FLUID MODELS FOR MULTI-SERVER QUEUES WITH ABANDONMENTS

by

Ward Whitt

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

February 6, 2004; Revision October 28, 2004

Abstract

Deterministic fluid models are developed to provide simple first-order performance descriptions for multi-server queues with abandonment under heavy loads. Motivated by telephone call centers, the focus is on multi-server queues with a large number of servers and non-exponential service-time and time-to-abandon distributions. The first fluid model serves as an approximation for the $G/GI/s+GI$ queueing model, which has a general stationary arrival process with arrival rate λ, independent and identically distributed (IID) service times with a general distribution, s servers and IID abandon times with a general distribution. The fluid model is useful in the overloaded regime, where $\lambda > s$, which is often realistic because only a small amount of abandonment can keep the system stable. Numerical experiments, using simulation for $M/GI/s+GI$ models and exact numerical algorithms for $M/M/s+M$ models, show that the fluid model provides useful approximations for steady-state performance measures when the system is heavily loaded. The fluid model accurately shows that steady-state performance depends strongly upon the time-to-abandon distribution beyond its mean, but not upon the service-time distribution beyond its mean.

The second fluid model is a discrete-time fluid model, which serves as an approximation for the $G_t(n)/GI/s+GI$ queueing model, having a state-dependent and time-dependent arrival process. The discrete-time framework is exploited to prove that properly scaled queueing processes in the queueing model converge to fluid functions as $s \to \infty$. The discrete-time framework is also convenient for calculating the time-dependent fluid performance descriptions.

Subject classifications: Queues, approximations: multi-server queues with abandonment. Queues, multichannel: approximation of non-Markovian multichannel queues with customer abandonment.

Area of Review: Stochastic Models.

Keywords: queues, multi-server queues, queues with customer abandonment, multi-server queues with customer abandonment, call centers, contact centers, deterministic fluid models, fluid limits, law of large numbers.

1. Introduction

Motivated by applications to telephone call centers and more general customer contact centers (with contact also made by other means, such as fax and email), there recently has been great interest in multi-server queues with a large number of servers and customer abandonment, e.g., see Gans et al. (2002), Garnett et al. (2002) and Mandelbaum and Zeltyn (2004). We believe that anyone interested in the behavior of these models should know about Figure 1 below. As we will explain in Section 3, Figure 1 depicts the steady-state behavior of a natural deterministic fluid model of a multi-server queue with abandonment under a heavy load, in particular, the general $G/GI/s + GI$ model. Figure 1 tells a remarkably simple story about performance, which is remarkably accurate under heavy loads (where the arrival rate exceeds the maximum possible service rate). In Figure 1, a major role is played by the general cumulative distribution functions (cdf's) of a service time, G, and of an abandon time, F. Even though the model being described is a deterministic fluid model, these stochastic model elements appear prominently. We start by explaining why systems with such heavy loads might be worth considering.

Most call centers can be classified into two types: (i) revenue-generating, and (ii) service-oriented. The revenue-generating call centers typically perform sales functions. For example, they may take customer orders, and have the opportunity to sell customers more goods. In contrast, the service-oriented call centers typically provide customer service, and generate only minimal revenue. For example, they may provide technical support. In both cases, the call centers are managed to meet service level requirements, but, naturally, there tends to be somewhat lower standards in the service-oriented call centers (with the exception of emergency-response centers, such as those answering 911 calls). Indeed, it is now common to outsource service-oriented call centers. Then the outsourcing contract contains a *service-level agreement* (SLA). The typical SLA contains performance targets such as (i) having less than 5% customer abandonment, and (ii) meeting a specified *service level*, i.e., answering $x\%$ of all calls that are eventually served within y seconds; common numbers are $x = 80\%$ and $y = 30$ seconds. With service-oriented call centers, and especially with outsourcing, there are penalties if the SLA is not met, but there tends to be little incentive to provide a much higher quality of service than stipulated in the SLA. In contrast, in revenue-generating call centers it may be worthwhile to aim to answer almost all calls immediately upon arrival, as discussed in Whitt (1999a).

With service-oriented call centers, it is often possible for a call-center operator to meet the

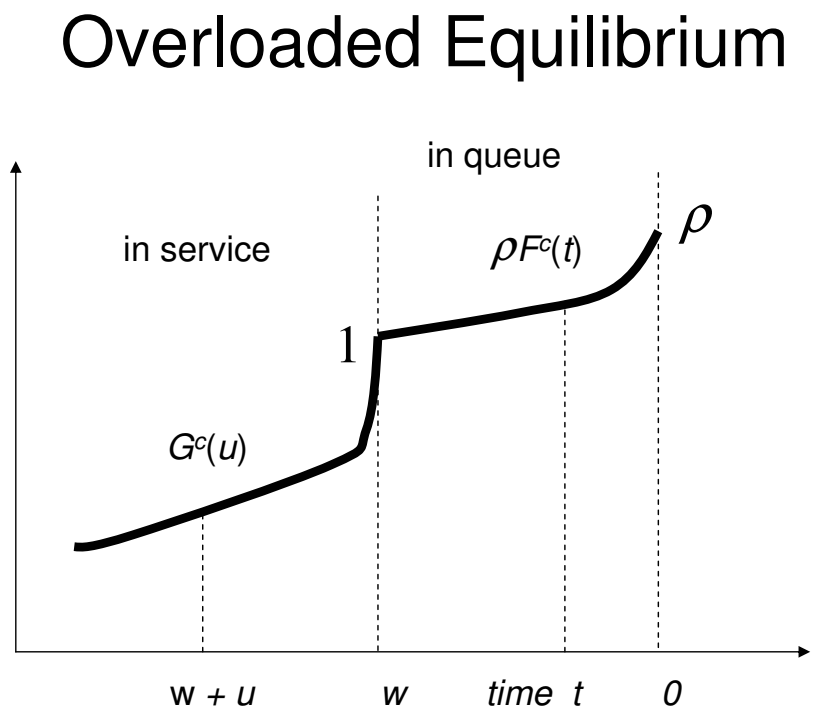

Figure 1: The steady-state distribution of fluid content in the $G/GI/s + GI$ fluid model with mean service time 1, arrival rate $\rho > 1$, service-time distribution G and abandon-time distribution F. Time increases to the left. The value at time t is the density of the fluid that has been in the system for length t, i.e., the remaining portion of the fluid that arrived t time units in the past. Fluid that does not abandon waits in queue until time w, after which it enters service. Entering fluid exits before time w by abandonment, and after time w by service completion. The corresponding queueing approximations are obtained by multiplying the fluid content by s.

SLA in a very efficient way. Given that the service representatives (or agents) do not need to be located near the callers, the call centers can be very large and handle calls from wide areas. *Given that the call centers are very large, the SLA often can be met in the overloaded regime.* The overloaded regime is very efficient, because the service representatives tend to be constantly busy. If that condition is considered too demanding, then the call center can be operated with each agent busy 95% of the time (or any other target percentage). Then the total number of agents is increased to compensate for the planned agent idle time. In the overloaded regime, all working agents can be busy effectively all the time.

What may not be so apparent is that it may be possible to meet the SLA in the overloaded regime. A key to accomplishing that goal is *customer abandonment*. Even a small level of customer abandonment can compensate for a slight excess in the arrival rate over the maximum possible service rate. To illustrate, we give an example. Our example is a standard multi-server queue with abandonments. However, we recognize that most call centers serve multiple classes of calls, using multiple classes of agents with different skills. Nevertheless, in our example here, and throughout this paper, we only consider the basic call-center model with a single class of calls handled by a single group of agents. Understanding the basic call-center model is a first step to understanding the more complicated multi-class models.

So, here is our example: Suppose that there are $s = 100$ agents (servers) each working at rate $\mu = 1$ (with time measured in units of mean call holding times). Then we might have an arrival rate of $\lambda = 102$. Necessarily, with those parameters, there must be at least $(2/102) \times 100 = 1.96\%$ abandonment, because the long-run rate in must equal the long-run rate out, by service or abandonment. Surprisingly perhaps, it turns out that the actual performance in that scenario may not be that bad. The actual level of abandonment will of course be higher than 1.96% because of stochastic fluctuations, but it need not be too high. It might be about 5%, while 80% of all served calls are answered within 30 seconds. Indeed, that is exactly what is predicted by the classical Erlang A model, the purely Markovian $M/M/s + M$ model, with Poisson arrival process, exponential service times, s servers and exponential abandon times (the final $+M$), when the mean service time and mean abandon time are both 5 minutes. As shown in Table 1 of Whitt (2005a), for that model with those parameters, exactly (within a small difference) 80% of served calls are answered within 30 seconds and 5% of the calls abandon.

The example just considered is only slightly overloaded, because the traffic intensity is only $\rho \equiv \lambda/s\mu = 1.02$. But the traffic intensity in call centers can easily be higher. When oper-

ating call centers are not managed well, they can become more heavily loaded. For example, absenteeism may make the traffic intensity much higher. Thus we want to understand the behavior of multi-server queues with abandonments in the overloaded regime. That led us in Whitt (2005b) to develop and evaluate fluid and diffusion approximations for the purely Markovian $M/M/s/r+M$ model based on many-server heavy-traffic stochastic-process limits in the *efficiency-driven* ED *limiting regime* (using terminology from Garnett et al. (2002)). (These are special cases of limits established by Mandelbaum and Pats (1995.) The ED limiting regime is characterized by having $s \to \infty$ and $\lambda \to \infty$, while holding ρ fixed with $\rho > 1$. As we should expect, the performance of the fluid approximation tends to improve as s and ρ increase. Whitt (2005b) shows that the fluid approximation for the $M/M/s/r+M$ model is a very crude approximation when $s = 100$ and $\rho = 1.02$, but that it is a remarkably good approximation when $s = 100$ and $\rho = 1.10$.

Our goal is to extend Whitt (2005b) by establishing corresponding results for the more general $G/GI/s+GI$ queueing model, which has a general stationary arrival process (the first G), independent and identically distributed (IID) service times with a general distribution (the first GI), s homogeneous servers working in parallel, unlimited waiting space and IID times for waiting customers to abandon if they have not yet begun service, again with a general distribution (the final $+GI$). We are especially interested in the impact upon performance caused by non-exponential service-time and abandon-time distributions, because statistical analysis of telephone-holding-time data and abandon-time data has shown that the probability distributions of both the service times and abandon times often are not nearly exponential; see Bolotin (1994) and Brown et al. (2002).

Since the general $G/GI/s+GI$ model is much more challenging than the Markovian $M/M/s/r+M$ model, it should come as no surprise that we are only partially successful in our attempt to extend Whitt (2005b). We do develop the desired deterministic fluid approximation for the $G/GI/s+GI$ model, but we have not yet justified it by establishing the supporting many-server heavy-traffic limit (corresponding to a functional law of large numbers). And we do not even propose a refined stochastic approximation, paralleling the diffusion approximation for the $M/M/s/r+M$ model, let alone establish the refined many-server heavy-traffic limit, justifying that refinement.

However, we do develop the deterministic fluid approximation, and we show that it provides useful insights, many of which are captured by Figure 1, as we will explain. Moreover, we conjecture that the fluid approximation can be justified by a many-server heavy-traffic

limit theorem, under the same assumptions as for Theorem 2.2 in Whitt (2005b). In fact, we also provide strong theoretical support for the deterministic fluid approximation by establishing such a many-server heavy-traffic limit for a related discrete-time model. Since the discrete-time model can be made arbitrarily close to the continuous-time model, we regard our limit theorem as providing the desired theoretical justification, from an applied engineering perspective. Nevertheless, it would be nice to prove the continuous-time theorem.

