OPERATIONS RESEARCH
Vol. 57, No. 1, January–February 2009, pp. 66–81
ISSN 0030-364X | EISSN 1526-5463 | 09 | 5701 | 0066

informs

DOI 10.1287/opre.1080.0533

The Impact of Delay Announcements in Many-Server Queues with Abandonment

Mor Armony
Stern School, New York University, New York, New York 10012, marmony@stern.nyu.edu

Nahum Shimkin
Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel, shimkin@ee.technion.ac.il

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027, ww2040@columbia.edu

This paper studies the performance impact of making delay announcements to arriving customers who must wait before starting service in a many-server queue with customer abandonment. The queue is assumed to be invisible to waiting customers, as in most customer contact centers, when contact is made by telephone, e-mail, or instant messaging. Customers who must wait are told upon arrival either the delay of the last customer to enter service or an appropriate average delay. Models for the customer response are proposed. For a rough-cut performance analysis, prior to detailed simulation, two approximations are proposed: (1) the equilibrium delay in a deterministic fluid model, and (2) the equilibrium steady-state delay in a stochastic model with fixed delay announcements. These approximations are shown to be effective in overloaded regimes, where delay announcements are important, by making comparisons with simulations. Within the fluid model framework, conditions are established for the existence and uniqueness of an equilibrium delay, where the actual delay coincides with the announced delay. Multiple equilibria can occur if a key monotonicity condition is violated.

Subject classifications: queues: applications, approximations, balking and reneging.
Area of review: Stochastic Models.
History: Received November 2006; revision received May 2007; accepted November 2007.
Published online in *Articles in Advance* September 17, 2008.

1. Introduction

We study the performance impact of making delay announcements in customer contact centers (telephone call centers) and other many-server service systems with invisible queues. (For general background on contact centers, see Brown et al. 2005 and Gans et al. 2003.) With invisible queues, delay announcements provide prospective customers with an estimate of the time they will have to wait before they can start service if they decide to enter the system, which they otherwise would not have. Making delay announcements is important because it is a relatively inexpensive way to improve the customer service experience. A maxim in the psychology of waiting is that "uncertain waits feel longer than known finite waits;" see Maister (1985) and the many papers that cite it; e.g., Carmon et al. (1995) and Durrande-Moreau (1999).

With high quality of service in many-server queues, delays tend to be negligible (as can be verified with the $M/M/s$ model), so that there is relatively little incentive to provide delay announcements. However, if the system can be overloaded for periods of time, then delays can become significant. We think that it is important to distinguish between two different cases: The first case is the ideal service scenario in which the service provider has the resources and the flexibility to respond quickly to adjust the staffing to meet unexpected high demand. In this ideal case, the common aim of an announcement is to explain the unusual circumstance and encourage the customer to remain because help will soon be on its way.

We are motivated by the less ideal second case, common in service-oriented (as opposed to revenue-generating) call centers, in which the service provider has limited ability to respond to unexpected high demand in the short run. One promising way to respond to this excess demand is to provide a call-back option, as in Armony and Maglaras (2004a, b), but an appropriate delay-announcement scheme may be a less costly "low-tech" way to achieve the same objective. We assume that the goal of the delay announcements, in addition to informing the customers, is to induce some customers to balk (leave immediately without waiting) or abandon earlier, hopefully to retry later when the system is more lightly loaded, and as a consequence reduce the delays of served customers, without significantly altering the number of customers that receive service. We demonstrate this important performance consequence of delay announcements through mathematical models of

many-server queues. We focus on overloaded regimes, where delay announcements are especially important.

We emphasize the equilibrium behavior associated with delay announcements, where the customers respond to the announcements, the system performance depends on the customer response, and the announcements depend on the system performance. Background on equilibrium behavior in queueing systems is provided by Hassin and Haviv (2003). The effect of customer delay *expectations* on the abandonment profile and resulting system equilibrium is studied in several papers, where the dependence of customer patience on expected wait is captured either through a rational decision model (Shimkin and Mandelbaum 2004 and references therein) or a descriptive behavior model (Zohar et al. 2002). The approach we take here is in the spirit of Zohar et al. (2002).

There is a substantial literature on delay announcements if we allow a broader class of models. For example, Duenyas and Hopp (1995), Spearman and Zhang (1999), and Plambeck (2004) study lead-time announcements in production systems. Whitt (1999a) considers the effect of delay announcement in a many-server ($M/M/s$) model, comparing the no-information case with reneging to the full-information case, where customers either balk immediately or remain in queue until served. Guo and Zipkin (2007) consider an $M/M/1$ system with only balking, under different levels of information. Both of these papers indicate the positive effect of delay information upon system performance. The effect of real-time delay estimates on a many-server queueing system with a call-back option is studied in Armony and Maglaras (2004a, b).

The DLS and FD Announcement Schemes. We primarily model the performance consequence of a delay announcement, but we also have some suggestions for the delay announcement itself. In particular, we propose two specific announcement schemes: (i) announcing the delay of the last customer to enter service (DLS), and (ii) making a fixed delay (FD) announcement, corresponding to a long-run average delay (appropriate for the time in question, allowing for a time-varying arrival rate). The DLS announcement is closely related to the longest waiting time of any customer in queue, which was used as an announcement in an Israeli bank studied by Mandelbaum et al. (2000) and mentioned as a candidate delay announcement by Nakibly (2002). We discuss motivation for DLS announcements further in §2.

We want to understand how DLS and FD announcements, or other natural delay announcements, will affect system performance. Accordingly, we model customer response to delay announcements. We do not examine data of any system with delay announcements, but we provide a modelling framework for looking at such data. Data revealing customer response to announcements are being analyzed by Feigin (2006). Some related laboratory experiments are described in Munichor and Rafaeli (2006).

As indicated before, we assume that each customer who cannot enter service immediately upon arrival is given an estimate w (a single number) of the waiting time before he can begin service. We are thinking of this announced delay being DLS or FD, but it could be obtained in other ways, e.g., by one of the alternative estimators in Whitt (1999b).

We model the customer response to the announced delay w by two functions: $B(w)$ and $F(t \mid w)$. Given a delay announcement of w, we assume that the customer balks with probability $B(w)$. We assume that B is a cumulative distribution function (c.d.f.), so that the customer is more likely to balk as the announced delay increases. If that customer does not balk, then that customer will abandon before time t if he has not begun service by that time with probability $F(t \mid w)$. We assume that $F(t \mid w)$ is a c.d.f. as a function of t with $F(0 \mid w) = 0$ for each w. Consistent with our focus on invisible queues, we assume that the reactions of successive customers are conditionally independent, given their delay announcements.

This model greatly generalizes the model of customer response to delay announcements proposed by Whitt (1999a). There, all abandonment without an announcement was replaced by balking with the announcement, and all distributions were assumed to be exponential, so that all models became Markovian. On the other hand, we could go further, as in Guo and Zipkin (2007), and derive our balking and abandonment functions $B(w)$ and $F(t \mid w)$ by considering customers maximizing their expected utility from service and waiting.

The Performance Impact of Delay Announcements. We aim to understand the performance impact of delay announcements in the setting of a conventional $M/GI/s{+}GI$ model. (However, DLS and FD announcements are appealing, in large part, because they apply much more generally.) The $M/GI/s{+}GI$ model has a Poisson arrival process (the M), s homogeneous servers working in parallel, unlimited waiting space, and the first-come first-served (FCFS) service discipline. The first GI means that successive service times are independent and identically distributed (i.i.d.) with a general c.d.f. G. The second GI (after the $+$) means that successive customers have i.i.d. times until they will abandon if service has not yet begun, again with a general c.d.f. F (where no announcement is given). The service times, times to abandon, and arrival process are assumed to be mutually independent.

For the $M/GI/s{+}GI$ model without delay announcements, the stochastic process representing the number of customers in the system as a function of time and other standard stochastic processes have proper limiting steady-state distributions for any arrival rate under minor regularity conditions because of customer abandonment. Under further regularity conditions, this should also be true with delay announcements and the i.i.d. customer response defined above, but with the addition of the customer response, the limiting steady-state behavior involves a complex equilibrium, as in Hassin and Haviv (2003)

and Shimkin and Mandelbaum (2004), because the average announced delay should agree with the average experienced delay. There are many open questions about system dynamics: (i) Under what conditions does there exist an equilibrium steady state? (ii) If there is an equilibrium, when is it unique? When can there be multiple equilibria? (iii) How do the stochastic processes evolve as a function of the initial conditions?

Simulation is ideally suited to analyze this delay-announcement problem, and we will use it here. But our purpose is to supplement simulation by developing more revealing and more efficient methods to approximately determine the equilibrium steady-state performance of the $M/GI/s+GI$ model with delay announcements. Approximations are needed because direct mathematical analysis is difficult. Even for the totally Markovian $M/M/s+M$ base model, a full-state description must include the announcements received and possibly the elapsed waiting times of each customer in queue.

Two Approximation Methods. We propose two methods to approximate the steady-state performance with delay announcements, to use in addition to simulation. Both approximation methods act as if all customers receive the same fixed deterministic delay announcement. For both methods, we find an equilibrium in the approximate model, where the expected steady-state delay coincides with the delay announcement. Those equilibrium delays for the approximate models are our proposed approximations for the expected steady-state delay with DLS and FD announcements.

The first approximation method is a *deterministic fluid model*, extending the fluid model approximation for the $G/GI/s+GI$ model in Whitt (2006). The fluid model is appealing because it is remarkably tractable. The fluid model provides useful insight here only in an overloaded regime, where the traffic intensity exceeds one, but that is when delay announcements are especially important.

