TWO FLUID APPROXIMATIONS FOR MULTI-SERVER QUEUES WITH ABANDONMENTS

by

Ward Whitt[1]
Columbia University

Abstract

Insight is provided into a previously developed $M/M/s/r + M(n)$ approximation for the $M/GI/s/r + GI$ queueing model by establishing fluid and diffusion limits for the approximating model. Fluid approximations for the two models are compared in the many-server efficiency-driven (overloaded) regime. The two fluid approximations do not coincide, but they are close.

Short title: Fluid Approximations

Keywords: multiserver queues with abandonment, queues with state-dependent rates, heavy-traffic limits, fluid limits, diffusion approximations, call centers.

June 2, 2004; Revision: August 24, 2004

[1]Ward Whitt, Department of Industrial Engineering and Operations Research, Columbia University, 500 West 120th Street, New York, NY 10027–6699, USA; Email: ward.whitt@columbia.edu

1. Introduction

The primary purpose of this short paper is to supplement and complement two recent papers on multi-server queues with abandonment. These two papers on multi-server queues with abandonment in turn were motivated by the desire to develop new tools to help analyze telephone call centers; for background, see Gans, Koole and Mandelbaum (2003).

First, in Whitt (2005a) we developed an algorithm for calculating approximations for all the standard steady-state performance measures in the $M/GI/s/r + GI$ model, having a Poisson arrival process, independent and identically distributed (IID) service times with a general distribution (the first GI), s servers, r extra waiting spaces, IID times to abandon before starting service with a general distribution (the $+GI$) and the first-come first-served (FCFS) service discipline. That algorithm is based on approximating the given $M/GI/s/r+GI$ model by an associated Markovian $M/M/s/r+M(n)$ model with state-dependent abandonment rates. It yields exact numerical results for the $M/M/s/r + M$ special case.

Second, in Whitt (2005c) we developed a deterministic fluid approximation for the general $G/GI/s/r + GI$ model, having an arrival process that is a general stationary point process, IID service times with a general distribution, s servers, r extra waiting spaces, IID times to abandon with a general distribution and the FCFS service discipline. That fluid approximation describes the transient behavior of the queueing system. The steady-state behavior of that fluid model serves as an approximation for the steady-state behavior of the queueing model. The fluid approximation becomes appropriate in the many-server heavy-traffic limit in which both the arrival rate and the number of servers are allowed to increase. The fluid approximation is especially interesting in the *efficiency-driven* (ED) *limiting regime*, in which the probability of eventually abandoning approaches a limit strictly between 0 and 1 as the the arrival rate and the number of servers approach infinity. Equivalently, the associated sequence of traffic intensities $\{\rho_s : s \geq 1\}$ approaches a limit $\rho > 1$. Indeed, it suffices to assume that the traffic intensity is held fixed with $\rho > 1$. The fluid approximation evidently is asymptotically correct in the ED many-server heavy-traffic limiting regime, but that is yet to be proved. In Whitt (2005c) strong supporting evidence is given by establishing the fluid limit in a discrete-time framework.

Given those two papers, we are interested in establishing an ED many-server heavy-traffic fluid limit for the $M/M/s/r + M(n)$ model to see if the approximation developed in Whitt (2005a) is asymptotically correct, i.e., to see if it agrees with the fluid approximation for the

$M/GI/s/r+GI$ special case in the ED regime. That would provide additional support for the approximation in Whitt (2005a), at least in the ED regime. We hasten to point out that there are two gaps in this program: First, establishing a fluid limit does not directly imply associated convergence of the steady-state distributions (invariant measures) and, second, convergence to the continuous-time fluid limit for the $M/GI/s+GI$ model has not yet been fully proved. But the issue addressed here is worth addressing.

In the present paper we establish the desired deterministic many-server heavy-traffic fluid limit, and a more general diffusion-process limit, for the $M/M/s/r+M(n)$ model. Unfortunately, however, we find that the two fluid approximations do not coincide, but they are sufficiently close that the new fluid limit nevertheless does provide positive support for the approximation in Whitt (2005a). More generally, the difference between the two fluid approximations can be used to help judge if the algorithm in Whitt (2005a) should produce good approximations in contemplated scenarios.

The present paper goes beyond that initial goal by establishing many-server heavy-traffic limits for the more general $M(n)/M(n)/s/r+M(n)$ model, having state-dependent arrival and service rates as well as state-dependent abandonment rates. The many-server heavy-traffic limits here extend the many-server heavy-traffic limits for the $M/M/s/r+M$ model in the ED limiting regime established in Whitt (2005b). Theorems 2.1, 2.2 and 2.3 there established a diffusion limit, a fluid limit, and limits for the steady-state distributions, respectively. That paper also presented numerical examples to show that the ED approximations can be useful for describing the performance of call centers that are providing low-to-moderate quality of service, and thus are experiencing substantial customer abandonment. Such low-to-moderate quality of service often occurs in service-oriented (non-revenue-generating) call centers. It is widely recognized that alternative quality-and-efficiency-driven (QED) many-server heavy-traffic limits yield useful approximations in a wide range of commonly occurring scenarios; see Garnett, Mandelbaum and Reiman (2002). The recent work is aimed at showing the ED approximations can also be useful.

As in Whitt (2005b), the stochastic-process limits established here can be viewed as consequences of corresponding results for more general state-dependent Markovian queues in Mandelbaum and Pats (1995); see Theorems 4.1 and 4.2 plus Section 5.3 there. Nevertheless, the alternative proofs here are appealing because the special cases considered here are much easier to treat directly. In the special cases considered here, the limit processes have no boundaries, so that it is not necessary to consider the reflection map at all. Instead, we use the relatively

simple argument in the seminal heavy-traffic paper on the $M/M/s$ model by Iglehart (1965), drawing upon Stone (1963). The main contribution here, though, is not general theory, but the new insight into the behavior of multi-server queues with abandonment gained by establishing the connections between the two papers Whitt (2005a, c).

