *floor* function, i.e., the greatest integer less than or equal to t. When $x < 0$, we need s to be sufficiently large to guarantee that $\lfloor s(1+q) + x\sqrt{s} \rfloor \geq 0$.

For any real number x and sequence of allowed values $\{x_s\colon s \geq 1\}$, the infinitesimal means are

$$\begin{aligned} m_s(x_s) &\equiv \lim_{h\to 0} E[(\mathbf{N}_s(t+h) - \mathbf{N}_s(t))/h \,|\, \mathbf{N}_s(t) = x_s] \\ &= \lim_{h\to 0} E\left[\frac{N_s(t+h) - N_s(t)}{h\sqrt{s}} \,\middle|\, N_s(t) = s + ((\rho-1)/\alpha)s + x_s\sqrt{s}\right] \\ &= \frac{\rho s - s - \alpha(((\rho-1)/\alpha)s + x_s\sqrt{s})}{\sqrt{s}} \\ &\to -\alpha x \equiv m(x) \quad \text{as } s \to \infty, \end{aligned}$$

and the infinitesimal variances are

$$\begin{aligned} \sigma_s^2(x_s) &\equiv \lim_{h\to 0} E[(\mathbf{N}_s(t+h) - \mathbf{N}_s(t))^2/h \,|\, \mathbf{N}_s(t) = x_s] \\ &= \lim_{h\to 0} E\left[\frac{(N_s(t+h) - N_s(t))^2}{hs} \,\middle|\, N_s(t) = s + ((\rho-1)/\alpha)s + x_s\sqrt{s}\right] \\ &= \frac{\rho s + s + \alpha(((\rho-1)/\alpha)s + x_s\sqrt{s})}{s} \\ &\to 2\rho \equiv \sigma^2(x) \quad \text{as } s \to \infty. \end{aligned}$$

It is well known that the OU diffusion has a normal steady-state distribution with variance equal to the infinitesimal variance divided by twice the state-dependent drift rate (e.g., see p. 218 of Karlin and Taylor 1981). □

The stochastic-process limit in Theorem 2.1 is often called a *functional central limit theorem* (FCLT) (e.g., see Whitt 2002). A simple consequence of the FCLT is a *functional weak law of large numbers* (FWLLN), which formalizes the heuristic discussion in (2.4)–(2.6). It is obtained simply by dividing by $\sqrt{s}$ before letting $s \to \infty$ in the setting of Theorem 2.1. To state the FWLLN, let

$$\widehat{\mathbf{N}}_s(t) \equiv \frac{N_s(t)}{s}, \quad t \geq 0. \tag{2.12}$$

COROLLARY 2.1 (FWLLN FOR THE ERLANG-A MODEL IN THE ED REGIME). *Under the conditions of Theorem* 2.1,

$$\widehat{\mathbf{N}}_s \Rightarrow \widehat{\mathbf{N}} \quad \text{in } D \quad \text{as } s \to \infty, \tag{2.13}$$

where

$$\widehat{\mathbf{N}}(t) = (1+q), \quad t \geq 0 \tag{2.14}$$

for q in (2.3).

PROOF. When we divide the scaled process in (2.7) by $\sqrt{s}$ and let $s \to \infty$, we obtain convergence in probability to the zero function by an application of a version of the continuous mapping theorem—Theorem 3.4.4 in Whitt (2002)—implying the result. □

It is also possible to establish a more general deterministic fluid approximation by just changing the initial conditions in Corollary 2.1. When we scale by dividing by s throughout, we obtain an *ordinary differential equation* (ODE) for the limit, which is useful for describing the transient behavior of the Erlang-A model. For a real number x, let $(x)^+ \equiv \max\{x, 0\}$ and $(x)^- \equiv \min\{x, 0\}$.

THEOREM 2.2 (ED FLUID LIMIT FOR THE ERLANG-A MODEL). *Consider the sequence of $M/M/s/r+M$ models specified above, satisfying* (2.1)–(2.3), *and let $\hat{\mathbf{N}}_s(t)$ be the scaled number in system in* (2.12). *If $\hat{\mathbf{N}}_s(0) \Rightarrow \mathbf{n}(0)$ as $s \to \infty$, where $\mathbf{n}(0)$ is a real number* (*deterministic*), *then*

$$\hat{\mathbf{N}}_s \Rightarrow \mathbf{n} \quad \text{in } D \quad \text{as } s \to \infty, \tag{2.15}$$

where $\mathbf{n}$ is a degenerate diffusion process with continuous piecewise-linear infinitesimal mean (*state-dependent drift*)

$$m(x) = (\rho - 1) - \alpha(x-1)^+ + (x-1)^- \tag{2.16}$$

and infinitesimal variance $\sigma^2(x) = 0$; i.e., $\mathbf{n}$ is the ODE

$$\begin{aligned} \dot{\mathbf{n}}(t) &\equiv \frac{d\mathbf{n}}{dt}(t) \\ &= (\rho - 1) - \alpha(\mathbf{n}(t) - 1)^+ + (\mathbf{n}(t) - 1)^- \end{aligned} \tag{2.17}$$

with initial value $\mathbf{n}(0)$.