First, for a general all-exponential stochastic model introduced in §5, where the customer response has exponential structure and the underlying queueing model is $M/M/s+M$, there is a simple equation any fluid equilibrium must satisfy; see (5.2). For a natural special case, there exists a unique equilibrium for the fluid model and it has a simple explicit formula; see (5.3). On the other hand, for the more general stochastic model, if regularity conditions are not imposed, there can be multiple equilibria for the approximating fluid model. We show that corresponding multiple equilibria hold in the corresponding stochastic models.

For the general fluid model, we prove that there exists a unique equilibrium delay under very general regularity conditions. That strongly supports our conjecture that a unique equilibrium steady-state delay exists with DLS and FD announcements under the same conditions. We also show how to perform a perturbation analysis of the fluid equilibrium delay to estimate its sensitivity to stochastic fluctuations (which are not considered directly). We are thus better able to understand the observed performance of the fluid model when compared to simulations.

For the fluid model, all served customers necessarily wait the same deterministic time. That fixed-delay property of the fluid model motivates using an FD announcement with the $M/GI/s+GI$ model as an approximation. (The fluid model proves its worth by that insight alone!) Our second approximation method is an *iterative numerical algorithm* (INA) for calculating the approximate steady-state performance in the $M/GI/s+GI$ model, based on Whitt (2005), assuming an FD announcement.

For the INA, we use the numerical algorithm for approximating the steady-state performance of the $M/GI/s+GI$ model in Whitt (2005). The first step in that algorithm is to approximate the given $M/GI/s+GI$ model by an associated Markovian $M/M/s+M(n)$ model with state-dependent abandonment rates. The second step is to numerically solve for the steady-state performance measures in the $M/M/s+M(n)$ model, which begins with the number in system, because that is a birth-and-death process. That same approximation applies with the FD announcements, because the delay announcements produce a new Poisson arrival process and a new time-to-abandon distribution, and thus a new $M/GI/s+GI$ model. Thus, for our model of customer response, the algorithm in Whitt (2005) applies with FD announcements just as without delay announcements; we just need to iteratively apply the algorithm to find the equilibrium FD announcement, where the announced fixed delay coincides with the expected conditional steady-state delay, given that the customer is served.

We could also have used alternative numerical algorithms in our INA, such as an exact numerical algorithm for the $M/M/s+GI$ model and heavy-traffic approximations developed by Zeltyn and Mandelbaum (2005). The main point, to be shown, is that the $M/M/s+GI$ model with an FD announcement yields a good approximation for the corresponding model with a state-dependent DLS delay announcement as well as with an FD announcement.

Contributions. We make several contributions in this paper. First, we suggest two specific single-number delay announcement schemes: FD and DLS. Second, we introduce a model of customer response to these announcements in the setting of invisible queues, based on the c.d.f.'s $B(w)$ and $F(t \mid w)$. Third, we use simulation to study the equilibrium behavior of these DLS and FD announcements in the $M/GI/s+GI$ model. The simulations show that the state-dependent DLS announcements are more effective having smaller variance. Fourth, we introduce and analyze two approximation methods for analyzing the equilibrium performance of the stochastic model with customer response: a deterministic fluid model, extending Whitt (2006), and an INA for approximately calculating the steady-state behavior, extending Whitt (2005), both based on using FD. Fifth, we obtain insight into the equilibrium behavior in this

setting. For example, we show how our model can be used to explore the consequences of biased announcements. Finally, we provide a framework for empirical research that will investigate the actual human response to delay announcements. More broadly, we advance research bridging the behavioral and quantitative-modelling traditions.

Organization of this Paper. We start in §2 by discussing the motivation for DLS and FD announcements. In §3, we describe the fluid model, both with and without delay announcements, focusing on the overloaded regime. In §4, we establish basic properties of the fluid model. In particular, we show that there exists a unique equilibrium fluid delay under very general regularity conditions. In §5, we introduce some all-exponential stochastic models, which we will consider in our numerical comparisons. In §6, we conduct experiments, comparing the fluid and INA approximation methods to simulation for the all-exponential stochastic models. In §7, we perform perturbation analysis of the all-exponential fluid model to understand how the fluid model performance is affected by ignoring stochastic fluctuations. In §8, we give an example of multiple equilibria that are possible when the regularity conditions are not satisfied. In §9, we make some concluding remarks. Additional material appears in an e-companion maintained by the journal (available at http://or.journal.informs.org/) and an online supplement (Armony et al. 2007). In the e-companion, we use the fluid model to study the impact of biased delay announcements, where the announcement is designed to differ from the actual delay. We also briefly discuss the consequence of increasing patience in response to delay announcements.

2. Motivation for the DLS and FD Announcements

When making delay announcements, we think that it is important to identify two cases, depending on the customer's ability to process information. A customer's ability to process information can be either quite low, as in a conventional telephone call, or quite high, as in service over the Internet, where information can be presented on the screen and read while waiting. With limited customer-information-processing ability, we may want to restrict the announcement to only a few numbers, perhaps only one.

In either case, it is not easy to make reliable delay estimates because future delays are inevitably uncertain. However, we contend that useful delay announcements can be made without great difficulty, even if limited to a single number. In particular, we propose the DLS and FD announcements as simple and robust single-number schemes. We focus especially on the DLS scheme.

Of course, no announcement at all is made if a customer can enter service immediately. Moreover, as observed by Hui and Tse (1996), delay announcements are more important when the delays are long, so that the actual announcement can be tuned to the estimated size of the delay. For example, when the delay is likely to be short, the customer might be told, "We should be able to serve you soon; the last customer to enter service waited less than one minute." On the other hand, when the delay is likely to be long, the customer might be told, "We are currently experiencing unexpected high demand; the last customer to enter service had to wait x minutes before beginning service. We will do our best to serve you without excessive delay, but you might want to try again later."

The DLS scheme is appealing for several reasons. First, the DLS announcements are transparent, directly communicating historical experience, so that customers are not left wondering how the estimate was made. Second, the DLS scheme extends directly to multiclass skill-based-routing scenarios; then, we can announce the delay of the last customer to enter service of that class. Finally, the DLS scheme makes no specific model assumptions. It can be used with conventional models without having to know the number of servers or the service-time distribution. The DLS scheme even allows for unconventional service mechanisms including heterogeneous servers, a random number of servers used per customer, and service interruptions (where service is conducted over several disjoint time intervals). Such phenomena commonly arise in contact centers, such as those providing technical support over multiple media. The DLS scheme also responds automatically to dynamic time-varying conditions.

We develop mathematical models showing that DLS announcements are effective, first, by having reasonable predictive power (e.g., low mean squared error) and, finally, by achieving the desired goal: causing some customers to leave earlier without receiving service and reducing the delays of served customers. For the overloaded model considered in §6, the DLS announcements reduce the average delay of served customers by 50%.

We also suggest FD as a simple alternative to the DLS announcements. With FD, the single-number announcement would be based on the average day of recent customers to complete service. For the fluid model, it turns out that FD coincides with DLS. Simulation results show that DLS is more accurate than FD, but they are quite close.

As the customer information-processing ability increases, we may want to announce additional information. Natural additions (or alternatives) are (i) the average delay among the last k customers to enter service for some $k > 1$ (which becomes our FD announcement if k is suitably large), and (ii) the estimated average delay computed by multiplying the current queue length times the average time between successive customer departures. For scenarios in which customer-information-processing ability is higher, we also propose a *vector extension* of the DLS announcement scheme: to provide an estimate of the probability distribution of the delay as well as a point estimate, we propose announcing the delays of the last k customers to enter service, in temporal order, for some $k \geqslant 1$ (in addition to a few summary statistics, such as the mean). For display

on a computer screen, we might display histograms of the last k delays for various k. A further extension would be a longer record of past experienced delays, together with the times that these customers entered service.

In addition, we might provide additional state information, such as the current queue length. However, we emphasize that even a full description of the current system state does not include the delay history we are advocating. As emphasized by Larson (1987) and Munichor and Rafaeli (2006), it is also helpful for customers to see that progress is being made. Thus, on a computer screen, customers could be shown the evolving queue and their place in it (we do not study the impact of such additional feedback here). In some circumstances, it may be good to also tell the customers the service times and/or response times (waiting times plus service times) of the last k customers to complete service. Here, we focus on delays before starting service, assuming that we announce a single number immediately upon arrival if a customer cannot enter service immediately.

The single-number assumption applies directly to the DLS scheme for $k = 1$, but applies more generally to the DLS scheme with $k > 1$ if we understand w to be a one-dimensional summary of the k-dimensional vector of delays, such as the median or the mean. However, it remains to investigate if, and how, customer response to a vector delay announcement can be accurately summarized by a single number.

For any given announcement scheme, we should ask how customers will interpret the announcement. The DLS scheme is so transparent that there should be little ambiguity in the customer's mind, except for the possibility that the customer may doubt whether the information is truthful. The interpretation is easier for the customer if he is also told the queue length and his position in it through time while he is waiting. Here, we assume that there is only a single-number announcement immediately upon arrival. That leaves the queue invisible, and makes it reasonable to assume that customer responses are mutually independent.

For non-DLS announcements, most customers should recognize that the announcement is only an estimate, necessarily being subject to error. Customers should learn how to interpret the announcements through experience. It is important to recognize, though, that the customer response is likely to depend on the way the announcement is made, beyond just the number w itself. Recalling Maister's (1985) propositions about waiting, we recommend that the announcement attempt to explain both what has happened and what action management recommends, as well as remove uncertainty and reduce anxiety (we do not consider such issues further here).