Here is how the rest of this paper is organized. First, in Section 2 we establish the the many-server heavy-traffic stochastic-process limits for the $M(n)/M(n)/s/r + M(n)$ model. Afterwards, in Section 3 we discuss associated approximations for the steady-state performance. That depends upon the existence and uniqueness of solutions to a fundamental fixed-point equation, ((2.2)). In general there can be multiple solutions, implying the existence of multiple asymptotic equilibrium points (as $s \to \infty$) even though there is always a unique limiting steady-state distribution for each s.

In Section 4 we briefly describe the $M/M/s/r+M(n)$ approximation for the $M/GI/s/r+GI$ model developed in Whitt (2005a). Next, in Section 5 we describe the fluid approximation for the $G/GI/s/r + GI$ model developed in Whitt (2005c). Finally, in Section 6 we compare the two fluid approximations in the ED regime.

For additional discussion about customer abandonment in queues, see Garnett, Mandelbaum and Reiman (2002) and Mandelbaum and Zeltyn (2004).

2. The Stochastic-Process Limits in the ED Limiting Regime

In the section we establish the stochastic-process limits for the $M(n)/M(n)/s/r + M(n)$ model with Markovian state-dependent arrival rates, service rates and abandonment rates. We consider a sequence of models indexed by the number of servers, s, and let $s \to \infty$.

For each $s \geq 1$, the model is characterized by one parameter and three functions. The parameter is the number of extra waiting spaces r_s, where $0 < r_s \leq \infty$. When $r_s < \infty$, we will let r_s be sufficiently large that it plays no role, asymptotically. The three functions are the arrival rate $\lambda_s \equiv \{\lambda_s(n) : 0 \leq n < s + r_s\}$, the (total) service rate $\mu_s \equiv \{\mu_s(n) : 1 < s + r_s + 1\}$ and the (total) abandonment rate $\delta_s \equiv \{\delta_s(n) : 1 \leq n < s + r_s + 1\}$. (We use $\equiv$ to denote equality by definition.) For example, $\lambda_s(n)$ is the arrival rate when there are n customers in the system, either being served or waiting. If $r_s < \infty$, then $\lambda_s(s + r_s) = 0$.

Let $N_s(t)$ be the number of customers in the system at time t. Let $N_s(0)$ be a random initial number of customers, specified independently of the evolution of the system after time 0 assumed to satisfy $0 \leq N_s(0) < r_s + 1$ with probability one. For each s, the stochastic process $\{N_s(t) : t \geq 0\}$ is a birth-and-death stochastic process with birth rates $\lambda_s(n)$ and death rates

$\mu_s(n) + \delta_s(n)$.

We assume that there are fixed functions $\hat{\lambda}$, $\tilde{\lambda}$, $\hat{\mu}$, $\tilde{\mu}$, $\hat{\delta}$ and $\tilde{\delta}$ such that for each positive real number x and each sequence $\{x_s : s \geq 1\}$, where sx_s is a nonnegative integer with $x_s \to x$ as $s \to \infty$,

$$\begin{aligned} \frac{\lambda_s(sx_s) - s\hat{\lambda}(x)}{\sqrt{s}} &\to \tilde{\lambda}(x) \ , \\ \frac{\mu_s(sx_s) - s\hat{\mu}(x)}{\sqrt{s}} &\to \tilde{\mu}(x) \ , \\ \frac{\delta_s(sx_s) - s\hat{\delta}(x)}{\sqrt{s}} &\to \tilde{\delta}(x) \quad \text{as} \quad s \to \infty \ . \end{aligned} \tag{2.1}$$

The general idea is that, asymptotically as $s \to \infty$, the number of customers in the system will concentrate at a point where the input rate equals the output rate. Thus, for each s, we look for a point $x_s > 1$ such that sx_s is an integer and $\lambda_s(sx_s) \approx \mu_s(sx_s) + \delta_s(sx_s)$. Because of the assumed behavior of the rate functions in (2.1), to capture the behavior asymptotically as $s \to \infty$, we seek $x > 0$ such that there is a solution to the *fundamental fixed-point equation*

$$\hat{\lambda}(x) = \hat{\mu}(x) + \hat{\delta}(x) \ . \tag{2.2}$$

Since we are primarily interested in the ED limiting regime in which the servers all tend to be busy, in our intended application we want $x > 1$, but in general we do not require it.

To establish convergence to a diffusion process in this setting, we form the normalized stochastic process

$$\mathbf{N}_s(t) \equiv \frac{N_s(t) - s\hat{x}}{\sqrt{s}}, \quad t \geq 0 \ , \tag{2.3}$$

for positive real number $\hat{x}$, which will turn out to be a solution to equation (2.2).

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 $Nor(m, \sigma^2)$ denote a random variable that is normally distributed with mean m and variance σ^2.

