EFFICIENCY-DRIVEN HEAVY-TRAFFIC APPROXIMATIONS FOR MANY-SERVER QUEUES WITH ABANDONMENTS

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027–6699

Abstract

Motivated by the desire to understand the performance of service-oriented call centers, which often provide low-to-moderate quality of service, this paper investigates the efficiency-driven (ED) limiting regime for many-server queues with abandonments. The starting point is the realization that, in the presence of substantial customer abandonment, call-center service-level agreements (SLA's) can be met in the ED regime, where the arrival rate exceeds the maximum possible service rate. Mathematically, the ED regime is defined by letting the arrival rate and the number of servers increase together so that the probability of abandonment approaches a positive limit. 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). Even though the probability of delay necessarily approaches 1 in the ED regime, the ED regime can be realistic because, due to the abandonments, the delays need not be excessively large.

This paper establishes ED many-server heavy-traffic limits and develops associated approximations for performance measures in the $M/M/s/r+M$ model, having a Poisson arrival process, exponential service times, s servers, r extra waiting spaces and exponential abandon times (the final $+M$). In the ED regime, essentially the same limiting behavior occurs when the abandonment rate α approaches 0 as when the number of servers s approaches ∞; indeed, it suffices to assume that $s/\alpha \to \infty$. The ED approximations are shown to be useful by comparing them to exact numerical results for the $M/M/s/r+M$ model obtained using an algorithm developed in Whitt (2003), which exploits numerical transform inversion.

Keywords: call centers, contact centers, queues, multiserver queues, queues with customer abandonment, multiserver queues with customer abandonment, Erlang A model, queues with state-dependent rates, heavy-traffic limits, efficiency-driven limiting regime,

February 6, 2004

1. Introduction

Recently there has been great interest in multiserver queues with a large number of servers, largely motivated by applications to telephone call centers; see Mandelbaum (2001) and Gans, Koole and Mandelbaum (2002). For the basic $M/M/s/\infty$ model with Poisson arrival process, independent and identically distributed (IID) exponential service times, s servers and unlimited waiting space, 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

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

where $0 < \beta < \infty$; see Halfin and Whitt (1981), Puhalskii and Reiman (2000), Jelenkovic, Mandelbaum and Momcilovic (2002) and Whitt (2002b,c). 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, Mandelbaum and Reiman (2002) show that the same QED limiting regime is also useful for multiserver queues with customer abandonment, specifically for the purely Markovian $M/M/s/\infty + M$ model with exponential abandonment times (the final $+M$); also see Whitt (2002c) for a generalization to $G/M/s/\infty + M$. Then, since 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 delay 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. First, 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 (SLA's) are being met. Second, computational results from computer simulations and numerical algorithms substantiate that performance targets can often be met when the arrival rate exceeds the maximum possible service rate. Specifically, in Whitt (2003) we developed a numerical algorithm for calculating approxima-

tions 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.

A specific example has 100 agents (servers) handling calls with mean holding (service) time and mean abandon time both equal to 5 minutes. The SLA might 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.

Those experiences do not rule out the QED limiting regime, but they suggest that an alternative limiting regime might also be worth considering. Thus, in this paper we investigate the many-server heavy-traffic regime in which the steady-state probability of abandonments approaches a limit strictly between 0 and 1. As shown by Garnett, Mandelbaum and Reiman (2002), such a limit occurs for the $M/M/s+M$ model 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 $s \to \infty$. Following Garnett, Mandelbaum and Reiman (2002), we call this the *efficiency-driven* (ED) *limiting regime*.

Without a finite waiting room or customer abandonment, the ED regime does not even arise. Then it is simply an overloaded regime, because the queue length explodes as time evolves. The overloaded regime is interesting primarily to describe transient behavior. However, even a small amount of abandonment can keep the system stable when the arrival rate exceeds the maximum possible serve rate. Even though the steady-state probability of delay then approaches 1, this alternative ED heavy-traffic limit can be realistic because the delays experienced need not be excessively 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. In-

deed, 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, in order 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 this 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 they can be established by the same arguments used in the seminal heavy-traffic paper on multiserver queues by Iglehart (1965). The main contribution here, we believe, is communicating that this efficiency-driven many-server heavy-traffic limiting regime is indeed worth serious attention. When this ED regime is present, 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.

Here is how the rest of this paper is organized. We begin in Section 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 Section 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 $\alpha \downarrow 0$ and, indeed, such a limit was established by Ward and Glynn (2003) (assuming fixed s). In Section 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 $\alpha \downarrow 0$, so that

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

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

In Section 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 (2003). (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.)

Motivated by Whitt (2003), in Section 6 we establish a stochastic-process limit for the queue-length stochastic process in $M(n)/M(n)/s/r+M(n)$ models in the ED regime. The Markovian $M(n)/M/s/r+M(n)$ model has state-dependent arrival rates, service rates and abandonment rates (denoted by the $M(n)$). We are interested in the $M(n)/M(n)/s/r+M(n)$ model, not only for its direct application, but because the $M/M/s/r+M(n)$ special case (without state-dependent arrivals) was proposed as an approximation for the $M/GI/s/r+GI$ model in Whitt (2003). Here we develop a two-step approximation for the $M/GI/s/r+GI$ model: first, approximate $M/GI/s/r+GI$ by $M/M/s/r+M(n)$ according to the procedure in Whitt (2003) and, second, use the diffusion approximation for $M/M/s/r+M(n)$ established in Section 6.

The paper ends in Section 7 with our conclusions. We conclude this introduction by mentioning a companion paper, Whitt (2004), in which 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 upon performance of a non-exponential service-time distribution and a non-exponential abandon-time distribution.

For additional discussion about customer abandonment in queues, see Brandt and Brandt (1999, 2002), Zohar, Mandelbaum and Shimkin (2002), Ward and Glynn (2003), Mandelbaum and Zeltyn (2003) and 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 ($r \leq \infty$). When $r < \infty$, we make r be sufficiently large that it is not a factor. 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.

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 assume that

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

and

$$r_s = rs, \quad \text{where} \quad r > \frac{\lambda - 1}{\alpha} \ . \tag{2.2}$$

Condition (2.1) determines the ED regime. Condition (2.2) ensures that, asymptotically, no customers are blocked (verified 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

$$N_s(t) \approx (1+q)s \ , \tag{2.3}$$

where

$$q \equiv \frac{\lambda - 1}{\alpha} \ . \tag{2.4}$$

Heuristically, we obtain (2.3) by finding the point where the input rate equals the output rate. Clearly that can occur only with all servers busy. 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

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

The solution to these equations is $x = q$ for q in (2.4).