PROOF. We first extend the process N_s to the entire real line by letting the birth rate be ρs and the death rate be 0 for negative integers, and by letting the

birth rate be 0 and the death rate be $s+\alpha\eta s$ for positive integers greater than ηs. With that construction the scaled process $\hat{\mathbf{N}}_s$ will never visit states outside the interval $[0,\eta]$, but at the same time will have no boundaries. The proof now is essentially the same as for Theorem 2.1. Now we need to calculate the infinitesimal means and variances when we scale by dividing by s instead of $\sqrt{s}$. Now we let x_s be a possible value of $\hat{\mathbf{N}}_s(t)$ for each s, such that $x_s \to x$ as $s\to\infty$. Now the limit for the infinitesimal means is

$$\begin{aligned} m_s(x_s) &\equiv \lim_{h\to 0} E\big[(\hat{\mathbf{N}}_s(t+h)-\hat{\mathbf{N}}_s(t))/h \mid \hat{\mathbf{N}}_s(t)=x_s\big] \\ &= \lim_{h\to 0} E\left[\frac{(N_s(t+h)-N_s(t))}{hs}\,\middle|\, N_s(t)=sx_s\right] \\ &= \begin{cases} \rho-1-\alpha(x_s-1), & x_s\ge 1, \\ \rho-x_s & x_s\le 1, \end{cases} \\ &= \begin{cases} \rho-1-\alpha(x_s-1), & x_s\ge 1, \\ \rho-1+(1-x_s), & x_s\le 1, \end{cases} \\ &\to \rho-1-\alpha(x-1)^+ +(x-1)^-. \end{aligned}$$

The limit for the infinitesimal variances is

$$\begin{aligned} \sigma_s^2(x_s) &\equiv \lim_{h\to 0} E\big[(\hat{\mathbf{N}}_s(t+h)-\hat{\mathbf{N}}_s(t))^2/h \mid \hat{\mathbf{N}}_s(t)=x_s\big] \\ &= \lim_{h\to 0} E\left[\frac{(N_s(t+h)-N_s(t))^2}{hs^2}\,\middle|\, N_s(t)=sx_s\right] \\ &= \begin{cases} \dfrac{\rho+1+\alpha(x_s-1)}{s}, & x_s\ge 1, \\ \dfrac{\rho+x_s}{s} & x_s\le 1, \end{cases} \\ &\to 0\equiv\sigma^2(x) \quad \text{as} \quad s\to\infty. \end{aligned}$$

It is well known that the degenerate OU diffusion (with 0 infinitesimal variance) is the ODE in (2.17). □

Remark 2.1 (Steady-State of the Fluid Limit). It is easy to see that the deterministic fluid limit function in Theorem 2.2, $\mathbf{n}(t)$, converges monotonically as $t\to\infty$ to its steady-state limit $\mathbf{n}(\infty)=q\equiv(\rho-1)/\alpha$.

For customary applications, we are primarily interested in approximations for the steady-state performance measures in the $M/M/s/r+M$ model. Such approximations can be generated heuristically from Theorem 2.1, but limits for the steady-state performance measures do not follow directly from Theorem 2.1. They do with additional arguments, however. They can also be established directly, starting from the steady-state distributions in the $M/M/s/r+M$ model. Here we apply Theorem 2.1.

Here are the performance measures we consider: $N_s(\infty)$, the steady-state number of customers in the system; $Q_s(\infty)$, the steady-state number of customers waiting in queue; $W_s(\infty)$, the steady-state waiting time (before beginning service or abandoning) of a typical customer (which has the same distribution as the virtual waiting time of an arrival at an arbitrary time, because of the Poisson arrival process); and $P_s(ab)$, the steady-state abandonment probability. Let S_s denote the event that a customer eventually is served; necessarily $P(S_s=0)=1-P(S_s=1)=P_s(ab)$. Let $(W_s(\infty)\mid S_s)$ denote a random variable with the conditional distribution of the waiting time given that the customer eventually will be served, i.e., $P((W_s(\infty)\mid S_s)\le x)\equiv P(W_s(\infty)\le x\mid S_s)$. For the Erlang-A model, it is well known that these steady-state quantities are well defined.

Theorem 2.3 (ED Heavy-Traffic Limit for Steady-State Quantities in the Erlang-A Model). *Consider the sequence of $M/M/s/r+M$ models specified above, satisfying* (2.1)–(2.3). *Then, as $s\to\infty$,*

$$\mathbf{N}_s(\infty)\equiv\frac{N_s(\infty)-s(1+q)}{\sqrt{s}}\Rightarrow\mathbf{N}(\infty)\stackrel{\mathrm{d}}{=}\mathrm{Nor}\left(0,\frac{\rho}{\alpha}\right), \tag{2.18}$$

$$\widehat{\mathbf{N}}_s(\infty)\equiv\frac{N_s(\infty)}{s}\Rightarrow\widehat{\mathbf{N}}(\infty)=1+q, \tag{2.19}$$

$$P(N_s(\infty)\le s)\to 0, \tag{2.20}$$

$$\mathbf{Q}_s(\infty)\equiv\frac{[N_s(\infty)-s]^+-sq}{\sqrt{s}}\Rightarrow\mathbf{Q}(\infty)\stackrel{\mathrm{d}}{=}\mathrm{Nor}\left(0,\frac{\rho}{\alpha}\right), \tag{2.21}$$

$$\widehat{\mathbf{Q}}_s(\infty)\equiv\frac{Q_s(\infty)}{s}\Rightarrow\widehat{\mathbf{Q}}(\infty)=q\equiv\frac{\rho-1}{\alpha}, \tag{2.22}$$

$$P_s(ab)\Rightarrow P(ab)\equiv\frac{\rho-1}{\rho}, \tag{2.23}$$

$$(W_s(\infty)\mid S_s)\Rightarrow w, \tag{2.24}$$

$$W_s(\infty)\Rightarrow W, \tag{2.25}$$

where w is the deterministic quantity

$$w\equiv\frac{1}{\alpha}\log_e(\rho)=-\frac{1}{\alpha}\log_e(1-P(ab))>0, \tag{2.26}$$

and W is the random variable with

$$P(W>x)=e^{-\alpha x},\quad 0\le x\le w,\quad \textit{and}\quad P(W>w)=0 \tag{2.27}$$