3. Fluid Model

In this section, we review the fluid model introduced in Whitt (2006) and develop an extension for delay announcements.

An Approximation for a Many-Server Queueing Model. 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): $A \equiv \{A(t):t \geqslant 0\}$ 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 c.d.f., and F is the time-to-abandon c.d.f. 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 $G(t) \equiv P(S \leqslant t)$ and $F(t) \equiv P(T \leqslant t)$ for $t \geqslant 0$. 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.

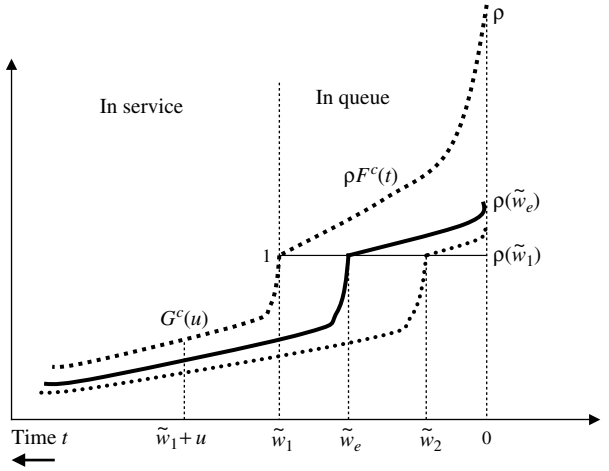
In this setting of the $G/GI/s+GI$ model with $\mu = 1$, the fluid model we use arises in the limit as

$$\lambda \to \infty \quad \text{and} \quad s \to \infty \quad \text{with } \rho \equiv \frac{\lambda}{s} \text{ held fixed.} \tag{3.1}$$

As the limit indicates, the fluid model is intended for scenarios with large s and λ. The parameter ρ defined in (3.1) is the traffic intensity in the original queueing model. It becomes the fluid arrival rate in the fluid model. The fluid model has been shown to be asymptotically correct in the limiting regime (3.1). There is a proviso, however: the asymptotic correctness has only been verified for a discrete-time analog of the general $G/GI/s+GI$ fluid model in Whitt (2006). Because 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 describes the evolution over time of the system, but we will only consider the steady-state behavior, under the condition that $\rho > 1$. 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. Whitt (2006) has shown that the fluid model provides a remarkably good approximation when λ and s are large and $\rho > 1$. For example, we might have $s = 100$ and $\rho = 1.2$ as in Table 1 of Whitt (2006). We anticipate that the fluid model will provide a useful approximation for queueing models with delay announcements in the same overloaded settings. However, we should be careful that the balking and abandonment not be so great that the system cease to be overloaded. The accuracy of the fluid approximation degrades when the system ceases to be heavily loaded.

Figure 1. Possible steady-state densities of fluid content: (i) without delay announcements (upper dotted curve), (ii) announcing the initial delay $\widetilde{w}_1$ (lower dotted curve), and (iii) with an equilibrium delay announcement $\widetilde{w}_e$ (solid curve).



The Steady-State Behavior Without Delay Announcements. Figure 1 depicts three possible steady-state distributions for the fluid model. Each curve in Figure 1 shows the steady-state density of fluid content that has been in the system for a period of length t as a function of t, where time t increases toward the left. It is not unreasonable to have t increase toward the left because t represents time in the *past*. We are looking at the system at one time in steady state. The plotted function at t represents current fluid content that arrived time t in the past.

From the fluid model, the approximate number of customers in the associated queueing system is obtained by multiplying by s: the arrival rate in the queueing system is ρs when the fluid arrival rate is ρ; the approximate queue length is $s\rho F^c(t)$ when the fluid queue content is $\rho F^c(t)$.

We start by focusing on the upper dotted curve, which depicts the steady-state behavior without any delay announcements. This fluid density is a deterministic function, but nevertheless the two model c.d.f.'s F and G play a prominent role in the description. At the right in Figure 1, we see a density of $\rho > 1$ at $t = 0$ for the upper dotted curve, which corresponds to the fluid arrival rate. Fluid abandons according to the c.d.f. F up until time $\widetilde{w}_1$ (subscript 1 denoting the first delay, without announcements). For $0 < t < \widetilde{w}_1$, a proportion $F(t)$ of the fluid that would have been in the system for t time units has abandoned, while the remaining proportion $F^c(t) \equiv 1 - F(t)$ remains in the system. The initial waiting time before served fluid enters service, $\widetilde{w}_1$, is determined by the requirement that

$$\rho F^c(\widetilde{w}_1) \leqslant 1 \quad \text{and} \quad \rho F^c(t) > 1 \quad \text{for } 0 \leqslant t < \widetilde{w}_1. \tag{3.2}$$

In (3.2), we allow the complementary c.d.f. (c.c.d.f.) F^c to have a discontinuity at $\widetilde{w}_1$. 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 $\widetilde{w}_1$. Thus, abandonment is occurring constantly at the rate $\rho - 1$.

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 to be able to clearly show the impact of delay announcements in this same scenario.

Although 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 with probability $F^c(t)$, provided that $0 < t < \widetilde{w}_1$. At time $\widetilde{w}_1$, customers enter service at rate 1 (because, at any given time, we assume that customers enter service before they consider abandoning). There could also be abandonment exactly at time $\widetilde{w}_1$ if the c.d.f. F has a jump at $\widetilde{w}_1$. The abandonment rate at time $\widetilde{w}_1$ is thus $\rho F^c(\widetilde{w}_1-) - 1$, assuming that $\rho F^c(\widetilde{w}_1-) > 1 \geqslant \rho F^c(\widetilde{w}_1)$. Hence, each customer abandons at time $\widetilde{w}_1$ with probability $F(\widetilde{w}_1-) - \rho^{-1}$, which will be zero unless F has a jump at $\widetilde{w}_1$.

The customer experience in service is described by the region of Figure 1 to the left of $t = \widetilde{w}_1$, i.e., for times $t > \widetilde{w}_1$. A proportion $G(u)$ of the fluid entering service after waiting for a time $\widetilde{w}_1$ will have completed service by time $\widetilde{w}_1 + u$. 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 $\widetilde{w}_1 + u$. The total fluid content in service at any time is

$$\int_{\widetilde{w}_1}^{\infty} G^c(u - \widetilde{w}_1)\, du = \int_0^{\infty} G^c(u)\, du = E[S] = 1; \tag{3.3}$$

the fluid content waiting in queue is

$$q \equiv \int_0^{\widetilde{w}_1} q(t)\, dt = \rho \int_0^{\widetilde{w}_1} F^c(u)\, du. \tag{3.4}$$

The expected or average waiting time for *all* fluid is

$$E[W_{all}] = \int_0^{\widetilde{w}_1} F^c(u)\, du = \frac{\widetilde{w}_1}{\rho} + \int_0^{\widetilde{w}_1} x dF(x) = \frac{q}{\rho}, \tag{3.5}$$

which of course is less than the waiting time $\widetilde{w}_1$ of the fluid that is served. (We remark that the corresponding formula (3.10) in Whitt 2006 is incorrect.) We regard W_{all} as a random variable because the experience of individual customers (atoms of fluid) is random.

Delay Announcements and Customer Response. We next consider making a delay announcement immediately

upon arrival to arriving customers if they must wait. We start by announcing the waiting time served customers have been experiencing without delay announcements, $\widetilde{w}_1$, which is the solution to (3.2). However, we now must consider the impact on customer behavior of making such an announcement. We assume that a proportion $B(\widetilde{w}_1)$ will balk in response to a delay announcement $\widetilde{w}_1$, where B is our balking c.d.f. We mention one possible form for the balking c.d.f.

DEFINITION 3.1 (INFORMATION-CONSISTENT BALKING). If $B^c(w) = F^c(w)$ for all $w \geqslant 0$, 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. We also have to specify how customers who decide to wait respond to the announcement. That is done via the conditional time-to-abandon c.d.f. $F(t \mid w)$, given any announced delay w. Because B already accounts for balking, we assume that $F(0 \mid w) = 0$ for all w. As before, we assume that 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, giving the experienced delay $d(w)$ associated with announcement w, if for each $w \geqslant 0$, either (i) $\rho B^c(w) \leqslant 1$ and $d(w) = 0$, or (ii) $\rho B^c(w) > 1$ and

$$\rho B^c(w) F^c(d(w) \mid w) \leqslant 1 \quad \text{and} \quad \rho B^c(w) F^c(t \mid w) > 1 \quad \text{for } 0 \leqslant t < d(w). \tag{3.6}$$

Because $F(\cdot \mid w)$ is assumed to be a c.d.f. for each $w \geqslant 0$, the response delay function d is well defined. Consequently, served fluid waits $\widetilde{w}_2 \equiv d(\widetilde{w}_1)$ in response to the first delay announcement $\widetilde{w}_1$. Thus, assuming that $\rho B^c(w) > 1$, the waiting fluid density (in queue) that has been in the system for time t becomes $\rho B^c(\widetilde{w}_1) F^c(t \mid \widetilde{w}_1)$ for $0 < t < \widetilde{w}_2$, where $\widetilde{w}_2 = d(\widetilde{w}_1)$ satisfies

$$\rho B^c(\widetilde{w}_1) F^c(\widetilde{w}_2 \mid \widetilde{w}_1) \leqslant 1 \quad \text{and} \quad \rho B^c(\widetilde{w}_1) F^c(t \mid \widetilde{w}_1) > 1 \quad \text{for } 0 \leqslant t < \widetilde{w}_2, \tag{3.7}$$

paralleling (3.2). Fluid enters service at rate 1 at time $\widetilde{w}_2$. Paralleling (3.4), the total queue content now is

$$q \equiv q(\widetilde{w}_1) = \int_0^{\widetilde{w}_2} q(t \mid \widetilde{w}_1)\, dt = \rho B^c(\widetilde{w}_1) \int_0^{\widetilde{w}_2} F^c(t \mid \widetilde{w}_1)\, dt. \tag{3.8}$$

The lower dotted curve in Figure 1 shows the steady-state fluid distribution in response to the initial announcement $\widetilde{w}_1$. The effective arrival rate is reduced from $\rho > 1$ to $\rho(\widetilde{w}_1) \equiv \rho B^c(\widetilde{w}_1)$ due to balking, and thereafter (provided that $\rho(\widetilde{w}_1) > 1$) abandonment occurs before time $\widetilde{w}_2$ at a slower rate. At time $\widetilde{w}_2$, the fluid density reaches level 1 and customers enter service at rate 1.

