THE IMPACT OF DELAY ANNOUNCEMENTS IN MANY-SERVER QUEUES WITH ABANDONMENT: SUPPLEMENTARY MATERIAL

by

Mor Armony
Stern School
New York University
marmony@stern.nyu.edu

Nahum Shimkin
EE Department
The Technion
shimkin@ee.technion.ac.il

Ward Whitt
IEOR Department
Columbia University
ww2040@columbia.edu

Abstract

This is a supplement to the main paper, having the same title. In this work we develop methods to study the impact upon steady-state performance of delay announcements made to arriving customers in a many-server queue with customer abandonment. We assume that the queue is not visible to waiting customers, as in most customer contact centers, when contact is made by telephone, email or instant messaging. We propose simple robust announcement schemes: (i) the delay of the last served (DLS) customer and (ii) a fixed delay (FD) announcement based on an appropriate long-run average delay. For any single-number delay announcement made immediately upon arrival, customers may balk or have new abandonment behavior as a function of the announced delay. We introduce a model of that customer response. To perform a rough-cut performance analysis, prior to detailed simulation, we introduce a fluid model, which provides an approximate and highly simplified description for large systems in an overloaded regime. In the fluid model, all customers are faced with the same delay and consequently can be given the same delay announcement. That property motivates considering a second approximation scheme: an equilibrium fixed delay announcement in the stochastic model, which we compute approximately using an iterative numerical algorithm (INA). We show that these two approximate descriptions of aggregate performance are effective by comparing to simulations. We simulate systems with state-dependent DLS announcements and we iterate over simulations using fixed delay announcements. Here we present additional material, supplementing the main paper.

May 16, 2007; updated June 3, 2008

1. Introduction

This is a supplement to the main paper, Armony et al. (2008), with the same title. Our purpose is to provide some additional supporting material, which was not possible to include in the main paper because of limited space. In places we are deliberately redundant in order to make this supplement easier to read.

The Main Paper. In the main paper we study the impact on steady-state performance of making delay announcements in a many-server queue with abandonments. We are thinking of applications with invisible queues, such as routinely occurs in a customer contact center. We consider single-number delay announcements. We propose two specific single-number announcement schemes: (i) the delay of the last served (DLS) customer and (ii) a fixed delay (FD) announcement, based on an appropriate long-run average delay.

To perform a rough-cut analysis of performance, prior to detailed simulation, we use a fluid model, extending Whitt (2006a), which provides an approximate and highly simplified description for large systems in an overloaded regime. In the fluid model, all customers are faced with the same delay and consequently can be given the same delay announcement. That structure of the fluid model motivates considering queueing models with FD announcements in order to approximate the aggregate performance of the queueing model with state-dependent announcements. We use an iterative numerical algorithm (INA) based on the numerical algorithm for approximately analyzing an $M/GI/s+GI$ model developed in Whitt (2005).

We evaluate the two approximation methods - the fluid model and the INA - by comparing the predicted performance with simulations of many-server queues with DLS and FD announcements. When we simulate with FD announcements, we iterate until the average delay of served customers coincides with the announced delay. The principal conclusions are: (1) the fluid model does indeed provide a convenient tool to describe aggregate performance and conduct further analysis, and (2) the INA provides a remarkably good approximation for the steady-state behavior of the queueing model, with both state-dependent DLS delay announcements and equilibrium fixed-delay announcements.

That initial investigation shows that the state-dependent announcements are more reliable than FD announcements, yielding smaller average absolute error and average squared error. The equilibrium FD announcement (making the actual delay equal to the announced delay) is approximately equal to the average of the state-dependent predictions, but the variation is greater with the fixed announcements.

Organization of this Supplement. We start in Sections 2 and 3 by expanding upon the description of the fluid model in Section 3 of the main paper. In Section 2 we describe the fluid model without any delay announcements. Then in Section 3 we describe the fluid model with delay announcements.

Next, in Sections 4 and 5 we present additional numerical results, complementing Section 6 of the main paper. In Section 4 we present numerical results for the all-exponential queueing model introduced in Section 5 of the main paper, but with lower arrival rates. Here we consider $\lambda = 120$ and $\lambda = 110$ instead of $\lambda = 140$. Since there are $s = 100$ servers, each with service rate $\mu = 1$, we are considering initial traffic intensities of 1.2 and 1.1 instead of 1.4. But the reduced arrival rates after balking are only 103.7 and 101.9, so the system ceases to be so overloaded after balking. In Section 5 we present additional numerical results for the simple all-exponential model considered in Section 6 of the main paper, Armony et al. (2007), with $\lambda = 140$. In particular, here we show the results of successive iterations, conducted to find the equilibrium steady-state performance measures.

In Section 6 we introduce an alternative model of customer response to a delay announcement w that falls in between the general model introduced in Section 3 of the main paper and the all-exponential model considered in Section 5 of the main paper. Specifically, here in Section 6 we consider conditional abandonment time distributions of the form

$$F^c(t|w) = \begin{cases} C^c(t), & 0 \le t \le w \ , \\ C^c(w)D^c(t-w), & t > w \ , \end{cases} \tag{1.1}$$

where $F^c(t|w) \equiv 1 - F(t|w)$ and C and D are cdf's with $C^c(0) = D^c(0) = 1$. In Section 6 we establish results paralleling the basic properties established in Section 5 of the main paper.

In Section 7 we study iterations in order to calculate the equilibrium performance. We do so in the setting of Section 6 here. In Section 8 we include additional structural results for the fluid model for more general after-announcement time-to-abandon cdf's, extending the results in Section 4 of the main paper. In Section 9 we draw conclusions.

2. The Fluid Model Without Announcements

An Approximation for a Many-Server Queueing Model. In this section we review the fluid model introduced in Whitt (2006a). Our starting point is the $G/GI/s+GI$ queueing model, which allows for a general stationary arrival process. It is specified by a model 4-tuple (A, s, G, F): $A \equiv \{A(t) : t \geq 0\}$ is the arrival process, understood to be a stationary point process with arrival rate λ, s is the number of servers, G is the service-time cumulative distribution function (cdf) and F is the time-to-abandon cdf. It is understood that there is an unlimited waiting room and the FCFS queue discipline is being used. Let S be a generic service time and let T be a generic time to abandon. Our assumptions mean that $G(t) \equiv P(S \leq t)$ and $F(t) \equiv P(T \leq t)$ for $t \geq 0$. Let $\mu^{-1} \equiv E[S]$ be the mean service time and $\theta^{-1} \equiv E[T]$ be the mean time to abandon, both assumed to be finite. For simplicity, and without loss of generality (by appropriately choosing the measuring units for time), we assume throughout this paper that the mean service time is $\mu^{-1} = 1$. So time is measured in units of mean service times.

In this setting of the $G/GI/s + GI$ model with $\mu = 1$, the fluid model we use arises in the limit as

$$\lambda \to \infty \quad \text{and} \quad s \to \infty \quad \text{with} \quad \rho \equiv \frac{\lambda}{s} \quad \text{held fixed} . \tag{2.1}$$

As the limit indicates, the fluid model is intended for scenarios with large s and λ. The parameter ρ defined in (2.1) is the traffic intensity in the original queueing model. It becomes the fluid arrival rate in the fluid model. The fluid model is characterized by the parameter 3-tuple (ρ, G, F). Even the way the parameters simplify going from the $G/GI/s + GI$ model with $\mu = 1$ to the fluid limit provides great insight. Note that the arrival process enters in only through its rate λ and, with the scaling in (2.1), is captured by the fluid arrival rate ρ. However, the full cdf's G and F remain relevant in the description of the fluid model, even though it becomes deterministic.

The fluid model has been shown to be asymptotically correct in the limiting regime (2.1). There is a proviso, however: The asymptotic correctness has only been verified for the Markovian $M/M/s + M$ special case in Garnett et al. (2002) and Whitt (2004) and a discrete-time analog of the general $G/GI/s + GI$ fluid model in Whitt (2006a). Since the time increments can be arbitrarily short in the discrete-time model, the discrete-time model can be made arbitrarily close to the continuous-time model. Thus the discrete-time proof suffices for practical engineering purposes, but it remains to directly treat the continuous-time model.

The fluid model has been shown to yield accurate approximations in overloaded scenarios through comparisons with exact numerical results obtained from numerical algorithms and simulations; see Whitt (2004, 2005,2006a). At the same time, the fluid model provides great simplification that makes it possible to investigate other more complicated questions. For example, the fluid model was already applied by Ren and Zhou (2006) to study outsourcing and by Whitt (2006b) to study staffing in the face of uncertain arrival rate and absenteeism. Here we again find that the fluid model is useful to gain insight about delay announcements.

The fluid model describes the evolution over time of the system, as discussed in Whitt (2006a). However, we will only consider the steady-state behavior of this fluid model. In addition, we will do so under the condition that $\rho > 1$, which we call the overloaded regime. Without customer abandonment, the system would be unstable when $\rho > 1$, and there would be no proper steady state, but with customer abandonment a proper steady-state distribution exists for the $G/GI/s+GI$ queueing model (under regularity conditions) and the limiting fluid model for all $\rho > 0$. Indeed, with customer abandonment, having $\rho > 1$ is quite natural.

The asymptotic regime we consider, having ρ held fixed with $\rho > 1$, is a special case of heavy traffic, often referred to as the efficiency-driven (ED) regime. The most common heavy-traffic limiting regime is the quality-and-efficiency-driven (QED) regime, which is characterized by having the traffic intensity approach 1 in the limit (at an appropriate rate); see Halfin and Whitt (1981), Garnett et al. (2002) and Zeltyn and Mandelbaum (2005). However, with customer abandonments, the ED regime also describes steady-state behavior and provides useful insight.

The Steady-State Behavior of the Fluid Model. We now describe the steady-state behavior of the fluid model without any system announcements. The steady-state distribution of the fluid content in this regime is depicted in Figure 1. Figure 1 shows the density of fluid content that has been in the system for a period of length t as a function of t, where time t advances toward the left. (The approximate number of customers in the queueing system is obtained by multiplying by s; e.g., the arrival rate in the queueing system is ρs when the fluid arrival rate is ρ.)

This fluid density is a deterministic function, but nevertheless the two model cdf's F and G play a prominent role in the description, especially the time-to-abandon cdf F. In particular, Figure 1 shows that the fluid density in queue is $q(t) \equiv \rho F^c(t)$, $0 \leq t \leq w$, where $F^c(t) \equiv 1 - F(t)$ is the complementary cdf (ccdf) of the abandonment cdf F. In steady state, the servers are all always busy, and the total fluid service rate is 1. The system is kept in steady state by having fluid enter service at rate 1 and fluid abandon at rate $\rho - 1$.