for w in (2.26), *which has expected value*

$$E[W]=\frac{P(ab)}{\alpha}=\frac{q}{\rho}. \tag{2.28}$$

Proof. For the first limit in (2.18), most of the work has been done by Theorem 2.1. To make use of Theorem 2.1, we can follow the argument in the proof of Theorem 4 in Halfin and Whitt (1981). We can deduce that the sequence of normalized steady-state random variables $\{\mathbf{N}_s(\infty):\ s\ge 1\}$ is tight by constructing upper and lower bounding processes

that have proper limits as $s \to \infty$ (see Halfin and Whitt 1981 for details). The tightness implies relative compactness by Prohorov's Theorem, Theorem 11.6.1 of Whitt (2002), thus every subsequence has a convergent sub-subsequence. We show convergence by showing that all convergent subsequences must have the same limit. Consider any convergent subsequence. That convergent subsequence can serve as the sequence of initial distributions in the conditions of Theorem 2.1. But because these particular initial distributions are stationary distributions, the limiting distribution must be a stationary distribution for the limiting OU diffusion process. The OU diffusion process has a unique stationary distribution, however. Thus, all convergent subsequences must have that normal stationary distribution as their limiting distribution. With that additional argument, Theorem 2.1 implies (2.18). The next limit (2.19) follows by dividing by $\sqrt{s}$ and letting $s \to \infty$, just as in Corollary 2.1. Then (2.20) is an immediate consequence. The limits for the scaled queue-length processes in (2.21) and (2.22) follow from (2.18) by continuous mapping theorems. To establish (2.23), note that in steady state, the servers are all busy asymptotically, by (2.20). Hence, after dividing by s, the service rate is asymptotically 1. Because the arrival rate is asymptotically ρ after dividing by s, the total abandonment rate after dividing by s necessarily is asymptotically $\rho - 1$ and $P_s(ab) \to P(ab) \equiv (\rho - 1)/\rho$.

We now turn to the waiting-time results. The waiting time for a customer that eventually will be served $(W_s(\infty) \mid S_s)$ is the first passage time to the zero state, starting with $Q_s(\infty)$ customers, if we turn off the arrival process directly after that arrival. With the arrival process turned off, $s^{-1}Q_s(t) \Rightarrow q(t)$ by the law of large numbers, where the limit $q(t)$ satisfies the ODE

$$\dot{q}(t) \equiv \frac{dq}{dt} = -1 - \alpha q(t) \tag{2.29}$$

with initial condition $q(0) = q$. Asymptotically as $s \to \infty$, at time t, the scaled queue-length process is being depleted by service completions at rate 1 and by abandonments at rate $\alpha q(t)$. Arguing more carefully, for any $\epsilon > 0$, the scaled number of departures in the interval $(t, t + \epsilon)$, given $Q_s(t)$, where $s^{-1}Q_s(t) \Rightarrow q(t)$, is asymptotically $(1 + \alpha q(t))\epsilon + o(\epsilon)$ as $s \to \infty$. Hence we indeed have (2.29). Next, it is easy to see that the unique solution to the ODE in (2.29) is

$$q(t) = \left(q + \frac{1}{\alpha}\right)e^{-\alpha t} - \frac{1}{\alpha}, \quad t \geq 0. \tag{2.30}$$

Thus, the waiting time of a customer that will eventually be served approaches the value w such that $q(w) = 0$. Solving $q(w) = 0$, we get

$$w = -\frac{1}{\alpha}\log_e\left(\frac{1}{1 + \alpha q}\right) = \frac{1}{\alpha}\log_e(\rho) \tag{2.31}$$

as given in (2.26). Given that served customers wait exactly w in the limit as $s \to \infty$, we immediately obtain (2.25) and (2.27). □

We observe that (2.28) is consistent with two exact relations for the $M/M/s + M$ model. First, by *Little's Law* or $L = \lambda W$, we have

$$E[Q_s(\infty)] = \lambda_s E[W_s(\infty)], \tag{2.32}$$

even without the $M/M/s + M$ assumptions, e.g., see Whitt (1991). Second, for the $M/M/s + M$ model, we can express the total abandonment rate in two ways, yielding

$$\lambda_s P_s(ab) = \alpha E[Q_s(\infty)] \tag{2.33}$$

or

$$P_s(ab) = \frac{\alpha}{\lambda_s}E[Q_s(\infty)] = \frac{\alpha}{s\rho}E[Q_s(\infty)]. \tag{2.34}$$

Combining (2.32) and (2.33), we obtain

$$P_s(ab) = \alpha E[W_s(\infty)]. \tag{2.35}$$

Thus, when we know any one of the three performance measures $P_s(ab)$, $E[Q_s(\infty)]$, or $E[W_s(\infty)]$, we know all three. Formula (2.28) shows that the relations remain valid in the limit.

3. ED Approximations

The main ED approximations for the $M/M/s/r + M$ model are the three simple approximations that follow directly from (2.22)–(2.26):

$$\begin{aligned} P_s(ab) &\approx P(ab) \equiv \frac{(\rho - 1)}{\rho}, \\ E[Q_s(\infty)] &\approx qs \equiv \frac{(\rho - 1)s}{\alpha}, \\ E[W_s(\infty) \mid S_s] &\approx w \equiv \frac{1}{\alpha}\log_e(\rho). \end{aligned} \tag{3.1}$$

It is important to recognize, however, that only the first simple approximation for the abandonment probability $P_s(ab)$ is generally valid beyond the Markovian $M/M/s/r + M$ model, as can be seen from Whitt (2005d). In particular, the approximations for the mean steady-state queue length and waiting time depend critically on the exponential distribution assumptions. These $M/M/s/r + M$ ED approximations can be very useful more generally, however, as rough approximations and to test whether an $M/M/s/r + M$ model is appropriate.