Theorem 2.1. (stochastic-process limit for the state-dependent model) *Consider the sequence of $M(n)/M(n)/s/r+M(n)$ models specified above, satisfying* (2.1). *Suppose that $\mathbf{N}_s(0) \Rightarrow \mathbf{N}(0)$ in $\mathbb{R}$ as $s \to \infty$, where $\mathbf{N}_s$ is the scaled process in* (2.3). *Assume that the fundamental fixed-point equation* (2.2) *has a solution, denoted by $\hat{x}$, and let the constant $\hat{x}$ appearing in* (2.3) *be*

such a solution. Assume that $r_s \geq s\zeta$ *for all* s*, where* $\hat{x} < \zeta$*, or* $r_s = \infty$ *for all* s*. Moreover, suppose that* (i) *the functions* $\hat{\lambda}$*,* $\hat{\mu}$ *and* $\hat{\delta}$ *appearing in* (2.1) *have continuous derivatives* $\hat{\lambda}'$*,* $\hat{\mu}'$ *and* $\hat{\delta}'$ *in the neighborhood of the point* $\hat{x}$ *and (ii) the functions* $\tilde{\lambda}$*,* $\tilde{\mu}$ *and* $\tilde{\delta}$ *appearing in* (2.1) *are continuous in the neighborhood of the point* $\hat{x}$*. Then*

$$\mathbf{N}_s \Rightarrow \mathbf{N} \quad in \quad D \quad as \quad s \to \infty \ , \tag{2.4}$$

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

$$m(y) = \tilde{\gamma} - \gamma y \tag{2.5}$$

for

$$\tilde{\gamma} \equiv \tilde{\lambda}(\hat{x}) - \tilde{\mu}(\hat{x}) - \tilde{\delta}(\hat{x}) \tag{2.6}$$

and

$$-\gamma \equiv \hat{\lambda}'(\hat{x}) - \hat{\mu}'(\hat{x}) - \hat{\delta}'(\hat{x}) \tag{2.7}$$

and infinitesimal variance

$$\sigma^2(y) = \sigma^2(0) = 2\hat{\lambda}(\hat{x}) \ . \tag{2.8}$$

If $\gamma > 0$*, then* $\mathbf{N}$ *is an Ornstein-Uhlenbeck* (OU) *diffusion process with*

$$\mathbf{N}(t) \Rightarrow \mathbf{N}(\infty) \stackrel{\mathrm{d}}{=} Nor(\tilde{\gamma}/\gamma, \hat{\lambda}(\hat{x})/\gamma) \quad as \quad t \to \infty \ . \tag{2.9}$$

Proof. Since N_s is a birth-and-death process and the limiting 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 (2.3) it suffices to show that the infinitesimal means and variances converge to the infinitesimal means and variance of the limit process.

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

$$y_s \equiv \frac{\lfloor s\hat{x} + y\sqrt{s} \rfloor - s\hat{x}}{\sqrt{s}} \ ,$$

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

To complete the proof, we exploit conditions (2.1) and (2.2) and apply Taylor's theorem to represent the functions $\hat{\lambda}$, $\hat{\mu}$ and $\hat{\delta}$ in the neighborhood of the point $\hat{x}$. Let $o(1)$ be a quantity that converges to 0 as $s \to \infty$.

For the infinitesimal means,

$$\begin{aligned}
m_s(y_s) &\equiv \lim_{h\to 0} E[(\mathbf{N}_s(t+h) - \mathbf{N}_s(t))/h | \mathbf{N}_s(t) = y_s] \\
&= \lim_{h\to 0} E[\frac{N_s(t+h) - N_s(t)}{h\sqrt{s}} | N_s(t) = \hat{x}s + \sqrt{s}y_s] \\
&= \frac{\lambda_s(\hat{x}s + y_s\sqrt{s}) - \mu_s(\hat{x}s + y_s\sqrt{s}) - \delta_s(\hat{x}s + y_s\sqrt{s})}{\sqrt{s}} \\
&= \frac{s\hat{\lambda}(\hat{x} + y_s/\sqrt{s}) + \sqrt{s}\tilde{\lambda}(\hat{x} + y_s/\sqrt{s}) - s\hat{\mu}(\hat{x} + y_s/\sqrt{s})}{\sqrt{s}} \\
&\quad - \frac{\sqrt{s}\tilde{\mu}(\hat{x} + y_s/\sqrt{s}) + s\hat{\delta}(\hat{x}s + y_s/\sqrt{s}) + \sqrt{s}\tilde{\delta}(\hat{x} + y_s/\sqrt{s})}{\sqrt{s}} + o(1) \\
&= \frac{s\hat{\lambda}(\hat{x}) + s\hat{\lambda}'(\hat{x})(y_s/\sqrt{s}) + \sqrt{s}\tilde{\lambda}(\hat{x} + y_s/\sqrt{s})}{\sqrt{s}} \\
&\quad - \frac{s\hat{\mu}(\hat{x}) + s\hat{\mu}'(\hat{x})(y_s/\sqrt{s}) + \sqrt{s}\tilde{\mu}(\hat{x} + y_s/\sqrt{s})}{\sqrt{s}} \\
&\quad - \frac{s\hat{\delta}(\hat{x}) + s\hat{\delta}'(\hat{x})(y_s/\sqrt{s}) + \sqrt{s}\tilde{\delta}(\hat{x} + y_s/\sqrt{s})}{\sqrt{s}} + o(1) \\
&\to \hat{\lambda}'(\hat{x})y + \tilde{\lambda}(\hat{x}) - \hat{\mu}'(\hat{x})y - \tilde{\mu}(\hat{x}) - \hat{\delta}'(\hat{x})y - \tilde{\delta}(\hat{x}) = \tilde{\gamma} - \gamma y \equiv m(y)
\end{aligned}$$

for $\tilde{\gamma}$ in (2.6) and γ in (2.7).