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

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

for q in (2.4). Let the initial state $N_s(0)$ be specified independently, so that the stochastic process $\{N_s(t) : t \geq 0\}$ 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 $N(m,\sigma^2)$ 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) – (2.2). *If* $\mathbf{N}_s(0) \Rightarrow \mathbf{N}(0)$ *as* $s \to \infty$, *then*

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

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

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

which has steady-state distribution

$$\mathbf{N}(\infty) \stackrel{\mathrm{d}}{=} N(0,\lambda/\alpha) \ . \tag{2.11}$$

Proof. Since 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 means and variances converge to the infinitesimal means and variance of the limit process. The infinitesimal means are

$$\begin{aligned} m_s(x) &\equiv \lim_{h\to 0} E[(\mathbf{N}_s(t+h) - \mathbf{N}_s(t))/h | \mathbf{N}_s(t) = x] \\ &= \lim_{h\to 0} E[\frac{(N_s(t+h) - N_s(t))}{h\sqrt{s}} | N_s(t) = s + \lfloor((\lambda-1)/\alpha)s + x\sqrt{s}\rfloor] \\ &= \frac{\lambda s - s - \alpha(\lfloor((\lambda-1)/\alpha)s + x\sqrt{s}\rfloor)}{\sqrt{s}} \quad \text{for} \quad s \quad \text{sufficiently large} \\ &\to -\alpha x \quad \text{as} \quad s \to \infty \ . \end{aligned}$$

The infinitesimal variances are

$$\begin{aligned}
\sigma_s^2(x) &\equiv \lim_{h\to 0} E[(\mathbf{N}_s(t+h) - \mathbf{N}_s(t))^2/h | \mathbf{N}_s(t) = x] \\
&= \lim_{h\to 0} E[\frac{(N_s(t+h) - N_s(t))^2}{hs} | N_s(t) = s + \lfloor((\lambda-1)/\alpha)s + x\sqrt{s}\rfloor] \\
&= \frac{\lambda s + s + \alpha(\lfloor((\lambda-1)/\alpha)s + x\sqrt{s}\rfloor)}{s} \quad \text{for} \quad s \quad \text{sufficiently large} \\
&\to 2\lambda \quad \text{as} \quad s \to \infty \ .
\end{aligned}$$

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 (2002a). A simple consequence of the FCLT is a functional weak law of large numbers (FWLLN), which formalizes the heuristic discussion in (2.3)–(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

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

Corollary 2.1. (FWLLN for the Erlang A model in the ED regime) *Under the conditions of Theorem* 2.1*,*

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

where

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

for q in (2.4)*.*

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 (2002a) – 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.

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) – (2.2)*, 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*

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

where $\mathbf{n}$ is a degenerate Ornstein-Uhlenbeck (OU) diffusion process with infinitesimal mean (state-dependent drift)

$$m(x) = -\alpha x \tag{2.16}$$

and infinitesimal variance

$$\sigma^2(x) = 0 \ ; \tag{2.17}$$

i.e., $\mathbf{n}$ is the ODE

$$\dot{\mathbf{n}}(t) = -\alpha \mathbf{n}(t) \tag{2.18}$$

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

Proof. The proof 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 the infinitesimal means are

$$\begin{aligned} m_s(x) &\equiv \lim_{h \to 0} E[(\hat{\mathbf{N}}_s(t+h) - \hat{\mathbf{N}}_s(t))/h | \hat{\mathbf{N}}_s(t) = x] \\ &= \lim_{h \to 0} E[\frac{(N_s(t+h) - N_s(t))}{hs} | N_s(t) = s + \lfloor ((\lambda - 1)/\alpha)s + xs \rfloor] \\ &= \frac{\lambda s - s - \alpha \lfloor ((\lambda - 1)/\alpha)s + x \rfloor}{s} \quad \text{for} \quad s \quad \text{sufficiently large} \\ &\to -\alpha x \quad \text{as} \quad s \to \infty \ . \end{aligned}$$

The infinitesimal variances are

$$\begin{aligned} \sigma_s^2(x) &\equiv \lim_{h \to 0} E[(\hat{\mathbf{N}}_s(t+h) - \hat{\mathbf{N}}_s(t))^2/h | \hat{\mathbf{N}}_s(t) = x] \\ &= \lim_{h \to 0} E[\frac{(N_s(t+h) - N_s(t))^2}{hs^2} | N_s(t) = s + \lfloor ((\lambda - 1)/\alpha)s + xs \rfloor] \\ &= \frac{\lambda s + s + \alpha(\lfloor ((\lambda - 1)/\alpha)s + xs \rfloor)}{s^2} \quad \text{for} \quad s \quad \text{sufficiently large} \\ &\to 0 \quad \text{as} \quad s \to \infty \ . \end{aligned}$$

It is well known that the degenerate OU diffusion (with 0 infinitesimal variance) is the ODE in (2.18). ∎

For customery 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. However, they do with additional arguments. 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) 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 $(W_s(\infty)|S_s)$ denote a random variable with the conditional distribution of the waiting time given that the customer eventually will be served, i.e., $P((W_s(\infty)|S_s) \le x) \equiv P(W_s(\infty) \le x|S_s)$. 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) – (2.2). *Then, as $s \to \infty$,*

$$\mathbf{N}_s(\infty) \equiv \frac{N_s(\infty) - s(1+q)}{\sqrt{s}} \Rightarrow \mathbf{N}(\infty) \stackrel{\mathrm{d}}{=} N(0, \frac{\lambda}{\alpha}) \ , \tag{2.19}$$

$$\hat{\mathbf{N}}_s(\infty) \equiv \frac{N_s(\infty)}{s} \Rightarrow \hat{\mathbf{N}}(\infty) = 1 + q \ , \tag{2.20}$$

$$P(N_s(\infty) \le s) \to 0 \ , \tag{2.21}$$

$$\mathbf{Q}_s(\infty) \equiv \frac{[N_s(\infty) - s]^+ - sq}{\sqrt{s}} \Rightarrow \mathbf{Q}(\infty) \stackrel{\mathrm{d}}{=} N(0, \frac{\lambda}{\alpha}) \ , \tag{2.22}$$

$$\hat{\mathbf{Q}}_s(\infty) \equiv \frac{Q_s(\infty)}{s} \Rightarrow \hat{\mathbf{Q}}(\infty) = q \equiv \frac{\lambda - 1}{\alpha} \ , \tag{2.23}$$

$$P_s(ab) \Rightarrow P(ab) \equiv \frac{\lambda - 1}{\lambda} = \frac{\rho - 1}{\rho} \ , \tag{2.24}$$

$$(W_s(\infty)|S_s) \Rightarrow w \ , \tag{2.25}$$

$$W_s(\infty) \Rightarrow W \ , \tag{2.26}$$

where w is the deterministic quantity

$$w \equiv \frac{1}{\alpha} \log_e(\lambda) = -\frac{1}{\alpha} \log_e(1 - P(ab)) > 0 \ , \tag{2.27}$$