We choose to focus on the conditional waiting time given that the arrival is eventually served, $(W_s(\infty) \mid S_s)$, but we also have results for the unconditional waiting time $W_s(\infty)$, from which we can obtain results for the conditional waiting time given that the arrival eventually abandons, $(W_s(\infty) \mid S_s^c)$, where S_s^c is the

complement of the event S_s. Theorem 2.3 implies that $(W_s(\infty) \mid S_s^c) \le (W_s(\infty) \mid S_s)$ in the limit as $s \to \infty$, so constraints on the distribution of $(W_s(\infty) \mid S_s)$ will tend to be somewhat more stringent than constraints on the distribution of $W_s(\infty)$, but these will differ relatively little when the abandonment probability is not great.

Combining the first approximation in (3.1) with the exact relation in (2.33), we obtain the approximation

$$E[W_s] = \frac{P_s(ab)}{\alpha} \approx \frac{P(ab)}{\alpha} = \frac{(\rho-1)}{\rho\alpha}. \tag{3.2}$$

We also obtain essentially the same approximation for $E[W_s(\infty) \mid S_s]$, based on an approximation for w,

$$\begin{aligned} E[W_s(\infty) \mid S_s] &\approx w = -\frac{1}{\alpha}\log_e(1-P(ab)) \\ &\approx \frac{P(ab)}{\alpha} = \frac{(\rho-1)}{\rho\alpha}. \end{aligned} \tag{3.3}$$

For approximation (3.3), we use the approximation $\log(1-x) \approx -x$ for small x, which is asymptotically correct (the ratio approaches 1) as $x \downarrow 0$.

From (2.21), we also obtain a normal approximation for the entire steady-state queue-length distribution, in particular,

$$Q_s(\infty) \approx \mathrm{Nor}\left(\frac{(\rho-1)s}{\alpha}, \frac{\rho s}{\alpha}\right). \tag{3.4}$$

A consequence is an approximation for the variance

$$\mathrm{Var}(Q_s(\infty)) \approx \frac{\rho s}{\alpha}. \tag{3.5}$$

An immediate practical insight to glean from the approximations for the steady-state queue length in (3.1), (3.4), and (3.5) is that the coefficient of variation (standard deviation divided by the mean) approaches 0 as $s \to \infty$ for all parameters $\rho > 1$ and $\alpha > 0$. Thus the crude deterministic analysis begins to tell more and more of the story as s increases. For example, the $4 \times 5 = 20$ cases in Table 1 in §5 have queue-length coefficients of variation ranging from 0.33 to 2.25. (And these values would be smaller if we looked at the conditional queue length given that it is positive.)

We next discuss refined heuristic approximations that do not follow directly from Theorem 2.3. An obvious simple refinement to (3.4) is

$$Q_s(\infty) \approx \mathrm{Nor}\left(\frac{(\rho-1)s}{\alpha}, \frac{\rho s}{\alpha}\right)^+, \tag{3.6}$$

where $x^+ \equiv \max\{0, x\}$. Let Φ and ϕ be the cumulative distribution function (cdf) and probability density function (pdf) of a standard normal random variable, respectively, i.e.,

$$\Phi(y) \equiv P(\mathrm{Nor}(0,1) \le x) \equiv \int_{-\infty}^{y} \phi(x)\,dx,$$
$$\text{where } \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{3.7}$$

Let Φ^c be the associated complementary cdf (ccdf), defined by $\Phi^c(x) \equiv 1 - \Phi(x)$, and let h be the associated hazard function

$$h(x) \equiv \frac{\phi(x)}{\Phi^c(x)}. \tag{3.8}$$

To simplify expressions, we use the following notation:

$$q \equiv \frac{\rho-1}{\alpha}, \qquad v \equiv \frac{\rho}{\alpha}, \quad \text{and}$$
$$\gamma \equiv \gamma(s,q,v) \equiv \frac{-qs}{\sqrt{vs}}. \tag{3.9}$$

We then obtain the associated approximations

$$P(Q_s(\infty) > 0) \approx P(\mathrm{Nor}(qs, vs) > 0) = \Phi^c(\gamma), \tag{3.10}$$

$$\begin{aligned} &E[Q_s(\infty) \mid Q_s(\infty) > 0] \\ &\quad \approx E[\mathrm{Nor}(qs, vs) \mid \mathrm{Nor}(qs, vs) > 0] \\ &\quad = qs + \sqrt{vs}\,h(\gamma), \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} &E[Q_s(\infty)^2 \mid Q_s(\infty) > 0] \\ &\quad \approx E[\mathrm{Nor}(qs, vs)^2 \mid \mathrm{Nor}(qs, vs) > 0] \\ &\quad = (qs)^2 + vs + 2q\sqrt{vs}\,h(\gamma) + vs\gamma h(\gamma) \end{aligned} \tag{3.12}$$

for q, v, and γ in (3.9). The conditional normal moments in (3.11) and (3.12) are standard (e.g., see Proposition 18.3 of Browne and Whitt 1995). Clearly, we can combine the last three formulas to obtain approximations for the mean $E[Q_s(\infty)]$ and variance $\mathrm{Var}[Q_s(\infty)]$.

Because the three performance measures $P_s(ab)$, $E[Q_s(\infty)]$, and $E[W_s(\infty)]$ are all related by the exact relations in (2.32)–(2.34), we can use a refined approximation for any one to obtain refined approximations for all three. Thus, we obtain the refined approximation for the abandonment probability

$$P_s(ab) \approx P(ab)\frac{E[Q_s(\infty)]}{(\rho-1)s/\alpha}. \tag{3.13}$$

Of course, approximations for $E[Q_s(\infty)]$ translate immediately into approximations for $E[W_s(\infty)]$ by virtue of Little's Law, (2.32).