From Figure 1, we see that the system has benefitted from the delay announcement because the lower dotted curve is below the upper dotted curve. The fluid throughput is still at the maximum value 1, but the waiting has been reduced. All customers who are served now wait $\widetilde{w}_2$ instead of $\widetilde{w}_1$. The abandoning customers wait less as well. The most impatient customers elect to balk when they get the delay message. The balking rate is $\rho B(\widetilde{w}_1)$. Those customers who decide not to balk abandon at a slower rate, but the remaining abandonments occur by time $\widetilde{w}_2$.

An Equilibrium Fluid Delay. However, there is a *consistency problem*. The announced delay for served customers, $\widetilde{w}_1$, is not consistent with the actual delay for served customers, $\widetilde{w}_2$, 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., $\widetilde{w}_e$ is an equilibrium delay if either (i) $\rho B^c(0) \leqslant 1$ and $\widetilde{w}_e = 0$, or (ii) $\rho B^c(0) > 1$ and

$$\rho B^c(\widetilde{w}_e) F^c(\widetilde{w}_e \mid \widetilde{w}_e) \leqslant 1 \quad \text{and} \quad \rho B^c(\widetilde{w}_e) F^c(t \mid \widetilde{w}_e) > 1 \quad \text{for } 0 \leqslant t < \widetilde{w}_e. \tag{3.9}$$

The solid curve in Figure 1 shows what might happen if we use an equilibrium delay announcement (depending on the detailed model elements). The equilibrium delay announcement $\widetilde{w}_e$ is less than the original delay $\widetilde{w}_1$ without an announcement, but it is greater than the delay $\widetilde{w}_2$ in response to the announced delay $\widetilde{w}_1$. We still achieve maximum throughput and we still reduce delays compared to what we achieve with no announcement at all, but we cannot do as well as the lower dotted curve, but that is understandable because the response delay $\widetilde{w}_2$ associated with announcement of $\widetilde{w}_1$ is inconsistent, which we regard as not sustainable (an assumption about human behavior). With the equilibrium delay $\widetilde{w}_e$, the effective arrival rate is reduced from $\rho > 1$ to $\rho(\widetilde{w}_e) = \rho B^c(\widetilde{w}_e) > 1$ due to balking, and thereafter abandonment occurs before time $\widetilde{w}_e$. At time $\widetilde{w}_e$, the fluid density reaches level 1 and all waiting customers enter service.

4. Basic Properties of the Fluid Model

We now establish the basic properties of the fluid model. Under regularity conditions, there exists a unique fluid equilibrium delay, but some care is needed.

EXAMPLE 4.1 (PATHOLOGICAL EXAMPLE). Suppose that $B(w) = F(w)$ for all w, so that we have information-consistent balking as in Definition 3.1. Suppose that

$\rho B^c(w^*) > 1$ for some $w^* > 0$. Moreover, suppose that we have an extreme abandonment response to the delay announcement: Suppose that $F^c(t \mid w) = 1$, $0 \leqslant t \leqslant w$ and $F^c(t \mid w) = 0$ for all $t > w$, implying that $d(w) = w$, i.e., that all customers who do not balk abandon precisely at time w, whatever is the announced wait w, provided $0 < w \leqslant w^*$, which we henceforth assume. We now use our previous assumption that, at time w, we allow customers to enter service first at time w. Then, customers enter service at rate 1 at time w, while the remaining customers abandon. That is, we have abandonment at rate $\rho B^c(w) - 1$ at time w. All customers wait precisely w, whether they get served or abandon. Customers enter service at rate 1 at time w, but every delay announcement w in the interval $[0, w^*]$ is an equilibrium delay announcement.

Condition 4.1 (Regularity Conditions). (a) $F^c(t \mid w) \equiv 1 - F(t \mid w)$ is a continuous strictly decreasing c.c.d.f. as a function of t with $F^c(0 \mid w) = 1$ for each $w \geqslant 0$.

(b) $B^c(w)F^c(w \mid w)$ is strictly decreasing in w.

(c) $B^c(w)$ is continuous in w.

(d) $F^c(t \mid w)$ is continuous in w for each t.

Theorem 4.1 (Existence and Uniqueness). *Consider the fluid model specified above. Assume that* $\rho B^c(0) > 1$.

(a) *If Condition* 4.1(a) *holds, then the delay response function* d *defined in Definition* 3.2 *satisfies*

$$\rho B^c(w)F^c(d(w) \mid w) = 1 \quad \textit{and} \quad \rho B^c(w)F^c(t \mid w) > 1 \quad \textit{for } 0 \leqslant t < d(w). \tag{4.1}$$

(b) *If, in addition, Condition* 4.1(b) *holds, then there is at most one equilibrium delay* $\widetilde{w}_e$*, as defined in Definition* 3.3*. If it exists, it satisfies* $\widetilde{w}_e > 0$*, and*

$$\rho B^c(\widetilde{w}_e)F^c(\widetilde{w}_e \mid \widetilde{w}_e) = 1 \quad \textit{and} \quad \rho B^c(\widetilde{w}_e)F^c(t \mid \widetilde{w}_e) > 1 \quad \textit{for } 0 \leqslant t < \widetilde{w}_e. \tag{4.2}$$

(c) *If Conditions* 4.1(a), (c)*, and* (d) *hold, then there exists at least one equilibrium delay* $\widetilde{w}_e$*.*

(d) *If all parts of Condition* 4.1 *hold, then there exists a unique equilibrium delay* $\widetilde{w}_e$*.*

Proof. For part (a), existence of a solution to the equation in (4.1) follows from the intermediate-value theorem, using the continuity in Condition 4.1(a). The inequality in (4.1) follows from the strict monotonicity in Condition 4.1(a). Part (b) is immediate from Condition 4.1(b) and the required equality in (4.1). As for the existence claim in (c), existence of a solution to the equation in (4.2) follows from Conditions 4.1(c) and (d) using the intermediate-value theorem. The required inequality in (4.2) is then satisfied by the strict monotonicity of $F^c(t \mid w)$ by Condition 4.1(a). Part (d) combines parts (b) and (c). □

The regularity conditions in Condition 4.1 ensuring a unique fluid equilibrium seem reasonable, but we should be cautious about human response. From a practical perspective, Condition 4.1(b) might be questioned. Violation of (b) can lead to multiple equilibria; see §8.

If Condition 4.1 holds, so that there exists a unique equilibrium fluid delay $\widetilde{w}_e$, that equilibrium delay $\widetilde{w}_e$ is easy to calculate. Assuming that $\rho B^c(0) > 1$, we can simply plot the strictly decreasing function $\rho B^c(w)F^c(w \mid w)$ and see where it equals one. Alternatively, we can perform bisection search.

As we discuss in the e-companion, we can also consider iteration to find the equilibrium fluid delay, but if we let $w_{k+1} = d(w_k)$, then we can get oscillation in the fluid model. However, under further regularity conditions, there is monotone convergence in the fluid model of the damped iteration $w_{k+1} = pd(w_k) + (1-p)w_k$ for $0 < p < 1$ if p is chosen small enough. Iterative schemes are not so important for the fluid model itself, but they can be very useful in practice as well as for the INA. In practice, the iterative process can be an important management tool for finding the right FD announcement. Thus, the results about iterative schemes for the fluid model provide valuable insight into corresponding iterative schemes for FD announcements, for both the actual system and stochastic models of it. The main insights are, first, that oscillations are possible, even when there exists a unique equilibrium fluid delay, and second, that oscillations can usually be avoided by using a damped iteration.

5. All-Exponential Stochastic Models

The analysis tools we develop apply to general probability distributions, but in our examples we consider all-exponential models. In doing so, we first assume that the base queueing model is $M/M/s{+}M$ with arrival rate λ, exponential service times having mean $\mu^{-1} = 1$, and exponential times to abandon with mean θ^{-1}. We assume that the associated traffic intensity satisfies $\rho \equiv \lambda/s > 1$. We then assume that the two customer-response functions B and $F(\cdot \mid w)$ are built from exponential c.d.f.'s as well.