For the infinitesimal variances,

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

It is well known that the limiting diffusion process is an OU process if $\gamma > 0$, and that process 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). Otherwise, the diffusion process does not possess a proper steady-state distribution. ∎

We are primarily interested in the ED limiting regime, where $\hat{x} > 1$. Complications occur in the QED limiting regime, where $\hat{x} = 1$. Theorem 2.1 then does not apply directly to typical

applications because the asymptotic rate functions $\hat{\mu}$ and $\hat{\delta}$ typically are not differentiable at 1. Mandelbaum and Pats (1995) address this more complicated situation. On the other hand, the case in which $\hat{x} < 1$ is also elementary, corresponding to the heavy-traffic limit for the infinite-server $M/M/\infty$ model, as in Iglehart (1965). That previous result is a special case of Theorem 2.1.

By a variation of the same reasoning, it is possible to establish a more general deterministic fluid approximation. We then scale more crudely by dividing by s instead of by $\sqrt{s}$. Instead of (2.1), we now assume that there are fixed functions $\hat{\lambda}$, $\hat{\mu}$ and $\hat{\delta}$ such that for each positive real number x and each sequence $\{x_s : s \geq 1\}$, where sx_s is a nonnegative integer with $x_s \to x$ as $s \to \infty$,

$$\begin{aligned} \frac{\lambda_s(sx_s)}{s} &\to \hat{\lambda}(x) \ , \\ \frac{\mu_s(sx_s)}{s} &\to \hat{\mu}(x) \ , \\ \frac{\delta_s(sx_s)}{s} &\to \hat{\delta}(x) \quad \text{as} \quad s \to \infty \ . \end{aligned} \tag{2.10}$$

Moreover, we assume that

$$\hat{\lambda}(0) > 0, \quad \hat{\mu}(0) = 0 \quad \text{and} \quad \hat{\delta}(0) = 0 \tag{2.11}$$

and, if $r_s < \infty$,

$$\hat{\lambda}(1+\zeta) = 0, \quad \hat{\mu}(1+\zeta) > 0 \quad \text{and} \quad \hat{\delta}(1+\zeta) \geq 0 \ , \tag{2.12}$$

where ζ is the limit of r_s/s as $s \to \infty$. A fundamental role is played by the *asymptotic total-drift function*

$$f(x) \equiv \hat{\lambda}(x) - \hat{\mu}(x) - \hat{\delta}(x) \tag{2.13}$$

We will obtain an *ordinary differential equation* (ODE) for the limit, which is useful for describing the transient behavior. To state the limit, we introduce the scaled process

$$\bar{\mathbf{N}}_s(t) \equiv \frac{N_s(t)}{s}, \quad t \geq 0 \ . \tag{2.14}$$

Now, since we scale by s, the process N_s need not be in the neighborhood of the point $s\hat{x}$, so we might encounter boundaries. We have assumed (2.11) and (2.12) to avoid the boundaries. They imply that $f(0) > 0$ and $f(\zeta) < 0$ if $0 < \zeta < \infty$.

Theorem 2.2. (fluid limit for the state-dependent model) *Consider the sequence of $M(n)/M(n)/s/r + M(n)$ models specified above, satisfying* (2.10)*, and let $\bar{\mathbf{N}}_s(t)$ be the scaled*

number in system in (2.14). *Assume that* $r_s \geq s\zeta$ *for all* s, *where* $\zeta \leq \infty$ *and* $\hat{x} < \zeta$. *Assume that the functions* $\hat{\lambda}$, $\hat{\mu}$ *and* $\hat{\delta}$ *appearing in* (2.10) *are continuous and satisfy* (2.11) *and* (2.12). *If* $\bar{\mathbf{N}}_s(0) \Rightarrow \mathbf{n}(0)$ *as* $s \to \infty$, *where* $\mathbf{n}(0)$ *is a real number* (*deterministic*) *satisfying* $0 < \mathbf{n}(0) < \zeta$, *then*

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

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

$$m(y) = f(y) \tag{2.16}$$

for f *in* (2.11) *and infinitesimal variance* $\sigma^2(y) = 0$; *i.e.,* $\mathbf{n}$ *is the ODE*

$$\dot{\mathbf{n}}(t) = f(\mathbf{n}(t)) \tag{2.17}$$

with initial value $\mathbf{n}(0)$. *If* (2.2) *has a unique solution, then*

$$\mathbf{n}(t) \to \mathbf{n}(\infty) \equiv \hat{x} \quad as \quad t \to \infty \ . \tag{2.18}$$

Proof. We first extend the scaled process $\mathbf{N}_s$ in (2.14) to the entire real line by letting (i) $\lambda_s(n) = \lambda_s(0)$, $\mu_s(n) = \mu_s(0) = 0$ and $\delta_s(n) = \delta_s(0) = 0$ for integers n with $n < 0$ and (ii) $\lambda_s(n) = \lambda_s(s + r_s) = 0$, $\mu_s(n) = \mu_s(s + r_s) > 0$ and $\delta_s(n) = \delta_s(s + r_s) \geq 0$ for integers n with $n > s + r_s$. With this construction, the process $\mathbf{N}_s$ will never visit negative values and will never exceed $s + r_s$. However, now the process $\mathbf{N}_s$ is defined on the whole real line, so there are no boundaries, and we can apply the argument of Theorem 2.1. For the final limit in (2.18), we use the fact that $f(0) > 0$, so that $f(x) > 0$ for $x < \hat{x}$, while $f(x) < 0$ for $x > \hat{x}$. ∎

As indicated in the introduction, Theorems 2.1 and 2.2 are also consequences of Theorems 4.1 and 4.2 of Mandelbaum and Pats (1995).