At the right in Figure 1 we see a density of ρ at $t = 0$, which corresponds to the fluid arrival rate. Fluid abandons according to the cdf F up until time w. For $0 < t < w$, a proportion

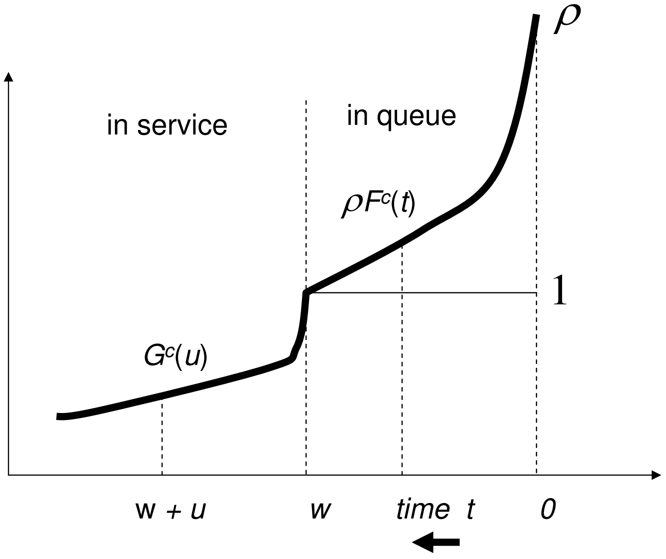

Figure 1: The steady-state density of fluid content as a function of total time t in system, with t increasing to the left, for the $G/GI/s+GI$ fluid model with service rate 1, arrival rate $\rho > 1$, service-time cdf G, time-to-abandon cdf F and no delay announcement.

$F(t)$ of the fluid that would have been in the system for t time units has abandoned, while the remaining proportion $F^c(t)$ remains in the system. The waiting time before served fluid enters service, w, is determined by the requirement that

$$\rho F^c(w) \le 1 \quad \text{and} \quad \rho F^c(t) > 1 \quad \text{for} \quad 0 \le t < w \; . \tag{2.2}$$

In (2.2) we allow the ccdf F^c to have a jump (discontinuity) at w. In doing so, we assume that ties are broken in favor of entering service: **Throughout this paper we assume that customers (or fluid) first enter service if possible and then afterwards the rest abandons.** Thus fluid enters service at rate 1 after waiting w. Abandonment is occurring constantly at the rate $\rho - 1$. The deterministic time w is determined so that the rate of fluid entering service is 1, which equals the maximum possible service rate. At time w, a proportion $(\rho - 1)/\rho$ of the arriving fluid has abandoned.

In Figure 1 we show the fluid arrival rate ρ being much higher than the maximum possible fluid service rate 1. In practice, we think of the fluid arrival rate being not so much higher.

We display a larger difference here in order to be able to clearly show the impact of a delay announcement in this same scenario in the figures in the next section.

While the density of fluid content is deterministic, we interpret the experience of individual customers or "atoms of fluid" as stochastic, regarding these as i.i.d. (The strong law of large numbers is acting behind the scenes to convert the individual independent actions into an overall system deterministic behavior.) Each "customer" abandons before time t with probability $F(t)$, while the customer remains in the system after time t probability $F^c(t)$, provided that $0 < t < w$. Each customer abandons at time w with probability $F(w) - 1$, which will be 0 unless F has a jump at w. Thus, as stated above, the waiting time (before entering service) is precisely w for all customers that do enter service.

The customer experience in service is described by the region of Figure 1 to the left of $t = w$, i.e., for times $t > w$. A proportion $G(u)$ of the fluid entering service at any time will have completed service by time $w + u$. Conversely, a proportion $G^c(u) \equiv 1 - G(u)$ will remain in service. Thus the fluid content density takes the value $G^c(u)$ at time $w + u$. The total fluid content in service at any time is

$$\int_w^\infty G^c(w - u)\, du = \int_0^\infty G^c(u)\, du = E[S] = 1 \; ; \tag{2.3}$$

the total queue content (fluid content waiting in queue) is

$$q \equiv \int_0^w q(t)\, dt = \rho \int_0^w F^c(u)\, du \; . \tag{2.4}$$

The total queue content q is strongly influenced by the time-to-abandon cdf F. The expected or average waiting time for *all* fluid is

$$E[W_{all}] = \int_0^w F^c(u)\, du = \frac{w}{\rho} + \int_0^w x dF(x) = \frac{q}{\rho} \; , \tag{2.5}$$

which of course is less than the waiting time w of the fluid that is served. (We remark that the corresponding formula (3.10) in Whitt (2006a) is incorrect.) We regard W_{all} as a random variable because the experience of individual customers (atoms of fluid) is random.

3. The Fluid Model with Delay Announcements

We now consider a modification of the fluid model in Section 2 to allow for announcements about anticipated delays to arriving customers. As in Section 2, we think of the fluid model arising in the limit (2.1), but we do not prove any limit theorems here.

From Section 2, it should be evident that there is little we can do about abandonment with the fluid model in the overloaded regime. The abandonment rate should be $\rho - 1$, and the throughput should be 1, which clearly is the maximum possible, provided that we remain in the overloaded regime. On the other hand there is considerable freedom in changing the customer waiting experience. Indeed, we can eliminate all waiting, while preserving maximum throughput: We can simply select a proportion $(\rho - 1)/\rho$ of all arrivals at every time and elect not to serve them. Upon arrival, we can send them an announcement that the system is overloaded, so that they will not be served at this time. They can be asked to call back later. That reduces the effective fluid arrival rate to 1, and everybody that remains then can enter service immediately. Those customers who cannot be served will leave immediately as well, so we achieve maximum throughput without any waiting at all. That strategy might well be deemed preferable to making no announcement at all, especially if the mechanism for choosing who gets served is fair and perceived to be fair.

The Initial Customer Response. Here, however, we consider what happens when we simply announce an anticipated delay. The fluid model provides great simplification, because the system state in fluid steady state is deterministic. Thus, if we make a state-dependent delay-announcement, then it is appropriate to give all customers the same announcement. Within the context of this fluid model, the impact of any announcement strategy is to simply change the arrival rate and the time-to-abandon cdf F.

We consider announcing to customers immediately upon arrival their anticipated waiting time w (before beginning service). In the context of the fluid model in Section 2, initially we would be informing all customers about the solution w to (2.2), which we call $\tilde{w}_b$ (b for *before* considering the customer response; in the main paper we shifted the notation to $\tilde{w}_1$, but we do not here). However, we now must consider the impact on customer behavior of making such an announcement. We assume that a proportion $B(\tilde{w}_b)$ will balk in response to a delay announcement $\tilde{w}_b$, where B is our balking cdf. We mention one possible form for the balking cdf.

Definition 3.1. (information consistent balking) *If $B^c(w) = F^c(w)$ for all $w \geq 0$, i.e., if a customer balks at an announced delay whenever that customer would have abandoned by that time without an announcement, then we say that we have information-consistent balking.*

Information-consistent balking is a natural assumption, but it might not hold. It is at least an important reference case. The actual fluid arrival rate, after removing the balking fluid will be $\rho(\tilde{w}_b) = \rho B^c(\tilde{w}_b)$. We will have failed to keep maximum throughput if $\rho(\tilde{w}_b) < 1$, because the throughput will be $min\{\rho(\tilde{w}_b), 1\}$.

We also have to specify how customers who decide to wait respond to the announcement. That is done via the conditional time-to-abandon cdf $F(t|w)$, given any announced delay w. Since B already accounts for balking, we assume that $F(0|w) = 0$ for all w. As before, we assume customers first enter service, and only abandon if that is not possible.

Definition 3.2. (response delay function) *A function $d : [0, \infty) \to [0, \infty)$ is a response delay function for the fluid model if, for each announced delay $w \geq 0$, either (i) $\rho B^c(w) \leq 1$ and $d(w) = 0$ or (ii) $\rho B^c(w) > 1$ and*

$$\rho B^c(w) F^c(d(w)|w) \leq 1 \quad \textit{and} \quad \rho B^c(w) F^c(t|w) > 1 \quad \textit{for} \quad 0 \leq t < d(w) \ . \tag{3.1}$$

Since $F(\cdot|w)$ is assumed to be a cdf for each $w \geq 0$, the response delay function d is well defined. Consequently, served customers (fluid) wait $\tilde{w}_a \equiv d(\tilde{w}_b)$ in response to the delay announcement $\tilde{w}_b$ (a for *after* the announcement $\tilde{w}_b$; in the main paper we shifted the notation from $\tilde{w}_a$ to $\tilde{w}_2$, but we do not here). Thus, assuming that $\rho B^c(w) > 1$, the waiting fluid density (in queue) that has been in the system for time t becomes $\rho B^c(\tilde{w}_b) F^c(t|\tilde{w}_b)$ for $0 < t < \tilde{w}_a$, where $\tilde{w}_a = d(\tilde{w}_b)$ satisfies

$$\rho B^c(\tilde{w}_b) F^c(\tilde{w}_a|\tilde{w}_b) \leq 1 \quad \text{and} \quad \rho B^c(\tilde{w}_b) F^c(t|\tilde{w}_b) > 1 \quad \text{for} \quad 0 \leq t < \tilde{w}_a \ , \tag{3.2}$$

paralleling (2.2). Fluid enters service at rate 1 at time $\tilde{w}_a$. Paralleling (2.4), the total queue content now is

$$q \equiv q(\tilde{w}_b) = \int_0^{\tilde{w}_a} q(t|\tilde{w}_b)\, dt = \rho B^c(\tilde{w}_b) \int_0^{\tilde{w}_a} F^c(t|\tilde{w}_b)\, dt \ . \tag{3.3}$$

To review, in the setting of Figure 1, we are announcing the delay of all served fluid, now called $\tilde{w}_b$. When we consider customer response, we will have the behavior shown in Figure 2. The new fluid density is shown by the solid-line curve, while the original fluid density – without the delay announcement – is now shown by the dashed-line curve. The effective arrival rate is reduced from $\rho > 1$ to $\rho(\tilde{w}_b) \equiv \rho B^c(\tilde{w}_b)$ due to balking, and thereafter (provided that $\rho(\tilde{w}_b) > 1$) abandonment occurs before time $\tilde{w}_a$ at a slower rate. At time $\tilde{w}_a$, the fluid density reaches (and possibly drops below) level 1 and customers enter service at rate 1.

From Figure 2, we see that the system has benefitted from the delay announcement. The fluid throughput is still at the maximum value 1, but the waiting has been reduced. All customers who are served now wait $\tilde{w}_a$ instead of $\tilde{w}_b$. The abandoning customers wait less as well. The most impatient customers elect to balk when they get the delay message. The

Direct Response to Delay Announcement

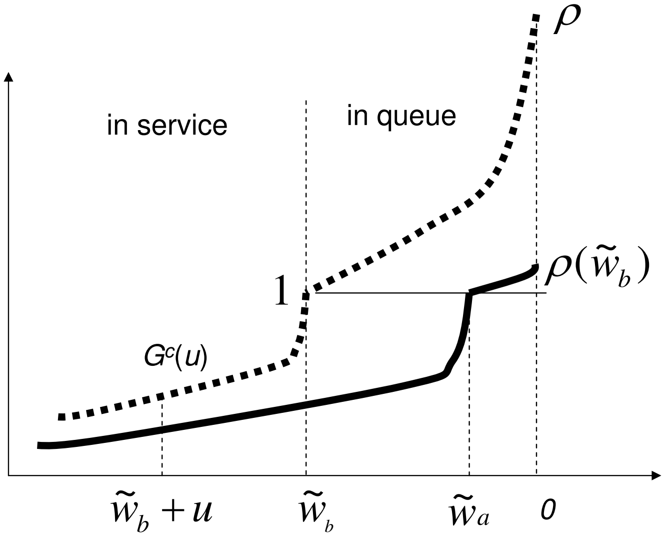

Figure 2: A possible steady-state density of fluid content in the same fluid model shown in Figure 1, but modified by announcing the initially observed delay $\tilde{w}_b$ to all customers immediately upon arrival.

balking rate is $\rho B(\tilde{w}_b)$. Those customers who decide not to balk abandon at a slower rate, but the remaining abandonments occur by time $\tilde{w}_a$.