As a general exponential form, we assume that the c.c.d.f. for balking is $B^c(w) \equiv 1 - B(w) = e^{-\beta w}$, $w \geqslant 0$ for some $\beta \geqslant 0$, and the conditional abandonment c.c.d.f. is

$$F^c(t \mid w) = \begin{cases} e^{-\gamma(w)t}, & 0 \leqslant t \leqslant w, \\ e^{-\gamma(w)w}e^{-\delta(w)(t-w)}, & t > w, \end{cases} \tag{5.1}$$

where $\gamma(w)$ and $\delta(w)$ are two component abandonment-rate functions, assumed to be continuous and positive in the announced delay w. Definition (5.1) allows for different customer abandonment behavior for times less than and greater than the announced delay w. One exponential with rate $\gamma(w)$ prevails up to time w, while a different exponential with rate $\delta(w)$ prevails afterwards. It also allows the pure-exponential special case in which $\delta(w) = \gamma(w)$

for all w; then, we have the exponential time-to-abandon distribution with rate $\gamma(w)$, a function of w.

The Equilibrium Delay Equation for the Fluid Model. Assuming that $\rho > 1$, we directly see that all possible equilibrium delays for the all-exponential fluid model must satisfy the equilibrium delay equation

$$w = \frac{\log(\rho)}{\beta + \gamma(w)}. \tag{5.2}$$

Note that the equilibrium delay Equation (5.2) is independent of the second abandonment rate function $\delta(w)$. As a consequence, any fluid equilibrium delay $\widetilde{w}_e$ itself is independent of the function $\delta(w)$.

If $\gamma(w)$ is nondecreasing as well as continuous, there necessarily exists a unique solution to (5.2) because the left side of (5.2) is linearly increasing in w, while the right side is necessarily nonincreasing in w, starting from $\log(\rho)/(\beta + \gamma(0)) > 0$ at $w = 0$. It is elementary to check that Condition 4.1 is satisfied for the all-exponential fluid model above, provided that the two rate functions $\gamma(w)$ and $\delta(w)$ are nondecreasing as well as continuous. Then, there is a well-defined delay-response function d and a unique equilibrium fluid delay $\widetilde{w}_e$.

From the equilibrium Equation (5.2), it is readily apparent how the fluid equilibrium delay $\widetilde{w}_e$ depends on the model parameters ρ and β and the function $\gamma(w)$, assuming that $\gamma(w)$ is nondecreasing. As a special case of Theorem 12.1 (in the electronic companion; see §10), we see that $\widetilde{w}_{1,e} > \widetilde{w}_{2,e}$ if $\gamma_1(w) < \gamma_2(w)$ for all $w > 0$.

The Simple All-Exponential Model. We will focus on the elementary special case in which both abandonment-rate functions $\gamma(w)$ and $\delta(w)$ are constant functions with $\gamma(w) \equiv \gamma$ and $\delta(w) \equiv \delta$; we call this case *the simple all-exponential model*. Condition 4.1 is, of course, satisfied for this special case. Then, we have the explicit *equilibrium fluid delay formula*

$$\widetilde{w}_e = \frac{\log(\rho)}{\beta + \gamma}, \tag{5.3}$$

again independent of δ.

Without announcements, we have the abandonment rate θ; with announcements, we have the balking rate β and the abandonment rate γ, up until time $\widetilde{w}_e$. For this elementary model, the deterministic fluid approximations are $\widetilde{w}_1 = \log(\rho)/\theta$ for the delay of all served fluid without an announcement and $\widetilde{w}_e = \log(\rho)/(\beta + \gamma)$ for the equilibrium delay of all served fluid with an announcement. Those simple equilibrium delay formulas show that announcements cause the average delay to be multiplied by the constant factor $\theta/(\beta + \gamma)$. We anticipate the parameters will be such that the multiplicative factor $\theta/(\beta + \gamma)$ is less than one. For example, that will be true if $\beta = \theta$, which occurs if we have information-consistent balking as defined in Definition 3.1. However, the main point is that we have a simple quantification, which is a useful reference point, both before and after performing more detailed analysis.

In the next section, we will show that the fluid approximation is remarkably accurate for the simple all-exponential model when $\delta = \gamma$, but *not* when δ differs significantly from γ. That can be explained by the discontinuity in the abandonment rate, right at the equilibrium point; we elaborate in §7. Fortunately, the equilibrium expected steady-state delay from the INA for the $M/GI/s{+}GI$ model with an FD announcement tends to provide a more accurate prediction.

6. Numerical Comparisons for the All-Exponential Models

In this section, we compare the two approximations—the fluid model and the INA—to simulations for the overloaded simple all-exponential stochastic model with constant abandonment rates $\gamma(w) \equiv \gamma$ and $\delta(w) \equiv \delta$.

The Iterative Numerical Approximation (INA). The INA method determines the approximate equilibrium expected steady-state delay for the associated $M/M/s{+}GI$ queueing model, assuming an FD announcement. The balking first reduces the arrival rate as a function of the announced delay from $\lambda = \rho s$ to $\lambda e^{-\beta w}$ as a function of the fixed announcement w. In the queueing models, we treat balking exactly because our balking mechanism is equivalent to an independent thinning of a Poisson process, which itself is a Poisson process.

Then, the abandonment distribution is the nonexponential distribution in (5.1) with two exponential components, again as a function of the announced delay w, where the two rates γ and δ are constants. The algorithm is applied iteratively, with the new FD announcement being the previously calculated expected conditional steady-state delay given that the customer is served, until the observed expected conditional steady-state delay differs only negligibly from the fixed announced delay.

The approximation method in Whitt (2005) approximates the $M/GI/s{+}GI$ model by an $M/M/s{+}M(n)$ model with state-dependent abandonment rate. We refer to that paper for a detailed explanation of the algorithm. Following (3.3) of Whitt (2005), we construct the state-dependent abandonment rate here out of the two individual exponential components, having rates γ and δ. In our first attempt to do so, we used the crude approximation with $r(k) = r(k^*)\gamma + (k - k^*)\delta$ for $k > k^*$, but we found that approximation led to multiple fixed points, caused by λw crossing over an integer point. Hence, we go beyond Whitt (2005) to carefully treat the boundary here. For that purpose, let $\lfloor x \rfloor$ be the greatest integer less than or equal to x. Here, we let the approximating state-dependent abandonment rate be $r(k)$,

where

$$r(k) = \begin{cases} k\gamma, & 1 \leqslant k \leqslant \lfloor \lambda w \rfloor, \\ r(\lfloor \lambda w \rfloor) + (\lambda w - \lfloor \lambda w \rfloor)\gamma \\ \quad + (\lfloor \lambda w \rfloor + 1 - \lambda w)\delta, & k = \lfloor \lambda w \rfloor + 1, \\ r(\lfloor \lambda w \rfloor + 1) + (k - \lfloor \lambda w \rfloor - 1)\delta, \\ \quad k \geqslant \lfloor \lambda w \rfloor + 2, \end{cases} \quad (6.1)$$

where λ in (6.1) is the arrival rate after balking, i.e., $\lambda e^{-\beta w}$, and w is the current FD announcement.

Comparisons with Simulations. We compare the two approximation procedures—the fluid approximation and the INA—to simulations. The simulation program was written in C. The simulation results are based on 100 independent replications of runs each with $1{,}000 \times \lambda$ (=140,000 here) arrivals. Data were collected after it was verified that initial conditions had only negligible effect on performance. The sample standard errors (standard deviation of the sample mean over the 100 replications) are shown in Table 2 below the estimates in parentheses.

We simulate both the specified $M/M/s{+}GI$ model with the DLS state-dependent announcements and the same model having fixed announcements. We iterate the simulations of the $M/M/s{+}GI$ model having fixed announcements until the long-run average delay for served customers coincides with the fixed announced delay.

We consider the simple all-exponential model with $s = 100$ servers, individual service rate $\mu = 1$, individual without-announcement abandonment rate $\theta = 1.0$, and arrival rate $\lambda = 140$. We let the balking rate be equal to the abandonment rate before an announcement ($\beta = \theta = 1.0$) to obtain information-consistent balking. We let the less-than-announcement-time abandonment rate γ for those who elect to wait be less than the abandonment rate without an announcement ($\gamma = 0.5 < 1.0 = \theta$). For the greater-than-announcement-time abandonment rate δ, we consider two cases: (i) one-parameter conditional abandonment, with $\delta = \gamma$, and (ii) two-parameter conditional abandonment, with $\delta \neq \gamma$. Assuming that customers will become more impatient if they have not been served by the announced time, in the second case we assume that $\delta > \gamma$. To take a challenging case, we let $\delta = 4$.

We display experimental results in Tables 1 and 2. Table 1 contains preliminary results describing performance of the two analytical methods—the fluid model and the numerical algorithm from Whitt (2005). We describe the performance with and without a single-number announcement, but we approximate equilibrium only by making the initial delay announcement be the fluid equilibrium delay $\widetilde{w}_e$. (We do not iterate in Table 1.) In response to that single announcement, the balking probability is $B(\widetilde{w}_e)$. That produces a new $M/M/s{+}GI$ model with common reduced arrival rate $\lambda B^c(\widetilde{w}_e)$. (Recall that there is only a single fluid approximation for both cases because formulas (5.2) and (5.3) are independent of δ.) In Table 1, we

Table 1. A comparison of the fluid approximations with numerical calculations of steady-state performance measures in the all-exponential model, without and with a delay announcement in the case $\lambda = 140$.