3. Approximations for the Steady-State Distribution

In this paper, we are primarily interested in applying the diffusion and fluid limits in Theorems 2.1 and 2.2 to generate approximations for the steady-state behavior of the queueing system. Since the stochastic process $\{N_s(t) : t \geq 0\}$ is a birth-and-death process, much is known about its limiting steady-state behavior (as $t \to \infty$). Under regularity conditions, which hold whenever $r_s < \infty$, there will exist a unique proper limiting steady-state distribution, which is also a stationary distribution. We assume that there does indeed exist a unique steady-state distribution for N_s for each s; let the random variable $N_s(\infty)$ have that steady-state distribution.

There will be no difficulty when there exists a unique solution to the fundamental fixed-point equation (2.2). If there does, then the main practical conclusion to draw from Theorems 2.1 and 2.2 is that, under the stated conditions, in steady state the number of customers in the system tends to be concentrate about the level $\hat{x}s$ for large s, where $\hat{x}$ is a solution of equation (2.2). The fluid limit in Theorem 2.2 concludes the error is of order $o(s)$ as $s \to \infty$, while the diffusion limit in Theorem 2.1 concludes the error is of order $O(\sqrt{s})$. The diffusion limit provides a finer description.

It is important to note, however, that in general the fundamental equation (2.2) need not have a solution and, if it does, the solution need not be unique. In a large class of settings there will exist a solution on account of the following elementary result.

Theorem 3.1. (existence of a solution to (2.2) for the state-dependent model) *Suppose that asymptotic arrival-rate function $\hat{\lambda}$ is a nonincreasing continuous function, while the asymptotic total-service-rate function $\hat{\mu}$ and the asymptotic total-abandonment-rate function $\hat{\delta}$ are both nondecreasing continuous functions, for all s sufficiently large. Suppose that*

$$\hat{\lambda}(0) > \hat{\mu}(0) + \hat{\delta}(0) \tag{3.1}$$

and $\hat{\delta}(x) \to \infty$ as $x \to \infty$. Then there exists at least one solution to equation (2.2).

Moreover, in a large class of settings the solution will be unique.

Theorem 3.2. (uniqueness of a solution to (2.2) for the state-dependent model) *In addition to the conditions of Theorem* 3.1*, suppose that $\hat{\mu} + \hat{\delta}$ is strictly increasing. Then there exists a unique solution to equation* (2.2).

For the standard application in Section 4, $\hat{\mu} + \hat{\delta}$ is strictly increasing because $\hat{\mu}$ is strictly increasing on the interval $[0, 1]$, but constant on the interval $(1, \infty)$, while $\hat{\delta}$ is strictly increasing on the interval $(1, \infty)$, but constant on the interval $[0, 1]$; i.e.,

$$\hat{\mu}(x) = x \wedge 1, \quad x \geq 0, \tag{3.2}$$

and

$$\hat{\delta}(x) = \eta(x - 1), \quad x \geq 1 \ , \tag{3.3}$$

where $\eta(x) = 0$ for $x \leq 0$, η is strictly increasing on $(0, \infty)$ and $\eta(x) \to \infty$ as $x \to \infty$.

We also want to draw attention to the possibility that the fundamental equation (2.2) could have more than one solution. The system would then have multiple stable points,

asymptotically as $s \to \infty$. Theorems 2.1 and 2.2 would then apply to all such solutions describing transient behavior that depends strongly upon the initial conditions. Since the stochastic process of interest $\{N_s(t) : t \geq 0\}$ is a birth-and-death process on a finite state space (assuming $r_s < \infty$) for each s, the process has a unique limiting steady-state distribution for each s. Nevertheless, as s increases, the process could tend to exhibit multi-stable behavior; i.e., the steady-state distribution for large s would tend to be approximately a mixture of point masses attached to the different equilibrium points. The stochastic process N_s would tend to remain a long time near one equilibrium point and then eventually move to another equilibrium point and spend a long time there.

Such anomalous behavior could arise if the natural monotonicity assumptions in Theorems 3.1 and 3.2 are violated. For example, the arrival rate could increase as the queue length increases if customers were somehow attracted to the queue. For example, customers in a store might think that there must be something worth waiting for if they see a line, and have a greater propensity to join the queue the longer it is.

Even more likely is the possibility that the total service rate might decrease when the congestion increases, perhaps because service efficiency declines due to fatigue caused by the higher workload. That phenomenon in call centers was noted by Sze (1984). However, it is not our purpose to explore multi-stability phenomena here.

Assuming that there exists unique solution to (2.2), and assuming that the conditions of Theorem 2.1 hold with $\gamma > 0$, we obtain the natural approximation for the steady-state number of customers in an $M(n)/M(n)/s/r + M(n)$ system by letting the actual system be term s in such a limit; i.e.,

$$N_s(\infty) \approx s\hat{x} + \sqrt{s}\mathbf{N}(\infty) \stackrel{\mathrm{d}}{=} Nor(s\hat{x} + \sqrt{s}(\tilde{\gamma}/\gamma), s\hat{\lambda}(\hat{x})/\gamma) \ , \tag{3.4}$$

where $\stackrel{\mathrm{d}}{=}$ means equal in distribution. That in turn implies that, for large s and t, $N_s(t)$ will tend to be of order $O(\sqrt{s})$ away from the level $s\hat{x}$. Of course, Theorem 2.1 also directly yields approximations for the transient behavior too.

Assuming that there exists unique solution to (2.2), and assuming that the conditions of Theorem 2.2 hold, we obtain the cruder approximation

$$N_s(\infty) \approx s\hat{x} \ . \tag{3.5}$$

This is the intended simple fluid approximation.

4. The $M/M/s/r + M(n)$ Approximation for $M/GI/s/r + GI$

As indicated in the introduction, we were motivated to consider the $M(n)/M(n)/s/r+M(n)$ 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 (2005a). Unlike the approximations developed in this paper above, that approximation is not based on any limit theorem.