An Equilibrium Fluid Delay. However, there is a **consistency problem**. The announced delay for served customers, $\tilde{w}_b$, is not consistent with the actual delay for served customers, $\tilde{w}_a$, after the customer response. With DLS announcements, we expect the average delay of served customers to nearly equal the average announced delay.

Definition 3.3. (equilibrium fluid delay) *An announced delay w is an **equilibrium delay** for the fluid model if $d(w) = w$, where d is the response delay function in Definition* 3.2*; i.e., $\tilde{w}_e$ is an equilibrium delay if either (i) $\rho B^c(0) \leq 1$ and $\tilde{w}_e = 0$ or (ii) $\rho B^c(0) > 1$ and*

$$\rho B^c(\tilde{w}_e) F^c(\tilde{w}_e|\tilde{w}_e) \leq 1 \quad \textit{and} \quad \rho B^c(\tilde{w}_e) F^c(t|\tilde{w}_e) > 1 \quad \textit{for} \quad 0 \leq t < \tilde{w}_e \ . \tag{3.4}$$

So how might Figure 2 change if we use an equilibrium-delay announcement? We might (depending on the detailed model elements) instead have the fluid-density plot shown in Figure 3. The equilibrium-delay announcement $\tilde{w}_e$ is less than the original delay $\tilde{w}_b$ before an announcement, but it is greater than the actual delay $\tilde{w}_a$ in response to the announced delay $\tilde{w}_b$. In the setting of Figure 3, we still achieve maximum throughput and we still reduce delays compared to what we achieve in Figure 1 with no announcement at all. However, we cannot do as well as in Figure 2, but that is understandable, because the actual delay $\tilde{w}_a$ associated with announcement of $\tilde{w}_b$ is inconsistent, and thus not sustainable (a behavioral assumption). The effective arrival rate is reduced from $\rho > 1$ to $\rho(\tilde{w}_e) = \rho B^c(\tilde{w}_e) > 1$ due to balking, and thereafter abandonment occurs before time $\tilde{w}_e$. At time $\tilde{w}_e$, the fluid density reaches level 1 and all waiting customers enter service. The dashed-line curves show the previous two cases for contrast.

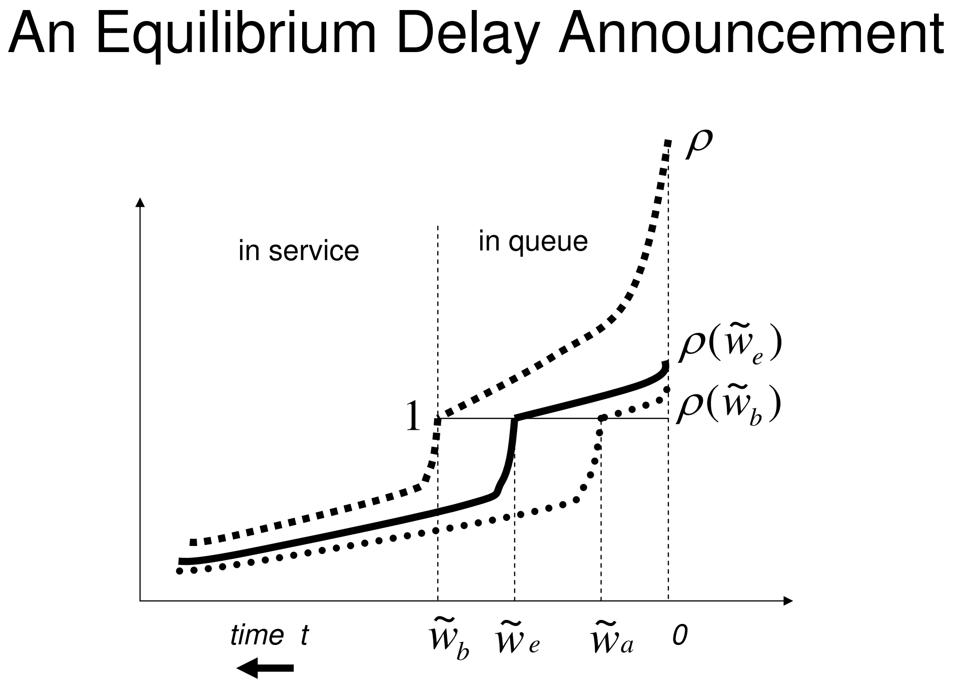

Figure 3: A possible steady-state density of fluid content in the same fluid model shown in Figures 1 and 2, but with an equilibrium-delay announcement $\tilde{w}_e$ to all customers immediately upon arrival.

4. Numerical Results for Lower Arrival Rates

In Section 6 of the main paper we provided numerical results, comparing aggregate system performance measures with state-dependent DLS announcements and FD announcements, obtained from simulations, to corresponding approximate aggregate performance measures obtained from the fluid-model approximation and the INA. We considered the all-exponential model in Section 5 of the main paper, which is the Erlang-A (or $M/M/s+M$) model before the announcement. In the case of FD announcements, the model becomes a $M/M/s+GI$ model, allowing a non-exponential time-to-abandon distribution after the announcement. The simulations are for both the case of an FD announcement and a state-dependent announcement, where the state-dependent announcement is the delay of the last customer to enter service.

In this section we present corresponding numerical results for other cases. In the main paper, we let the arrival rate be $\lambda = 140$, which is quite a heavy load. That led to a reduced arrival rate of 111.9 after balking, which still leaves the system in an overloaded regime. In contrast, here we consider initial arrival rates of 120 and 110 and a higher balking rate of $\beta = 2.0$ instead of $\beta = 1.0$. These lower arrival rates and higher balking rates here yield lower reduced arrival rates (after balking) of only 103.7 and 101.9, which can be considered normal loading, given that we have customer abandonment.

We now describe our experiment in more detail: We consider the all-exponential model with $s = 100$ servers and individual service rate $\mu = 1.0$. We also let the individual abandonment rate before an announcement be $\theta = 1.0$. When we make the delay announcement, we then use the all-exponential model with $\beta = 2.0$, $\gamma = 0.5$ and $\delta = 4.0$. We make the balking rate be greater than the abandonment rate before an announcement ($\beta = 2.0 > 1.0 = \theta$), but we make the abandonment rate for those who elect to wait be less than the abandonment rate before an announcement ($\gamma = 0.5 < 1.0 = \theta$). For the initial delay announcement, we use the equilibrium delay $\tilde{w}_e$ in the fluid model. We assume that the balking probability is $B(\tilde{w}_e) = 1 - e^{-\beta \tilde{w}_e}$ in both the queueing models and the fluid model. By the independent-thinning property of Poisson processes, that produces an $M/M/s+GI$ model, having IID times to abandon with a non-exponential distribution, with common reduced arrival rate $\rho B^c(\tilde{w}_e)$ after taking account of balking.

As noted in the main paper, the simulation program was written in C. As in the main paper, the simulation results are based on 100 independent replications of runs each with

$1,000 \times \lambda$ ($= 120,000$ and $= 110,000$ here, respectively) arrivals. Data were collected after it was verified that initial conditions had only negligible effect on performance. The sample standard error (standard deviation of the sample mean over the 100 replications) are shown below the estimates in parentheses.

Paralleling Table 1 in the main paper, in Tables 1 and 2 here we display performance results for the system before and after an initial announcement in the cases $\lambda = 120$ and $\lambda = 110$. We do not yet describe equilibrium behavior in Tables 1 and 2 here (except for the fluid model). Tables 5 and 6 here contain the main results for the equilibrium behavior. In Tables 1 and 2 we make a single announcement: we always announce the equilibrium fluid delay. That puts the fluid model in equilibrium, but not the stochastic model.

To understand the practical implications, suppose that the mean service time is 10 minutes. Without a delay announcement, the expected waiting times for served customers in those two cases is, respectively 0.179 and 0.101, which corresponds to 1.79 minutes or 107.4 seconds and 1.01 minutes or 60.6 seconds. In contrast, the equilibrium delay in the fluid model for these two scenarios is, respectively, 0.073 and 0.038. With a mean service time of 10 minutes, these equilibrium delays are 43.7 seconds and 22.8 seconds, respectively. When we announce these equilibrium delays, the balking rate is 16.3 and 8.1, respectively, yielding reduced arrival rates of 103.7 and 101.9, respectively.

Tables 1 and 2 compare numerical approximations to the fluid approximations. Given those arrival rates and parameters, we can calculate essentially all steady-state performance measures for the resulting full $M/M/100 + GI$ model, first for the fluid approximation and second for the two queueing models. As in the main paper, we consider two cases for the after-announcement-time abandonment rate δ: (i) $\delta = \gamma = 0.5$ and (ii) $\delta = 4.0 > 0.5 = \gamma$. The numerical algorithm is exact for the $M/M/s + M$ model.

These performance measures given in Tables 1 and 2 show that the accuracy of the fluid approximation decreases as the load decreases, but the accuracy remains excellent before the announcement. The fluid model performs poorly after the announcement in Tables 1 and 2, but it performs much better when compared to the equilibrium behavior of the stochastic models, as we will show in Tables 5 and 6.