We now consider refined approximations for $P(W_s(\infty) \le x \mid S_s)$, the cdf of the conditional steady-state waiting time until beginning service, given that

the customer is served. Because the waiting time tends to be the sum of a relatively large number of service times, we apply an approximation based on the law of large numbers together with the fluid approximation in (2.26), saying that

$$P(W_s(\infty) \le x \mid S_s, Q_s(\infty) = qs) \approx 1 \quad \text{if } -\frac{1}{\alpha}\log_e\left(\frac{1}{1+\alpha q}\right) \le x \qquad (3.14)$$

and 0 otherwise. However,

$$-\frac{1}{\alpha}\log_e\left(\frac{1}{1+\alpha q}\right) \le x \quad \text{if and only if } q \le \frac{1}{\alpha}(e^{\alpha x}-1). \qquad (3.15)$$

Thus, we obtain the approximation

$$\begin{aligned} P(W_s(\infty) \le x \mid S_s) &\approx P\left(Q_s(\infty) \le \frac{s}{\alpha}(e^{\alpha x}-1)\right) \\ &\approx P\left(\mathrm{Nor}(sq, sv) \le \frac{s}{\alpha}(e^{\alpha x}-1)\right) \\ &= \Phi\left(\frac{[s/\alpha(e^{\alpha x}-1) - sq]}{\sqrt{sv}}\right) \end{aligned} \qquad (3.16)$$

for v in (3.9).

4. Slow Abandonments in the ED Regime: The Limit as $s/\alpha \to \infty$

From Equations (3.4) and (3.6), we see that the normal approximations for the steady-state queue length $Q_s(\infty)$ depend on the parameters s and α only through the ratio s/α. Thus, it is natural to consider limits in which we let $\alpha \to 0$ instead of $s \to \infty$ and, indeed, that has already been done by Ward and Glynn (2003). However, they did not emphasize the ED regime, where ρ is fixed with $\rho > 1$. Instead, they emphasize the limiting regime in (1.2).

We go further for the ED limiting regime here by establishing limits as $s/\alpha \to \infty$, allowing either $s \to \infty$ or $\alpha \to 0$ or both. For the special case in which only $\alpha \to 0$, we recover Part 4 of Theorem 1 in Ward and Glynn (2003); for the special case in which only $s \to \infty$, we recover Theorem 2.1.

We start by defining scaled processes, indexed by both s and α,

$$\mathbf{Y}_{s,\alpha}(t) \equiv \sqrt{\frac{\alpha}{s}}\left[N_{s,\alpha}(t/\alpha) - s - \frac{(\rho-1)s}{\alpha}\right] \quad \text{for } t \ge 0. \qquad (4.1)$$

We now state the result, omitting the proof because it is just like the proof of Theorem 2.1.

Theorem 4.1 (The ED Limit as $s/\alpha \to \infty$). *Consider the $M/M/s/r+M$ models defined in §2, satisfying* (2.1) *and* (2.3). *If $\mathbf{Y}_{s,\alpha}(0) \Rightarrow \mathbf{Y}(0)$ in $\mathbb{R}$ as $s/\alpha \to \infty$, where $\mathbf{Y}_{s,\alpha}$ is the scaled process in* (4.1), *then*

$$\mathbf{Y}_{s,\alpha} \Rightarrow \mathbf{Y} \quad \text{in } D \quad \text{as } s/\alpha \to \infty, \qquad (4.2)$$

where $\mathbf{Y}$ is an OU diffusion process with infinitesimal mean $m(x) = -x$ and infinitesimal variance $\sigma^2(x) = 2\rho$.

From Theorem 4.1, we obtain the same approximation for the steady-state queue length $Q_s(\infty)$ as before, i.e., as in (3.4) and (3.6).

The fact that the ratio s/α plays such a critical role invites an explanation. We would like to identify the fundamental role it plays, in the spirit of the traffic intensity, $\rho = \lambda/s\mu$. In fact, the role of the ratio s/α is brought out by the limits. Indeed, we see the role of s/α from the first heuristic approximation developed in (2.4)–(2.6), based on finding the point where the input rate matches the output rate. That informal reasoning tends to make sense if either s is large or α is small or both. We also see that, in first order, the queue length should be proportional to s/α.

The important role of the ratio s/α is further brought out by the stochastic-process limits in Theorems 2.1 and 4.1. Even from the basic limit in Theorem 2.1, we can see the role of s and α. First, the role of s is clear from the scaling in (2.7). We see that the mean and variance of $N_s(t)$ should both be approximately proportional to s. Next, the role of α is first seen by its contribution to q in (2.3) and the way it affects the centering term in (2.7). We then see that the state-dependent drift of the limiting OU diffusion process is proportional to α, which implies that the variance of the steady-state distribution is inversely proportional to α. Combining those insights, we see that both the mean and variance of the steady-state queue length should be approximately proportional to s/α. The proof of Theorem 4.1 shows that it suffices to let $s/\alpha \to \infty$.

We can also provide a more informal direct argument. If we consider a waiting customer, that customer tends to abandon at rate α and tends to move toward service at rate $s\mu = s$. Thus s/α is the ratio of two rates: the rate a customer moves toward entering service, and the rate the customer tends to abandon. When the ratio s/α is large, the tendency to abandon is less, so that the queue size is likely to be larger. In other words, the key ratio s/α can be thought of as the tendency to be served instead of abandon. For given proportion of abandonment, estimated by $\rho - 1$, a higher ratio s/α causes larger queues, in particular, a larger value of sq, pushing us more into the ED regime.