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 (2005a), 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 the special $M/M/s/r + M(n)$ case, we choose the approximating exponential service-time distribution by simply matching the mean of the given service-time distribution. We choose the total-abandonment-rate function δ_s to approximate the behavior in the $M/GI/s/r+GI$ model with IID abandon times having abandon-time cdf F. We assume that the cdf F is absolutely continuous with a probability density function (pdf) f; i.e., we assume that $F(x) = \int_0^x f(y)\,dy$ for all $x > 0$. We then work with the hazard-rate (or failure-rate) function

$$h(x) \equiv \frac{f(x)}{1 - F(x)}, \quad x \geq 0 \ . \tag{4.1}$$

We think of the pdf f as being continuous and positive on the entire nonnegative real line, but that is not required.

The key approximation in Whitt (2005a) is an approximation for the abandonment rate of a customer who is j^{th} from the end (back end) of a queue (necessarily of length at least j):

$$\alpha_s(j) \approx h(j/\lambda_s) \ . \tag{4.2}$$

(Here, of course, the arrival rate is constant.)

We get approximation (4.2) 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, given the state of the $M/M/s/r + M(n)$ model at any time, we do not know how long customers have waited, so we estimate it. Thus, as an approximation, we estimate that there have been j arrivals since the time a customer who is j^{th} from the end of the queue arrived. (In this step, we are ignoring abandonments, which tend to occur at a lower rate than arrivals.) Given that customers arrive at rate λ_s, the expected time between successive arrivals is $1/\lambda_s$. Combining these two approximations, 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 (4.2). The approximation may seem terribly crude, but numerical comparisons indicate that it is remarkably accurate.

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

$$\delta_s(k) \approx \sum_{j=1}^{k-s} \alpha_s(j) \quad \text{for} \quad k > s \ , \tag{4.3}$$

with $\delta_s(k) = 0$ if $k \le s$, because in this application we are assuming customers only abandon before beginning service. As indicated in Whitt (2005a), if the density f were not smooth, then we might instead let

$$\alpha_s(j) \approx \lambda_s \int_{(j-1)/\lambda_s}^{j/\lambda_s} h(t)\, dt \ . \tag{4.4}$$

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

$$\delta_s(k) \approx \lambda_s \int_0^{(k-s)/\lambda_s} h(t)\, dt = -\lambda_s \log_e F^c((k-s)/\lambda_s) \quad \text{for} \quad k > s \ , \tag{4.5}$$

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

The special $M/M/s/r + M(n)$ case considered here starts with (4.5). In addition, we assume that $\lambda_s(k) = s\rho$ for all k and $\mu_s(k) = k \wedge s = \min\{k, s\}$ for $k \ge 0$. (That implies we are assuming that the mean service time is 1.) Since here we are evaluating the approximation in the ED limiting regime, we assume that $\rho > 1$. As a consequence, in the special case the assumptions in both (2.1) and (2.10) are satisfied, with $\tilde{\lambda}(x) = \tilde{\mu}(x) = \tilde{\delta}(x) = 0$ forall x. Indeed, $\delta_s(xs) = s\hat{\delta}(x)$ for all $x > 0$ and $s > 0$, where

$$\hat{\delta}(x) = -\rho \log_e F^c((x-1)/\rho) \quad \text{for all} \quad x > 1 \ , \tag{4.6}$$

and $\hat{\delta}(x) = 0$ for $0 \le x \le 1$, with $F^c(x) \equiv 1 - F(x)$ being the complementary cdf (ccdf).

As a consequence, the fundamental fixed-point equation (2.2) becomes

$$F^c((\hat{x}-1)/\rho) = e^{-(\rho-1)/\rho} \ , \tag{4.7}$$

where $\rho > 1$. In this setting, there clearly exists a unique solution $\hat{x} > 1$ to the fundamental fixed-point equation, because the right side is a number strictly between 0 and 1, while the left side is a continuous function on the interval $(1, \infty)$ decreasing from 1 at $x = 1$ toward 0 as $x \to \infty$.

The drift rate in the limiting OU diffusion process obtained from Theorem 2.1 is

$$\gamma = \hat{\delta}'(\hat{x}) = h((\hat{x}-1)/\rho), \tag{4.8}$$

where h is the hazard-rate function in (4.1) and $\hat{x}$ is the unique solution to the fundamental fixed-point equation (2.2) in this context, i.e., to (4.8).

Remark 4.1. *The $M/M/s/r + M$ Special Case.* We now show that Theorem 2.1 here is consistent with Theorem 2.1 in Whitt (2005a) for the $M/M/s/r + M$ special case. 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 (2.2) thus becomes (4.8) with $F^c(x) = e^{-\alpha x}$, which implies that

$$\hat{x} - 1 = \frac{\rho - 1}{\alpha} \ , \tag{4.9}$$

just as in Theorem 2.1 of Whitt (2005a). Since $h(t) = \alpha$ for all t, the state-dependent drift is

$$\gamma = \hat{\delta}'((\hat{x} - 1)/\rho) = \alpha \ , \tag{4.10}$$

again just as in Theorem 2.1 of Whitt (2005a). ∎

5. The Fluid Approximation for $G/GI/s/r + GI$

In this section we describe the equilibrium behavior of the fluid approximation for the general $G/GI/s/r + GI$ model; for the full time-dependent behavior, see Whitt (2005c). The fluid approximation directly approximates the scaled process $N_s(t)/s$ and related quantities; we obtain the desired approximation for $N_s(t)$ by undoing the scaling.