Both the INA and the simulation program provide the basis for an efficient algorithm to find the equilibrium delay in the context of the queueing model. We can start by applying the fluid model, using the fluid equilibrium delay as the initial candidate fixed. Then we can iterate in the manner described above, to converge to the equilibrium delay in the queueing *all-expon. model with* $\lambda = 120$, $s = 100$, $\mu = \theta = 1.0$, $\beta = 2.0$, $\gamma = 0.5$ *and two cases for* δ

	before announcement		*after announcement* $\tilde{w}_e = 0.073$		
performance measure	exact	fluid	numer. ($\delta = 0.5$)	numer. ($\delta = 4.0$)	fluid
announced delay	–	–	0.073	0.073	0.073
balking rate	0.0	0.0	16.3	16.3	16.3
arr. rate (after balk)	120.0	120.0	103.7	103.7	103.7
abandon rate	20.2	20.0	5.4	6.3	3.7
throughput rate	99.8	100.0	98.3	97.4	100.0
$P(B)$	0.000	0.000	0.136	0.136	0.136
$P(A)$	0.168	0.167	0.045	0.053	0.031
$P(B \cup A)$	0.168	0.167	0.181	0.189	0.167
$P(W > 0 \| S \cup A)$	0.97	1.00	0.76	0.53	1.00
$E[Q]$	20.1	20.9	10.9	4.5	8.6
$E[W \| S]$	0.179	0.182	0.105	0.0417	0.073
$SD[W \| S]$	0.097	0.00	0.106	0.048	0.000
$E[W \| A]$	0.113	0.138	0.105	0.076	0.070
$E[W; S \cup A]$	0.168	0.174	0.105	0.044	0.072
$P(W \leq 0.05 \| S)$	0.098	0.00	0.406	0.624	0.00
$P(W \leq 0.073 \| S)$	0.146	0.00	0.477	0.738	0.50
$P(W \leq 0.1 \| S)$	0.220	0.00	0.559	0.856	1.00
$P(W \leq 0.2 \| S)$	0.596	1.00	0.810	0.997	1.00

Table 1: A comparison of the fluid approximations with numerical calculations of steady-state performance measures in the all-exponential model, before and after the delay announcement in the case $\lambda = 120$. Two cases are used for the after-announcement-time abandonment rate: (i) $\delta = 0.5 = \gamma$ and (ii) $\delta = 4.0 > 0.5 = \gamma$. In all cases the constant fluid-equilibrium announcement $\tilde{w}_e = 0.073$ is used, without iteration.

all-expon. model with $\lambda = 110$, $s = 100$, $\mu = \theta = 1.0$, $\beta = 2.0$,$\gamma = 0.5$ *and two cases for* δ

	before announcement		*after announcement*		
performance measure	exact	fluid	numer. ($\delta = 0.5$)	numer. ($\delta = 4.0$)	fluid
announced delay	–	–	0.038	0.038	0.038
arr. rate (after balk)	110.0	110.0	101.9	101.9	101.9
balking rate	0.0	0.0	8.1	8.1	8.1
abandon rate	10.9	10.0	4.7	5.8	1.9
throughput rate	99.1	100.0	97.2	96.1	100.0
$P(B)$	0.000	0.000	0.074	0.074	0.074
$P(A)$	0.099	0.091	0.042	0.051	0.017
$P(A \cup B)$	0.099	0.091	0.116	0.127	0.091
$P(W > 0 \vert S \cup A)$	0.84	1.00	0.68	0.49	1.00
$E[Q]$	10.9	10.2	8.6	2.73	3.8
$E[W \vert S]$	0.101	0.095	0.084	0.025	0.038
$SD[W \vert S]$	0.085	0.00	0.106	0.037	0.000
$E[W \vert A]$	0.101	0.113	0.097	0.056	0.038
$E[W; S \cup A]$	0.099	0.093	0.085	0.027	0.038
$P(W \leq 0.038 \vert S)$	0.298	0.00	0.589	0.718	0.50
$P(W \leq 0.05 \vert S)$	0.342	0.00	0.636	0.777	1.00
$P(W \leq 0.1 \vert S)$	0.541	1.00	0.801	0.946	1.00
$P(W \leq 0.2 \vert S)$	0.862	1.00	0.965	0.999	1.00

Table 2: A comparison of the fluid approximations with numerical calculations of steady-state performance measures in the $M/M/100 + M$ model before and after the delay announcement in the case $\lambda = 110$. Two cases are used for the after-announcement-time abandonment rate: (i) $\delta = 0.5 = \gamma$ and (ii) $\delta = 4.0 > 0.5 = \gamma$. In all cases the constant fluid-equilibrium announcement $\tilde{w}_e = 0.038$ is used, without iteration.

model. We apply this iteration scheme, with both the numerical algorithm and simulation, to find the equilibrium-fixed-delay announcement for the two cases $\lambda = 120$ and $\lambda = 110$ in Tables 3 and 4. We see that the INA based on Whitt (2005) is quite close to the simulation.

In addition to the previous notation, W_a denotes the announced waiting time, which is itself a random variable with state-dependent announcements.

Iteration Results for $\lambda = 120$, $s = 100$, $\mu = \theta = 1.0$, $\beta = 2.0$, $\gamma = 0.5$ *and* $\delta = 4.0$

	simulation			*numerical algorithm*			
performance measure	Iter. 1	Iter. 2	Iter. 3	Iter. 1	Iter. 2	Iter. 3	Iter. 4
announced delay	0.073	0.0426	0.048	0.073	0.0417	0.0479	0.0473
arr. rate (after balk)	103.7	110.4	108.9	103.7	110.4	109.0	109.17
$P(B)$	0.136 (0.00011)	0.082 (0.000085)	0.092 (0.000080)	0.136	0.080	0.092	0.090
$P(A)$	0.053 (0.00023)	0.097 (0.00030)	0.089 (0.00028)	0.053	0.098	0.088	0.089
$E[Q]$	4.7 (0.018)	5.6 (0.016)	5.5 (0.015)	4.5	5.5	5.4	5.40
$E[W\|S]$	0.0426 (0.00016)	0.048 (0.00014)	0.048 (0.00013)	0.0417	0.0479	0.0473	0.0475
$SD[W\|S]$	0.0491 (0.00007)	0.047 (0.000061)	0.047 (0.000056)	0.048	0.046	0.046	0.046
$E[W\|A]$	0.084 (0.000093)	0.070 (0.000070)	0.073 (0.000065)	0.076	0.064	0.067	0.067
$SD[W\|A]$	0.0411 (0.000059)	0.034 (0.000063)	0.035 (0.000057)	0.040	0.035	0.036	0.036
$E[W - W_a\|S]$	-0.030 (0.00016)	0.0055 (0.00014)	-0.000027 (0.00013)				
$E[\|W - W_a\|\|S]$	0.051 (0.000067)	0.039 (0.000048)	0.040 (0.000041)				
$E[\|W - W_a\|^2\|S]$	0.0033 (0.0000063)	0.0022 (0.0000067)	0.0022 (0.0000052)				
$Prob(W > W_a)$	0.257 (0.0004)	0.484 (0.0012)	0.444 (0.0011)	0.263	0.497	0.449	0.455

Table 3: Iterations toward equilibrium with $\lambda = 120$: A comparison between the numerical calculations of steady-state performance measures in the all-exponential model with FD announcements (INA) to simulation estimates with FD announcements in the case $\lambda = 120$. The abandonment rate before the announcement is $\gamma = 0.5$, while the abandonment rate after the announcement is $\delta = 4.0$. The initial announcement is the the fluid-equilibrium $\tilde{w}_e = 0.073$.

We are primarily interested in comparing the equilibrium steady-state performance with announcements to the corresponding approximate equilibrium steady-state performance based on the fluid model and the INA. We do that for $\lambda = 120$ and $\lambda = 110$ in Tables 5 and 6.

The example considered in the main paper was quite heavily loaded, with $\lambda = 140$. In

Iteration Results for $\lambda = 110$, $s = 100$, $\mu = \theta = 1.0$, $\beta = 2.0$, $\gamma = 0.5$ *and* $\delta = 4.0$

	simulation			*numerical algorithm*		
performance measure	Iter. 1	Iter. 2	Iter. 3	Iter. 1	Iter. 2	Iter. 3
announced delay	0.038	0.0257	0.027	0.038	0.0251	0.0267
arr. rate (after balk)	101.9	104.5	104.2	101.9	104.6	104.28
$P(B)$	0.074 (0.000079)	0.050 (0.000073)	0.053 (0.000063)	0.074	0.049	0.052
$P(A)$	0.052 (0.000195)	0.069 (0.00027)	0.067 (0.00027)	0.051	0.070	0.067
$E[Q]$	2.82 (0.0096)	3.03 (0.0109)	3.01 (0.011)	2.73	2.97	2.95
$E[W\|S]$	0.0257 (0.000087)	0.027 (0.000096)	0.027 (0.00010)	0.0251	0.0267	0.0266
$SD[W\|S]$	0.037 (0.00006)	0.037 (0.000065)	0.037 (0.000068)	0.037	0.036	0.037
$E[W\|A]$	0.062 (0.000073)	0.054 (0.000062)	0.055 (0.000066)	0.056	0.050	0.051
$SD[W\|A]$	0.031 (0.000053)	0.0298 (0.000062)	0.0299 (0.000068)	0.032	0.031	0.031
$E[W - W_a\|S]$	-0.012 (0.000087)	0.0013 (0.000096)	-0.00011 (0.00010)			
$E[\|W - W_a\|\|S]$	0.035 (0.000029)	0.030 (0.0000396)	0.030 (0.000039)			
$E[\|W - W_a\|^2\|S]$	0.00156 (0.0000032)	0.0014 (0.00000498)	0.0014 (0.000005)			
$Prob(W > W_a)$	0.294 (0.00092)	0.392 (0.0012)	0.381 (0.0012)	0.283	0.378	0.366

Table 4: Iterations toward equilibrium with $\lambda = 110$: A comparison between the numerical calculations of steady-state performance measures in the all-exponential model with FD announcements (INA) to simulation estimates with FD announcements in the case $\lambda = 110$. The abandonment rate before the announcement is $\gamma = 0.5$, while the abandonment rate after the announcement is $\delta = 4.0$. The initial announcement is the the fluid-equilibrium $\tilde{w}_e = 0.038$.

contrast, Tables 5 and 6 consider examples with lower arrival rates – 120 and 110 – and higher balking probability – $\beta = 2.0$ instead of $\beta = 1.0$ – but still with $\delta = 4.0 > 0.5 = \gamma$. Since these models are also overloaded before the announcement, it should come as no surprise that the approximations are effective before considering the announcement. As we should anticipate, Tables 5 and 6 show that the accuracy of the INA approximations for the equilibrium steady-state delay $E[W|S]$ after the announcement decreases as the load decreases, with the error increasing from 4% for $\lambda = 140$ to 16% for $\lambda = 120$ and 18% for $\lambda = 110$, but surprisingly the error in the fluid approximation actually declines, with the error decreasing from 33% for $\lambda = 140$ to 28% for $\lambda = 120$ and 15% for $\lambda = 110$.