and W is the random variable with

$$P(W > x) = e^{-\alpha x}, \quad 0 \le x \le w, \quad \textit{and} \quad P(W > w) = 0 \tag{2.28}$$

for w in (2.27)*, which has expected value*

$$E[W] = \frac{P(ab)}{\alpha} = \frac{q}{\lambda}. \tag{2.29}$$

Proof. For the first limit in (2.19), 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 (2002a); 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 since these particular initial distributions are stationary distributions, the limiting distribution must be a stationary distribution for the limiting OU diffusion process. However, the OU diffusion process has a unique stationary distribution. Thus all convergent subsequences must have that normal stationary distribution as their limiting distribution. With that additional argument, Theorem 2.1 implies (2.19). The next limit (2.20) follows by dividing by $\sqrt{s}$ and letting $s \to \infty$, just as in Corollary 2.1. Then (2.21) is an immediate consequence. The limits for the scaled queue-length processes in (2.22) and (2.23) follow from (2.19) by continuous mapping theorems. To establish (2.24), note that in steady state the servers are all busy asymptotically, by (2.21). Hence, after dividing by s, the service rate is asymptotically 1. Since the arrival rate is asymptotically λ after dividing by s, the abandonment rate necessarily is asymptotically $\lambda - 1$ and $P_s(ab) \to P(ab) \equiv (\lambda - 1)/\lambda$.

We now turn to the waiting-time results. The waiting time for a customer that eventually will be served $(W - s(\infty)|S_s)$ is the first passage time to the zero state, starting with $Q_s(\infty)$ customers, if we turn off the arrival process right 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

$$\dot{q}(t) \equiv \frac{dq}{dt} = -1 - \alpha q(t) \tag{2.30}$$

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 (2.30). Next, it is easy to see that the unique solution to the ODE in (2.30) is

$$q(t) = (q + \frac{1}{\alpha})e^{-\alpha t} - \frac{1}{\alpha}, \quad t \geq 0 \ . \tag{2.31}$$

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

$$w = -\frac{1}{\alpha}\log_e(\frac{1}{1+\alpha q}) = \frac{1}{\alpha}\log_e(\lambda) \ , \tag{2.32}$$

as given in (2.27). Given that served customers wait exactly w in the limit as $s \to \infty$, we immediately obtain (2.26) and (2.28). ∎

We observe that (2.29) is consistent with two exact relations for the $M/M/s+M$ model. First, by *Little's law* or $L = \lambda W$, we have

$$E[Q_s(\infty)] = \lambda_s E[W_s(\infty)] \ , \tag{2.33}$$

even without the $M/M/s+M$ assumptions; e.g., see Whitt (1991). Second, for the $M/M/s+M$ model we have

$$P_s(ab) = \alpha E[W_s(\infty)] \ . \tag{2.34}$$

Combining (2.33) and (2.34), we obtain

$$P_s(ab) = \frac{\alpha}{\lambda_s}E[Q_s(\infty)] = \frac{\alpha}{s\lambda}E[Q_s(\infty)] \ . \tag{2.35}$$

Thus, when we know any one of these three important performance measures, we know all three. Formula (2.29) 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 (2.23)–(2.27):

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

Combining the first approximation in (3.1) with the exact relation in (2.34), we obtain the approximation

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

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

$$E[W_s(\infty)|S_s] \approx w = -\frac{1}{\alpha}\log_e(1 - P(ab)) \approx \frac{P(ab)}{\alpha} = \frac{(\lambda - 1)}{\lambda\alpha} \ . \tag{3.3}$$

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

$$Q_s(\infty) \approx N(\frac{(\lambda - 1)s}{\alpha}, \frac{\lambda s}{\alpha}) \ . \tag{3.4}$$

An important consequence is an approximation for the variance:

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

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. In particular, the approximations for the mean steady-state queue length and waiting time depend critically upon the model 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 next discuss refined heuristic approximations that do not follow directly from Theorem 2.3. A simple refinement to (3.4) is

$$Q_s(\infty) \approx N(\frac{(\lambda - 1)s}{\alpha}, \frac{\lambda s}{\alpha})^+ \ , \tag{3.6}$$

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.,

$$\Phi(y) \equiv P(N(0,1) \le x) \equiv \int_{-\infty}^{y} \phi(x)\, dx \quad \text{where} \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \ . \tag{3.7}$$

Let Φ^c be the associated complementary cdf (ccdf), defined by $\Phi^c(x) \equiv 1 - \Phi(x)$. We then obtain the associated approximations:

$$P(Q_s(\infty) > 0) \approx P(N(qs, vs) > 0) = \Phi^c(-qs/\sqrt{vs}) \ , \tag{3.8}$$

$$E[Q_s(\infty)|Q_s(\infty) > 0)] \approx E[N(qs, vs)|N(qs, vs) > 0)] = qs + \sqrt{vs}\frac{\phi(-qs/\sqrt{vs})}{\Phi^c(-qs/\sqrt{vs})} \tag{3.9}$$

and

$$\begin{aligned} E[Q_s(\infty)^2|Q_s(\infty) > 0) &\approx E[N(qs, vs)^2|N(qs, vs) > 0) \\ &= (qs)^2 + vs + 2q\sqrt{vs}\frac{\phi(-qs/\sqrt{vs})}{\Phi^c(-qs/\sqrt{vs})} \\ &\quad +(vs)\frac{\frac{-qs}{\sqrt{vs}}\phi(\frac{-qs}{\sqrt{vs}})}{\Phi^c(\frac{-qs}{\sqrt{vs}})} \ , \end{aligned} \tag{3.10}$$

where

$$q \equiv \frac{\lambda - 1}{\alpha} \quad \text{and} \quad v \equiv \frac{\lambda}{\alpha} \ . \tag{3.11}$$

The conditional normal moments in the last display 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 $Var[Q_s(\infty)]$.

Since the three performance measures $P_s(ab)$, $E[Q_s(\infty)]$ and $E[W_s(\infty)]$ are all related by the exact relations in (2.33)–(2.35), 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

$$P_s(ab) \approx P(ab)\frac{E[Q_s(\infty)]}{(\lambda - 1)s/\alpha} \ . \tag{3.12}$$

Of course, approximations for $E[Q_s(\infty)]$ translate immediately into approximations for $E[W_s(\infty)]$ by virtue of Little's law, (2.33).