All-exponential model with $\lambda = 140$, $s = 100$, $\mu = \theta = \beta = 1.0$, $\gamma = 0.5$, and two cases for δ

	Without an announcement		With one announcement $\widetilde{w}_e = 0.224$		
Performance measure	Exact	Fluid	Numer. ($\delta = 0.5$)	Numer. ($\delta = 4.0$)	Fluid
Initial arrival rate	140.0	140.0	140.0	140.0	140.0
Balking rate	0.0	0.0	28.1	28.1	28.1
Reduced arrival rate	140.0	140.0	111.9	111.9	111.9
Abandon rate	40.0	40.0	12.1	12.2	11.9
Throughput rate	100.0	100.0	99.8	99.7	100.00
$P(B)$	0.000	0.000	0.201	0.201	0.201
$P(A)$	0.286	0.286	0.086	0.087	0.085
$P(A \cup B)$	0.286	0.286	0.287	0.288	0.286
$\text{Prob}(W > 0 \mid A \cup S)$	1.000	1.000	0.956	0.938	1.000
$E[Q]$	40.0	40.0	24.3	17.3	23.7
$E[W \mid S]$	0.332	0.336	0.225	0.157	0.224
$SD[W \mid S]$	0.0997	0.000	0.134	0.088	0.000
$E[W \mid A]$	0.172	0.148	0.150	0.137	0.111
$E[W; B^c]$	0.286	0.286	0.217	0.155	0.212
$P(W \leqslant 0.1 \mid S)$	0.008	0.000	0.192	0.273	0.000
$P(W \leqslant 0.2 \mid S)$	0.092	0.000	0.443	0.647	0.000
$P(W \leqslant 0.224 \mid S)$	0.140	1.000	0.512	0.754	1.000
$P(W \leqslant 0.4 \mid S)$	0.757	1.000	0.897	1.000	1.000

Notes. Two cases are used for the greater-than-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 $\widetilde{w}_e = 0.224$ is used as the announcement, without iteration.

Table 2. A comparison between INA, iterative simulations with fixed delay announcements, and simulations with DLS announcements in the all-exponential model.

Equilibrium fixed delay announcements vs. state-dependent 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-dependent		Equilibrium fixed			State-dependent	
Perf. measure	INA	Sim.		DLS		INA	Sim.		DLS	
Announcement	0.225	0.225		last		0.1616	0.155		last	
Reduced arr. rate	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)
$E[W \mid S]$	0.225	0.225	(0.00079)	0.226	(0.00031)	0.1616	0.155	(0.00021)	0.169	(0.00026)
$SD[W \mid S]$	0.134	0.133	(0.00038)	0.091	(0.00017)	0.066	0.066	(0.00013)	0.072	(0.00012)
$E[W \mid A]$	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)
$E[W - W_a \mid S]$		0.00096	(0.00079)	0.011	(0.000025)		0.00050	(0.00021)	0.0057	(0.000014)
$E[\lvert W - W_a \rvert \mid S]$		0.108	(0.00033)	0.055	(0.000081)		0.052	(0.00010)	0.039	(0.000047)
$E[\lvert W - W_a \rvert^2 \mid S]$		0.018	(0.00010)	0.0050	(0.000016)		0.0044	(0.000017)	0.0025	(0.0000056)
$\text{Prob}(W > W_a)$	0.456	0.367	(0.0018)	0.418	(0.00028)	0.523	0.477	(0.00097)	0.470	(0.00023)

Note. Two cases are used for the after-announcement-time abandonment rate: $\delta = \gamma = 0.5$ and $\delta = 4.0 > 0.5 = \gamma$.

often condition on the event S that the customer is served or the event A that a customer abandons. Let B here be the event that a customer balks. Let $E[W \mid B^c]$ be the conditional expected delay for those who do not balk, and let $E[W; B^c] = E[W \mid B^c]P(B^c)$.

Because the traffic intensity is $\rho = 1.4$, the system is significantly overloaded, so we expect close agreement between the fluid model and the exact numerical computation without a delay announcement (which is exact because the model is $M/M/100+M$), and indeed, that is what we see. With the fluid model, the equilibrium delay is reduced from 0.336 without an announcement to $\widetilde{w}_e = 0.224$ with the equilibrium fluid announcement. With that equilibrium delay, the reduced arrival rate after balking is 111.9, so that the system remains overloaded after the announcement.

From Table 1, we see that the performance predictions for the fluid model agree very closely with those from the numerical algorithm in the case $\delta = \gamma = 0.5$, but are not nearly so close when $\delta = 4.0 > 0.5 = \gamma$. The balking probability is necessarily the same, and the abandonment probability is very close, but the mean queue length $E[Q]$ and the mean waiting time of served customers $E[W \mid S]$ differ considerably when $\delta > \gamma$.

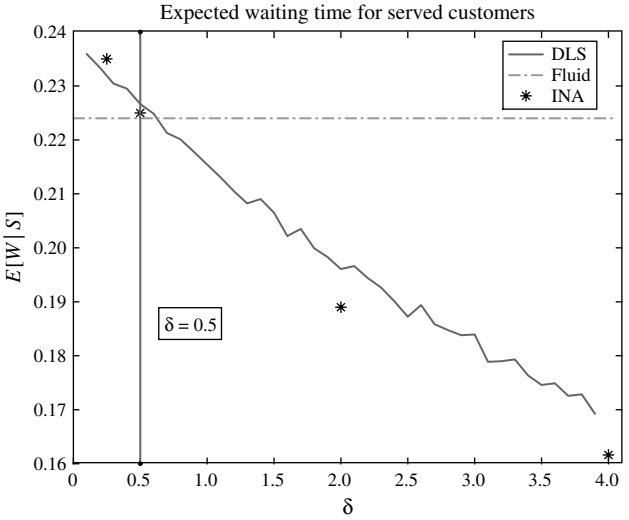
Table 2 displays corresponding equilibrium results for the INA and simulations for the two specific cases discussed in Table 1: (i) $\delta = 0.5 = \gamma$ and (ii) $\delta = 4.0 > 0.5 = \gamma$. In addition to the previous notation, W_a denotes the announced waiting time, which is itself a random variable with state-dependent announcements. We perform two different simulations, both involving equilibrium behavior. First, we iterate simulations in which we make FD announcements, iterating until the fixed announcement agrees with the long-run average delay of a served customer, and second, we simulate with DLS announcements. The iterative simulation is directly verifying the accuracy of the INA. In Armony et al. (2007), we show the numerical results for each of the iterations used to obtain the equilibrium fixed delays. In these cases, no more than four iterations were required.

Tables 1 and 2 show, first, that the two approximation methods—the fluid model and INA—are both remarkably accurate (with less than 1% error) when $\delta = \gamma = 0.5$, and second, that the INA is quite accurate (with 4% error) in the second case when $\delta = 4.0 > 0.5 = \gamma$. However, the fluid approximation is not nearly so accurate (with 33% error) in the second case when $\delta = 4.0 > 0.5 = \gamma$. We saw in Equation (5.3) that the fluid equilibrium is independent of the second abandonment rate δ. Evidently, that is a shortcoming of the fluid approximation for this model. Thus, as stated before, the fluid approximation should be regarded as only a crude approximation.

On the other hand, the INA is remarkably effective. The results for the INA agree quite closely with both associated iterative simulations with FD announcements and DLS simulations. The comparison with the iterative simulation further substantiates the accuracy of the approximation in Whitt (2005). The comparison with the DLS simulation shows that the INA does indeed predict the aggregate DLS performance remarkably well.

From Table 2, we also see that the state-dependent DLS announcements yield more reliable predictions than the FD announcements because the the expected absolute difference $E[\lvert W - W_a \rvert \mid S]$ and the expected squared

Figure 2. The effect of the assumed greater-than-announcement-time abandonment rate δ on the expected delays for both DLS announcements and the two approximations—the fluid model and INA—for the current example with $\gamma = 0.5$.



difference $E[|W - W_a|^2 \mid S]$ are smaller with the DLS announcements.

The poor performance of the fluid model when $\delta = 4.0 > 0.5 = \gamma$ led us to investigate more carefully the dependence of performance on the greater-than-announcement-time abandonment rate δ. Consistent with Table 2 and intuition, $E[W \mid S]$ decreases for DLS announcements as δ increases. Through simulations, we found that the fluid approximation agrees most closely with the actual performance for DLS announcements when $\delta = \gamma$. Figure 2 provides more detail for the special case $\gamma = 0.5$. Unlike the fluid model approximation, we found that INA is consistently quite accurate when $\delta \neq \gamma$. For example, INA estimates $E[W \mid S]$ as 0.235 when $\delta = 0.25$ and as 0.189 when $\delta = 2.0$.

The example considered in Tables 1 and 2 is quite heavily loaded. We consider examples with lower arrival rates (120 and 110) and higher balking probability ($\beta = 2.0$ instead of $\beta = 1.0$), but still with $\delta = 4.0 > 0.5 = \gamma$ in Tables 5 and 6 of Armony et al. (2007). Because 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, the accuracy of the INA approximation 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 $\lambda = 110$.

7. Insight into the Performance of the Fluid Approximation

From Figure 2, we see that the fluid approximation $\widetilde{w}_e$ for the equilibrium delay is accurate when $\delta = \gamma = 0.5$, but is quite inaccurate when $\delta \neq \gamma$. When $\delta < 0.5 = \gamma$, the fluid approximation *underestimates* the simulated value, but when $\delta = 4.0 > 0.5 = \gamma$, the fluid approximation *overestimates* the simulated value. The fluid approximation is 0.224, while the simulated values for equilibrium fixed delay and last experienced delay are 0.155 and 0.169, respectively. For these cases, the fluid model overestimates the actual value by about 30%. This is unexpected because we are accustomed to fluid approximations underestimating the actual stochastic values, because the extra variability ignored in the deterministic fluid approximation tends to increase congestion.