State-Dependent Announcements Versus Equilibrium-Fixed-Delay Announcements

$\lambda = 120$ and $s = 100$	*simulation*		*approximations*	
performance measure	state-dep.	equil. fixed	equil. fixed	fluid
announced delay	last	0.048	0.0473	0.073
arr. rate (after balk)	107.8	108.9	109.17	103.7
$P(B)$	0.102 (0.00027)	0.092 (0.000080)	0.090	0.136
$P(A)$	0.071 (0.00012)	0.089 (0.00028)	0.089	0.31
$E[Q]$	6.2 (0.015)	5.5 (0.015)	5.4	8.6
$E[W\|S]$	0.057 (0.00015)	0.048 (0.00013)	0.0475	0.073
$SD[W\|S]$	0.0507 (0.00008)	0.047 (0.000056)	0.046	0.000
$E[W\|A]$	0.068 (0.0001)	0.073 (0.000065)	0.067	0.070
$SD[W\|A]$	0.0462 (0.00009)	0.035 (0.000057)	0.036	
$E[W - W_a\|S]$	0.0065 (0.00001)	-0.000027 (0.00013)		
$E[\|W - W_a\|\|S]$	0.022 (0.00004)	0.040 (0.000041)		
$E[\|W - W_a\|^2\|S]$	0.001 (0.000003)	0.0022 (0.0000052)		
$Prob(W > W_a)$	0.460 (0.0004)	0.444 (0.0011)	0.455	

Table 5: A comparison of equilibrium steady-state performance measures for the case $\lambda = 120$: state-dependent announcements (the delay of the last customer to enter service), estimated by simulation, versus equilibrium-fixed-delay announcements using three different methods: (1) simulation, (2) the numerical $M/M/s+GI$ algorithm (INA), and (3) the fluid approximation. The all-exponential model is used with arrival rate $\lambda = 120$, service rate $\mu = 1$, abandonment rate before the announced time $\gamma = 0.5$, and after-announcement-time abandonment rate $\delta = 4.0$.

As stated in the main paper, we conclude that all approximations perform remarkably well when $\delta = \gamma$. We also conclude that the INA approximation is quite accurate under heavy loading, but the accuracy decreases as the load decreases. Finally, we conclude that the fluid approximation is a useful rough approximation.

State-Dependent Announcements Versus Equilibrium-Fixed-Delay Announcements

$\lambda = 110$ and $s = 100$	*simulation*		*approximations*	
performance measure	state-dep.	equil. fixed	equil. fixed	fluid
announced delay	last	0.027	0.0267	0.038
arr. rate (after balk)	103.4	104.2	104.28	101.9
$P(B)$	0.0599 (0.00018)	0.053 (0.000063)	0.052	0.074
$P(A)$	0.0499 (0.00012)	0.067 (0.00027)	0.067	0.017
$E[Q]$	3.56 (0.010)	3.01 (0.011)	2.95	3.8
$E[W\|S]$	0.033 (0.000098)	0.027 (0.00010)	0.0266	0.038
$SD[W\|S]$	0.042 (0.00007)	0.037 (0.000068)	0.037	
$E[W\|A]$	0.056 (0.00010)	0.055 (0.000066)	0.051	0.030
$SD[W\|A]$	0.041 (0.000088)	0.0299 (0.000068)	0.031	
$E[W - W_a\|S]$	0.0050 (0.000012)	-0.00011 (0.00010)		
$E[\|W - W_a\|\|S]$	0.015 (0.000038)	0.0301 (0.000039)		
$E[\|W - W_a\|^2\|S]$	0.00064 (0.0000021)	0.0014 (0.000005)		
$Prob(W > W_a)$	0.379 (0.00064)	0.381 (0.0012)	0.366	

Table 6: A comparison of aggregate performance measures for the case $\lambda = 110$: state-dependent announcements (the delay of the last customer to enter service), estimated by simulation, versus equilibrium-fixed-delay announcements using three different methods: (1) simulation, (2) the numerical $M/M/s + GI$ algorithm (INA), and (3) the fluid approximation. The all-exponential model is used with arrival rate $\lambda = 110$, service rate $\mu = 1$, abandonment rate before the announced time $\gamma = 0.5$, and after-announcement-time abandonment rate $\delta = 4.0$.

5. Additional Details for Arrival Rate $\lambda = 140$

In this section we present additional details for the model considered in Section 6 of Armony et al. (2007), where $\lambda = 140$. In Table 2 of Armony et al. (2007) we showed simulation results comparing the equilibrium steady-state performance with state-dependent DLS announcements to equilibrium-fixed-delay announcements, both the numerical algorithm (INA) and iterated simulation. We repeat Table 2 of the main paper as Table 7 here.

State-Dependent Announcements Versus Equilibrium-Fixed-Delay Announcements

Simulation Results for $\lambda = 140$, $s = 100$, $\mu = \theta = \beta = 1.0$, $\gamma = 0.5$ and two cases for δ

	$\delta = 0.5 = \gamma$			$\delta = 4.0 > 0.5 = \gamma$		
	equilibrium-fixed		*state-dep.*	*equilibrium-fixed*		*state-dep.*
performance measure	numerical algorithm	sim.	DLS	numerical algorithm	sim.	DLS
announcement	0.225	0.225	last	0.1616	0.155	last
arr. rate (after)	111.8	111.8	112.1	119.11	119.8	118.6
$P(B)$	0.201	0.201 (0.000091)	0.199 (0.00022)	0.149	0.144 (0.00010)	0.153 (0.00022)
$P(A)$	0.086	0.087 (0.00028)	0.086 (0.000092)	0.137	0.143 (0.00028)	0.132 (0.00013)
$E[Q]$	24.3	24.3 (0.084)	24.2 (0.030)	18.8	18.5 (0.025)	19.4 (0.027)
$E[W\|S]$	0.225	0.225 (0.00079)	0.226 (0.00031)	0.1616	0.155 (0.00021)	0.169 (0.00026)
$SD[W\|S]$	0.134	0.133 (0.00038)	0.091 (0.00017)	0.066	0.066 (0.00013)	0.072 (0.00012)
$E[W\|A]$	0.150	0.149 (0.00040)	0.129 (0.00019)	0.137	0.145 (0.00010)	0.136 (0.00017)
$E[W_a]$	0.224	0.224	0.226 (0.00032)	0.162	0.155	0.169 (0.00026)
$E[W - W_a\|S]$		0.00096 (0.00079)	0.011 (0.000025)		0.00050 (0.00021)	0.0057 (0.000014)
$E[\|W - W_a\|\|S]$		0.108 (0.00033)	0.055 (0.000081)		0.052 (0.00010)	0.039 (0.000047)
$E[\|W - W_a\|^2\|S]$		0.018 (0.00010)	0.0050 (0.000016)		0.0044 (0.000017)	0.0025 (0.0000056)
$Prob(W > W_a)$	0.456	0.367 (0.0018)	0.418 (0.00028)	0.523	0.477 (0.00097)	0.470 (0.00023)

Table 7: A comparison between the numerical calculations of equilibrium steady-state performance measures in the all-exponential model with FD announcements (INA) to simulation estimates for both FD announcements and state-dependent DLS delay announcements. Two cases are used for the after-announcement-time abandonment rate: $\delta = \gamma = 0.5$ and $\delta = 4.0 > 0.5 = \gamma$.

all-expon. model with $\lambda = 140$*,* $s = 100$*,* $\mu = \theta = \beta = 1.0$*,* $\gamma = 0.5$ *and* $\delta = 4.0$

performance measure	iteration 2	iteration 3	iteration 4
announced delay	0.157	0.161	0.1616
arrival rate after balking	119.7	119.2	119.11
balking rate	20.3	20.8	20.89
abandon rate	19.8	19.28	19.20
throughput rate	99.9	99.02	99.01
$Prob$(Balk)	0.145	0.149	0.149
$Prob$(Aban)	0.141	0.138	0.137
$Prob$(Aban or Balk)	0.286	0.287	0.286
$Prob$(Delay|Not Balk)	0.981	0.980	0.979
$E[Q]$	18.7	18.78	18.77
$E[W\|Served]$	0.161	0.1616	0.1616
$SD[W\|Served]$	0.065	0.066	0.066
$E[W\|Aban]$	0.136	0.1364	0.1365
$E[W;\text{all, Not Balk}]$	0.157	0.1575	0.1576
$P(W \leq 0.1\|Served)$	0.169	0.172	0.172
$P(W \leq announced\|Served)$	0.431	0.450	0.454
$P(W \leq 0.2\|Served)$	0.721	0.710	0.708
$P(W \leq 0.4\|Served)$	1.000	1.000	1.000

Table 8: Iterations $2-4$ of FD with the numerical algorithm from Whitt (2005) until the constant announced delay matches the actual conditional expected delay given that the customer is served, in the $M/M/100+GI$ model with $\lambda = 140$.

In Table 8 we show iterations $2-4$ to arrive at the equilibrium delay announcement with the numerical queueing algorithm from Whitt (2005) in the case $\delta = 4.0$. The first iteration is shown in Table 1 of the main paper. It starts with the announcement $E[W|S] = 0.224$ and obtains the steady-state expected delay $E[W|S] = 0.157$ that is the announcement in iteration 2. In the subsequent table, Table 9 we show the corresponding iterations performed using simulation with the same model. We see close agreement between Tables 8 and 9.

Finally, in Table 10 we present additional details for Example 7.1 in the main paper, which investigates all-exponential models with $\gamma(w) = \delta(w)$ for all w and $\gamma(w)$ linear. Specifically, we let $\delta(w) = \gamma(w) = \gamma_0 + \gamma_1 w$, where the (constant) parameters γ_0 and γ_1 are chosen so that the different cases all have the same equilibrium fluid delay, $\tilde{w}_e = 0.224$. At that equilibrium fluid delay, we let $\gamma(\tilde{w}_e) \equiv \gamma_0 + \gamma_1 \tilde{w}_e = 0.5$ for all our choices of the parameters γ_0 and γ_1. Table 10 displays simulation results for $\gamma_0 = 0.0, 0.25, 0.45$ and 0.50. The corresponding values of the other parameter are: $\gamma_1 = 2.232, 1.116, 0.2232$ and 0.000. The case of $\gamma_0 = 0.5$ and $\gamma_1 = 0.0$ coincides with the previous simulation results in Table 2 of the main paper. Here we compare these simulation results to the fluid approximation from Table 1 of the main paper.

all-expon. model with $\lambda = 140$*,* $s = 100$*,* $\mu = \theta = \beta = 1.0$*,* $\gamma = 0.5$ *and* $\delta = 4.0$

performance measure	iteration 1	iteration 2	iteration 3
announced delay	0.224	0.153	0.155
$Prob$(Balk)	0.201 (0.000092)	0.142 (0.00010)	0.144 (0.00010)
$Prob$(Aban)	0.087 (0.00026)	0.145 (0.00033)	0.143 (0.00028)
$E[Q]$	17.1 (0.041)	18.4 (0.028)	18.4 (0.025)
$E[W\|Served]$	0.153 (0.00036)	0.155 (0.00023)	0.155 (0.00021)
$E[W\|Aban]$	0.148 (0.00023)	0.145 (0.00012)	0.145 (0.00010)
$P(W > W_a)$	0.197 (0.00094)	0.485 (0.0011)	0.477 (0.00097)

Table 9: Three iterations of FD with simulation until the constant announced delay matches the actual conditional expected delay given that the customer is served, in the $M/M/100 + GI$ model with $\lambda = 140$.