The work here is closely related to previous work on infinite-server queues, especially heavy-traffic limits for these models; see Duffield and Whitt (1997), Glynn and Whitt (1991) and Krichagina and Puhalskii (1997). In relation to these papers, our main contribution here is to focus on customer abandonments. These papers illustrate stronger mathematical results we hope to obtain for our model.

Organization of the rest of this paper. Here is how the rest of this paper is organized: In Section 2 we develop the deterministic fluid approximation for the $G/GI/s + GI$ queueing model with large s. In Section 3 we determine its steady-state behavior. We prove that the fluid model has a unique steady-state distribution, and fully characterize that steady-state distribution (depicted in Figure 1 for the interesting overloaded case).

In Sections 4 and 5 we present numerical examples to show that the $G/GI/s + GI$ fluid model provides useful results. We compare the fluid approximation to exact results for $M/GI/s+GI$ queueing models obtained from computer simulations and numerical algorithms. In Section 4 we consider overloaded $M/GI/s + GI$ queueing models with non-exponential service-time and time-to-abandon distributions, investigating the impact of these distributions. We show that steady-state performance is significantly affected by the time-to-abandon distribution beyond its mean but not by the service-time distribution beyond its mean. In Section 5 we consider more examples, showing how the approximation performs as a function of the load.

In Section 6 we establish the many-server heavy-traffic limit, with convergence to a deterministic fluid process, for a family of discrete-time $G_t(n)/GI/s + GI$ queueing models. These models are more general than the $G/GI/s + GI$ queueing models in Section 2, because they have arrival processes that are both time-dependent and state-dependent. From the theoretical perspective, a major innovation in our analysis is to approach the convergence problem in discrete time. By working in discrete time, the proof becomes a relatively elementary recursive application of the weak law of large numbers (WLLN). The discrete-time framework

is also convenient for computation of the approximating fluid performance measures. Just as the proof can be done recursively, so can the discrete-time computations, making it easy to compute time-dependent generalizations of Figure 1.

In Section 7 we briefly discuss steady-state solutions for $G(n)/GI/s + GI$ fluid models, which have state-dependent, but not-time-dependent, input. When the arrival rate is state-dependent, we can easily have multiple steady-state regimes for the fluid model.

2. The G/GI/s+GI Fluid Model

In this section we develop the $G/GI/s + GI$ fluid model. We start with the associated $G/GI/s + GI$ queueing model.

The $G/GI/s+GI$ queueing model. The $G/GI/s+GI$ queueing model is a general multi-server queue with customer abandonment. Customers arrive according to a general stationary arrival process (the initial G) with arrival rate λ. Each arriving customer enters service immediately upon arrival if there is a server available. If the servers are all busy, the arriving customer waits in queue. Customers are served in order of their arrival by the first available server. Waiting customers may also elect to abandon. We assume that each customer has a random abandon time, after which he will abandon if he has not yet begun service. Once service starts, the customer remains until service is provided. There are no retrials; abandoning customers leave without affecting future arrivals.

The two $GI's$ in the notation mean that the service times and abandon times come from two independent sequences of independent and identically distributed (IID) random variables, which are independent of the arrival process. In many applications, such as telephone call centers, customers cannot see the queue (the case of invisible queues), and thus do not know the experience of other customers, so that it is natural to assume that abandon times are IID.

Given that service times and abandon times come from independent IID sequences, their stochastic behavior is determined by the distribution of single times. Let S and T denote a generic service time and abandon time, respectively. Let G and F be the service-time and abandon-time cdf's, which we assume have probability density functions (pdf's) g and f, respectively, i.e,,

$$G(x) \equiv P(S \le x) = \int_0^x g(u)\,du \quad \text{and} \quad F(x) \equiv P(T \le x) = \int_0^x f(u)\,du \quad \text{for} \quad x \ge 0\ . \quad (2.1)$$

We assume that the pdf's g and f are strictly positive on the positive halfline. Let G^c and F^c denote the associated complementary cdf's (ccdf's), defined by $G^c(x) \equiv 1 - G(x)$ and

$F^c(x) \equiv 1 - F(x)$. We assume that the mean service time is 1, so we have

$$ES = \int_0^\infty xg(x)\,dx = \int_0^\infty G^c(x)\,dx = 1\ . \tag{2.2}$$

That choice is without loss of generality, because we are free to choose the measuring units for time. We elect to measure time in units of mean service times.

A sequence of queueing models. We now start to develop the fluid approximation. Even though we do not establish a stochastic-process limit in this section, it is helpful to formulate the conjectured limit, in order to better understand how the fluid approximation is related to the queueing model. To do so, we consider a sequence of queueing systems indexed by s, where $s \to \infty$. (For background, it may be helpful to review Theorem 2.2 of Whitt (2005a), which establishes the corresponding (more elementary) result for the $M/M/s/r+M$ queueing model.)

Let the associated family of arrival processes be defined by simple scaling, i.e.,

$$A_s(t) = A(\lambda_s t), \quad t \geq 0\ , \tag{2.3}$$

for a fixed initial rate-1 arrival process $A \equiv \{A(t) : t \geq 0\}$, where λ_s is the arrival rate is model s. To justify the stochastic-process limit, we assume that A satisfies a functional weak law of large numbers (FWLLN), i.e.,

$$\left\{\frac{A(nt)}{n} : t \geq 0\right\} \Rightarrow \{t : t \geq 0\} \quad \text{as} \quad n \to \infty \quad \text{in} \quad D\ , \tag{2.4}$$

where $D \equiv D([0,\infty), \mathbb{R})$ is the usual function space endowed with one of the Skorohod topologies and $\Rightarrow$ denotes convergence in distribution; e.g., see Whitt (2002).

We now introduce stochastic processes to describe the system content. Let $B_s'(t,y)$ be the number of customers in service at time t that have been in service for time less than or equal to y, and let $Q_s'(t,y)$ be the number of customers in queue at time t that have been in queue for time less than or equal to y, for $y > 0$. Then form the scaled processes

$$B_s(t,y) \equiv \frac{B_s'(t,y)}{s} \quad \text{and} \quad Q_s(t,y) \equiv \frac{Q_s'(t,y)}{s} \tag{2.5}$$

for $t \geq 0$ and $y \geq 0$. In (2.5) we are scaling the vertical spatial dimension, but not the horizontal time dimension (thinking of Figure 1). This scaling is the same as for fluid approximations in infinite-server queues; e.g., see Duffield and Whitt (1997), Glynn and Whitt (1991), Krichagina and Puhalskii (1997) and Chapter 10 of Whitt (2002).

We now want to consider the many-server heavy-traffic limiting regime in which $s \to \infty$ and $\lambda_s \to \infty$. Since the mean service time is 1, the traffic intensity in model s is $\rho_s = \lambda_s/s$. We thus let $s \to \infty$, assuming that

$$\lambda_s = s\rho \quad \text{and} \quad \rho_s = \rho \quad \text{for all} \quad s \ , \tag{2.6}$$

for some initial fixed ρ and $s \to \infty$.

Formulation of the stochastic-process fluid limit requires some care. One approach is to exploit the function space $D_2 \equiv D([0,\infty) \times [0,\infty), \mathbb{R})$ with a two-dimensional parameter space and an appropriate generalization of one of the standard Skorohod topologies, as in Neuhaus (1971) and Straf (1971). Even when the limit is a continuous function, some care is needed because the sample paths of the converging processes are not continuous. Convergence of a sequence in D_2 with this topology reduces to uniform convergence over compact sets (u.o.c.) when the limit is a continuous function; the extra complexity is to achieve proper measurability. For more on stochastic processes with two-dimensional parameters, see Sections 1.10-1.15 of Csörgő and Révész (1981) and Krichagina and Puhalskii (1997). Convergence in D_2 to a continuous limit implies pointwise convergence for all argument pairs (t, y).

A proper formulation of the stochastic-process limit requires a careful specification of the initial conditions. Because of the generality of the $G/GI/s + GI$ queueing model, the initial conditions can lead to complications. It should suffice to assume appropriate convergence of the initial conditions. A special case is starting out empty. We will simply state our conjecture for that special case, with the understanding that the result should extend, with appropriate qualifications. We will also conjecture a stochastic refinement.

Conjecture 2.1. (stochastic-process fluid limit and stochastic refinement) *Consider a sequence of initially empty $G/GI/s + GI$ models satisfying the assumptions above, including* (2.1)–(2.6). *Then*

$$\begin{aligned} &\{(B_s(t,y), Q_s(t,y)) : t \geq 0, y \geq 0\} \Rightarrow \{(B(t,y), Q(t,y)) : t \geq 0, y \geq 0\} \\ &\{\sqrt{s}\,[(B_s(t,y), Q_s(t,y)) - (B(t,y), Q(t,y))] : t \geq 0, y \geq 0\} \\ &\qquad \Rightarrow \{[\mathbf{B}(t,y), \mathbf{Q}(t,y)] : t \geq 0, y \geq 0\} \quad \textit{in} \quad D_2 \times D_2 \ , \end{aligned} \tag{2.7}$$

where D_2 is the function space specified above, $D_2 \times D_2$ is the associated product space, B and Q are continuous deterministic functions of (t, y) specified below, with $B(0, y) = Q(0, y) = 0$ for all $y > 0$, and $[\mathbf{B}, \mathbf{Q}] \equiv \{[\mathbf{B}(t,y), \mathbf{Q}(t,y)] : t \geq 0, y \geq 0\}$ is a random element of $D_2 \times D_2$.

The deterministic functions $B(t,y)$ and $Q(t,y)$ serve as the direct fluid approximation for the scaled stochastic processes $B_s(t,y)$ and $Q_s(t,y)$ in (2.5). From (2.5), we unscale to get the final fluid approximation for the queueing model:

$$B_s^{'}(t,y) \approx sB(t,y) \quad \text{and} \quad Q_s^{'}(t,y) \approx sQ(t,y) \ . \tag{2.8}$$

There also is a refined approximation stemming from the stochastic refinement, e.g.,

$$Q_s^{'}(t,y) \approx sQ(t,y) + \sqrt{s}\mathbf{Q}(t,y) \ . \tag{2.9}$$

The conjectured stochastic limit $[\mathbf{B}, \mathbf{Q}]$ should be related to the limit processes described in Theorems 2 and 3 of Krichagina and Puhalskii (1997), which in turn are related to the Kiefer process; see Csörgő and Révész (1981).

The Corresponding fluid model. We now develop the $G/GI/s + GI$ fluid model. The adjective "$G/GI/s + GI$" is somewhat of a misnomer, because s and the G arrival process no longer appear in the fluid model, but we use the adjective because we are thinking of the application to the original queueing model. The resulting deterministic fluid approximation has three model elements: (ρ, G, F). Consistent with familiar fluid approximations for single-server queues, the fluid approximation depends upon the arrival process A_s only through its rate λ_s, and λ_s is replaced by ρ because of the scaling. The number of servers, s, disappears because we scale by s. Given extensive experience with fluid models associated with single-server queues, we might think that the service-time cdf G and the time-to-abandon cdf F could be replaced by their mean values. That is not so.

In the fluid model, if $\rho < 1$, then the input rate of fluid is less than the maximum possible service rate of fluid, which is 1, and we speak of the model as being *underloaded*. On the other hand, if $\rho > 1$, then the input rate of fluid is greater than the maximum possible service rate of fluid, and we speak of the model as being *overloaded*. These correspond to the QD and ED limiting regimes, respectively. The QED limiting regime occurs when $\rho = 1$ (and further conditions are satisfied). We call the case $\rho = 1$ the *balanced* case.