We now consider refined approximations for $P(W_s(\infty) \le x|S_s)$, the cdf of the conditional steady-state waiting time until beginning service, given that the customer is served. Since 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 (2.27), saying that

$$P(W_s(\infty) \le x|S_s, Q_s(\infty) = qs) \approx 1 \quad \text{if} \quad -\frac{1}{\alpha}\log_e(\frac{1}{1 + \alpha q}) \le x \tag{3.13}$$

and 0 otherwise. However,

$$-\frac{1}{\alpha}\log_e(\frac{1}{1 + \alpha q}) \le x \quad \text{if and only if} \quad q \le \frac{1}{\alpha}(e^{\alpha x} - 1) \ . \tag{3.14}$$

Thus, we obtain the approximation

$$\begin{aligned} P(W_s(\infty) \le x|S_s) &\approx P(Q_s(\infty) \le \frac{s}{\alpha}(e^{\alpha x} - 1)) \\ &\approx P(N(sq, sv) \le \frac{s}{\alpha}(e^{\alpha x} - 1)) \\ &= \Phi(\frac{[\frac{s}{\alpha}(e^{\alpha x} - 1) - sq]}{\sqrt{sv}}) \end{aligned} \tag{3.15}$$

for v in (3.11).

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 $\alpha \downarrow 0$ instead of $s \uparrow \infty$, and indeed that has already been done by Ward and Glynn (2002). However, they did not emphasize the overloaded regime, where ρ is fixed with $\rho > 1$.

We go further here by establishing limits as $s/\alpha \to \infty$, allowing either $s \uparrow \infty$ or $\alpha \downarrow 0$, or both. For the special case in which only $\alpha \downarrow 0$, we recover the result by Ward and Glynn (2002); for the special case in which only $s \uparrow \infty$, we recover Theorem 2.1.

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

$$\mathbf{Y}_{s,\alpha}(t) \equiv \sqrt{\frac{\alpha}{s}}[N_{s,\alpha}(t/\alpha) - s - \frac{(\lambda-1)s}{\alpha}], \quad \text{for} \quad t \geq 0 \ . \tag{4.1}$$

Theorem 4.1. (the ED limit as $s/\alpha \to \infty$) *Consider the $M/M/s/r + M$ models defined in Section 2, satisfying (2.1) and (2.2). 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*

$$\mathbf{Y}_{s,\alpha} \Rightarrow \mathbf{Y} \quad \textit{in} \quad D \quad \textit{as} \quad s/\alpha \to \infty \ , \tag{4.2}$$

where $\mathbf{Y}$ is an OU diffusion process with infinitesimal mean

$$m(x) = -x \tag{4.3}$$

and infinitesimal variance

$$\sigma^2(x) = 2\lambda \ . \tag{4.4}$$

Proof. Just as in Theorem 2.1, we can apply Stone (1963). Here, the infinitesimal means are

$$\begin{aligned}
m_{s,\alpha}(x) &\equiv \lim_{h \to 0} E[(\mathbf{Y}_{s,\alpha}(t+h) - \mathbf{Y}_{s,\alpha}(t))/h | \mathbf{Y}_{s,\alpha}(t) = x] \\
&= \lim_{h \to 0} \sqrt{\alpha/s} E[(N_{s,\alpha}((t+h)/\alpha) - N_{s,\alpha}(t/\alpha))/h | N_{s,\alpha}(t/\alpha) = s + \lfloor((\lambda-1)s/\alpha) + (x\sqrt{s/\alpha})\rfloor] \\
&= \frac{(\lambda-1)\sqrt{s}}{\sqrt{\alpha}} - \sqrt{\alpha/s}(\lfloor((\lambda-1)s/\alpha) + (x\sqrt{s/\alpha})\rfloor) \quad \text{for all} \quad \text{sufficiently large} \quad s/\alpha \\
&\to -x \quad \text{as} \quad s/\alpha \to \infty \ .
\end{aligned}$$

The infinitesimal variances are

$$\begin{aligned}
\sigma^2_{s,\alpha}(x) &\equiv \lim_{h\to 0} E[(\mathbf{Y}_\alpha(t+h) - \mathbf{Y}_\alpha(t))^2/h | \mathbf{Y}_\alpha(t) = x] \\
&= \lim_{h\to 0} (\alpha/s) E[(N_{s,\alpha}((t+h)/\alpha) - N_{s,\alpha}(t/\alpha))^2]/h | N_{s,\alpha}(t/\alpha) = s + \lfloor ((\lambda-1)s/\alpha) + x\sqrt{s/\alpha} \rfloor] \\
&= \lambda + 1 + (\alpha/s)(\lfloor ((\lambda-1)s/\alpha) + x \rfloor) \quad \text{for all} \quad \text{sufficiently large} \quad s/\alpha \\
&\to 2\lambda \quad \text{as} \quad s/\alpha \to \infty \ .
\end{aligned}$$

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).

5. Numerical Comparisons

In this section we evaluate the ED approximations in Section 3, based on the limits in Sections 2 and 4, by comparing them to exact numerical results for the Erlang A model, using the algorithm described in Whitt (2003). We vary s with both the ratio s/α and the limiting abandonment rate $(\lambda_s - s)/s = \lambda - 1$ held fixed. We consider two values for the ratio s/α, 1000 and 100; we consider two values for the abandonment rate $\lambda - 1$, 0.10 and 0.02.

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, $E[W_s(\infty)|S_s]$. It should be denoted that the first two, $P_s(ab)$ and $E[Q_s(\infty)]$ are connected by the exact relation (2.35), 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 (3.8) and (3.9). The refined approximation for the standard deviation of the queue length is obtained by combining (3.8)–(3.10). The refined approximation for the probability of abandonment, $P_s(ab)$, is obtained from (3.12), 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 the abandonment rate $\lambda - 1$ increases. (Consistent with intuition, that tends to be true more generally.) As should be expected, the approximations are very accurate when $s/\alpha = 1,000$ and $\lambda - 1 = 0.10$, but the approximations are crude when $s/\alpha = 100$ and $\lambda - 1 = 0.02$. In the first case, the simple predicted mean steady-state queue length is 100, while in the second case the simple predicted mean steady-state queue length is only 2.0.

	performance measures as a function of s with s/α fixed						
	$s/\alpha = 1,000$ and $\lambda - 1 = 0.10$						
	approximations		exact with s servers				
perf. meas.	*simple*	*refined*	*1*	*10*	*100*	*1,000*	*10,000*
$P_s(ab)$	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
$sE[W_s(\infty)\|S_s]$	95.31	–	94.92	94.90	94.85	94.82	94.81
	$s/\alpha = 1,000$ and $\lambda - 1 = 0.02$						
	approximations		exact with s servers				
perf. meas.	*simple*	*refined*	*1*	*10*	*100*	*1,000*	*10,000*
$P_s(ab)$	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
$sE[W_s(\infty)\|S_s]$	19.79	–	33.2	32.0	29.2	24.6	20.9
	$s/\alpha = 100$ and $\lambda - 1 = 0.10$						
	approximations		exact with s servers				
perf. meas.	*simple*	*refined*	*1*	*10*	*100*	*1,000*	*10,000*
$P_s(ab)$	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
$sE[W_s(\infty)\|S_s]$	9.53	–	11.9	11.1	10.1	9.4	9.2
	$s/\alpha = 100$ and $\lambda - 1 = 0.02$						
	approximations		exact with s servers				
perf. meas.	*simple*	*refined*	*1*	*10*	*100*	*1,000*	*10,000*
$P_s(ab)$	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
$sE[W_s(\infty)\|S_s]$	1.98	–	8.03	6.88	4.90	3.00	2.13