$M/GI/100/200 + GI$ *model with* $\lambda = 140$ *and* $E[T] = 1.0$

all-exponential model with $\gamma(w) = \delta(w) = \gamma_0 + \gamma_1 w$

where $\gamma_0 + \gamma_1 \tilde{w}_e = 0.5$ *with* $\tilde{w}_e = 0.224$*, for four values of* γ_0

Perf. Measure	$\gamma_0 = 0.0$	$\gamma_0 = 0.25$	$\gamma_0 = 0.45$	$\gamma_0 = 0.50$	fluid approx.
$P(Balk)$	0.194	0.196	0.198	0.201	0.201
$P(Aban)$	0.089	0.087	0.087	0.087	0.085
$E[Q]$	23.8	23.9	24.2	24.2	23.7
$E[W\|Served]$	0.221	0.222	0.226	0.226	0.224
$E[W - W_a]$	0.012	0.012	0.011	0.011	–
$E[\|W - W_a\|]$	0.056	0.056	0.055	0.055	–
$E[(W - W_a)^2]$	0.0052	0.0051	0.0051	0.0050	–

Table 10: Fluid approximations compared to simulations of DLS Delay Announcements in the $M/M/100 + GI$ model with $\lambda = 140$.

6. Alternative Customer-Response Model

Much is possible if we allow very general cdf's $F(t|w)$. Obviously there are uncountably many cdf's $F(t|w)$ (indexed by w), each a function of t. So it is helpful to consider some concrete structural assumptions.

In Sections 5 and 6 of the main paper we focused attention on the special case of all-exponential models. Here we consider a generalization of that model. In this Section, we assume that the family of conditional cdf's $F(t|w)$ is characterized by two fixed cdf's C and D, one describing the abandonment behavior **before** the announced time w, and the other describing the abandonment behavior **after** the announced time w. We might anticipate that an announcement w will make many customers balk instead of wait. But, for those who elect to wait, we expect the abandonment rate to be less up to time w, and perhaps that behavior after not balking can be characterized by a single cdf $C(t)$. Similarly, customers may abandon at a faster rate after the announcement time w if they have not yet been served at that time, and that too might be characterized by a fixed cdf $D(t)$. But below we do not make any specific assumptions about these cdf's C and D.

What we propose is displayed in the introduction here in equation (1.1). The general case is discussed in the main paper.

Theorem 6.1. (existence and uniqueness) *Consider the fluid model specified above.*

(a) Suppose that $B^c(w)C^c(w)$ is strictly decreasing in w. Then there is at most one equilibrium delay satisfying Definition 3.3 in the main paper, denoted by $\tilde{w}_e$.

(b) Suppose, in addition, that C^c is strictly decreasing and $B^c(w)C^c(w)$ is a continuous function of w. If

$$\rho B^c(0)C^c(0) = \rho > 1 \quad \textit{and} \quad \lim_{w\to\infty} \rho B^c(w)C^c(w) < 1 \ , \tag{6.1}$$

then there exists a unique equilibrium delay $\tilde{w}_e$, satisfying $\tilde{w}_e > 0$ and

$$\rho B^c(\tilde{w}_e)C^c(\tilde{w}_e) = 1 \ . \tag{6.2}$$

Proof. Existence of an equilibrium delay in (b) follows from the intermediate-value theorem, using (6.1), which ensures a solution to the first equation in (4.5) of the main paper. ▪

For a related result on the existence and uniqueness of the equilibrium in an $M/M/s+GI$ queue, see Zohar et. al. (2002). To avoid pathologies, we introduce the following regularity condition.

Condition 6.1. (regularity condition) *Assume that $\rho > 1$ and the ccdf's $F^c(w)$, $B^c(w)$, $C^c(w)$ and $D^c(w)$ are all continuous and strictly decreasing with*

$$F^c(0) = B^c(0) = C^c(0) = D^c(0) = 1 \tag{6.3}$$

and

$$F^c(w) \to 0, \qquad C^c(w) \to 0 \quad \textit{and} \quad D^c(w) \to 0 \quad \textit{as} \quad w \to \infty \ . \tag{6.4}$$

Clearly Condition 6.1 implies both conditions in Theorem 6.1. Moreover, Condition 6.1 implies Condition 5.1 of the main paper. Thus we can apply Theorem 6.1 to conclude that now, under Condition 6.1 there exists a unique equilibrium delay.

Corollary 6.1. (existence and uniqueness) *Under the regularity conditions in Condition* 6.1*, there exists a unique equilibrium delay $\tilde{w}_e$.*

We can characterize the equilibrium delay as a **quantile**. We say that x is the q^{th} quantile of a continuous and strictly increasing cdf H if $H(x) = q$.

Corollary 6.2. (quantile characterization) *Let X and Y be random variables with cdf's B and C, respectively. Under the regularity conditions in Condition* 6.1*, the unique equilibrium delay $\tilde{w}_e$ is the $[(\rho-1)/\rho]^{\mathrm{th}}$ quantile of the cdf of $X \wedge Y \equiv \min\{X, Y\}$.*

Proof. Note that the ccdf of $X \wedge Y$ is $B^c(t)C^c(t)$. Then note that $\rho B^c(w)C^c(w) = 1$ as required in (6.2), if and only if

$$P(X \wedge Y \leq w) = 1 - B^c(w)C^c(w) = \frac{\rho - 1}{\rho} \ . \quad \blacksquare$$

7. Iterations and Convergence

In this section, we assume that we make an initial delay announcement w_0. Then we observe the actual steady-state delay w_1 of those customers who are served, in the context of the fluid model. We then make the latter our delay announcement, and see the actual steady-state delay w_2 of those customers served with announcement w_1. We continue in this way, looking at the actual steady-state delay w_{k+1} of those customers served, given delay announcement w_k, for $k \geq 0$. This iteration scheme is a natural way to compute the equilibrium delay, but it does not actually correspond to a natural evolution of the system over time, because the system with DLS announcements would not be able to reach steady state before adjustment. For that reason, we have left this material out of the main paper.

At the end of this section we also consider an alternative damped iteration scheme. Such alternative iterative schemes may be of considerable practical importance, because the direct iteration scheme can lead to oscillations, as we will show. These iteration schemes are useful because they provide ways to determine the equilibrium delay.

For the specified iterative scheme, we seek conditions for the **convergence** of the sequence $\{w_k : k \ge 0\}$ to a limit w_∞ and for that limit to be the unique equilibrium-delay announcement $\tilde{w}_e$. We might naturally consider starting with $w_0 = \tilde{w}_b$, the actual delay of served fluid with no delay announcement, but we would like to say that $w_k \to \tilde{w}_e$ as $k \to \infty$ for all initial announced delays w_0.

For further discussion, let $d(w)$ denote the actual delay associated with announcement w, i.e., the response delay function defined in Definition 4.2 of the main paper. We are interested in the structure of the function d. For example, we are interested in the possible **monotonicity** of d and the possible convergence of the resulting sequence $w_{k+1} = d(w_k)$, $k \ge 0$. For example, we would like to conclude that $w_{k+1} \equiv d(w_k) > w_k$ for all $k \ge k_0$, for some initial k_0. We start by describing some **pathological behavior**.

Example 7.1. (oscillation with excessive balking) Suppose that $\rho(w) = \rho B^c(w) < 1$ for all $w > \tilde{w}$, where $\tilde{w} < \tilde{w}_1$ with $\rho B^c(0) C^c(0) D^c(\tilde{w}_1) = 1$. If $w_0 > \tilde{w}$, then $w_{2k-1} = 0$ and $w_{2k} = \tilde{w}_1$ for all $k \ge 1$. However, for any balking cdf B satisfying Condition 6.1, this pathology will not occur if $D^c(t)$ decreases rapidly enough in t.

Unlike the last example, our next example requires that we drop Condition 6.1.

Example 7.2. (reluctance to balk or abandon before the announced delay) Suppose that customers are reluctant to balk. Specifically, suppose that $\rho B^c(w) > 1 + \epsilon > 1$ for all $w \ge 0$, contrary to condition (6.4). Moreover, suppose that customers are reluctant to abandon before the announced delay. Specifically, suppose that $C^c(t) = 1$ for all t. Then $w_k \to \infty$ as $k \to \infty$. Indeed, $w_k \ge w_0 + k\delta$, where $D^c(\delta) = \epsilon$. The reason is that $w_{k+1} > w_k + \delta$ for all k, as is easily verified. ■

Example 7.3. (convergence to $\tilde{w}_b$) Let us reconsider Example 5.1 in the main paper with $B = F$ and $C^c(t) = 1$ for all t, but now let $D^c(t)$ be continuous and strictly decreasing instead of having $D^c(t) = 0$ for all t. Suppose in addition that

$$D^c(t) < \frac{F^c(t+w)}{F^c(w)} \quad \text{for all} \quad t \quad \text{and} \quad w \ . \tag{7.1}$$

Then it is easy to see that there is no equilibrium delay and

$$w_{k+1} > w_k \quad \text{for all} \quad k \geq 0 \quad \text{and} \quad w_k \uparrow \tilde{w}_b \quad \text{as} \quad k \to \infty \tag{7.2}$$

provided that $w_0 < \tilde{w}_b$. ▪

Now we state properties of the iteration, many of which can be verified directly. Necessary and sufficient conditions for monotone convergence are given in part (f). While these conditions can be satisfied, they also can easily be violated.

Theorem 7.1. (properties of the iteration) *Consider the fluid model with Condition* 6.1.

(a) For each delay announcement w with $0 \leq w < \tilde{w}_e$, the actual delay $d(w) = w'$ satisfies

$$\rho B^c(w) C^c(w) D^c(w' - w) = 1 \ , \tag{7.3}$$

so that $w < d(w) \leq w + d(0)$.

(b) For each delay announcement w with $w > \tilde{w}_e$, the actual delay $d(w) = w'$ satisfies $0 \leq d(w) < \tilde{w}_e$.

(c) If $w_2 > w_1 > \tilde{w}_e$ and $d(w_1) > 0$, then $d(w_1) > d(w_2)$.

(d) If $w_1 < w_2 < \tilde{w}_e$ and

$$D^c(w - w_1) - D^c(w - w_2) \leq B^c(w_1)C^c(w_1) - B^c(w_2)C^c(w_2) \quad \textit{for all} \quad w > w_2 \ , \tag{7.4}$$

then $d(w_1) < d(w_2)$.

(e) If $d(w) < \tilde{w}_e$ for all w with $0 \leq w < \tilde{w}_e$, then

$$w_k < w_{k+1} \equiv d(w_k) < \tilde{w}_e \quad \textit{for all} \quad k \geq 1 \quad \textit{and} \quad w_k \uparrow \tilde{w}_e \quad \textit{as} \quad k \to \infty \ . \tag{7.5}$$

(f) We have $d(w) < \tilde{w}_e$ for all $0 \leq w < \tilde{w}_e$, as needed in part (e) *above, if and only if*

$$\rho B^c(w) C^c(w) D^c(\tilde{w}_e - w) < 1 \quad \textit{for all} \quad 0 \leq w < \tilde{w}_e \ . \tag{7.6}$$

Proof. For part (e), we need to show that the established limit of w_k must indeed be $\tilde{w}_e$. Suppose that $w_k \uparrow w_\infty < \tilde{w}_e$. But we must have $d(w_\infty) > w_\infty$. However, $d(w)$ is continuous in w at w_∞. That would imply that $\lim_{k \to \infty} w_k = d(w_\infty) > w_\infty$, which is a contradiction. So the only possible limit is $\tilde{w}_e$. ▪

A concrete case where condition (7.6) is satisfied appears in Theorem 7.2 (a) below. A concrete case where condition (7.6) is violated and the conclusion also is violated appears in Theorem 7.2 (b).