5. Numerical Comparisons

In this section, we evaluate the ED approximations in §3, based on the limits in §§2 and 4, by compar-

Table 1 A Comparison of Approximations with Exact Numerical Values for Several Performance Measures in the $M/M/s+M$ Model

	Performance measures as a function of s with s/α fixed						
	Approximations		Exact with s servers				
Perf. meas.	Simple	Refined	1	10	100	1,000	10,000
$s/\alpha = 1{,}000$ and $\rho - 1 = 0.10$							
$P_s(ab)$	0.0909	0.0909	0.0910	0.0910	0.0909	0.0909	0.0909
$E[Q_s(\infty)]$	100.0	100.0	100.1	100.1	100.0	100.0	100.0
$SD(Q_s(\infty))$	33.17	33.1	33.0	33.0	33.1	33.1	33.1
$sE[W_s(\infty) \mid S_s]$	95.31	—	94.92	94.90	94.85	94.82	94.81
$s/\alpha = 1{,}000$ and $\rho - 1 = 0.02$							
$P_s(ab)$	0.0196	0.0247	0.0329	0.0318	0.0291	0.0246	0.0210
$E[Q_s(\infty)]$	20.0	25.2	33.6	32.4	29.7	25.1	21.4
$SD(Q_s(\infty))$	31.9	24.9	23.5	23.9	24.5	25.0	24.7
$sE[W_s(\infty) \mid S_s]$	19.79	—	33.2	32.0	29.2	24.6	20.9
$s/\alpha = 100$ and $\rho - 1 = 0.10$							
$P_s(ab)$	0.0909	0.0995	0.1148	0.1087	0.0992	0.0927	0.0911
$E[Q_s(\infty)]$	10.0	10.95	12.6	12.0	10.9	10.2	10.0
$SD(Q_s(\infty))$	10.5	8.99	8.6	8.8	9.1	9.2	9.2
$sE[W_s(\infty) \mid S_s]$	9.53	—	11.9	11.1	10.1	9.4	9.2
$s/\alpha = 100$ and $\rho - 1 = 0.02$							
$P_s(ab)$	0.0196	0.0499	0.0792	0.0686	0.0499	0.0316	0.0223
$E[Q_s(\infty)]$	2.00	5.10	8.08	7.00	5.09	3.23	2.28
$SD(Q_s(\infty))$	10.1	6.57	6.83	6.93	6.68	5.85	5.14
$sE[W_s(\infty) \mid S_s]$	1.98	—	8.03	6.88	4.90	3.00	2.13

ing them to exact numerical results for the Erlang-A model, using the algorithm described in Whitt (2005c). We vary s with both the ratio s/α and the limiting abandonment rate $(\lambda_s - s)/s = \rho - 1$ held fixed. We consider two values for the ratio s/α, 1,000 and 100; we consider two values for the scaled total abandonment rate $\rho - 1$, 0.10 and 0.02. The two cases with $s/\alpha = 100$ are the two cases discussed in the introduction.

We consider four different performance measures: the probability of abandonment, $P_s(ab)$; the mean steady-state queue length, $E[Q_s(\infty)]$; the standard deviation of the steady-state queue length, $SD(Q_s(\infty))$; and the conditional mean steady-state waiting time given that the customer eventually will be served, $E[W_s(\infty) \mid S_s]$. It should be noted that the first two, $P_s(ab)$ and $E[Q_s(\infty)]$, are connected by the exact relation (2.34) so that they have the same relative error. It is interesting to see the actual values, however.

In Table 1 we display two different approximations: first, the simple approximation, and, second, a refined approximation. The simple approximations are given in (3.1) and (3.5). The refined approximation for the mean queue length is obtained by combining (3.10) and (3.11). The refined approximation for the standard deviation of the queue length is obtained by combining (3.10)–(3.12). The refined approximation for the probability of abandonment, $P_s(ab)$, is obtained from (3.13), using the refined approximation for the mean queue length.

From Table 1 we see that, as expected, the quality of the results improve as the ratio s/α increases and as the scaled total abandonment rate $\rho - 1$ increases. Consistent with intuition, that evidently is a general property. We have found that a good way to estimate, in advance, the overall quality of the ED approximations is to look at the *product* of these quantities s/α and $\rho - 1$, which is just the approximate steady-state queue length, i.e.,

$$E[Q_s(\infty)] \approx sq = \frac{(\rho - 1)s}{\alpha} = (\rho - 1) \times \frac{s}{\alpha}. \tag{5.1}$$

In the four cases here, sq is 100, 20, 10, and 2. Accordingly, the approximations are accurate when $s/\alpha = 1{,}000$, $\rho - 1 = 0.10$, and $sq = 100$, but the approximations are crude when $s/\alpha = 100$, $\rho - 1 = 0.02$, and $sq = 2$. Importantly, the approximations appear useful in the middle two cases in which $sq = 20$ and $sq = 10$.

Very roughly, the percentage errors in the approximations can be estimated by the *reciprocal* of sq, i.e., the percentage errors should be of order $1/sq$. For example, the approximation errors should be about 10% when $sq = 10$.

Except possibly in the last case with $s/\alpha = 100$ and $\rho - 1 = 0.02$, Table 1 shows that the four performance measures do not vary greatly with s, over a very wide range, when s/α and $\rho - 1$ are held fixed. The mean steady-state queue length is approximately proportional to s/α, but independent of s for fixed s/α. On the other hand, by Little's Law, the mean steady-state

waiting time is approximately inversely proportional to s for fixed s/α.

Except in the best case with $s/\alpha = 1{,}000$ and $\rho - 1 = 0.10$, the refinements are much better than the simple approximations. Nevertheless, we feel the simple approximations are especially useful for making quick rough estimates. The difference between the refined approximation and the simple approximation gives a good idea of the accuracy of the simple approximation.