Table 1: A comparison of approximations with exact numerical values for several performance measures in the $M/M/s+M$ model: 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 scaled mean conditional steady-state waiting time given that the customer is eventually served, $sE[W_s(\infty)|S_s]$. The results are displayed as a function of s for fixed ratio s/α and fixed abandonment rate $\lambda - 1$ (and thus fixed limiting abandonment probability $P(ab)$). Four cases are considered: all combinations of $s/\alpha = 100$ and $1,000$ and $\lambda - 1 = 0.02$ and 0.10.

Except possible in the last case with $s/\alpha = 100$ and $\lambda - 1 = 0.02$, Table 1 shows that the four performance measures do not vary greatly with s, over a very wide range, when s/α and $\lambda - 1$ 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 $\lambda - 1 = 0.10$, the refinements are much better than the simple approximations. Nevertheless, we feel the simple approximations are especially useful for making very 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, $E[W_s(\infty)|S_s]$. 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.

In order to evaluate call-center performance, there is great interest in *service level*. It is standard to require that $x\%$ of all calls be answered within y seconds, for values such as $x = 80\%$ and $y = 30$ seconds. Thus we are especially interested in having an approximation for the conditional cdf $P(W_s(\infty) \le x|S_s)$ for appropriate values of x. Since 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 $P(W_s(\infty) \le x|S_s)$ in (3.15) performs as a function of θ for arguments of the form $x = \theta P(ab)/\alpha$ in Table 2. Here we vary both s and θ, keeping the quantities s/α and $\lambda - 1$ fixed. We consider the same four cases of s/α and $\lambda - 1$ 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 quite well for the higher arguments needed to determine staffing in order to meet SLA's.

$P(W_s(\infty) \le pP(ab)/\alpha|S)$ as a function of s and θ

			$s/\alpha = 1,000$ and $\lambda - 1 = 0.10$				
case				number of servers s			
θ	*1*	*10*	*100*	*1,000*	*approx.*	*10,000*	*100,000*
0.00	0.00011	0.00033	0.128	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 $\lambda - 1 = 0.02$				
case				number of servers s			
θ	*1*	*10*	*100*	*1,000*	*approx.*	*10,000*	*100,000*
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 $\lambda - 1 = 0.10$				
case				number of servers s			
θ	*1*	*10*	*100*	*approx.*	*1,000*	*10,000*	*100,000*
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 $\lambda - 1 = 0.02$				
case				number of servers s			
θ	*1*	*10*	*100*	*approx.*	*1,000*	*10,000*	*100,000*
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

Table 2: A comparison of the normal approximation for the conditional waiting-time cdf given that a customer is served, $P(W_s(\infty) \le x|S)$, with exact numerical values in the $M/M/s + M$ model. The arguments x considered are $x = \theta P(ab)/\alpha$ as a function of θ and s for fixed ratio s/α and fixed $\lambda - 1$. Four cases are considered: all combinations of $s/\alpha = 100$ and $1,000$ and $\lambda - 1 = 0.02$ and 0.10.

6. A Markovian Model with State-Dependent Abandonment

In this section we extend Section 2 by establishing a stochastic-process limit for the number in system in the $M(n)/M(n)/s/r+M(n)$ model with Markovian state-dependent arrival rates, service rates and abandonment rates. For previous work on queues with state-dependent rates, see Brandt and Brandt (1999, 2002), Mandelbaum and Pats (1995) and references therein.

As before, we consider a sequence of models indexed by the number of servers, s, and let $s \to \infty$. However, now the arrival rate λ_s, total service rate μ_s and total abandonment rate δ_s are functions of the number of customers in the system for each s. We now allow abandonments by customers in service. We make the following regularity assumption: We assume that there are continuous functions $\hat{\lambda}$, $\hat{\mu}$ and $\hat{\delta}$ such that

$$\begin{aligned} \frac{\lambda_s(xs) - s\hat{\lambda}(x)}{\sqrt{s}} &\to 0 , \\ \frac{\mu_s(xs) - s\hat{\mu}(x)}{\sqrt{s}} &\to 0 , \\ \frac{\delta_s(xs) - s\hat{\delta}(x)}{\sqrt{s}} &\to 0 \quad \text{as} \quad s \to \infty . \end{aligned} \tag{6.1}$$

As before, considering ED regimes, we expect the number in system with s servers to concentrate around a level above s where the arrival rate equals the sum of the service rate and the abandonment rate. Thus we look for solutions $x > 1$ to the equation

$$\hat{\lambda}(x) = \hat{\mu}(x) + \hat{\delta}(x) . \tag{6.2}$$

(Essentially the same approach applies when $x \le 1$.)

We are motivated to consider the more general $M(n)/M(n)/s/r + M(n)$ model because the $M/M/s/r+M(n)$ special case was proposed as an approximation for the $M/GI/s/r+GI$ model in Whitt (2003). The Markovian $M/M/s/r+M(n)$ model is much more tractable than the $M/GI/s/r + GI$ model because, in the Markovian model, the number of customers in the system over time is a birth-and-death process. In Whitt (2003), further approximations are proposed to describe the experience of individual customers, starting with a more careful analysis of which customers abandon when an abandonment occurs. For that special case, we choose δ_s to approximate the behavior of the $M/GI/s/r+GI$ model with IID abandon times having abandon-time cdf F. We assume that the cdf F has a probability density function f and work with the hazard-rate function

$$h(x) \equiv \frac{f(x)}{1 - F(x)}, \quad x \ge 0 . \tag{6.3}$$

In particular, the key approximation in Whitt (2003) is an approximation for the abandonment rate of a customer who is j^{th} from the end of a queue (necessarily of length at least j):

$$\alpha_s(j) \approx h(j/\lambda_s) \ . \tag{6.4}$$

We get approximation (6.4) by first recognizing that, in the actual $M/GI/s/r + GI$ model, any customer's abandonment rate would be exactly $h(t)$ if he had been waiting for time t. The problem is that we do not know t, so we must estimate it. Given that customers arrive at rate λ_s, the expected time between successive arrivals is $1/\lambda_s$. Thus, as an approximation, we estimate that a customer who is j^{th} from the end of the queue has been waiting for time j/λ_s. That gives us approximation (6.4).