There are some difficulties in general: First, we may have $d(w) > \tilde{w}_e$ for $w < \tilde{w}_e$ and, in the event that occurs, we may fail to have strict monotonicity of the two-stage iteration $d^{(2)}(w) \equiv d(d(w))$, and consequently the announced delay sequence might oscillate or diverge. We illustrate by next giving an example in which the two-stage iteration operator $d^{(2)}(w)$ has multiple fixed points.

Example 7.4. (multiple fixed points for the two-stage iteration operator) In this example we give an example of cycling around the equilibrium delay $\tilde{w}_e$ instead of convergence to it. In particular, we show that the two-stage iteration operator $d^{(2)}(w) \equiv d(d(w))$ has multiple fixed points. This example uses linear functions with slope -1. In particular, $\rho F^c(t) = \rho B^c(t) = \rho - t$ for $0 \le t \le \rho$. We also have $\rho B^c(w) F^c(t|w) = \rho - 2t$, which we point out is not consistent with (1.1), because we cannot define the cdf D independent of w, as is done there. The cyclic behavior is shown in Figure 4. In this example, $w_{2k} = w_0$ and $w_{2k-1} = w_1 = d(w_0)$ for all $k \ge 1$. Such cycling will occur in this linear example (with lines of slope -1) for each announced delay w with $0 < w < \tilde{w}_n$ except for the equilibrium delay $\tilde{w}_e = \tilde{w}_n/2$.

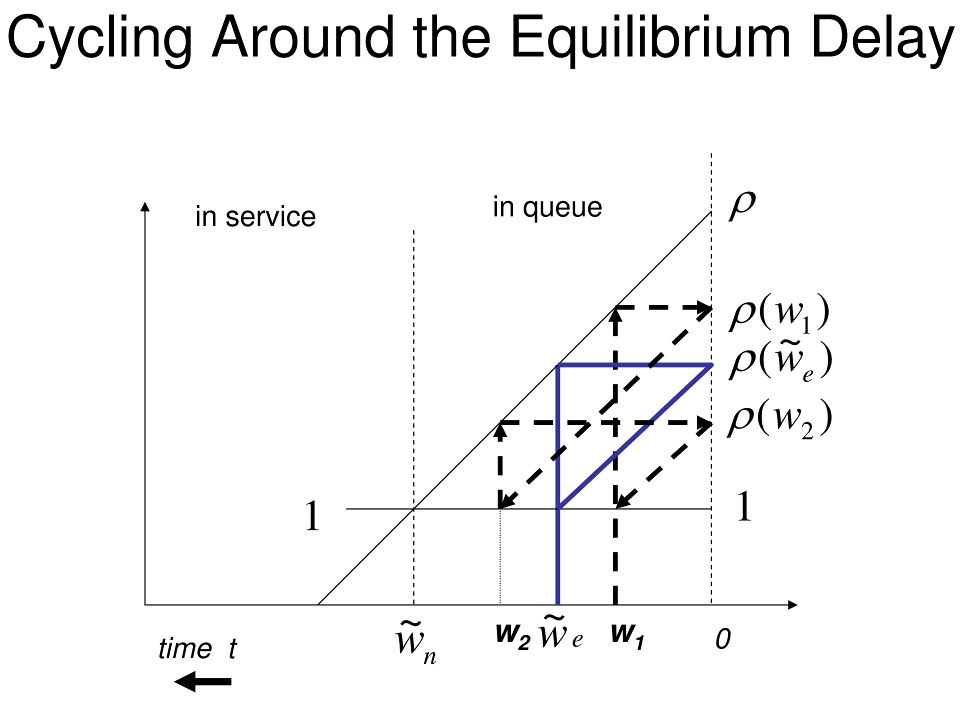

Figure 4: Cycling around the equilibrium delay: The announced delay becomes the actual delay every other iteration.

We now consider iteration and convergence for the all-exponential model. We show that there is bad oscillating behavior when $\delta = \gamma < \beta$ in part (b). let $[x]^+ \equiv \max\{x, 0\}$.

Theorem 7.2. (iteration and convergence for the all-exponential model) *Consider the fluid model above under the all-exponential condition* (6.1) *in the main paper.*

(*a*) *Assume that $\delta \ge \beta + \gamma$. Then the delay associated with announcement w is*

$$d(w) = \frac{\log \rho + (\delta - \gamma - \beta) w}{\delta} \quad \textit{for} \quad 0 < w \le \tilde{w}_e \ , \tag{7.7}$$

which has the property that $w < d(w) < \tilde{w}_e = \log \rho/(\beta + \gamma)$, while $d(w) = 0$ for $w \ge \tilde{w}_b$, $0 < d(w) < \tilde{w}_e$ for $\tilde{w}_e \le w \le \tilde{w}_b$ and $d(0) = \tilde{w}_1 = \log \rho/\delta < \tilde{w}_e$. As a consequence,

$$\tilde{w}_e > w_{k+1} \equiv d(w_k) > w_k \ge \tilde{w}_1 > 0 \quad \textit{for all} \quad k \ge 2 \tag{7.8}$$

and $w_k \equiv d^{(k)}(w_0) \to \tilde{w}_e$ as $k \to \infty$.

(*b*) *Assume that $\delta = \gamma < \beta$. Then*

$$d(w) = \left[\frac{\log(\rho)}{\gamma} - \left(\frac{\beta}{\gamma}\right) w\right]^+ = \left[\tilde{w}_e - \left(\frac{\beta}{\gamma}\right)(w - \tilde{w}_e)\right]^+ \ , \tag{7.9}$$

$$d^{(2n)}(w) = \begin{cases} 0, & w \le \tilde{w}_e (1 - (\gamma/\beta)^{2n}) \ , \\ \frac{\log(\rho)}{\gamma}, & w \ge \tilde{w}_e (1 + (\gamma/\beta)^{2n-1}) \ , \\ \tilde{w}_e + (w - \tilde{w}_e) \left(\frac{\beta}{\gamma}\right)^{2n}, & t > w \ , \end{cases} \tag{7.10}$$

Consequently, for all $w < \tilde{w}_e$,

$$d^{(2n)}(w) = 0 \quad \textit{and} \quad d^{(2n+1)}(w) = \frac{\log(\rho)}{\gamma} \tag{7.11}$$

for all n sufficiently large; for all $w > \tilde{w}_e$,

$$d^{(2n+1)}(w) = 0 \quad \textit{and} \quad d^{(2n)}(w) = \frac{\log(\rho)}{\gamma} \tag{7.12}$$

for all n sufficiently large.

Proof. (a) Note that $w' \equiv d(w)$ is characterized by $\rho(w') F^c(w'|w) = 1$, which for the all-exponential model can be expressed as $\rho e^{-\beta w} e^{-\gamma w} e^{-\delta(w' - w)} = 1$, $0 < w \le \tilde{w}_e$, or, equivalently (by taking logarithms), (7.7). The other formulas are elementary. ■

Remark 7.1. (damped iteration) In order to avoid oscillations, it may be desirable to use a **damped iteration**. We can let the successive announced delays be defined recursively by

$$w_{k+1} = pd(w_k) + (1-p)w_k = w_k + p(d(w_k) - w_k) \tag{7.13}$$

for some constant p with $0 < p \leq 1$. Instead of part (f) of Theorem 7.1, with the damped iteration we have $d(w) < \tilde{w}_e$ for all $0 \leq w < \tilde{w}_e$, as needed in part (e) above, if and only if $\rho B^c(w)C^c(w)D^c((\tilde{w}_e - w)/p) < 1$ for all $0 \leq w < \tilde{w}_e$. As $D^c(w)$ is strictly decreasing, it is easy to verify that this condition will always be satisfied for p small enough. We establish a result for the general conditional time-to-abandon cdf in the next section.

8. More Results for a General Time-to-Abandon CDF

In this section we present a few additional results for the more general time-to-abandon cdf functions $F(t|w)$ considered in Section 5 of the main paper.

We consider the model of Section 3, with arrival rate $\rho(w)$ and conditional time-to-abandon cdf $F(t|w)$ in response to the announced delay w. The equilibrium delay $\tilde{w}_e$ is given by Definition 3.1 of the paper, namely

$$\rho(\tilde{w}_e)F^c(\tilde{w}_e|\tilde{w}_e) = 1 \ , \tag{8.1}$$

with

$$\rho(\tilde{w}_e)F^c(t|\tilde{w}_e) > 1 \quad \text{for} \quad 0 < t < \tilde{w}_e \ . \tag{8.2}$$

We shall employ the following regularity condition. It is stronger than Condition 5.1 in the main paper, containing some features used to establish iteration results.

Condition 8.1. (regularity condition) *Assume that*

(a) $\rho(w)F^c(w|w)$ *is strictly decreasing in* w.

(b) $\rho(w)$ *is continuous in* w, *and* $F^c(t|w)$ *is continuous in both arguments.*

(c) For each $w \geq 0$, $F^c(t|w)$ *is continuous and strictly decreasing in* t, *and* $F^c(t|w) \leq K(t)$ *for some function* $K(t)$ *with* $lim_{t\to\infty}K(t) = 0$.

(d) $\rho(w)F^c(t|w)$ *is strictly decreasing in* w, *for all* $t \geq 0$ *and* $w \geq t$.

(e) $\rho(0)F^c(0|0) > 1$, *and* $\lim_{w\to\infty} \rho(w)F^c(w|w) < 1$.

Corresponding to Theorem 4.1 of the paper, we have:

Theorem 8.1. (existence and uniqueness) *Consider the fluid model specified above.*

(a) Suppose that Condition 8.1(a) *holds. Then there is at most one equilibrium delay* $\tilde{w}_e$*. It is characterized by* $\rho(\tilde{w}_e)F^c(\tilde{w}_e|\tilde{w}_e) = 1$.

(b) Suppose Condition 8.1 *holds. Then there exists a unique equilibrium delay* $\tilde{w}_e$*.*

Proof. Part (a) is immediate from the required equality in (8.1). As for the existence claim in (b), existence of a solution to equation (8.1) follows from 8.1(b,e) using the intermediate-value theorem, and the required inequality in (8.2) is then satisfied by the strict monotonicity of $F^c(t|w)$ as per Condition 8.1(c). ■

We assume henceforth that our model satisfies Condition 8.1.

Consider next the iterative algorithm: $w_{k+1} = d(w_k)$.