This phenomenon can be understood by, first, recognizing that the actual waiting times for served customers should fluctuate around the equilibrium expected value $E[W \mid S]$, and second, by analyzing the consequences of such fluctuations. Consistent with the numerical results in the last section, we show in this section that the stochastic fluctuations about the fluid equilibrium should cause no problem when $\delta = \gamma$, but the fluid equilibrium delay should significantly overestimate the simulated value when $\delta > \gamma$.

Perturbation Analysis. We accomplish this goal by considering the impact of a small perturbation of the equilibrium announcement in the all-exponential fluid model in §5 when $\gamma(w) = \gamma$ and $\delta(w) = \delta$. Recall that the fluid equilibrium wait is $\widetilde{w}_e = \log(\rho)/(\beta + \gamma)$. First, we consider the case of delays greater than $\widetilde{w}_e$. For that purpose, suppose that the actual delay is $\widetilde{w}_e + \epsilon$ instead of $\widetilde{w}_e$, where $\epsilon > 0$, so that we announce $\widetilde{w}_e + \epsilon$ as well. If we work with the fluid model, using (3.6) and (5.1), then the experienced delay for a served customer, $d(\widetilde{w}_e + \epsilon)$, will satisfy

$$\rho e^{-\beta(\widetilde{w}_e+\epsilon)} e^{-\gamma d(\widetilde{w}_e+\epsilon)} = 1, \tag{7.1}$$

which leads to

$$d(\widetilde{w}_e + \epsilon) = \frac{\log(\rho) - \beta(\widetilde{w}_e + \epsilon)}{\gamma} = \widetilde{w}_e - \frac{\beta\epsilon}{\gamma} < \widetilde{w}_e. \tag{7.2}$$

The situation is more interesting when we suppose that the actual delay is $\widetilde{w}_e - \epsilon$ for $\epsilon > 0$, and announce $\widetilde{w}_e - \epsilon$ as well. When $\delta = \gamma$, the reasoning in (7.1) applies, whether ϵ is positive or negative, but in this case the experienced delay is greater by the same difference $(\beta/\gamma)\epsilon$. As a consequence, if the true delay distribution is symmetric around $\widetilde{w}_e$, then the two errors will tend to cancel.

However, the situation is very different when $\delta \neq \gamma$. In this second situation when we announce $\widetilde{w}_e - \epsilon$, (3.6) and (5.1) produce the equation

$$\rho e^{-\beta(\widetilde{w}_e-\epsilon)} e^{-\gamma(\widetilde{w}_e-\epsilon)} e^{-\delta(d(\widetilde{w}_e-\epsilon)-\widetilde{w}_e+\epsilon)} = 1, \tag{7.3}$$

which leads to

$$d(\widetilde{w}_e - \epsilon) = \frac{\log(\rho)}{\delta} + \frac{(\delta-\beta-\gamma)(\widetilde{w}_e-\epsilon)}{\delta} = \widetilde{w}_e - \epsilon\frac{(\delta-\beta-\gamma)}{\delta}. \quad (7.4)$$

From (7.2) and (7.4), we see that if $\delta > \beta+\gamma$, then $d(\widetilde{w}_e+\epsilon) < \widetilde{w}_e$, both when $\epsilon > 0$ and $\epsilon < 0$, so we can anticipate that stochastic fluctuations of any kind will make the actual equilibrium wait less than the fluid approximation $\widetilde{w}_e$. That partially explains the results for $\delta > \gamma$. (On the other hand, if $\delta < \beta+\gamma$, then $d(\widetilde{w}_e-\epsilon) > \widetilde{w}_e$.)

For more general fluid models, we see that we have the good behavior above provided that the conditional time-to-abandon c.d.f. $F(t \mid w)$ has a continuous derivative as a function of t at $t = w = \widetilde{w}_e$. This requirement is quite natural and should be expected to hold in practice. The example with $\delta = 4.0 > 0.5 = \gamma$ is actually less likely to occur. The following example shows that the good behavior for $\delta = \gamma$ extends to the more general model with nondecreasing functions $\gamma(w)$ and $\delta(w)$ when $\delta(w) = \gamma(w)$ for all w.

Example 7.1 (Linear Abandonment Rates). To investigate other all-exponential models with $\delta(w) = \gamma(w)$ for all w, we considered the special case of linear functions: $\gamma(w) = \gamma_0 + \gamma_1 w$, where γ_0 and γ_1 are positive constants. To relate to the simple all-exponential model, we chose the constants γ_0 and γ_1 to have the same equilibrium $\widetilde{w}_e = 0.224$ as when $\gamma(w) = \gamma = 0.5$. That dictates that we satisfy the equation $\gamma_0 + \gamma_1(0.224) = 0.5$. Accordingly, we considered the following three cases: (i) $(\gamma_0, \gamma_1) = (0.000, 2.232)$, (ii) $(\gamma_0, \gamma_1) = (0.250, 1.116)$, and (iii) $(\gamma_0, \gamma_1) = (0.450, 0.223)$. For these three cases, we obtained the following simulation estimates for DLS announcements: $E[W \mid S] = 0.2208$, 0.2216, and 0.2262, respectively. The errors in the fluid approximations are all less than 2%.

Quantifying the Impact of Stochastic Fluctuations. We now consider how to approximately quantify the impact. To do so, as a rough approximation, we suppose that the actual delay is normally distributed with mean $\widetilde{w}_e$ and standard deviation σ_e. First, we can apply (7.2) and (7.4) to obtain

$$d(\widetilde{w}_e + \sigma_e N(0,1)) \approx \widetilde{w}_e + d^+(\widetilde{w}_e)\sigma_e N(0,1)^+ + d^-(\widetilde{w}_e)\sigma_e N(0,1)^-, \quad (7.5)$$

where $d^+(x)$ and $d^-(x)$ are the right and left derivatives of d at x, $(x)^+ \equiv \max\{x, 0\}$ and $(x)^- \equiv -\min\{x, 0\} \geqslant 0$. Next, recalling that $E[|N(0,1)|] = \sqrt{2/\pi} \approx 0.8$, we can apply (7.5) to obtain the associated numerical estimate

$$E[d(\widetilde{w}_e + \sigma_e N(0,1))] \approx \widetilde{w}_e - 0.8\frac{\sigma_e}{2}\left(\frac{\beta}{\gamma} + \frac{\delta-\beta-\gamma}{\delta}\right). \quad (7.6)$$

We can check to see if this is consistent with our numerical example in Table 2. For that example, we had $\beta = 1$ and $\gamma = 0.5$. We had two cases for δ: $\delta = \gamma = 0.5$ and $\delta = 4.0$. From (7.6), we see that the adjustment is zero, so that $d(\widetilde{w}_e + \sigma_e N(0,1)) \approx \widetilde{w}_e$ if $\delta = \gamma$, which is consistent with our numerical results.

The other case with $\delta = 4.0$ leads to the approximation $d(\widetilde{w}_e + \sigma_e N(0,1)) \approx \widetilde{w}_e - 1.05\sigma_e$, but we have yet to determine the standard deviation. Suppose that we use the simulation estimate for the standard deviation, using the announcement of the delay of the last customer to be served. Then, we get the estimate $\sigma_e \approx SD(W \mid S) = 0.072$. That yields the detailed approximation $\tau \approx 0.224 - 0.076 = 0.148$. That produces an estimate that is 12% too small, compared to the original fluid approximation $\widetilde{w}_e = 0.224$, which is 32% too large.

Of course, to be able to make a priori predictions, we need to produce an estimate for the standard deviation σ_e, without exploiting simulation results. More generally, the nonlinear behavior of the abandonment rate at the announcement time is likely to make the actual distribution non-Gaussian. Better quantifying the impact of stochastic fluctuations remains a problem for future research.

8. Multiple Equilibria

As observed in §4, the general conditions ensuring a unique fluid equilibrium in Condition 4.1 seem quite natural, but we should be cautious about human response. For the general all-exponential model in (5.1), if the abandonment rate $\gamma(w)$ fails to be nondecreasing, there can be multiple solutions to the equilibrium Equation (5.2). It is not entirely unreasonable to have $\gamma(w)$ decreasing over subintervals because more customers will elect to balk as w increases by our assumed exponential balking c.d.f. B. It is possible that the customers who decide to wait in response to a delay announcement w, instead of balk, tend to be the more patient customers as that announcement w increases; the less patient customers may already have balked. If $\gamma(w)$ is indeed decreasing over subintervals, then it is possible for there to exist multiple equilibria.

To illustrate, we consider an example, which is chosen to be easy to analyze rather than realistic. Suppose that

$$\begin{aligned} &\gamma(w) = 4.0, \quad 0 \leqslant w < 0.10, \\ &\gamma(w) = 7.5 - 35w, \quad 0.10 \leqslant w < 0.20, \\ &\gamma(w) = 0.5, \quad t > 0.20. \end{aligned} \quad (8.1)$$

We have constructed $\gamma(w)$ to be constant over the two subintervals $[0, 0.10)$ and $[0.20, \infty)$, linear and decreasing in the interval $[0.10, 0.20)$, and continuous overall. It is elementary to see that the fluid model has three equilibria, with one in each region: the three fluid equilibria are $\widetilde{w}_e = 0.0672$, $\widetilde{w}_e = 0.193$, and $\widetilde{w}_e = 0.224$. The abandonment rates at these three equilibria are, respectively, $\gamma(0.0672) = 4.0$, $\gamma(0.193) = 0.7395$, and $\gamma(0.224) = 0.5$. The associated fluid queue contents are $q(0.672) = 0.077$, $q(0.193) = 0.180$, and $q(0.224) = 0.237$. One may multiply by $s = 100$ to get the associated approximating queue lengths.