As before, for the initial queueing model, we assume that the individual service rate is 1 and that the arrival rate is $s\rho$ for $\rho > 1$, which puts us in the ED limiting regime. The key model elements are the service-time cdf G and the abandon-time cdf F. Let G^c and F^c be the associated ccdf's. We assume that the arrival process is a general stationary point process with a well defined rate, with that rate being $s\rho$, where $\rho > 1$.

We scale by dividing the number in system for each s by s and letting $s \to \infty$. Our final approximation for the steady-state number of customers in the system is obtained by unscaling, i.e.,

$$N_s(\infty) \approx s(1 + q^F), \tag{5.1}$$

where q^F is the queue content (amount of fluid waiting before starting service), which is given in (5.4) below.

The fluid approximation for the equilibrium behavior in the ED limiting regime (without undoing the scaling) is depicted in Figure 1. It is significant that the fluid approximation depends on the two cdf's G and F (or ccdf's G^c and F^c), but not on the stochastic structure

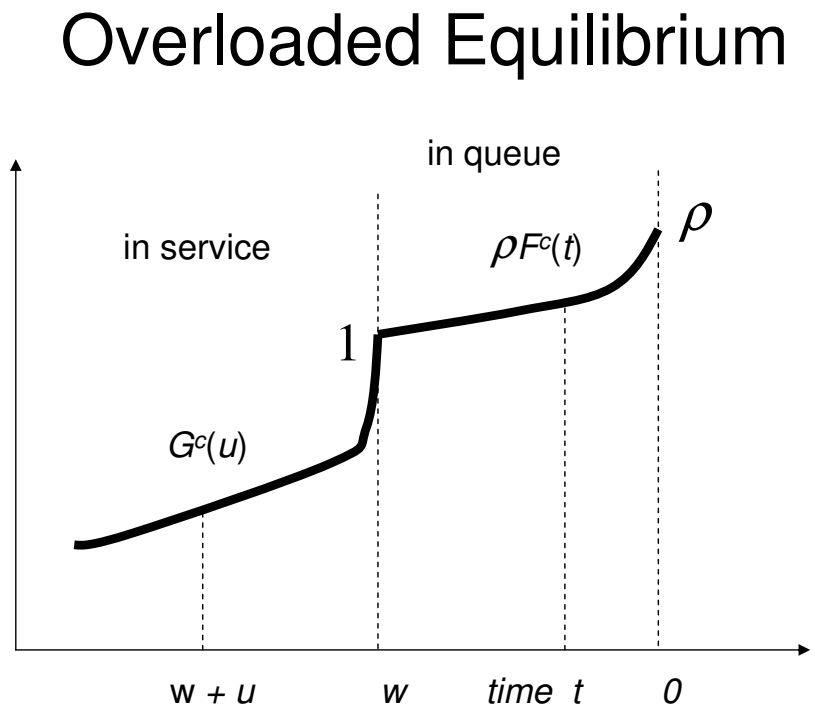

Figure 1: The steady-state distribution of fluid content in the $G/GI/s + GI$ fluid model with individual service rate 1, traffic intensity $\rho > 1$, service-time distribution G and abandon-time distribution F. The figure plots the density of fluid content that has been in the system for time t. Time increases to the left.

of the arrival process (beyond its rate). In Figure 1, time appears on the horizontal axis, increasing toward the left, while queue content (scaled number of customers) appears on the vertical axis. Specifically, the value at time t is the density of the fluid that has been in the system for exactly length t, i.e., the remaining portion of the fluid that arrived t time units in the past. Fluid arrives at rate ρ and a proportion $F(t)$ of that fluid abandons by time t. Fluid that does not abandon waits in queue until time w, after which it is in service. Entering fluid exits before time w by abandonment, and after time w by service completion. In particular, the general fluid approximation has, first, all customers who do not abandon waiting exactly time w and, second, a proportion $F(t)$ of arrivals abandoning before time t after arrival, for $0 < t < w$. Moreover, in equilibrium for the fluid approximation, all servers are busy and fluid abandons at an overall rate $\rho - 1$.

The density of fluid content that has been waiting in queue for a length t is

$$q(t) = \rho F^c(t), \quad 0 \le t \le w, \quad \text{and} \quad q(t) = 0, \quad t > w, \tag{5.2}$$

where the constant waiting time before starting service, w, is the solution to the equation

$$F(w) = \frac{\rho - 1}{\rho} \ . \tag{5.3}$$

The total fluid content waiting in queue is

$$q^F = \int_0^w q(t)\, dt = \rho \int_0^w F^c(t)\, dt \ . \tag{5.4}$$

Since the servers are all busy in the fluid model, we can apply (5.4) to obtain the desired fluid approximation in (5.1).

Remark 5.1. *The $M/M/s/r+M$ Special Case.* We now observe that the fluid approximation in (5.1), (5.3) and (5.4) here is consistent with both Theorem 2.1 in Whitt (2005a) and Theorem 2.1 here for the $M/M/s/r + M$ special case, continuing Remark 4.1. If the abandon-time cdf F is exponential with mean $1/\alpha$, then $F^c(x) = e^{-\alpha x}$ and equation (5.3) becomes $1 - e^{-\alpha w} = (\rho - 1)/\rho$. Then equation (5.4) becomes

$$q^F = \rho \int_0^w e^{-\alpha t}\, dt = \rho \frac{(1 - e^{-\alpha w})}{\alpha} = \frac{\rho - 1}{\alpha} \ , \tag{5.5}$$

just as in (4.10).■

However, more generally, we see that q^F in (5.4) need not coincide with $\hat{x} - 1$ obtained as the solution to (4.8). In support of the $M/M/s/r + M(n)$ approximation for the $M/GI/s/r + GI$

model in Whitt (2005a), though, we see that the service-time cdf G beyond its mean has played no role in the fluid approximation for the $G/GI/s/r+GI$ model. The service-time cdf G only plays a role in describing how long fluid in service has been in service. Let $b(t)$ denote the density of fluid that has been in service for a length of time t. Equilibrium for the fluid approximation has $b(t) = G^c(t)$, $t \geq 0$.