As a consequence of the spatial scaling, individual customers shrink down into "quanta" of fluid, but the length of time each customer (or quantum of fluid) spends in the system is unchanged because there is no time scaling. If the fluid queue contains content Q, then that corresponds to a queue content of Qs in the associated queueing model with s servers, each working at rate 1. However, the length of time that the fluid spends in the system is

unchanged. For example, if a quantity of fluid enters service at time 0, then a proportion $G(t)$ of that fluid will have been served by time t, while the remaining proportion $G^c(t)$ will remain in service, having been in service for time t. The original IID assumption for the service times leads us to treat the different quanta of fluid as independent. Formally, we apply the law of large numbers to deduce that a proportion $G(t)$ of the customers starting service at time 0 will complete service by time t. Similarly, if a quantity of fluid enters the queue at time 0, then a proportion $F(t)$ of that fluid will have abandoned by time t, while the remaining proportion $F^c(t)$ will remain in queue, provided that it does not move into service.

We now describe how the fluid model evolves. The deterministic functions $B(t, y)$ and $Q(t, y)$ appearing as limits in Conjecture 2.1 fully describe the state of the fluid system: For each $t \geq 0$ and $y > 0$, $B(t, y)$ is the amount of fluid in service at time t that has been in service for time less than or equal to y, while $Q(t, y)$ is the amount of fluid in queue at time t that has been in queue for time less than or equal to y. We assume that these functions are integrable with densities b and q; i.e.,

$$B(t, y) = \int_0^y b(t, y)\, dy \quad \text{and} \quad Q(t, y) = \int_0^y q(t, y)\, dy \ . \tag{2.10}$$

Let $B(t) \equiv B(t, \infty)$ be the total fluid content in service at time t and let $Q(t) \equiv Q(t, \infty)$ be the total fluid content in queue at time t.

First, as indicated above, input of fluid occurs at constant rate ρ. To describe service and abandonment, we work with the hazard (or failure) rates of the service-time and abandon-time distributions, which are well defined because of assumption (2.1). (Here we use the assumption that the pdf's are strictly positive.) These hazard-rate functions are defined as usual by

$$h_s(x) \equiv \frac{g(x)}{G^c(x)} \quad \text{and} \quad h_a(x) \equiv \frac{f(x)}{F^c(x)} \quad \text{for} \quad x \geq 0 \ . \tag{2.11}$$

Clearly, $h_s(x)$ is the conditional rate of service for a customer that has been in service for a length of time x, conditional on that customer not having been served previously. Similarly, $h_a(x)$ is the conditional rate of abandonment for a customer that has been in queue for a length of time x, conditional on that customer not having abandoned previously.

The total service rate (actually performed) at time t is

$$\sigma(t) \equiv \int_0^\infty b(t, x) h_s(x)\, dx, \quad t \geq 0, \tag{2.12}$$

for h_s in (2.11), while the total abandonment rate at time t is

$$\alpha(t) \equiv \int_0^\infty q(t, x) h_a(x)\, dx, \quad t \geq 0, \tag{2.13}$$

for h_a in (2.11).

Fluid in service that is not served remains in service, leading to the *first fundamental evolution equation*

$$b(t+u, x+u) = b(t,x)\frac{G^c(x+u)}{G^c(x)} \quad \text{for all} \quad x \geq 0, \quad t \geq 0 \quad \text{and} \quad u > 0 \ . \tag{2.14}$$

Similarly, fluid waiting in queue that does not abandon, and does not move into service, remains in queue, leading to the *second fundamental evolution equation*

$$q(t+u, x+u) = q(t,x)\frac{F^c(x+u)}{F^c(x)} \quad \text{for all} \quad x \geq 0, \quad t \geq 0 \quad \text{and} \quad u > 0 \ , \tag{2.15}$$

provided that the content has not moved into service.

If the queue is not empty, fluid moves into service at time t at rate $\sigma(t)$ to exactly match the rate of service completion. Fluid moves into service from the front of the queue (the fluid that has been waiting the longest), while new input joins the end of the queue. Thus, at time t, there will be a queue boundary $w(t)$ such that

$$q(t,x) = 0 \quad \text{for all} \quad x > w(t) \ ; \tag{2.16}$$

otherwise the queue content evolves as described in (2.15). In the transient regime, as time evolves the boundary $w(t)$ will evolve as well.

We also have new input at the specified rate ρ. Since that new input goes at the end of the queue whenever the queue is nonempty, we have

$$q(t,0) = \rho \quad \text{for all} \quad t \quad \text{such that} \quad w(t) > 0 \quad (\text{if} \quad Q(t) > 0) \ . \tag{2.17}$$

On the other hand, if the queue is empty and the servers are not all busy, then instead we have

$$b(t,0) = \rho \quad \text{if} \quad B(t) < 1 \ . \tag{2.18}$$

Finally, there is the case in which the queue is empty but the servers are all busy:

$$b(t,0) = \sigma(t) \wedge \rho \quad \text{and} \quad q(t,0) = \rho - (\sigma(t) \wedge \rho) \quad \text{if} \quad B(t) = 1 \quad \text{and} \quad Q(t) = 0 \ , \tag{2.19}$$

where $x \wedge y \equiv \min\{x, y\}$.

It remains to determine whether or not the specification above uniquely determines well-defined deterministic functions B and Q with the properties above; we conjecture that it does. It also remains to do so for appropriate initial conditions.

Conjecture 2.2. (existence and uniqueness for the deterministic fluid process) *Under the assumptions above, there exists a unique pair of continuous functions $B(t,y)$ and $Q(t,y)$ satisfying the description in* (2.10)–(2.19) *above and $B(0,y) = Q(0,y) = 0$ for all $y > 0$.*

Given that the deterministic fluid process is indeed well defined, it also remains to develop an algorithm to compute the functions $B(t,y)$ and $Q(t,y)$ for various initial conditions. We will establish such an algorithm and make partial progress toward proving both Conjectures 2.1 and 2.2 by establishing discrete-time analogs of these conjectures in Section 6.

3. Steady State of the Fluid Model

In this section we describe the steady-state behavior of the fluid model just developed. Since the original $G/GI/s + GI$ queueing model is well known to possess a unique steady-state distribution under minor regularity conditions, it is natural to expect that the same is true for our $G/GI/s + GI$ fluid model, and we show that is the case (without requiring to assume either Conjecture 2.1 or Conjecture 2.2). The following theorem establishes the existence of a unique steady state for the fluid model, and describes the steady-state fluid content. We will describe it in terms of the vector of elements $(q, b, \sigma, \alpha, Q, B)$, defined as above, except we delete the argument t. In steady-state, the values are independent of t.

To describe the steady-state fluid content in the overloaded regime, we use the stationary-excess cdf associated with the abandon-time cdf F, defined by

$$F_e(t) \equiv \frac{1}{m_a} \int_0^t F^c(u)\, du \ , \tag{3.1}$$

where m_a is the mean abandon time, assuming that $m_a < \infty$. (The integral representation holds even if $m_a = \infty$.)

Theorem 3.1. (steady state of the $G/GI/s+GI$ fluid model) *The $G/GI/s+GI$ fluid model specified above with model data (ρ, G, F) has a unique steady state described by the vector $(b, q, \sigma, \alpha, Q, B)$, whose character depends on whether $\rho \le 1$ or $\rho > 1$.*

(a) (underloaded and balanced cases: $\rho \le 1$)

If $\rho \le 1$, then

$$\sigma = B = \rho, \quad \alpha = Q = 0, \quad \text{and} \quad b(x) = \rho G^c(x), \quad x \ge 0 \ . \tag{3.2}$$

(b) (overloaded case: $\rho > 1$)

If $\rho > 1$, then

$$\sigma = 1 \quad \alpha = \rho - 1, \quad Q > 0, \quad B = 1 \ , \tag{3.3}$$

$$b(x) = G^c(x), \quad x \geq 0 \ , \tag{3.4}$$

$$q(x) = \rho F^c(x), \quad 0 \leq x \leq w \quad \textit{and} \quad q(x) = 0, \quad x > w \ , \tag{3.5}$$

where w is the solution of the equation

$$F(w) = \frac{\alpha}{\rho} = \frac{\alpha}{1+\alpha} \ . \tag{3.6}$$

The total queue content is

$$Q = \int_0^w q(x)\,dx = \rho \int_0^w F^c(x)\,dx = \rho m_a F_e(w) \ , \tag{3.7}$$

where F_e is the stationary-excess cdf associated with F defined in (3.1).

Proof. We must have the rate into service equal the rate of service completion, which implies that $b(0) = \sigma$. Then the basic evolution equation (2.14) implies that $b(x) = \sigma G^c(x)$ for $x \geq 0$, which, by integrating implies in general that $B = \sigma$. Since the total input rate must equal the total output rate, we must also have $\rho = \sigma + \alpha$. Since we must have $Q = 0$ when $B < 1$ and we must have $B = 1$ when $Q > 0$, we must have $\alpha = 0$ when $\rho < 1$ and we must have $\sigma = 1$ when $\rho > 1$. That leaves the balanced case in which $\rho = 1$.

In the balanced case, since the total input rate must equal the total output rate, if $\alpha > 0$, then $\sigma = B < 1$, which would imply that $Q = 0$, which would imply that $\alpha = 0$, a contradiction. Hence, we must have $\alpha = 0$ and $\sigma = 1$ in the balanced case, which yields $B = 1$ and $Q = 0$.

It remains to describe the queue content in the overloaded regime. We have determined that $\sigma = 1$ and $\alpha = \rho - 1$ in the overloaded regime. At the same time, the abandonment rate when $Q > 0$ is

$$\begin{aligned} \alpha &= \int_0^w q(x) h_a(x)\,dx = \int_0^w q(0) F^c(x) \frac{f(x)}{F^c(x)}\,dx = \int_0^w \rho F^c(x) \frac{f(x)}{F^c(x)}\,dx \\ &= \int_0^w \rho f(x)\,dx = \rho F(w) = (1+\alpha) F(w) \ , \end{aligned} \tag{3.8}$$

implying formula (3.6). Alternatively, we can find two expressions for the flow into service, obtaining

$$1 = \sigma = b(0) = q(w) = \rho F^c(w) \ , \tag{3.9}$$

which also leads to (3.6). ∎

Figure 1 in Section 1 shows the overloaded steady-state regime established in Theorem 3.1. We now present the corresponding figure for the somewhat less interesting underloaded regime: Figure 2. Since the servers are not all utilized in the underloaded regime, the fluid model in the underloaded regime corresponds to a fluid model for an infinite-server queue; see Duffield and Whitt (1997) and Krichagina and Puhalskii (1997). As in Duffield and Whitt (1997), we here emphasize the importance of the customer ages, i.e., the length of time that they have been in the system.

Implications. The fluid model and its steady-state distribution depend on the triple (ρ, G, F). Important insight into the $G/GI/s + GI$ queueing model is gained simply by looking at what features of the original model appear in the fluid model. Nothing at all is lost from the service times and abandon times. Because of the IID (GI) assumptions in the original $G/GI/s + GI$ model, the service times and abandon times are fully specified by the cdf's G and F, and these cdf's appear as an integral part of the $G/GI/s + GI$ fluid model. That is very different from most single-server fluid models, where the distributions appear only through their mean values. The fluid model we introduce here is significant in large part because the full service-time distribution G and the full abandon-time distribution F play important roles.

In stark contrast, however, the number of servers, s, and a description of the (possibly complex) stochastic behavior of the arrival process do not appear at all. The arrival process and the number of servers only affect the overall rates. The number of servers, s, does not appear, because we scale by s, measuring only relative to s. As we said above, the arrival rate is relative to the maximum possible total service rate, which we have stipulated is 1 in the fluid model. The arrival process only appears via the scaled arrival rate ρ, so almost all of any detailed description of the arrival process plays no role.

This phenomenon is not hard to understand: When we consider very large s, there tends to be a *separation of time scales*. As s increases, arrivals and service completions occur more and more rapidly (in a fast time scale), while the experience of individual customers (as characterized by their abandon times and service times) remains unchanged (in a slow time scale). See Sections 2.4.2, 9.8, 10.3 and 10.4 of Whitt (2002) for further discussion about separation of time scales.