The weakest approximation in Table 1 is clearly the simple approximation for the mean conditional waiting time, $E[W_s(\infty) \mid S_s]$. The results suggest that it would be better to just focus on the unconditional expected waiting time, $E[W_s(\infty)]$, and use Little's Law with the approximations for the mean queue length. Then we obtain the same accuracy as the mean queue length. In summary, we regard the numerical results in Table 1 as strong evidence that the approximations can be very useful.

To evaluate call center performance, there is great interest in *service level*. It is standard to require that y% of all calls be answered within x seconds for values such as $y = 80\%$ and $x = 30$ seconds. Thus, we are especially interested in having an approximation for the conditional cdf $P(W_s(\infty) \le x \mid S_s)$ for appropriate values of x. In our numerical examples, the mean waiting times vary from one example to the next, so natural values of x change from one example to the next. Thus, we will standardize by expressing our service-level cut-off as a constant θ multiple of the mean waiting time. Because the mean waiting time can be roughly approximated by (3.2), it is natural to look at arguments of the form $x = \theta P(ab)/\alpha$ for various values of θ, centered around 1. We thus examine how the approximation for $P(W_s(\infty) \le x \mid S_s)$ in (3.16) performs in Table 2 for arguments $x = \theta P(ab)/\alpha$ for various values of θ.

Here, we vary both s and θ, keeping the quantities s/α and $\rho - 1$ fixed. We consider the same four cases of s/α and $\rho - 1$ as in Table 1. We make comparisons with exact results for all powers of 10 ranging from $s = 1$ to $s = 100{,}000$.

The approximation for the waiting-time cdf in Table 2 is not as accurate as the approximations in Table 1. The approximations are pretty good for higher values of θ, e.g., for $\theta \ge 1$, but not for very small values, especially when the argument is very small, e.g., when $\theta = 0$. It is significant that the conditional waiting-time cdf approximation does work well for

Table 2 A Comparison of the Normal Approximation for the Conditional Waiting-Time cdf Given That a Customer Is Served, $P(W_s(\infty) \le x \mid S_s)$, with Exact Numerical Values in the $M/M/s + M$ Model

	$P(W_s(\infty) \le \theta P(ab)/\alpha \mid S_s)$ as a function of s and θ						
	Number of servers s						
Case θ	1	10	100	1,000	Approx.	10,000	100,000
$s/\alpha = 1{,}000$ and $\rho - 1 = 0.10$							
0.00	0.00011	0.00033	0.00082	0.00105	0.0020	0.00115	0.00116
0.75	0.199	0.200	0.200	0.200	0.199	0.200	0.200
1.00	0.452	0.453	0.453	0.453	0.445	0.453	0.453
1.25	0.725	0.725	0.725	0.725	0.725	0.725	0.725
1.50	0.905	0.905	0.905	0.905	0.907	0.905	0.905
$s/\alpha = 1{,}000$ and $\rho - 1 = 0.02$							
0.0	0.014	0.046	0.128	0.261	0.265	0.370	0.406
0.5	0.166	0.194	0.265	0.380	0.376	0.474	0.504
1.0	0.331	0.354	0.411	0.503	0.497	0.578	0.603
2.0	0.643	0.643	0.686	0.735	0.732	0.775	0.788
4.0	0.957	0.959	0.962	0.968	0.972	0.973	0.974
$s/\alpha = 100$ and $\rho - 1 = 0.10$							
0.0	0.026	0.078	0.158	0.182	0.212	0.227	0.228
0.5	0.205	0.253	0.325	0.313	0.373	0.385	0.387
1.0	0.417	0.451	0.504	0.482	0.540	0.549	0.550
2.0	0.788	0.800	0.820	0.817	0.833	0.836	0.836
4.0	0.995	0.995	0.996	0.998	0.996	0.996	0.996
$s/\alpha = 100$ and $\rho - 1 = 0.02$							
0.0	0.061	0.186	0.422	0.422	0.625	0.735	0.763
0.5	0.132	0.257	0.470	0.460	0.671	0.770	0.794
1.0	0.199	0.313	0.511	0.499	0.696	0.787	0.810
2.0	0.329	0.425	0.590	0.578	0.745	0.822	0.841
4.0	0.563	0.626	0.733	0.727	0.834	0.884	0.897

Note. The arguments x considered are $x = \theta P(ab)/\alpha$ as a function of θ and s for fixed ratio s/α and fixed $\rho - 1$.

Table 3 **The Probability of Initially Being Delayed, $1 - P(W = 0; S_s)$, in Each of the Cases Considered Previously**

Case					
$\rho - 1$	s/α	sq	s	$P(S_s)$	$1 - P(W = 0; S_s)$
0.10	1,000	100	1	0.9090	0.99990
			10	0.9090	0.9997
			100	0.9091	0.9998
			1,000	0.9091	0.9990
			10,000	0.9091	0.9989
0.02	1,000	20	1	0.9671	0.9865
			10	0.9682	0.9554
			100	0.9709	0.8757
			1,000	0.9754	0.7454
			10,000	0.9790	0.6378
0.10	100	10	1	0.8852	0.9744
			10	0.8903	0.9306
			100	0.9008	0.8577
			1,000	0.9073	0.8077
			10,000	0.9089	0.7937
0.02	100	2	1	0.9208	0.9438
			10	0.9314	0.8268
			100	0.9501	0.5991
			1,000	0.9684	0.3947
			10,000	0.9777	0.2814

the higher arguments needed to determine staffing to meet SLAs.