Theorem 8.2. (properties of the iteration)

(a) For each delay announcement w *with* $0 \leq w < \tilde{w}_e$*, we have* $d(w) > w$*.*

(b) For each delay announcement w *with* $w > \tilde{w}_e$*, we have* $d(w) < \tilde{w}_e$*.*

(c) If $w_2 > w_1 > \tilde{w}_e$ *then* $d(w_2) \leq d(w_1)$*, where the inequality is strict if* $d(w_1) > 0$*.*

(d) Suppose that $\rho(w)F^c(t|w)$ *is strictly decreasing in* w *for each* $t \geq 0$*. If* $w_2 > w_1$ *then* $d(w_2) \leq d(w_1)$*, where the inequality is strict if* $d(w_1) > 0$*.*

(e) If $d(w) < \tilde{w}_e$ *for all* w *with* $0 \leq w < \tilde{w}_e$*, then* $w_k \uparrow \tilde{w}_e$ *as* $k \to \infty$*.*

(f) We have $d(w) < \tilde{w}_e$ *for all* $0 \leq w < \tilde{w}_e$*, as needed in part* (e) *above, if and only if*

$$\rho(w)F^c(\tilde{w}_e|w) < 1 \quad \textit{for all} \quad 0 \leq w < \tilde{w}_e \ . \tag{8.3}$$

Proof. (a) Recall that the equilibrium delay $\tilde{w}_e$ satisfies $\rho(\tilde{w}_e)F^c(\tilde{w}_e|\tilde{w}_e) = 1$. By Condition 8.1(a) we have that $\rho(w)F^c(w|w) > 1$ for $w < \tilde{w}_e$. Therefore, by monotonicity of $F^c(t|w)$ the solution $d(w)$ to equation in (4.2) of the main paper satisfies $d(w) > w$.

(b) From $\rho(\tilde{w}_e)F^c(\tilde{w}_e|\tilde{w}_e) = 1$ and Condition 8.1(d) it follows that $\rho(w)F^c(\tilde{w}_e|w) = 1$. The monotonicity of $F^c(t|w)$ in t now implies that $d(w) < w$.

(c) If $d(w_1) = 0$ then $\rho(w_1)F^c(0|w_1) \leq 1$. Condition 8.1(d) thus implies that $\rho(w_2)F^c(0|w_2) < 1$, so that $d(w_2) = 0$. If $d(w_1) > 0$ then $\rho(w_1)F^c(d(w_1)|w_1) = 1$. Note that $d(w_1) < w_1$ and $d(w_2) < w_1$ by part (b) of this Theorem. Condition 8.1(d) then implies that $\rho(w_2)F^c(d(w_1)|w_2) < 1$. Thus, $d(w_2) < d(w_1)$.

(d) The proof is similar to that of part (c).

(e) The proof is identical to part (e) of Theorem 5.1 in the paper. Continuity of the solution $d(w)$ to $\rho(w)F^c(d(w)|w) = 1$ follows from the continuity and monotonicity properties in Condition 8.1(b,c).

(f) Immediate by the equilibrium condition and the strict monotonicity of $F^c(t|w)$, as per Condition 8.1(c). ∎

We next consider the damped iteration scheme mentioned in Remark 5.1.

Theorem 8.3. (monotone convergence of the damped iteration) *Consider the iteration* $w_{k+1} = d_p(w_k)$*, where*

$$d_p(w) = w + p(d(w) - w) \tag{8.4}$$

and $0 \leq p < 1$*.*

(a) If $d_p(w) < \tilde{w}_e$ *for all* $0 \leq w < \tilde{w}_e$*, then* $w_k \uparrow \tilde{w}_e$ *as* $k \to \infty$*.*

(b) The condition of part (a) holds if and only if

$$\rho(w)F^c\left(w + \frac{\tilde{w}_e - w}{p}|w\right) < 1 \quad \textit{for all} \quad 0 \leq w < \tilde{w}_e\ . \tag{8.5}$$

(c) Suppose that $\alpha \triangleq \frac{d}{dt}F^c(t|\tilde{w}_e)|_{t=\tilde{w}_e} < 0$ *and* $\beta \triangleq \frac{d}{dw}F^c(\tilde{w}_e|w)|_{w=\tilde{w}_e} < \infty$*. Then condition* (8.5) *is satisfied for all* p *small enough.*

Proof. (a) The proof of this part is similar to that of Theorem 7.1(e).

(b) By (8.4), $d_p(w) < \tilde{w}_e$ is equivalent to $d(w) < w + (\tilde{w}_e - w)/p$. But since there exists a unique response function $d(w)$ by Theorem 5.1 (a) of the main paper, by monotonicity of $F^c(t|w)$ in t the last inequality is equivalent to the inequality in (8.5).

(c) Define $g_p(w) = \rho(w)F^c(w + (\tilde{w}_e - w)/p|w)$. Note that $g_p(\tilde{w}_e) = 1$ by the equilibrium condition, and that

$$\frac{d}{dw}g_p(w)|_{w=\tilde{w}_e} \leq \rho(\tilde{w}_e)\left(\alpha + \frac{p-1}{p}\beta\right)\ . \tag{8.6}$$

Since $\alpha < 0$ then the latter is positive for some p_1 small enough. Therefore $g_{p_1}(w) < 1$ on some interval $\tilde{w}_e - \epsilon < w < \tilde{w}_e$ with $\epsilon > 0$. Since g_p is monotone increasing in p, the same holds for all $p \leq p_1$. On the other hand, for $w \leq \tilde{w}_e - \epsilon$ we have that

$$g_p(w) \leq \rho(0)F^c\left(\frac{\epsilon}{p}|w\right)$$

which uniformly converges to 0 as $p \downarrow 0$, by the bound in Condition 8.1(c). If follows that for p small enough, say $p \leq p_2$, we have that $g_p(w) < 1$ for all $w \leq \tilde{w}_e - \epsilon$. Thus, Condition 8.5 holds for all $p \leq \min\{p_1, p_2\}$. ∎

9. Conclusions

To investigate the impact on aggregate system performance of making delay announcements, either state-dependent or fixed, we proposed the fluid model approximating the $G/GI/s+GI$ queueing model with customer abandonment, based on Whitt (2006a), and the iterative numerical algorithm (INA) based on the approximate numerical algorithm for the $M/GI/s+GI$ model with an FD announcement, based on Whitt (2005). We showed how these approaches can be adapted to treat delay announcements.

Simulation experiments in which we announce the delay of the last customer to enter service show that these approximate descriptions of aggregate performance, based on fixed delay (FD) announcements, are quite accurate in an overloaded regime. For describing the aggregate performance, the numerical algorithm provides greater accuracy, but the fluid model yields simple closed-form expressions that make it possible to analyze other related issues, as illustrated by the short investigation of biased announcements.

In the context of the fluid model, we established conditions for the existence of a unique equilibrium delay announcement and provided conditions for a natural iteration to converge to that equilibrium delay. We gave explicit formulas in the case of the all-exponential model. We showed that multiple fluid equilibria can exist if regularity conditions are not satisfied.

Many interesting directions for research remain. For example, it remains to investigate and model actual customer response to announcements in practice, continuing work in progress by Feigin (2005). For practical application, we need to determine the critical model elements: the balking cdf B and the conditional time-to-abandon cdf $F(t|w)$. If we use model (1.1), then we need to determine the two-abandonment cdf's C and D appearing in (1.1) as well as the basic queueing model elements. We need to better understand how customers actually do respond to different kinds of announcements. With that information, hopefully the present paper will help to understand the impact of customer behavior upon aggregate system performance, including throughput, abandonment rate, queue lengths and delays. As a consequence, hopefully contact centers will be able to improve both efficiency and the quality of service provided, giving customers a better service experience, without expending unnecessary resources.

References

Armony M. and C. Maglaras C. 2004a. On customer contact centers with a call-back option: customer decisions, routing rules and system design. *Operations Research* 52(2), 271–292.

Armony M. and C. Maglaras C. 2004b. Contact centers with a call-back option and real-time delay information. *Operations Research* 52(4), 527–545.

Armony, M., N. Shimkin and W. Whitt. 2008. The impact of delay announcements in many-server queues with abandonments. *Operations Research*, forthcoming. Available at http://columbia.edu/∼ww2040.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Statist. Assoc.* 100, 36–50.

Carmon, Z., J. G. Shanthikumar and T. Carmon. 1995. A psychological perspective on service segmentation models: the significance of accounting for customers' perceptions of waiting and service. *Management Science* 41, 1806-1815.

Duenyas, I. and W. Hopp. 1995. Quoting lead times. *Management Science* 41, 43–57.

Durrande-Moreau, A. 1999. Waiting for service: ten years of empirical research. *International Journal of Service Industry Management* 10, 171–189.

Feigin, P. 2006. Analysis of customer patience in a bank call center. Working paper in preparation, The Technion, Haifa, Israel.

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

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

Guo, P. and P. Zipkin 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* 53 (6) 962–970.

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

Hassin, R. and M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer.

Hui, M. and D. Tse. 1996. What to tell customers in waits of different lengths: an integrative model of service evaluation. *J. Marketing* 60, 81–90.

Ibrahim, R. E. and W. Whitt. 2008a. Real-time delay estimation based on delay history in the $GI/M/s$ queue. *Manufacturing Service Oper. Management*, forthcoming. Available at http://columbia.edu/∼ww2040.

Ibrahim, R. E. and W. Whitt. 2008b. Real-time delay estimation in overloaded multiserver queues with abandonment. Working paper, Columbia University, New York City. Available at http://columbia.edu/∼ww2040.

Larson, R. C. 1987. Perspectives on queues: social justice and the psychology of queueing. *Operations Research* 35, 895–905.

Maister, D. H. 1985. The psychology of waiting lines. In *The Service Encounter*, J. A. Czepiel, R. M. Solomon and F.C. Surprenant (eds.), Lexington Books, Lexington, MA, 113–123.

Mandelbaum, A. and 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.

Munichor, N. and A. Rafaeli. 2006. Numbers or apologies? Customer reactions to tele-waiting time fillers. *J. Applied Psychology* 92 (2) 511–518.

Nakibly, E. 2002. *Predicting Waiting Times in Telephone Service Systems*, MS thesis, the Technion, Haifa, Israel.

Plambeck, E. L. 2004. Optimal leadtime differentiation via diffusion approximations. *Operations Research* 52, 213–228.

Ren, Z. J. and Y.-P. Zhou. 2006. Call center outsourcing: coordinated staffing level and service quality. *Management Science*, forthcoming.

Shimkin, N. and A. Mandelbaum. 2004. Rational abandonment from tele-queues: nonlinear waiting cost with heterogeneous preferences. *Queueing Systems* 47, 117–146.

Spearman, M. and R. Zhang. 1999. Optimal lead time policies. *Management Science* 45, 290–295.

Whitt, W. 1999a. Improving service by informing customers about anticipated delays. *Management Science* 45, 192–207.

Whitt, W. 1999b. Predicting queueing delays. *Management Science* 45, 870–888.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50, 1449–1461.

Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Science* 51, 221–235.

Whitt, W. 2006a. Fluid models for multi-server queues with abandonments. *Operations Research* 54, 37–54.

Whitt, W. 2006b. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* 15 (1), 88–102.

Zeltyn S. and A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems* 51 (3/4) 361–402.

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