Especially interesting is the steady state in the overloaded regime. Unlike the standard single-server model, the overloaded $G/GI/s + GI$ fluid model provides a useful approximation for steady-state behavior of a multi-server queue with abandonments. When the number of

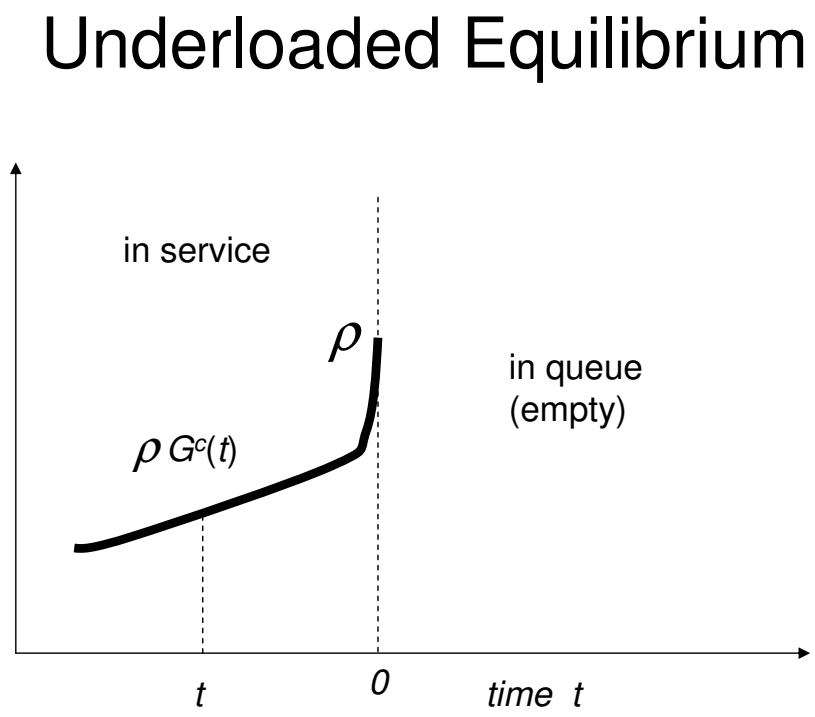

Figure 2: The steady-state distribution of fluid content in the $G/GI/s + GI$ fluid model with service rate 1, arrival rate $\rho \leq 1$, service-time distribution G and abandon-time distribution F. The value at time t is the density of the fluid that has been in the system for length t, i.e., the remaining portion of the fluid that arrived t time units in the past. Since the system is underloaded, the queue is empty, and all new input goes immediately into service. The density at 0 is the arrival rate ρ. The rate of service completion is $\sigma = b(0) = \rho$.

servers is large, a stable solution can be obtained with $\rho > 1$, because a small amount of abandonment can compensate for the excess arrival rate. Indeed, this regime exhibits the ultimate efficiency: The servers are all working at full 100% utilization, and yet no customer waits more than w. In steady state, all customers who are served wait exactly w before starting service. In a call center in the overloaded regime, it is possible to meet a target service level without sacrificing utilization: Instead of answering 80% of all calls within 20 seconds, we can answer 100% within 20 seconds, by making w be just below 20 seconds.

The fluid model provides important insights into the way performance in the $G/GI/s+GI$ queueing model depends upon the two component distributions: the service-time distribution G and the abandon-time distribution F. We summarize these insights in the following corollary.

Corollary 3.1. (dependence of performance on model elements) *The steady-state of the $G/GI/s+GI$ fluid model established in Theorem* 3.1 *has the following properties:*

(*i*) *The performance descriptions ρ, α, σ and B depend upon the two cdf's G and F only through their means.*

(*ii*) *The performance descriptions w, Q and q depend upon the cdf G only through its mean, but upon the cdf F beyond its mean.*

(*iii*) *The performance description b depends upon the cdf F only through its mean (via ρ), but upon the cdf G beyond its mean.*

Corollary is consistent with the numerical experiments in Whitt (2005a), supporting the approximation for the $M/GI/s/r+GI$ queueing model developed there. Additional insight is contained in Mandelbaum and Zeltyn (2004).

The boundary point w in equation (3.6) represents the waiting time for all served fluid. Let W be the average waiting time for *all* fluid. Clearly, W satisfies

$$W = \frac{w}{1+\alpha} + \frac{\alpha}{1+\alpha}\int_0^w xf(x)\,dx\ . \tag{3.10}$$

It is not difficult to see that W and Q are related by the classical relation $L = \lambda W$. Given Mandelbaum and Zeltyn (2004), it is also interesting to see how W is related to the steady-state probability of abandonment, $P(ab) = \alpha/(1+\alpha)$. Just as in Mandelbaum and Zeltyn (2004), we provide support for an approximately linear relation, but *we show that exact linearity only holds in the exponential case.* We say that f is asymptotically equivalent to g as $x \to \infty$, and write $f(x) \sim g(x)$ as $x \to \infty$, if $f(x)/g(x) \to 1$ as $x \to \infty$.

Corollary 3.2. (relations for the overall average waiting time in the fluid model) *In the overloaded regime in the setting of Theorem 3.1,*

$$Q = \rho W \tag{3.11}$$

for Q in (3.7), W in (3.10) and $\rho = 1 + \alpha$. Moreover,

$$P(ab) = \frac{\alpha}{1+\alpha} = \frac{W}{m_a}\frac{F(w)}{F_e(w)} \sim W h_a(0) \quad as \quad \rho \downarrow 1 \ . \tag{3.12}$$

Since $F = F_e$ if and only if F is exponential, we have

$$P(ab) = \frac{W}{m_a} \quad for\ all \quad \rho > 1 \tag{3.13}$$

if and only if F is exponential.

Proof. Using integration by parts in the second integral in (3.10), we obtain (3.11) for Q defined in (3.7). From (3.11) and (3.7), we obtain

$$W = \int_0^w F^c(t)\, dt = m_a F_e(w) = m_a F(w)\frac{F_e(w)}{F(w)} = m_a P(ab)\frac{F_e(w)}{F(w)} \ . \tag{3.14}$$

Solving for $P(ab)$ gives the two equalities in (3.12). Using Taylor series expansions, we get

$$F_e(w) = f_e(0)w + o(w) \quad \text{and} \quad F(w) = f(0)w + o(w) \quad \text{as} \quad w \downarrow 0 \ , \tag{3.15}$$

where $o(w)/w \to 0$ as $w \to 0$. Hence $F(w)/F_e(w) \to f(0)/f_e(0)$ as $w \downarrow 0$. By (3.6), $w \downarrow 0$ as $\rho \downarrow 1$. Finally, we have $F(t) = F_e(t)$ if and only if F is an exponential distribution. As ρ varies from 1 to ∞, $P(ab) = \alpha/(1+\alpha)$ varies from 0 to 1, which in turn yields all possible w as solutions of equation (3.7). ∎

We have observed that the service-time cdf G has negligible impact on the steady-state behavior of the $G/GI/s + GI$ fluid model. In contrast, the service-time cdf G can have a significant impact on the transient behavior of the $G/GI/s + GI$ fluid model. As shown in (2.12), the instantaneous service rate $\sigma(t)$ depends on both the density $b(t,x)$ and the service-time hazard function $h_s(x)$. It is only in steady-state that we obtain the steady-state behavior $\sigma = B$. However, if the service-time distribution is exponential, then the hazard rate is constant. Then $\sigma(t) = B(t)$ for all density functions $b(t,x)$. That is not true for other service-time distributions.

4. Numerical Examples with Non-Exponential Distributions

In this section we start examining how the steady-state description of the $G/GI/s+GI$ fluid model works as an approximation for the steady-state performance of queueing models in the overloaded regime, when the input rate is greater than the maximum possible service rate. As before, we will assume that the mean service time is always 1. We will consider $M/GI/s+GI$ queueing models, which have a Poisson arrival process and non-exponential service-time and time-to-abandon distributions. We have two main objectives here: (i) verify that the fluid approximation is sufficiently accurate and (ii) verify that the structural conclusions in Corollary 3.1 provide insight for overloaded queueing models.

We consider two different non-exponential probability distributions with mean 1: the Erlang E_2 distribution, which has squared coefficient of variation (SCV, variance divided by the square of its mean) $c^2 = 0.5$ and the lognormal $LN(1, 4)$ distributions with SCV $c^2 = 4$. An E_2 distribution has a single parameter, the mean, which we have fixed at 1. A lognormal distribution is fully characterized by its mean and variance; we use $LN(m, v)$ to denote a lognormal distribution with mean m and variance v. An Erlang distribution is less variable than an exponential distribution with the same mean, while a lognormal distribution with SCV $c^2 = 4$ is more variable than an exponential distribution with the same mean (by various measures, one being the SCV).

We let each of these two non-exponential distributions play the role of the service-time distribution and the time-to-abandon distribution, in all possible combinations, so that there are four cases in all. We focus on the overloaded regime, letting $\rho = 1.2$. Since we are thinking of a large number of servers, such as $s \geq 100$, we let $s = 100$, but to see how the approximation performs with a smaller number of servers, we also consider the case $s = 20$. In both cases we let $\rho = 1.2$, so that the approximating fluid model is the same for the corresponding models (when they have the same distributions F and G). For the simulations, we actually consider systems with a finite waiting room, but we let the waiting room be so large that it does not affect the result. In particular there is waiting space for 200 additional customers. (We verify that the waiting room has no effect by also considering the case of 300.)

We display simulation results for $s = 100$ and $s = 20$ in Tables 1 and 2. All simulation experiments are based on ten independent replications of runs each having five million arrivals. The independent replications make it possible to reliably estimate confidence intervals using the t-statistic. For all simulation estimates, we show the half-width of 95% confidence intervals.

To define the performance measures we examine, let S_s be the event that a typical customer eventually will be served, as a function of the number of servers s; let A_s be the event that a typical customer abandons before starting service; let W_s be the steady-state waiting time (before beginning service or abandoning, whichever happens first) for a typical customer (conditional on the arrival not being blocked); let N_s be the steady-state number of customers in the system at an arbitrary time; and let $Q_s \equiv \max\{0, N_s - s\}$ be the steady-state queue length at an arbitrary time. We estimate, the mean, the variance and the SCV of both Q_s and $(W_s|S_s$, the conditional waiting time, given that the customer is served.

We primarily consider five steady-state performance measures in the queueing models: the steady-state probability a customer abandons, $P(A_s)$, the mean steady-state number in queue, $E[Q_s]$, the mean steady-state number in the system, $E[N_s]$, the steady-state probability of experiencing no delay (entering service immediately upon arrival), $P(W_s = 0)$, and the expected conditional steady-state waiting time given that a customer is served, $E[W_s|S_s]$. These performance measures are approximated by the fluid-model performance measures in the obvious way:

$$\begin{aligned} P(A_s) &\approx \frac{\alpha}{\rho} = \frac{\alpha}{1+\alpha}, \\ E[Q_s] &\approx sQ, \\ E[N_s] &\approx s(1+Q), \\ P(W_s = 0) &\approx 0, \\ E[W_s|S_s] &\approx w\ , \end{aligned} \tag{4.1}$$

where α, Q and w are defined in terms of the fluid model elements ρ and F according to equations (3.3), (3.6) and (3.7). We use binary search to solve the equation for w in (3.6). Given w, we calculate Q by numerically integrating the integral in (3.7). Both numerical steps are elementary (done with matlab). We apply the relations $E[Q_s] = \lambda_s E[W_s]$ and

$$E[W_s] = P(S_s)E[W_s|S_s] + P(A_s)E[W_s|A_s] \tag{4.2}$$

to calculate the approximation for $E[W_s|A_s]$. The fluid approximations for the variances $Var(Q_s)$ and $Var(W_s|S_s)$ and the associated SCV's are of course 0. It is convenient to look at the SCV's because they quantify the level of variability independent of scale (the mean).