When considering whether to use QED or ED approximations, it is instructive to look at the abandonment probability and the probability of initially being delayed. In the QED (ED) regime, the first should be relatively small (large), while the second should be relatively large (small). We expect the ED approximations to be reasonably accurate when the probability of initially being delayed is greater than or equal to 0.80. To illustrate, we plot the probability of not being served immediately, $1 - P(W = 0; S_s)$, for the cases considered previously in Table 3. For the two examples in the introduction, this probability of initially being delayed is 0.5991 and 0.8577, respectively. We would thus expect the ED approximation to be very crude in the first case, but relatively good in the second case, as we observed.

6. Conclusions

In this paper, we established ED many-server heavy-traffic limits for Markovian queues with customer abandonments, specifically for the $M/M/s/r + M$ model. Many-server limiting regimes involve a sequence of queueing systems indexed by the number of servers, s, in which both s and the arrival rate λ_s are allowed to increase without bound, while the exponential service-time and abandon-time distributions are held fixed. Within the context of the many-server heavy-traffic limiting regimes for queues with abandonments, the ED regime can be characterized in two equivalent ways: (1) assume in addition that the probability of abandonment, $P_s(ab)$, converges to a limit strictly between 0 and 1, or (2) assume in addition that the traffic intensity, ρ_s, approaches a limit ρ with $\rho > 1$.

We also developed direct and refined approximations based on the ED many-server heavy-traffic limits. We conclude that the ED many-server heavy-traffic approximations can be very useful to describe the performance of many-server queues with substantial abandonments. The first approximations for key performance measures in the $M/M/s/r + M$ model (Erlang-A model) in (3.1)–(3.5), obtained directly from the limits, are remarkably simple. The heuristic refinements in (3.6)–(3.16) are also not too complicated. Tables 1 and 2 show that the approximations are quite accurate when the probability of initially being delayed is not too small, e.g., when $1 - P(W = 0; S_s) \geq 0.80$ or the approximate mean queue length $sq \equiv (\rho - 1)s/\alpha$ is not too small, e.g., when $sq \geq 10$. In extreme cases, as when $1 - P(W = 0; S_s) \geq 0.99$ or when $sq \geq 100$, the ED approximations will be extremely accurate, and clearly better than the QED approximations. However, for less extreme cases when $\rho > 1$, we can expect the QED approximations to perform better than the ED approximations. For such cases, the great appeal of the ED approximations is not their accuracy but their simplicity. They permit back-of-the-envelope calculations. They can greatly help understand the performance of call centers providing low-to-moderate quality of service. To estimate the quality of the ED approximations in advance, we suggest looking at $sq = (\rho - 1)s/\alpha$.

Both the theory (Theorem 4.1) and the numerical comparisons (Table 1) show that key parameters, determining both (i) the performance of the $M/M/s/r + M$ model in the ED regime, and (ii) the accuracy of the ED approximations, are the ratio s/α and the scaled abandonment rate $(\lambda_s - s)/s = \rho - 1$. Indeed, in §4, we show that the ED many-server heavy-traffic limit holds when $s/\alpha \to \infty$, which can occur if either $\alpha \downarrow 0$ or $s \uparrow \infty$ or both.

Acknowledgments

The author is grateful to Avishai Mandelbaum of the Technion, Andrew Ross of Lehigh University, and anonymous referees for helpful suggestions. The author was supported by National Science Foundation Grant DMS-02-2340.

References

Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd ed. John Wiley and Sons, New York.

Brandt, A., M. Brandt. 1999. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Eval.* **35** 1–18.

Brandt, A., M. Brandt. 2002. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s+GI$ system. *Queueing Systems* **41** 73–94.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J. Dshalalow, ed. *Advances in Queueing*. CRC Press, Boca Raton, FL, 463–480.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 208–227.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Iglehart, D. L. 1965. Limit diffusion approximations for the many server queue and the repairman problem. *J. Appl. Probab.* **2** 429–441.

Jelenkovic, P., A. Mandelbaum, P. Momcilovic. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* **47** 53–69.

Karlin, S., H. M. Taylor. 1981. *A Second Course in Stochastic Processes.* Academic Press, New York.

Mandelbaum, A. 2001. Call centers. Research bibliography with abstracts. Industrial Engineering and Management, Technion, Haifa, Israel.

Mandelbaum, A., G. Pats. 1995. State-dependent queues: Approximations and applications. F. P. Kelly, R. J. Williams, eds. *Stochastic Networks*, Institute for Mathematics and Its Applications, Vol. 71. Springer-Verlag, New York, 239–282.

Mandelbaum, A., S. Zeltyn. 2004. The impact of customers' patience on delay and abandonment: Some empirically-driven experiments with the $M/M/n+G$ queue. *OR Spektrum* **26** 377–411.

Puhalskii, A. A., M. I. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probob.* **32** 564–595.

Stone, C. 1963. Limit theorems for random walks, birth and death processes and diffusion processes. *Illinois J. Math.* **4** 638–660.

Ward, A. R., P. W. Glynn. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* **43** 103–128.

Whitt, W. 1991. A Review of $L = W$ and Extensions. *Queueing Systems* **9** 235–268.

Whitt, W. 1992. Correction note on $L = W$. *Queueing Systems* **12** 431–432.

Whitt, W. 1999. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Oper. Res. Lett.* **24** 205–212.

Whitt, W. 2002. *Stochastic-Process Limits.* Springer-Verlag, New York.

Whitt, W. 2005a. A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* Forthcoming.

Whitt, W. 2005b. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* Forthcoming.

Whitt, W. 2005c. Engineering solution of a basic call-center model. *Management Sci.* Forthcoming.

Whitt, W. 2005d. Fluid models for multiserver queues with abandonments. *Oper. Res.* Forthcoming.

Whitt, W. 2005e. Two fluid approximations for multi-server queues with abandonments. *Oper. Res. Lett.* Forthcoming.

Zohar, E., A. Mandelbaum, N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.* **48** 566–583.