The associated approximation for the total abandonment rate when there are k customers in the system is then

$$\delta_s(k) \approx \sum_{j=1}^{k-s} \alpha_s(j) \quad \text{for} \quad k > s \ , \tag{6.5}$$

with $\delta_s(k) = 0$ if $k \le s$, because we are assuming customers only abandon before beginning service. As indicated in Whitt (2003), if the density f were not smooth, then we might instead let

$$\alpha_s(j) \approx \lambda_s \int_{(j-1)/\lambda_s}^{j/\lambda_s} h(t)\, dt \ . \tag{6.6}$$

Then the approximate total abandonment when there are k customers in the system would be

$$\delta_s(k) \approx \lambda_s \int_0^{(k-s)/\lambda_s} h(t)\, dt = -\lambda_s \log_e F^c((k-s)/\lambda_s) \quad \text{for} \quad k > s \ , \tag{6.7}$$

and $\delta_s(k) = 0$ for $k \le s$.

The special case considered here starts with (6.7). As in previous sections, we assume that $\lambda_s(k) = s\lambda$ for all k and $\mu_s(k) = k \wedge s = \min\{k, s\}$ for $k \ge 0$. (In this section, $\mu_s(k)$ is the *total* service rate when there are k customers in the system.) As a consequence, the special case in (6.7) satisfies (6.1). Indeed, $\delta_s(xs) = s\hat{\delta}(x)$ for all $x > 0$ and $s > 0$, where

$$\hat{\delta}(1+x) = -\lambda \log_e F^c(x/\lambda) \quad \text{for all} \quad x > 0 \tag{6.8}$$

and $\hat{\delta}(x) = 0$ for $x \le 0$.

Our goal here related to Whitt (2003) is to establish a heavy-traffic stochastic-process limit in the ED regime for the $M/M/s/r + M(n)$ model with the state-dependent abandonment rates δ_s satisfying (6.7). Moreover, we want to see how the associated FWLLN (the solution to equation (6.2)) is related to the fluid approximation for the $M/GI/s/r + GI$ model with the

same abandon-time cdf F, discussed in Whitt (2004), because that is evidently asymptotically correct for the $M/GI/s/r+GI$ model. (In Whitt (2004) a proof is given for a discrete-time analog.) We would like to have the approximation in Whitt (2003) be asymptotically correct in the ED heavy-traffic limit, but we find that the FWLLN for the $M/M/s/r+M(n)$ model, assuming (6.7), is different from the fluid approximation for the associated $M/M/s/r+GI$ model. But the difference is often very small, so the two separate fluid models do provide support for the approximation in Whitt (2003).

As in Section 2, we focus on the number in system at time t, $N_s(t)$, and work with the scaled process $\mathbf{N}_s$ defined in (2.7). As in Section 2, let the initial state $N_s(0)$ be specified independently, so that the stochastic process $\{N_s(t) : t \geq 0\}$ is Markov.

Theorem 6.1. (stochastic-process limit for the state-dependent model in the ED regime) *Consider the sequence of $M(n)/M(n)/s/r+M(n)$ models specified above, satisfying* (6.1). *Suppose that*

$$\mathbf{N}_s(0) \Rightarrow \mathbf{N}(0) \quad in \quad \mathbb{R} \quad as \quad s \to \infty \ , \tag{6.9}$$

where $\mathbf{N}_s$ is the scaled process in (2.7) *and the constant $1+q$ appearing there is a solution to equation* (6.2). *Assume that $r_s = sr$, where $1+q < r \leq \infty$. Moreover, suppose that the functions $\hat{\lambda}$, $\hat{\mu}$ and $\hat{\delta}$ appearing in* (6.1) *have continuous derivatives $\hat{\lambda}'$, $\hat{\mu}'$ and $\hat{\delta}'$ in the neighborhood of the point $1+q$. Then*

$$\mathbf{N}_s \Rightarrow \mathbf{N} \quad in \quad D \quad as \quad s \to \infty \ , \tag{6.10}$$

where $\mathbf{N}$ is a diffusion process with infinitesimal mean

$$m(x) = -\zeta x \tag{6.11}$$

for

$$-\zeta \equiv \hat{\lambda}'(1+q) - \hat{\mu}'(1+q) - \hat{\delta}'(1+q) \tag{6.12}$$

and infinitesimal variance

$$\sigma^2(x) = 2\hat{\lambda}(1+q) \ . \tag{6.13}$$

If $\zeta > 0$, then $\mathbf{N}$ is an OU process with steady-state distribution

$$\mathbf{N}(\infty) \stackrel{\mathrm{d}}{=} N(0, \lambda/\zeta) \ . \tag{6.14}$$

For the special case of the $M/M/s/r+M(n)$ model with δ_s in (**??**), *the constant q satisfies the equation*

$$F^c(q/\lambda) = e^{-(\lambda-1)/\lambda} \ , \tag{6.15}$$

where $F^c \equiv 1 - F$ is the abandon-time ccdf and the drift rate is

$$\zeta = \hat{\delta}'(1+q) = h(q/\lambda) \ , \tag{6.16}$$

where h is the failure-rate function associated with the abandon-time cdf F in (6.3).

Proof. The proof is a modification of the proof of Theorem 2.1. Again N_s is a birth-and-death process and the limit process has no boundaries, so we can apply Stone (1963). Given that result, it suffices to show that the infinitesimal means and variances converge. Let $o(1)$ be a quantity that converges to 0 as $s \to \infty$. We exploit conditions (6.1) and (6.2) and apply Taylor's theorem to represent the functions $\hat{\lambda}$, $\hat{\mu}$ and $\hat{\delta}$ in the neighborhood of the point $1+q$.

For the infinitesimal means,

$$\begin{aligned}
m_s(x) &\equiv \lim_{h\to 0} E[(\mathbf{N}_s(t+h) - \mathbf{N}_s(t))/h | \mathbf{N}_s(t) = x] \\
&= \lim_{h\to 0} E[\frac{N_s(t+h) - N_s(t)}{h\sqrt{s}} | N_s(t) = (1+q)s + \lfloor x\sqrt{s} \rfloor] \\
&= \frac{\lambda_s((1+q)s + \lfloor x\sqrt{s}\rfloor) - \mu_s((1+q)s + \lfloor x\sqrt{s}\rfloor) - \delta_s((1+q)s + \lfloor x\sqrt{s}\rfloor)}{\sqrt{s}} \\
&= \frac{s\hat{\lambda}((1+q) + x/\sqrt{s}) - s\hat{\mu}((1+q) + x/\sqrt{s}) - s\hat{\delta}((1+q)s + x/\sqrt{s})}{\sqrt{s}} + o(1) \\
&= \frac{s\hat{\lambda}(1+q) + s\hat{\lambda}'(1+q)(x/\sqrt{s}) - s\hat{\mu}(1+q) - s\hat{\mu}'(1+q)(x/\sqrt{s})}{\sqrt{s}} \\
&\qquad - \frac{s\hat{\delta}(1+q) - s\hat{\delta}'(1+q)(x/\sqrt{s})}{\sqrt{s}} + o(1) \\
&\to \hat{\lambda}'(1+q)x - \hat{\mu}'(1+q)x - \hat{\delta}'(1+q)x = -\zeta x
\end{aligned}$$

for ζ in (6.12).