From Tables 1 and 2, we see that the fluid approximation is remarkably accurate in these heavily loaded scenarios, with the quality of the approximations improving as s increases. The

$M/GI/100/200 + GI$ *model with* $\lambda = 120$ *and* $E[T] = 1.0$

	E_2 *time-to-abandon cdf* *service cdf*			$LN(1,4)$ *time-to-abandon cdf* *service cdf*		
Perf. Meas.	E_2	$LN(1,4)$	approx.	E_2	$LN(1,4)$	approx.
$P(A_s)$	0.16653 ±0.00035	0.16683 ±0.00060	0.16667 –	0.1678 ±0.00023	0.1696 ±0.00054	0.16667 –
$E[Q_s]$	40.25 ±0.057	39.56 ±0.097	41.11 –	14.51 ±0.018	14.52 ±0.043	14.63 –
$Var(Q_s)$	139.6 ±0.69	221.6 ±1.09	0.00 –	61.1 ±0.18	81.5 ±0.30	0.00 –
$SCV(Q_s)$	0.086	0.142	0.00	0.290	0.387	0.000
$E[N_s]$	140.3 ±0.057	139.5 ±1.22	141.11 –	114.4 ±0.019	114.2 ±0.47	114.6 –
$P(W_s = 0)$	0.00046 ±0.00006	0.0068 ±0.00035	0.00000 –	0.032 ±0.00037	0.065 ±0.00077	0.000 –
$E[W_s\|S_s]$	0.353 ±0.00051	0.343 ±0.00094	0.365 –	0.126 ±0.00017	0.125 ±0.00040	0.131 –
$Var(W_s\|S_s)$	0.0097 ±0.000058	0.0176 ±0.000087	0.0000 –	0.0046 ±0.000014	0.0066 ±0.000027	0.0000 –
$SCV(W_s\|S_s)$	0.078	0.149	0.000	0.290	0.422	0.000
$E[W_s\|A_s]$	0.247 ±0.00025	0.261 ±0.00041	0.231 –	0.095 ±0.00008	0.103 ±0.00014	0.077 –

Table 1: A comparison of the fluid approximations with simulation estimates of steady-state performance measures in $M/GI/100/200 + GI$ models under heavy load, specifically for $\lambda = 120$. The mean time to abandon is $E[T] = 1$. The two distributions Erlang (E_2) with $c^2 = 1/2$ and lognormal LN(1,4) with $c^2 = 4$ are used in all four combinations. The half-width of the 95% confidence interval is given for each simulation estimate.

$M/GI/20/200 + GI$ *model with* $\lambda = 24$ *and* $E[T] = 1.0$

	E_2 *time-to-abandon cdf* *service cdf*			$LN(1,4)$ *time-to-abandon cdf* *service cdf*		
Perf. Meas.	E_2	$LN(1,4)$	approx.	E_2	$LN(1,4)$	approx.
$P(A_s)$	0.1747	0.1838	0.16667	0.1914	0.1993	0.16667
	±0.00031	±0.00034	–	±0.00022	±0.00044	–
$E[Q_s]$	7.7	7.6	8.2	3.15	3.26	2.93
	±0.013	±0.013	–	±0.0036	±0.0081	–
$Var(Q_s)$	25.2	33.3	0.00	9.6	12.0	0.00
	±0.04	±0.07	–	±0.10	±0.037	–
$SCV(Q_s)$	0.425	0.577	0.00	0.967	1.13	0.000
$E[N_s]$	27.5	27.2	28.2	22.56	22.48	22.93
	±0.015	±0.15	–	±0.0052	±0.0092	–
$P(W_s = 0)$	0.068	0.126	0.00000	0.210	0.254	0.000
	±0.0004	±0.0007	–	±0.00047	±0.00064	–
$E[W_s\|S_s]$	0.322	0.307	0.365	0.129	0.130	0.131
	±0.0005	±0.0005	–	±0.00016	±0.00035	–
$Var(W_s\|S_s)$	0.042	0.061	0.0000	0.0166	0.0227	0.0000
	±0.00006	±0.00013	–	±0.000025	±0.00010	–
$SCV(W_s\|S_s)$	0.405	0.647	0.000	1.00	1.34	0.000
$E[W_s\|A_s]$	0.309	0.351	0.231	0.139	0.159	0.077
	±0.0003	±0.0004	–	±0.00008	±0.00022	–

Table 2: A comparison of the fluid approximations with simulation estimates of steady-state performance measures in $M/GI/20/200+GI$ models under heavy load, specifically for $\lambda = 24$. The mean time to abandon is $E[T] = 1$. The two distributions Erlang (E_2) with $c^2 = 1/2$ and lognormal (LN(1,4) with $c^2 = 4$ are used in all four combinations. (Everything is the same as in Table 1 except the number of servers and the arrival rate have been divided by 5.) The half-width of the 95% confidence interval is given for each simulation estimate.

numerical results indicate that conclusions (i) and (ii) of Corollary 3.1 apply to the queueing models. For example, the abandonment probability $P(A_s)$ is quite near the fluid value 0.166667 in all cases. Note that the mean queue length, $E[Q_s]$ and the conditional mean waiting times, $E[W_s|S_s]$ and $E[W_s|A_s]$, depend strongly upon the time-to-abandon distribution, but not upon the service-time distribution. (All the service-time and time-to-abandon distributions have mean 1.) This property of the fluid model accurately describes the steady-state behavior of these queueing systems. The weakest approximations are for the quantities that are approximated by 0, such as $P(W = 0)$ and the SCV's.

5. Numerical Examples with Different Loads

We now consider additional numerical examples, focusing more on the quality of the approximation when the queueing system is less heavily loaded, but we are only consider the overloaded case with $\rho > 1$.

Now we often consider exponential distributions. In the special case of an exponential time-to-abandon distribution, it is easy to solve for the fluid parameters w and Q explicitly. Letting m_a denote the mean abandon time, the formulas are

$$w = m_a \log_e(1 + \alpha) \quad \text{and} \quad Q = m_a\alpha \ . \tag{5.1}$$

Like Figure 1, these simple approximations for the overloaded $M/M/s + M$ model provide useful quick approximations. Even if they are not incredibly accurate, they provide useful reference points to think about call-center performance.

We now compare the fluid approximations to the exact values in the $M/M/s+GI$ queueing models. As before, the mean service time is always 1. For the purely Markovian $M/M/s + M$ model, we use the exact numerical algorithm in Whitt (2005a). For $M/M/s + GI$ queueing models with non-exponential abandon-time distributions, we use simulations, just as in Section 4 above.

The results of our numerical comparisons are given in Table 3. The first rows display comparisons between the fluid approximation in (4.1) and computer simulations for non-exponential abandon-time distributions. The two non-exponential abandon-time distributions we consider here are $E_2(m)$ and $LN(m, v)$, where m denotes the mean and v denotes the variance..

We should emphasize that our experiments involve realistic scenarios that might actually occur in call centers meeting their SLA's. The $M/M/s+GI$ examples all have $s = 100$ servers

and arrival rate $\lambda = 102$, yielding $\rho = 1.02$ instead of $\rho = 1.20$ in Section 4.

In Table 3 we also look at examples with larger s and/or greater overload (a larger value of ρ. In particular, we also consider $M/M/s + M$ examples with ($s = 100$, and $\lambda = 110$), and ($s = 1000$ and $\lambda = 1020$). When we do, we see that the fluid model looks more impressive. The lower rows in Table 3 compare the fluid approximation in (4.1) to exact numerical results for the Markovian $M/M/s + M$ queueing model, where the exponential time-to-abandon distribution with mean m is denoted by $M(m)$.

As shown in Whitt (2005a), the abandon-time distribution beyond its mean has a significant impact upon performance, so that in general the $M/M/s+GI$ model is not well approximated by the Markovian $M/M/s+M$ model with an exponential time-to-abandon distribution having the same mean as the given cdf F. Indeed, that is clearly shown in Table 3. In most cases involving the non-exponential time-to-abandon distributions, the fluid approximation for the $M/M/s + GI$ performance is superior to the exact $M/M/s + M$ results for the corresponding model having exponential abandon times with the same mean as F.

The lognormal $LN(4, 4)$ example is particularly striking: The simulated mean queue length and conditional mean wait are $EQ_s = 118.1$ and $E[W_s|S_s] = 1.15$, but the corresponding results for exponential abandon-time distribution with the same mean 4 are $EQ_s \approx 14.8$ and $E[W_s|S_s] \approx 0.145$. The actual values are about eight times the values predicted by the $M/M/s + M$ model with an $M(4)$ abandon-time distribution. In contrast, the corresponding fluid approximations are $EQ_s \approx 137.4$ and $E[W_s|S_s] \approx 1.35$, an error of only about 17%.

Consistent with intuition, Table 3 shows that the quality of the fluid approximation improves as s increases and as the ρ increases. The congestion increases as the mean time to abandon, m_a, increases and as the excess arrival rate, α, increases. Consider $s = 100$ with an exponential abandon-time distribution: Table 3 shows that the fluid approximation performs badly when the arrival rate is relatively low, $\lambda = 102$, and the mean time to abandon is relatively low, $m_a = 0.1$. On the other hand, the fluid approximation performs spectacularly well when the arrival rate is relatively high, $\lambda = 110$, and the mean time to abandon is relatively large, $m_a = 10.0$.

6. The Discrete-Time Fluid Limit

In this section we establish a stochastic-process fluid limit in discrete time, which supports the fluid approximation in Section 2. The mathematics is greatly simplified by working in discrete time. In discrete time, the proof is relatively transparent because it can be recursive.

$M/M/s/+GI$ model with different abandon-time cdf's

case	$P(A_s)$		$E[Q_s]$		$E[W_s\|S_s]$	
$s = 100$, $\lambda_s = 102$	*sim.*	*approx.*	*sim.*	*approx.*	*sim.*	*approx.*
$E_2(0.25)$	0.0530	0.0196	3.07	2.69	0.0284	0.0265
	$\pm$0.00029	–	$\pm$0.014	–	$\pm$0.00012	–
$E_2(1)$	0.0378	0.0196	11.8	10.9	0.113	0.106
	$\pm$0.00032	–	$\pm$0.075	–	$\pm$0.00072	–
$E_2(4)$	0.0236	0.0196	41.6	43.0	0.407	0.425
	$\pm$0.00036	–	$\pm$0.44	–	$\pm$0.0042	–
$LN(1,1)$	0.0376	0.0196	11.4	12.9	0.109	0.127
	$\pm$0.00032	–	$\pm$0.071	–	$\pm$0.00067	–
$LN(4,4)$	0.0206	0.0196	118.1	137.4	1.15	1.35
	$\pm$0.00029	–	$\pm$0.75	–	$\pm$0.0073	–
$LN(4,64)$	0.0349	0.0196	14.9	14.0	0.145	0.131
	$\pm$0.00030	–	$\pm$0.095	–	$\pm$0.00091	–
$s = 100$, $\lambda = 102$	*exact*	*approx.*	*exact*	*approx.*	*exact*	*approx.*
$M(0.1)$	0.0713	0.0196	0.73	0.20	0.0062	0.0020
$M(0.25)$	0.0637	0.0196	1.62	0.50	0.0148	0.0049
$M(1)$	0.0499	0.0196	5.09	2.0	0.0490	0.0198
$M(4)$	0.0363	0.0196	14.8	8.0	0.1455	0.0792
$M(10)$	0.0292	0.0196	29.7	20.0	0.292	0.1980
$M(100)$	0.0200	0.0196	204.5	200.0	2.020	1.9802
$s = 100$, $\lambda = 110$	*exact*	*approx.*	*exact*	*approx.*	*exact*	*approx.*
$M(0.1)$	0.1189	0.0909	1.31	0.20	0.0106	0.0095
$M(0.25)$	0.1112	0.0909	3.06	0.50	0.0268	0.0238
$M(1)$	0.0992	0.0909	10.9	10.0	0.1007	0.0953
$M(4)$	0.0919	0.0909	40.4	40.0	0.3811	0.3812
$M(10)$	0.0909	0.0909	100.0	100.0	0.9485	0.9531
$s = 1000$, $\lambda = 1020$	*exact*	*approx.*	*exact*	*approx.*	*exact*	*approx.*
$M(0.1)$	0.0316	0.0196	3.22	2.00	0.0031	0.0020
$M(0.25)$	0.0290	0.0196	7.40	5.00	0.0071	0.0049
$M(1)$	0.0246	0.0196	25.1	20.0	0.0246	0.0198
$M(4)$	0.0210	0.0196	85.8	80.0	0.0846	0.0792
$M(10)$	0.0198	0.0196	202.8	200.0	0.2003	0.1980

Table 3: A comparison of the overloaded $G/GI/s + GI$ fluid approximation with simulations in the $M/M/s + GI$ queueing model with different abandon-time distributions and with exact numerical results in the Markovian $M/M/s + M$ queueing model having an exponential abandon-time distribution. The number of servers, s, and the arrival rate $\lambda > s$ are indicated at the left. The time-to-abandon distributions considered are Erlang of order 2 with mean m, denoted by $E_2(m)$, lognormal with mean m and variance v, denoted by $LN(m, v)$ and exponential with mean m, denoted by $M(m)$.