6. A Comparison of the Two Fluid Approximations

In this final section we do further analysis to compare (i) the fluid approximation for the $M/M/s/r + M(n)$ approximation to the $M/GI/s/r + GI$ model and (ii) the direct fluid approximation for the $M/GI/s/r+GI$ model. To have similar notation, let q^M (M for Markov) denote the fluid approximation for the scaled queue content (waiting in queue before starting service) in the $M/M/s/r+M(n)$ approximation to the $M/GI/s/r+GI$ model; i.e., $q^M = \hat{x}-1$ for $\hat{x}$ the solution to (4.8). Our goal now is to compare q^M to q^F in (5.4) above. Let other quantities associated with the two models be designated by superscripts M and F.

We have already observed that in general q^M does not coincide with q^F. In any contemplated scenario, we can calculate q^M and q^F to judge how close the $M/M/s/r + M(n)$ approximation is likely to be. To establish more general connections, we first change notation, writing $\epsilon \equiv \rho - 1$, so that we can focus on the comparison for ρ close to 1, which corresponds to 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). Up to now, it has not mattered which customers abandon in the Markovian models. With that assumption, the waiting time of all customers, served or not, is the same, and by Little's law ($L = \lambda W$) must be $w^M = q^M/\rho$. Thus Combining this with (4.8), we obtain the equation

$$F(w^M) = 1 - e^{-(\rho-1)/\rho} = 1 - e^{-\epsilon/(1+\epsilon)} \ . \tag{6.1}$$

Equation (6.1) is convenient, because it is easy to compare to equation (5.3), which with the change of notation becomes

$$F(w^F) = \frac{\epsilon}{1+\epsilon} \ . \tag{6.2}$$

First, from equations (6.1) and (6.2), we easily see that in all cases $w^F \neq w^M$, even in the $M/M/s/r+M$ model, where $q^F = q^M$, as shown in Remark 5.1. (That is not surprising, since we are treating the abandonments differently.) However, if we expand the exponential in (6.1),

then we obtain

$$1 - e^{-\epsilon/(1+\epsilon)} = \frac{\epsilon}{1+\epsilon} - \frac{\epsilon^2}{2(1+\epsilon)^2} + \frac{\epsilon^3}{6(1+\epsilon)^3} + O(\epsilon^4) \quad \text{as} \quad \epsilon \downarrow 0 \ . \tag{6.3}$$

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 $w^F = g(1/(1+\epsilon))$ and $w^M = g(1-e^{-\epsilon/(1+\epsilon)})$. Using a Taylor series expansion, we get

$$w^M \approx w^F - g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2} \ . \tag{6.4}$$

From formulas (6.1)–(6.2), we also have the inequalities $w^F \le q^F \le w^F(1+\epsilon)$, while

$$w^F(1+\epsilon) - g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)} \approx q^M = w^M(1+\epsilon) \le w^F(1+\epsilon) \ . \tag{6.5}$$

We then have the bounds

$$\begin{aligned} |q^F - w^F(1+\epsilon)| &\le \epsilon w^F \\ |q^M - w^F(1+\epsilon)| &\le g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2} \\ |q^M - q^F| &\le \max\{\epsilon w^F, g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2}\} \ . \end{aligned} \tag{6.6}$$

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 (5.3) and (5.4), we see that in this case $q^F = \epsilon - \frac{\epsilon^2}{2(1+\epsilon)}$, while, from (4.8),

$$q^M = (1+\epsilon)(1-e^{-\epsilon/(1+\epsilon)}) = \epsilon - \frac{\epsilon^2}{2(1+\epsilon)} + \frac{\epsilon^3}{6(1+\epsilon)^2} + O(\epsilon^4) \ , \tag{6.7}$$

so that

$$q^M - q^F = \frac{\epsilon^3}{6(1+\epsilon)^2} + O(\epsilon^4) \ . \tag{6.8}$$

For example, if $\epsilon = 0.1$ ($\rho = 1.1$), then $q^F = 0.09545$, while $q^M = 0.09559$ and $q^M - q^F \approx 0.0001377$. There is a difference of only about 0.1%. That is much closer than predicted by the bounds in (6.6), because $w^F(1+\epsilon) = \epsilon = 0.1$, $\epsilon w^F = \epsilon^2/(1+\epsilon) = 0.0091$, $g'(x) = 1$, $0 < x < 1$, and $g'(\epsilon/(1+\epsilon))\frac{\epsilon^2}{2(1+\epsilon)^2} = \frac{\epsilon^2}{2(1+\epsilon)^2} = \frac{0.01}{2.42} = 0.0041$. ∎

References

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

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, Review and Research Prospects. *Manufacturing and Service Opns. Mgmt.* 5, 79–141.

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

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

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

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. 2004. The impact of customers patience on delay and abandonment: some empirically-driven experiments with the $M/M/n + G$ queue. *OR Spectrum* 26, 377–411.

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

Sze, D. Y. 1984. A queueing model for telephone operator staffing. *Operations research* 32, 229–249.

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

Whitt, W. 2005a. Engineering solution of a basic call-center model. *Management Science*, to appear. Available at http://columbia.edu/∼ww2040.

Whitt, W. 2005b. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, to appear. Available at http://columbia.edu/∼ww2040.

Whitt, W. 2005c. Fluid models for multiserver queues with abandonments. *Operations research*, to appear. Available at http://columbia.edu/∼ww2040.