For the infinitesimal variances,

$$\begin{aligned}
\sigma_s^2(x) &\equiv \lim_{h\to 0} E[(\mathbf{N}_s(t+h) - \mathbf{N}_s(t))^2/h | \mathbf{N}_s(t) = x] \\
&= \lim_{h\to 0} E[\frac{(N_s(t+h) - N_s(t))^2}{hs} | N_s(t) = (1+q)s + \lfloor x\sqrt{s} \rfloor] \\
&= \frac{\lambda_s((1+q)s + \lfloor x\sqrt{s}\rfloor) + \mu_s((1+q)s + \lfloor x\sqrt{s}\rfloor) + \delta_s((1+q)s + \lfloor x\sqrt{s})}{\sqrt{s}\rfloor} \\
&= \frac{s\hat{\lambda}((1+q) + x/\sqrt{s}) + s\hat{\mu}((1+q) + x/\sqrt{s}) + s\hat{\delta}((1+q)s + x/\sqrt{s})}{s} + o(1) \\
&\to 2\hat{\lambda}(1+q) \ .
\end{aligned}$$

As before, the OU diffusion has a normal steady-state distribution with variance equal to the infinitesimal variance divided by twice the state-dependent drift rate. ∎

An analog of Corollary 2.1 clearly holds here too. It is also possible to establish analogs of Theorem 2.2 and 2.3 too, but we do not. However, without additional regularity conditions, the behavior of the steady-state distributions can be much more complicated. We have not yet imposed conditions strong enough to have a unique solution to the fundamental "net-input" equation (6.2). In general there could be multiple "equilibrium points" about which the process $\mathbf{N}_s$ would tend to fluctuate. However, in many applications the solution to (6.2) will be unique; sufficient conditions are easy to provide. In particular, in our application to obtain ED limits for the $M/M/s/r + M(n)$ model, it suffices to have $\hat{\delta}$ be strictly increasing (as well as continuous) in x for $x > 1$.

For the $M/M/s/r + M(n)$ model, assuming that (6.2) has a unique solution q, we can obtain approximations just as in Section 3. Paralleling (3.6), a natural approximation for the steady-state queue length in the $M/M/s/r + M(n)$ model is

$$Q_s(\infty) \approx N(qs, \lambda s/\zeta)^+ \ , \tag{6.17}$$

for λ in (2.1), q in (6.15) and ζ in (6.16). Assuming that there are points a and b with $0 \leq a < b \leq \infty$ such that $F^c(a) = 1$, $F^c(b) = 0$ and F^c is strictly decreasing on the interval (a, b), there necessarily is a unique solution q to equation (6.15). It suffices for the density f to be strictly positive on the interval (a, b) and be 0 elsewhere.

We thus see that the crucial quantities to understand the ED behavior of the $M/M/s/r + M(n)$ model are, first, the solution q to equation (6.15) (which will be unique under the regularity condition just mentioned) and, second, the OU drift parameter ζ, which is approximately $h(q/\lambda)$. Of course, for typical values of s, if the hazard rate function is not nearly constant in the neighborhood of q/λ, then the approximation method in Whitt (2003) is likely to yield better numerical results.

Remark 6.1. *The $M/M/s/r+M$ Special Case.* We now show that Theorem 6.1 is consistent with Theorem 2.1 in Section 2. If the abandon-time cdf F is exponential with mean $1/\alpha$, then $F^c(x) = e^{-\alpha x}$ and the failure-rate function is $h(t) = \alpha$ for all t. Equation (6.2) thus becomes (6.15) with $F^c(x) = e^{-\alpha x}$, which implies that

$$q = \frac{\lambda - 1}{\alpha} \ , \tag{6.18}$$

just as in Theorem 2.1. Since $h(t) = \alpha$ for all t, the state-dependent drift is

$$\hat{\delta}'(q/\lambda) = \alpha \ , \tag{6.19}$$

again as in Theorem 2.1. ∎

Remark 6.2. *Comparing Approximations.* To gain insight into the $M/M/s/r + M(n)$ approximation for the $M/GI/s/r + GI$ model proposed in Whitt (2003), we now compare the fluid approximation for the $M/GI/s/r + GI$ model in Whitt (2004) to the solution for q here in (6.15), which we denote by q^M, using the superscript M to denote "Markov." In contrast, let q^F denote the corresponding fluid approximation from Whitt (2004).

We do not intend to discuss the fluid approximation for the $M/GI/s/r + GI$ queue in Whitt (2004) in detail; we only summarize the results. From Whitt (2004), we have

$$q^F = \lambda \int_0^{w^F} F^c(x)\, dx \ , \tag{6.20}$$

where

$$F(w^F) = \frac{\lambda - 1}{\lambda} \ . \tag{6.21}$$

The quantity w^F is the time spent in system by served fluid, analogous to w in (2.27).

Comparing equations (6.20) and (6.21) to equation (6.15), we see that in general q^M need not coincide with q^F. To make further connections, we first change notation, writing

$$\epsilon = \lambda - 1 \ , \tag{6.22}$$

so that we can focus on the comparison for small ϵ. We then make an additional simplifying assumption for the Markovian model: We assume for the $M/M/s/r + M$ model that all abandonments are from the front of the queue (by the customers who have been there the longest). With that assumption, the waiting time of all customers, served or not, is the same, and must be

$$w^M = q^M/\lambda \ . \tag{6.23}$$

Combining equations (6.15) and (6.23), we obtain the equation

$$F(w^M) = 1 - e^{-(\lambda-1)/\lambda} = 1 - e^{-\epsilon/(1+\epsilon)} \ . \tag{6.24}$$

Equation (6.24) is convenient, because it is easy to compare to equation (6.21), which with the change of notation, becomes

$$F(w^F) = \frac{\epsilon}{1+\epsilon} \ . \tag{6.25}$$

First, from equations (6.24) and (6.25), we easily see that $w^F \neq w^M$. However, if we expand the exponential in (6.24), then we obtain

$$1 - e^{-\epsilon/(1+\epsilon)} = \frac{\epsilon}{1+\epsilon} - \frac{\epsilon^2}{2(1+\epsilon)^2} + \frac{\epsilon^3}{6(1+\epsilon)^3} + O(\epsilon^4) \quad \text{as} \quad \epsilon \downarrow 0 \ . \tag{6.26}$$