The form of the associated fluid limit in continuous time is also readily apparent from the discrete-time limit, but it remains to prove Conjecture 2.1.

To make the connections to continuous time, we assume that all events take place in the discrete time scale $\{k\delta : k \geq 0\}$ for some small $\delta > 0$. We often will drop the δ and use integer arguments, with the understanding that the interval between successive time epochs is δ.

Since multiple events can take place at the same time epoch, we need to specify the order of events at each time epoch. We are able to do the construction and proof for any order, but we do need to specify what order we are using. We assume that, first, customers in service are served; second, waiting customers in queue move into service (from the front of the queue, in order of arrival); third, some waiting customers elect to abandon; and finally we add new arrivals (to the end of the queue).

We also consider a more general model here, letting the arrival process be both state-dependent and time-dependent. The model we consider can be denoted by $G_t(n)/GI/s + GI$. The waiting room may be either finite or infinite; the case of a finite waiting room arises as the special case in which the state-dependent arrival rate becomes zero whenever the number of customers in the queue reaches some level.

As in Section 2, we assume that the service times and abandon times come from independent sequences of IID random variables with general distributions. As in Section 2, let S be a generic service time and let T be a generic abandon time. Here we assume that the possible values these random variables are positive-integer multiples of the small positive δ. We define the probability mass functions (pmf's) and associated cdf's and ccdf's by letting

$$g(k) \equiv P(S = k\delta), \quad G(k) \equiv \sum_{j=1}^{k} g(j), \quad G^c(k) = 1 - G(k) \ , \tag{6.1}$$

$$f(k) \equiv P(T = k\delta), \quad F(k) \equiv \sum_{j=1}^{k} f(j) \quad \text{and} \quad F^c(k) = 1 - F(k) \tag{6.2}$$

for $k \geq 0$, where $g(0) \equiv f(0) \equiv 0$ and

$$\sum_{k=1}^{\infty} f(k) \equiv \sum_{k=1}^{\infty} g(k) \equiv 1 \ . \tag{6.3}$$

As in Section 2, we assume that $ES = 1$, but because the discrete time unit is δ, we have

$$\sum_{k=0}^{\infty} kg(k) = \sum_{k=0}^{\infty} G^c(k) = \delta^{-1} > 1 \ . \tag{6.4}$$

Let a subscript s indicate that the random variable is associated with the system having s servers, with the understanding that we will be letting $s \to \infty$. Let $b_s(n, k)$ be the number

of busy servers at time $n\delta$ that are serving customers that have been in service precisely for time $k\delta$. Let $q_s(n,k)$ be the number of customers in queue at time $n\delta$ that have been in queue precisely for time $k\delta$. Let $\sigma_s(n)$ be the number of service completions at time epoch $n\delta$ and let $\alpha_s(n)$ be the total number of abandonments at time epoch $n\delta$. We assume that customers are served in order of arrival ($FCFS$) by the first available server. Customers enter service whenever a server is available, so that the system is *work-conserving*; i.e., letting

$$B_s(n) \equiv \sum_{k=0}^{\infty} b_s(n,k) \quad \text{and} \quad Q_s(n) \equiv \sum_{k=1}^{\infty} q_s(n,k) \ , \tag{6.5}$$

we assume that $Q_s(n) = 0$ whenever $B_s(n) < s$, and that $B_s(n) = s$ whenever $Q_s(n) > 0$. Given that $b_s(n,k)$ and $q_s(n,k)$ are nonnegative, that condition can be summarized by the equation

$$(1 - (B_s(n)/s))Q_s(n) = 0 \quad \text{for all} \quad n \quad \text{and} \quad s \ . \tag{6.6}$$

Since $B_s(n) \le s$ and since the time unit is actually δ, the maximum long-run service-completion rate in integer time is δ.

The new arrivals at time epoch $n\delta$ become $q_s(n,0)$; we do not include these new arrivals in $Q_s(n)$. In time epoch $(n+1)\delta$, these new arrivals possibly move into service and possibly abandon.

Let $a_s(n)$ count the number of arrivals at time epoch $n\delta$, in the system with s servers. We allow $a_s(n)$ to depend on the history of the system up to time epoch $n\delta$. We make an assumption about the limiting behavior of $a_s(n)$, which depends upon the limiting behavior of $B_s(n)$ and $Q_s(n)$. (The time epoch is n in both cases because arrivals at time epoch n occur after the other events at epoch n.) Let $\Rightarrow$ denote convergence in distribution, which is equivalent to convergence in probability for deterministic limits. Here we assume that

$$\frac{a_s(n)}{s} \Rightarrow \lambda(n, B(n) + Q(n)) \quad \text{as} \quad s \to \infty \tag{6.7}$$

for all $n \ge 0$, where $\lambda(n,t)$ is a nonnegative real-valued (deterministic) function of a nonnegative-integer argument n and a nonnegative-real argument t, whenever

$$\frac{B_s(k)}{s} \Rightarrow B(k) \quad \text{and} \quad \frac{Q_s(k)}{s} \Rightarrow Q(k) \quad \text{as} \quad s \to \infty \tag{6.8}$$

for all k, $0 \le k \le n$, and for all $n \ge 0$, where $B(k)$ and $Q(k)$ are deterministic for all k.

If we initialize the system properly, then we can establish the desired fluid limit recursively by successive applications of the weak law of large numbers (WLLN).

Theorem 6.1. (the discrete-time fluid limit) *Consider the discrete-time $G_t(n)/GI/s + GI$ model specified above, with arrivals satisfying* (6.7). *Suppose that, for each s, the system is initialized with workload characterized by nonnegative-integer-valued stochastic processes $\{b_s(0,k) : k \geq 0\}$ and $\{q_s(0,k) : k \geq 1\}$, where*

$$B_s(0) \equiv \sum_{k=0}^{\infty} b_s(0,k) \leq s \quad \textit{and} \quad Q_s(0) \equiv \sum_{k=1}^{\infty} q_s(0,k) < \infty \ , \tag{6.9}$$

with

$$(1 - (B_s(0)/s))Q_s(0) = 0 \tag{6.10}$$

for each s w.p.1. Suppose that

$$\frac{b_s(0,k)}{s} \Rightarrow b(0,k) \quad \textit{for} \quad k \geq 0 \quad \textit{and} \quad \frac{q_s(0,k)}{s} \Rightarrow q(0,k) \quad \textit{for} \quad k \geq 1 \tag{6.11}$$

as $s \to \infty$, where $b(0,\cdot)$ and $q(0,\cdot)$ are deterministic functions. Moreover, suppose that for each $\epsilon > 0$ and $\eta > 0$, there exists an integer k_0 such that

$$P(\sum_{k=k_0}^{\infty} \frac{b_s(0,k)}{s} > \epsilon) < \eta \quad \textit{and} \quad P(\sum_{k=k_0}^{\infty} \frac{q_s(0,k)}{s} > \epsilon) < \eta \ . \tag{6.12}$$

Then, first, condition (6.12) *holds for all n, i.e., for each n, $\epsilon > 0$ and $\eta > 0$, there exists an integer k_0 such that*

$$P(\sum_{k=k_0}^{\infty} \frac{b_s(n,k)}{s} > \epsilon) < \eta \quad \textit{and} \quad P(\sum_{k=k_0}^{\infty} \frac{q_s(n,k)}{s} > \epsilon) < \eta \ . \tag{6.13}$$

And, second, as $s \to \infty$,

$$\begin{aligned}
\frac{b_s(n,k)}{s} &\Rightarrow b(n,k), \\
\frac{B_s(n)}{s} \equiv \frac{\sum_{k=0}^{\infty} b_s(n,k)}{s} &\Rightarrow B(n) \equiv \sum_{k=0}^{\infty} b(n,k), \\
\frac{q_s(n,k)}{s} &\Rightarrow q(n,k), \\
\frac{Q_s(n)}{s} \equiv \frac{\sum_{k=1}^{\infty} q_s(n,k)}{s} &\Rightarrow Q(n) \equiv \sum_{k=1}^{\infty} q(n,k), \\
\frac{\sigma_s(n)}{s} &\Rightarrow \sigma(n), \\
\frac{\alpha_s(n)}{s} &\Rightarrow \alpha(n)
\end{aligned} \tag{6.14}$$

for each $k \geq 0$ and $n \geq 1$, where (b, q, σ, α) is a vector of deterministic functions characterized as follows for $n \geq 1$. For each n, $0 \leq B(n) \leq 1$, $Q(n) \geq 0$ and $(1 - B(n))Q(n) = 0$. As we go from time $n-1$ to n, there are two cases, depending on whether $B(n-1) = 1$ or $B(n-1) < 1$.