To relate the quantities w^F and w^M, assume that the abandon-time cdf F has a positive density f. Then the cdf F is continuous and strictly increasing, so that it has an inverse, say $g \equiv F^{-1}$. Then

$$w^F = g(1/(1+\epsilon)) \quad \text{and} \quad w^M = g(1 - e^{-\epsilon/(1+\epsilon)}) \ . \tag{6.27}$$

Using a Taylor series expansion, we get

$$w^M \approx w^F - g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2} \ . \tag{6.28}$$

From formulas (6.20) – (6.24), we also have the inequalities

$$w^F \le q^F \le w^F(1+\epsilon) \tag{6.29}$$

while

$$w^F(1+\epsilon) - g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)} \approx q^M = w^M(1+\epsilon) \le w^F(1+\epsilon) \ . \tag{6.30}$$

We then have the bounds

$$|q^F - w^F(1+\epsilon)| \le \epsilon w^F \ , \tag{6.31}$$

$$|q^M - w^F(1+\epsilon)| \le g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2} \tag{6.32}$$

and

$$|q^M - q^F| \le \max\{\epsilon w^F, g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2}\} \ . \tag{6.33}$$

■

Example 6.1. *The case of a Uniform Abandon-Time Distribution.* Suppose that the abandon-time distribution is uniformly distributed on the interval $[0,1]$, so that the abandon-time cdf is $F(x) = x$, $0 \le x \le 1$. From (6.20) and (6.15), we see that in this case

$$q^F = \epsilon - \frac{\epsilon^2}{2(1+\epsilon)} \ , \tag{6.34}$$

while

$$q^M = (1+\epsilon)(1 - e^{-\epsilon/(1+\epsilon)}) = \epsilon - \frac{\epsilon^2}{2(1+\epsilon)} + \frac{\epsilon^3}{6(1+\epsilon)^2} + O(\epsilon^4) \ , \tag{6.35}$$

so that

$$q^M - q^F = \frac{\epsilon^3}{6(1+\epsilon)^2} + O(\epsilon^4) \ . \tag{6.36}$$

For example, if $\epsilon = 0.1$, then $q^F = 0.09545$, while $q^M = 0.09559$ and $q^M - \mathbf{q}^F \approx 0.0001377$. There is a difference of only about 0.1%. That is much closer than predicted by the bounds in (6.31)–(6.33), because $w^F(1+\epsilon) = \epsilon = 0.1$, $\epsilon w^F = \epsilon^2/(1+\epsilon) = 0.0091$, $g'(x) = 1$, $0 < x < 1$, and $g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2} = \frac{\epsilon^2}{2(1+\epsilon)^2} = \frac{0.01}{2.42} = 0.0041$. ■

7. Conclusions

This paper establishes ED many-server heavy-traffic limits for Markovian queues with customer abandonments. 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 service-time distribution is 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: (i) Assume in addition that the probability of abandonment, $P_s(ab)$ converges to a limit strictly between 0 and 1, or (ii) Assume in addition that the traffic intensity ρ remains fixed with $\rho > 1$.

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 (a.k.a Erlang A model) in (3.1)–(3.5), obtained directly from the limits, are remarkably simple. The heuristic refinements in (3.6)–(3.15) are also not too complicated. Tables 1 and 2 show that the approximations are quite accurate. Compared to QED approximations, 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.

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 $(\lambda_s - s)/s = \lambda - 1$. Indeed, in Section 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.

In the final section, Section 6, we extend the ED many-server heavy-traffic limit to $M(n)/M(n)/s/r + M(n)$ models, which have state dependent arrival rates, service rates and abandonment rates. By combining the $M/M/s/r + M(n)$ approximation for the $M/GI/s/r + GI$ model in Whitt (2003) with the ED approximations for the $M/M/s/r + M(n)$ model in Section 6, we obtain new ED diffusion approximations for the $M/GI/s/r + GI$ model too. The resulting simple approximations for performance measures in the $M/GI/s/r + GI$ model can hardly compete with the numerical results in Whitt (2003), but the simple formulas provide insight and again can serve as the basis for back-of-the-envelope calculations. For example, we see that changing the abandon-time distribution from exponential (M) to non-exponential

(GI) leads to a new normal approximation for the steady-state queue length, (6.17), with the key parameters q and ζ determined by (6.15) and (6.16), instead of (3.6). Consistent with the numerical results in Whitt (2003), both the mean and the variance of the normal random variable change when the abandon-time distribution is changed from exponential to non-exponential, even when the mean time to abandon is unchanged.

References

Billingsley, P. 1999. *Convergence of Probability Measures*, second edition, Wiley, New York.

Brandt, A., M. Brandt. 1999. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation* 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. and W. Whitt, 1995. Piecewise-linear diffusion processes. *Advances in Queueing*, J. Dshalalow, ed., CRC Press, Boca Raton, FL, 463–480.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, Review and Research Prospects. *Manufacturing and Service Opns. Mgmt.* 5, to appear.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Opns. Mgmt.*, 4, 208-227.

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

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

Jelenkovic P., A. Mandelbaum, P. Momcilovic. 2002. Heavy Traffic Limits for Queues with Many Deterministic Servers. Columbia University.

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

Mandelbaum, A. 2001. Call centers (centres): research bibliography with abstracts. Faculty of Industrial Engineering and Management, Technion, Haifa.

Mandelbaum, A. and G. Pats. 1995. State-dependent queues: approximations and applications. In *Stochastic Networks*, IMA Volumes in Mathematics, F. P. Kelly and R. J. Williams, eds., Springer, 239–282.

Mandelbaum, A., S. Zeltyn. 2003. The impact of customers patience on delay and abandonment: some empirically-driven experiments with the $M/M/n+G$ queue. The Technion, Israel.

Puhalskii, A. A., M. I. Reiman. 2000. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Prob.* 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. and 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. (Plus correction note, 12, 431–432.)

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

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

Whitt, W. 2002b. A diffusion approximation for the $G/GI/n/m$ queue. Department of Industrial Engineering and Operations Research, Columbia University. *Operations Res.*, to appear.

Whitt, W. 2002c. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. Department of Industrial Engineering and Operations Research, Columbia University.

Whitt, W. 2003. Engineering solution of a basic call-center model. Department of Industrial Engineering and Operations Research, Columbia University.

Whitt, W. 2004. Fluid models for multiserver queues with abandonments. Department of Industrial Engineering and Operations Research, Columbia University.

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