Case 1: $B(n-1) = 1$**.**

In this first case, after epoch $n-1$ *asymptotically all servers are busy and in general there may be a positive queue. In this case,*

$$\sigma(n) = \sum_{k=1}^{\infty} b(n-1, k-1) \frac{g(k)}{G^c(k-1)} \ , \tag{6.15}$$

$$b(n,k) = b(n-1,k-1)\frac{G^c(k)}{G^c(k-1)}, \quad k \geq 1 \ , \tag{6.16}$$

$$b(n,0) = \min\left\{\sigma(n), Q(n-1) + q(n-1,0)\right\} \ , \tag{6.17}$$

$$B(n) = \sum_{k=0}^{\infty} b(n,k) \ , \tag{6.18}$$

$$q(n,k) = 0 \quad \textit{for} \quad k \geq c_n + 2, \quad q(n, c_n+1) = (1-p_n) q(n-1, c_n) \frac{F^c(c_n+1)}{F^c(c_n)} \ , \tag{6.19}$$

where c_n *and* p_n *are determined by*

$$\sum_{k=c_n+1}^{\infty} q(n-1,k) \leq \sigma(n) < \sum_{k=c_n}^{\infty} q(n-1,k) \tag{6.20}$$

and

$$p_n = \frac{\sigma(n) - \sum_{k=c_n+1}^{\infty} q(n-1,k)}{q(n-1,c_n)} \ , \tag{6.21}$$

$$\alpha(n) = \sum_{k=0}^{c_n-1} q(n-1,k) \frac{f(k+1)}{F^c(k)} + (1-p_n) q(n-1,c_n) \frac{f(c_n+1)}{F^c(c_n)} \ , \tag{6.22}$$

$$q(n,k) = q(n-1,k-1)\frac{F^c(k)}{F^c(k-1)} \quad \textit{for} \quad 1 \leq k \leq c_n \ , \tag{6.23}$$

$$Q(n) = \sum_{k=1}^{\infty} q(n,k) \ , \tag{6.24}$$

$$q(n,0) = \lambda(n, B(n) + Q(n)) \ . \tag{6.25}$$

Case 2: $B(n-1) < 1$**.**

In this second case, after epoch $n-1$ *asymptotically all servers are not busy so that there is no queue. As in the first case, equations* (6.15)*,* (6.16) *and* (6.18) *hold. Instead of* (6.17)*,*

$$b(n,0) = \min\left\{\sigma(n) + 1 - B(n-1), q(n-1,0)\right\} \ . \tag{6.26}$$

Then

$$q(n,k) = 0 \quad \textit{for all} \quad k \geq 2 \quad \textit{and} \quad q(n,1) = \left(q(n-1,0) - b(n,0)\right)^+ \ . \tag{6.27}$$

Finally, $q(n,0)$ *is just as in* (6.25)*.*

Proof. We apply mathematical induction on n, starting with $n = 0$. The convergence of $B_s(n)$ and $Q_s(n)$ in (6.14) for $n = 0$ can be proved using conditions (6.11) and (6.12). Those limits together with the assumed limit (6.7) for the arrival process imply that the limit for the arrival process $q_s(n, 0) = a_s(n)$ in (6.7) holds for $n = 0$.

Now we go on to treat higher values of n. We first observe that the regularity property in (6.12) extends to higher n, as indicated in (6.13). Given the limits for the summands, that follows from the inequalities

$$\sum_{k=k_0+n}^{\infty} b_s(n,k) \le \sum_{k=k_0}^{\infty} b_s(0,k) \tag{6.28}$$

and

$$\sum_{k=k_0+n}^{\infty} q_s(n,k) \le \sum_{k=k_0}^{\infty} q_s(0,k) \ , \tag{6.29}$$

which hold $w.p.1$ for all s, n and k.

Suppose that the limits in (6.14) have been established for $0 \le k \le n-1$. We show that the limits also hold for n. To be concrete, we first consider Case 1 in which $B(n-1) = 1$. We first consider the service process for those customers in service. Let $\sigma_s(n, k-1)$ be the number of customers served at time epoch $n\delta$ who had been in service for time $(k-1)\delta$ at time epoch $(n-1)\delta$. For each s, $n \ge 1$ and $k \ge 1$, we can represent $\sigma_s(n, k-1)$ and $b_s(n,k)$ as random sums of IID Bernoulli random variables; in particular,

$$\sigma_s(n,k-1) = \sum_{i=1}^{b_s(n-1,k-1)} X_i \tag{6.30}$$

and

$$b_s(n,k) = \sum_{i=1}^{b_s(n-1,k-1)} (1 - X_i) \ , \tag{6.31}$$

where X_i assumes the value 1 if the i^{th} customer among those in service at time epoch $(n-1)\delta$ that have been in the system for time $(k-1)\delta$ is served at epoch $n\delta$, and assumes the value 0 otherwise. Thus $\{X_i : i \ge 1\}$ is a sequence of IID random variables with

$$P(X_i = 1) = 1 - P(X_i = 0) = \frac{g(k)}{G^c(k-1)} \ . \tag{6.32}$$

Given the established limit for $b_s(n-1, k-1)$, we can apply the WLLN to obtain convergence for $\sigma_s(n, k-1)$ and $b_s(n,k)$ at epoch $n\delta$:

$$\begin{aligned} \frac{\sigma_s(n,k-1)}{s} &= \frac{b_s(n-1,k-1)}{s} \frac{\sum_{i=1}^{b_s(n-1,k-1)} X_i}{b_s(n-1,k-1)} \\ &\Rightarrow \quad b(n-1,k-1)EX_i = b(n-1,k-1)\frac{g(k)}{G^c(k-1)} \end{aligned} \tag{6.33}$$

and

$$\frac{b_s(n,k)}{s} = \frac{b_s(n-1,k-1)}{s}\frac{\sum_{i=1}^{b_s(n-1,k-1)}(1-X_i)}{b_s(n-1,k-1)}$$
$$\Rightarrow \quad b(n-1,k-1)(1-EX_i) = b(n-1,k-1)\frac{G^c(k)}{G^c(k-1)} \tag{6.34}$$

as $s \to \infty$.

We apply these limits in (6.33) and (6.34) to establish the limits for the sum $\sigma_s(n)$ in (6.14). To obtain convergence for the sum, we use the inequality $\sigma_s(n,k) \leq b_s(n-1,k-1)$ and inequality (6.28), which hold $w.p.1$ for all s, n and k. These inequalities with the first regularity condition (6.12) ensure convergence of $\sigma_s(n)$ given the established convergence of the summands $\sigma_s(n,k-1)$.

Given that the number of served customers in epoch $n\delta$ is $\sigma_s(n)$, the number of customers to move into service from the queue at that time epoch is the minimum of $\sigma_s(n)$ and the queue length at that point, which is $Q_s(n-1) + q_s(n-1,0)$. Thus, we have the limit

$$\frac{b_s(n,0)}{s} \Rightarrow b(n,0) \quad \text{as} \quad s \to \infty \tag{6.35}$$

for $b(n,0)$ in (6.17).

We next apply the limits established for $b_s(n,k)$ in (6.34) and (6.35) and the first regularity condition in (6.13) to deduce that

$$\frac{B_s(n)}{s} \Rightarrow B(n) \quad \text{as} \quad s \to \infty\ . \tag{6.36}$$

We next need to establish the asymptotic effect on the queue caused by the customers moving into service. For each s, we have the analog of equations (6.19)– (6.21), specifically,

$$q_s(n,k) = 0 \quad \text{for} \quad k \geq c_{s,n}+2, \quad q_s(n,c_{s,n}+1) = (1-p_{s,n})q_s(n-1,c_{s,n})\frac{F^c(c_{s,n}+1)}{F^c(c_{s,n})}\ , \tag{6.37}$$

where $c_{s,n}$ and $p_{s,n}$ are determined by

$$\sum_{k=c_{s,n}+1}^{\infty} q_s(n-1,k) \leq \sigma_s(n) < \sum_{k=c_{s,n}}^{\infty} q_s(n-1,k) \tag{6.38}$$

and

$$p_{s,n} = \frac{\sigma_s(n) - \sum_{k=c_{s,n}+1}^{\infty} q_s(n-1,k)}{q_s(n-1,c_{s,n})}\ , \tag{6.39}$$

We now divide by s in (6.38) and let $s \to \infty$. There are two cases: (i) when $0 < p_n < 1$ and (ii) when $p_n = 0$. In the first case, when $0 < p_n < 1$, we have $p_{s,n} \to p_n$ and $P(c_{s,n} = c_n) \to 1$ as $s \to \infty$, implying that $q_s(n,k)/s \Rightarrow q(n,k)$ for $k \geq c_n + 1$. In the second case, when $p_n = 0$, we

cannot claim that the index $c_{s,n}$ determined by (6.38) converges in probability to the limiting index c_n. However, even though the index c_n is not continuous in the amount served $\sigma(n)$ and the queue-length function $q(n-1,k)$, the resulting queue-length $q(n,k)$ is. In this second case, we have $P(c_{s,n} \in \{c_n, c_{n-1}\}) \to 1$ as $s \to \infty$, and we still have $q_s(n,k)/s \Rightarrow q(n,k)$ for $k = c_n + 2$ and $k = c_n + 1$. When $c_{s,n} = c_{n-1}$, the limit is determined by the abandonment from queue, discussed next.

We now go on to treat the customers remaining in the queue. The argument is similar to the treatment of the customers in service in (6.30)–(6.32). Let $\alpha_s(n,k-1)$ be the number of customers abandoning at time epoch $n\delta$ who had been in queue for time $(k-1)\delta$ at time epoch $(n-1)\delta$. For each s, $n \geq 1$ and $k \geq 1$, we can represent $\alpha_s(n,k-1)$ and $q_s(n,k)$ as random sums of IID Bernoulli random variables; in particular,

$$\alpha_s(n,k-1) = \sum_{i=1}^{q_s(n-1,k-1)} Y_i \tag{6.40}$$

and

$$q_s(n,k) = \sum_{i=1}^{q_s(n-1,k-1)} (1-Y_i) \ , \tag{6.41}$$

where $\{Y_i : i \geq 1\}$ is a sequence of IID random variables with

$$P(Y_i = 1) = 1 - P(Y_i = 0) = \frac{f(k)}{F^c(k-1)} \ . \tag{6.42}$$

Given the established limit for $q_s(n-1,k-1)$, we can apply the WLLN to obtain convergence for $\alpha_s(n,k-1)$ and $q_s(n,k)$ at epoch n for $0 \leq k \leq c_n$:

$$\begin{aligned} \frac{\alpha_s(n,k-1)}{s} &= \frac{q_s(n-1,k-1)}{s} \frac{\sum_{i=1}^{q_s(n-1,k-1)} Y_i}{q_s(n-1,k-1)} \\ &\Rightarrow \quad q(n-1,k-1)EY_i = q(n-1,k-1)\frac{f(k)}{F^c(k-1)} \end{aligned} \tag{6.43}$$

and

$$\begin{aligned} \frac{q_s(n,k)}{s} &= \frac{q_s(n-1,k-1)}{s} \frac{\sum_{i=1}^{q_s(n-1,k-1)} (1-Y_i)}{q_s(n-1,k-1)} \\ &\Rightarrow \quad q(n-1,k-1)(1-EY_i) = q(n-1,k-1)\frac{F^c(k)}{F^c(k-1)} \end{aligned} \tag{6.44}$$

as $s \to \infty$.

We next apply the limits established for $q_s(n,k)$ and the second regularity condition in (6.13) to deduce that

$$\frac{Q_s(n)}{s} \Rightarrow Q(n) \quad \text{as} \quad s \to \infty \ . \tag{6.45}$$

Finally, given the limits for $B_s(n)$ and $Q_s(n)$ in (6.36) and (6.45), and assumption (6.7) for the arrival process, we have the limit

$$\frac{q_s(n,0)}{s} = \frac{a_s(n)}{s} \Rightarrow \lambda(n, B(n) + Q(n)) = q(n,0) \quad \text{as} \quad s \to \infty \ . \tag{6.46}$$

We now turn to Case 2, in which $B(n-1) < 1$. In this second case, after time epoch $(n-1)\delta$ asymptotically all servers are not busy, so that the queue is necessarily empty except for the new arrivals characterized by $q(n-1,0)$. The service process limits are as in Case 1. We need a new argument to move new customers into service, but the convergence $b_s(n,0)/s \Rightarrow b(n,0)$ for $b(n,0)$ in (6.26) follows from the previously established limits for $\sigma_s(n)$, $B_s(n-1)$ and $q_s(n-1,0)$. As a consequence, we then obtain the limit $q_s(n,k)/s \Rightarrow q(n,k)$ for $k \geq 1$ with $q(n,k)$ in (6.27). Finally, we obtain the limit for $q_s(n,0)$ just as before. ∎

We conclude this section by observing that we can compute the limiting fluid functions in Theorem 6.1 recursively by proceeding in the same order as the proof. Indeed, if we had directly considered the continuous-time model, it would be natural to introduce the associated discrete-time model in order to calculate the desired deterministic fluid functions. That corresponds to the elementary Euler method for solving differential and integral equations.

7. Equilibria Without Time-Dependence

In this section we describe the steady-state (or equilibrium) behavior of the fluid limit in Section 6 under the condition that there is no time-dependence in the arrival process, i.e., so that $\lambda(n,x) = \lambda(x)$ for all $n \geq 0$. We do still allow state-dependence, however.

We say that the deterministic fluid process characterized by the vector of functions (b, q, λ) has an equilibrium, denoted by (b^*, q^*, λ^*) without the time argument n, if

$$(b(n,k), q(n,k), \lambda(n, B(n) + Q(n))) = (b^*(k), q^*(k), \lambda^*) \quad \text{for all} \quad k \geq 0 \quad \text{and} \quad n \geq 0 \ , \tag{7.1}$$

when

$$(b(0,k), q(0,k), \lambda(0, B(0) + Q(0))) = (b^*(k), q^*(k), \lambda^*) \quad \text{for all} \quad k \geq 0 \ . \tag{7.2}$$

If (7.1) holds, then necessarily we also have

$$(B(n), Q(n), \sigma(n), \alpha(n)) = (B^*, Q^*, \sigma^*, \alpha^*) \quad \text{for all} \quad n \geq 1 \ , \tag{7.3}$$

where these associated quantities are defined in terms of the elements above by equations (6.15), (6.18), (6.20)– (6.22) and (6.24)–(6.27), depending upon the case (whether $B(n) < 1$ or $B(n) = 1$).

It is not difficult to see that the fluid model has an equilibrium under very general conditions. In fact, with a state-dependent arrival rate, the fluid model may well have multiple equilibria.

Theorem 7.1. (existence of equilibria for the discrete-time fluid process) *Consider the discrete-time fluid limit for the $G_t(n)/GI/s+GI$ model established in Theorem* 6.1 *in the case that the arrival rate is independent of time n; i.e., so that $\lambda(n,x) = \lambda(x)$ for all n. As before, let the mean service time be $ES = 1$, so that we have* (6.4).

(*a*) (**underloaded and balanced cases**)

For each x, $0 < x \leq 1$, such that $\lambda(x) = \delta x$, the fluid limit in Theorem 6.1 *has an underloaded (or balanced if $x = 1$) equilibrium (b^*, q^*, λ^*) with*

$$B^* = x, \quad Q^* = 0, \quad \sigma^* = \delta x, \quad \alpha^* = 0, \quad \lambda^* = \delta x \quad \textit{and} \quad b^*(k) = \delta x G^c(k), \quad k \geq 0 \ . \quad (7.4)$$

(*b*) (**overloaded case**)

For each vector of numbers, $(Q^*, \alpha^*, c^*, p^*)$, where $Q^* > 0$, $\alpha^* > 0$, c^* is a nonnegative integer and $0 \leq p^* < 1$ such that

$$\begin{aligned}
\lambda(1+Q^*) &= \delta + \alpha^* \ , \\
Q^* &= \sum_{k=0}^{c^*} (\delta+\alpha^*) F^c(k) + (1-p^*)(\delta+\alpha^*) F^c(c^*+1) \ , \\
\alpha^* &= \sum_{k=0}^{c^*} (\delta+\alpha^*) f(k) + (1-p^*)(\delta+\alpha^*) f(c^*+1) \ , \\
p^* &= \frac{\delta - (\delta+\alpha^*) F^c(c^*+1)}{(\delta+\alpha^*) F^c(c^*)} \ ,
\end{aligned} \quad (7.5)$$

the fluid limit in Theorem 6.1 *has an overloaded equilibrium (b^*, q^*, λ^*) with*

$$\begin{aligned}
B^* &= 1, \quad Q^* > 0, \quad \sigma^* = \delta, \quad \alpha^* = \lambda^* - \delta > 0 \ , \\
b^*(k) &= \delta G^c(k), \quad k \geq 0 \ , \\
q^*(k) &= (\delta+\alpha^*) F^c(k), \quad 0 \leq k \leq c^* \ , \\
q^*(c^*+1) &= (1-p^*)(\delta+\alpha^*) F^c(c^*+1) \ , \\
q^*(k) &= 0, \quad k \geq c^*+2 \ .
\end{aligned} \quad (7.6)$$

(*c*) (**no state-dependence**)

A unique equilibrium exists if $\lambda(x) = \lambda(0)$ for all $x > 0$. The underloaded, balanced and overloaded cases arise if $\lambda(0) < \delta$, $\lambda(0) = \delta$ or $\lambda(0) > \delta$.

Proof. In general, from (6.16), we have

$$b^*(k) = b^*(k-1)\frac{G^c(k)}{G^c(k-1)} = b^*(0)G^c(k) \ . \tag{7.7}$$

Given (7.7), we obtain $B^* = b^*(0)\delta^{-1}$ by (6.4) and $\sigma^* = b^*(0)$.

We first consider the underloaded and balanced cases. Letting $B^* = x$ and $Q^* = 0$, we obtain $\lambda^* = \lambda(B^* + Q^*) = \delta x$, which implies that $b^*(0) = \delta x$, which in turn implies that $\sigma^* = \delta x$ and $B^* = x$. Thus the alleged equilibrium indeed is an equilibrium.

We now go on to consider the overloaded case. The equilibrium we construct will have

$$B^* = 1, \quad Q^* > 0 \quad \text{and} \quad \lambda^* = \lambda(B^* + Q^*) = \delta + \alpha^* \ . \tag{7.8}$$

We again have (7.7) above, but now (7.8) implies that $b^*(0) = \delta$. Since $\sigma^* = b^*(0)$, we must have $\sigma^* = \delta$ too. By the first equation in (7.5),

$$\alpha^* = \lambda(B^* + Q^*) - \delta = \lambda^* - \delta \ . \tag{7.9}$$

Given that the equations in (7.5) hold, we then obtain the equations in (7.6).

Finally, we consider part (c). First, if $\lambda(0) \leq \delta$, then we can apply the underloaded or balanced case above, starting with $\lambda(x) = \delta x$ for $x = \lambda(0)/\delta$, to construct a fully specified equilibrium, which therefore must be unique

On the other hand, if $\lambda(0) > \delta$, then we can apply the overloaded case, letting $\alpha^* = \lambda(0) - \delta$. Given α^*, we find p^* and c^* by using the last two equations in (7.5). Then we obtain $q*$, Q^* and the other elements. Again we have constructed a fully specified equilibrium, so that it is necessarily unique. ∎

We have not yet established convergence of the fluid limit to the fluid equilibrium as time evolves when there is no time dependence. It seems evident that such convergence holds for all initial conditions when the arrival rate is constant, but we do not yet have a proof. We do have the following partial result about the convergence of the steady-state distributions in the queueing model.

Theorem 7.2. (convergence of steady-state queueing distributions) *Consider the discrete-time $G/GI/s + GI$ queueing model indexed by s in Theorem* 6.1 *when the arrival rate λ_s is constant (proportional to s), being neither state-dependent not time-dependent. Suppose that queueing model has a limiting steady-state distribution as $n \to \infty$ for each s, characterized by $b_s(\infty, k)$ for $k \geq 0$ and $q_s(\infty, k)$ for $k \geq 1$. Suppose that scaled versions of the steady-state*

distributions converge, i.e.,

$$\frac{b_s(\infty,k)}{s} \Rightarrow b(\infty,k) \quad \textit{and} \quad \frac{q_s(\infty,k)}{s} \Rightarrow q(\infty,k) \tag{7.10}$$

for all k, where $b(\infty,k)$ and $q(\infty,k)$ are deterministic functions, and suppose that the analog of regularity condition (6.12) *holds, i.e.,*

$$P(\sum_{k=k_0}^{\infty} \frac{b_s(\infty,k)}{s} > \epsilon) < \eta \quad \textit{and} \quad P(\sum_{k=k_0}^{\infty} \frac{q_s(\infty,k)}{s} > \epsilon) < \eta \ . \tag{7.11}$$

Then these limits $b(\infty,k)$ and $q(\infty,k)$ coincide with the unique equilibrium of the associated $G/GI/s+GI$ fluid model established in Theorem 7.1.

Proof. The assumptions for $b_s(\infty,\cdot)$ and $q_s(\infty,\cdot)$ above imply that they satisfy the initial conditions in Theorem 6.1. Hence we have the convergence established in Theorem 6.1 with these initial conditions. However, since this initial fluid process holds for all times n, the limit must actually be an equilibrium distribution. Since Theorem 7.1 implies that there is a unique equilibrium for the fluid model, the limits $b(\infty,k)$ and $q(\infty,k)$ must indeed coincide with that equilibrium. ∎

For the underloaded case, it is easy to see that the fluid process converges monotonically to its equilibrium when the system starts out empty. We state the elementary result without proof.

Theorem 7.3. (monotonic convergence starting out empty in the underloaded regime) *Consider the discrete-time $G/GI/s+GI$ fluid model, where the arrival rate is constant, being neither time-dependent nor state-dependent, with $\lambda \le \delta$. If $b(0,k) = 0$ for all $k \ge 0$ and $q(0,k) = 0$ for all $k \ge 1$, then the fluid process converges monotonically to the unique equilibrium established in Theorem* 7.1*; i.e.,*

$$b(n,k) = \lambda G^c(k), 0 \le k \le n, \quad \textit{and} \quad b(n,k) = 0 \quad \textit{for all} \quad k > n \ , \tag{7.12}$$

so that

$$B(n) \uparrow B^* \equiv \frac{\lambda}{\delta} \quad \textit{and} \quad \sigma(n) \uparrow \sigma^* \equiv \lambda \tag{7.13}$$

as $n \to \infty$.

The transient behavior is more complicated otherwise, even when the system starts empty (in the overloaded regime). However, as we observed before, it is easy to numerically compute the transient descriptions of the fluid model.

We conclude this section by observing that the steady-state behavior of the continuous-time fluid model established in Theorem 3.1 appears as the limit of the unique equilibrium in Theorem 7.1 associated with constant arrival rate λ (after changing the notation to rho) as $\delta \downarrow 0$. (Of course, that does not provide a proof of Conjecture 2.1.) To get the densities in Theorem 3.1, we must of course divide by δ in the discrete-time model with step size δ. Thus, Theorems 6.1 and 7.1 provide strong support for the $G/GI/s + GI$ fluid model in Section 2. It would be nice to apply them to prove Conjecture 2.1. And it would be nice to establish a functional-central-limit-theorem refinement, paralleling Krichagina and Puhalskii (1997).

8. Acknowledgments

The author is grateful to Columbia University undergraduate Margaret Pierson for writing the $M/GI/s/r+GI$ simulation program and performing the simulation experiments, which we draw on to make our numerical comparisons in Sections 4 and 5. The author was supported by National Science Foundation Grant DMS-02-2340.

References

Bolotin, V. 1994. Telephone circuit holding-time distributions. In *Proceedings of the International Teletraffic Congress, ITC* 14, J. Labetoulle and J. W. Roberts (eds.), North-Holland, Amsterdam, 125-134.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2002. Statistical analysis of a telephone call center: a queueing-science perspective. Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Csörgő, M., P. Révész (1981) *Strong Approximations in Probability and Statistics*, Academic Press.

Duffield, N. G. and W. Whitt. 1997. Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems* 26, 69–104.

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.

Glynn, P. W. and W. Whitt. 1991. A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Prob.* 23, 188–209.

Krichagina, E. V. and A. A. Puhalskii. 1997. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* 25, 235–280.

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.

Neuhaus, G. 1971. On weak convergence of stochastic processes with multidimensional time parameter. *Ann. Math. Statist.* 42, 1285–1295.

Straf, M. L. 1971. Weak convergence of stochastic processes with several parameters. *Proceedings Sixth Berkeley Symp. Math. Stat. Prob.* 2, 187–221.

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

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.