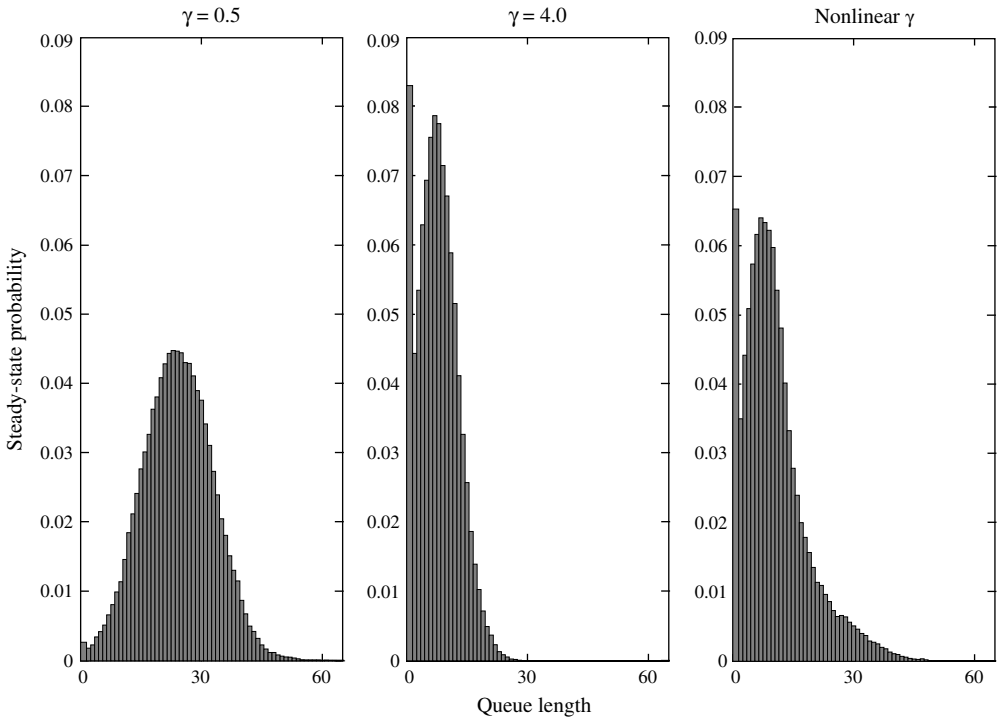
Corresponding to each of these three fluid equilibria, we find an INA equilibrium by iteratively applying the numerical algorithm for the $M/M/s+M$ model. The three INA equilibria occur at 0.0643, 0.1933, and 0.225. The associated equilibrium abandonment rates are 4.0, 0.7345, and 0.5. The associated equilibrium queue lengths are 7.89, 21.2, and 23.7. These three INA equilibria are bonafide equilibria for the case of FD announcements, as confirmed by simulations. With FD announcements, the system manager would thus have a choice of equilibrium delay announcements. Presumably, the one yielding the lowest delay should be used.

On the other hand, we conjecture that there exists a unique equilibrium with DLS announcements, but with the steady-state distributions influenced by the abandonment-rate function. The intuition is that the adaptive DLS announcements, together with stochastic fluctuations, should lead to a full range of experienced delays over time, and thus announcements, preventing the system from "getting stuck" in the region of any fixed announcement equilibrium. That means that the DLS steady-state behavior will not be like any one fluid equilibrium, but will in some sense reflect all of them. We illustrate in Figure 3 by plotting histograms estimating the density of the steady-state queue-length distribution associated with DLS announcements estimated from simulation in three cases: (i) $\gamma = 0.5$, (ii) $\gamma = 4.0$, and (iii) nonlinear $\gamma(w)$ in (8.1). For constant abandonment rate, we have the $M/M/s+M$ model for which the queue-length distribution is asymptotically normally distributed around the fluid equilibrium. We see the decidedly different skewed steady-state distribution for the nonlinear abandonment rate in (8.1). Sample paths of the queue-length process also show excursions in the lower-announcement and higher-announcement regions; see the e-companion.

9. Conclusions

We introduced two specific single-number delay announcement schemes intended for heavily-loaded invisible multi-server queues: (i) the delay of the last customer to enter service (DLS), and (ii) the equilibrium fixed delay (FD) announcement, a fixed deterministic announced delay chosen to coincide with the mean steady-state delay. We emphasized the equilibrium behavior associated with such delay announcements, where the customers respond to the announcements, the system performance depends on the customer response, and the announcements depend on the system performance. We also introduced a modelling framework to study the equilibrium behavior of these delay announcements. The starting point is our modelling of customer response through the balking and abandonment functions $B(w)$ and $F(t \mid w)$. Given that model, we showed that simulation can be used to evaluate the steady-state performance of delay announcements within conventional queueing models, such as the many-server $M/GI/s+GI$ model, where we iteratively apply the simulation to find the equilibrium behavior associated with FD announcements. We can thus determine the performance impact of these delay announcements. The simulation experiments showed

Figure 3. Histograms of the steady-state queue-length distribution for the all-exponential model with $\delta(w) = \gamma(w)$ for all w, in three cases: (i) $\gamma = 0.5$, (ii) $\gamma = 4.0$, and (iii) nonlinear $\gamma(w)$ in (8.1).



that the delay announcements in overloaded regimes can significantly reduce the delays of served customers without adversely affecting the number of customers receiving service.

Our 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 DLS predictions, but the variation is greater.

We also developed mathematical models and analysis techniques to provide additional insight into the performance impact of these delay announcements. Specifically, we introduced two methods to describe the approximate performance with these delay announcements: (i) a deterministic fluid model, extending the fluid model in Whitt (2006), and (ii) an iterated numerical algorithm (INA), based on Whitt (2005), assuming the use of an FD announcement. We conducted simulation experiments to evaluate the accuracy of these approximations. These approximations were found to be remarkably accurate considering that they are only simplified rough descriptions of a very complicated system.

We showed that the fluid model is sufficiently tractable to obtain solid theoretical results. For example, Theorem 4.1 provides general conditions for the existence of a unique equilibrium delay in the fluid model, while formula (5.3) provides an explicit formula for the equilibrium delay in the simple all-exponential model. In §8, we showed that the fluid model is effective at predicting important qualitative behavior, such as the multiplicity of equilibrium points in the case of FD announcements. In the e-companion, we show that the fluid model can also be applied to predict the performance impact of biased announcements. We have thus shown that the fluid model provides both important insight and a useful means to perform "back-of-the-envelope" performance calculations.

There are many important directions for future research. First, we need to study the actual human response to delay announcements, following Brown et al. (2005), Feigin (2006), and Munichor and Rafaeli (2006). To what extent is the customer-response model based on the balking and abandonment functions $B(w)$ and $F(t \mid w)$ justified? And what properties do these functions satisfy? With such empirical studies in mind, it is also natural to investigate to what extent these balking and abandonment functions $B(w)$ and $F(t \mid w)$ arise via individual customers maximizing their expected utility from service and waiting, as postulated by Guo and Zipkin (2007). If that view is appropriate, then we should consider equilibrium analysis in that framework.

Second, we need to systematically investigate the effectiveness of alternative real-time delay estimators based on recent system state. A study of alternative delay estimators based on recent delay history in the $GI/M/s$ model, without considering customer response, has been conducted by Ibrahim and Whitt (2008). That study supports the use of DLS, but more work is needed, including the investigation of more complex models involving equilibrium behavior. We have seen that the DLS delay announcements can be quite effective for a single $M/GI/s{+}GI$ model. We need to test the performance of the DLS in more complex multiskill environments, typical of modern call centers.

Finally, we want to obtain theoretical results for DLS and related announcements in actual queueing models, paralleling the theoretical results for the fluid model obtained in this paper. To repeat what we said at the outset, there are many open questions about system dynamics: (i) Under what conditions does there exist an equilibrium steady-state behavior for the actual system with DLS announcements and the postulated customer response? (ii) If there is an equilibrium, when is it unique? When can there be multiple equilibria? (iii) How do the stochastic processes evolve as a function of the initial conditions?

We would like to be able to conclude that there exists a unique equilibrium delay for the DLS announcement scheme under general regularity conditions. However, it is natural to first seek easier asymptotic results. As a first step, we could try to demonstrate asymptotic accuracy of DLS and asymptotic validity of the fluid model in the efficiency-driven many-server heavy-traffic limiting regime, as in Whitt (2006). We seek asymptotic support as provided by Armony and Maglaras (2004a, b) for their call-back scheme.

10. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Acknowledgments

This research was supported by grant no. 2002112 from the United States–Israel Binational Science Foundation (BSF). Ward Whitt was also supported by NSF grant DMI-0457095.

References

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

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

Armony, M., N. Shimkin, W. Whitt. 2007. The impact of delay announcements in many-server queues with abandonment: Supplementary material. http://columbia.edu/∼ww2040.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, 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, T. Carmon. 1995. A psychological perspective on service segmentation models: The significance of accounting for customers' perceptions of waiting and service. *Management Sci.* **41** 1806–1815.

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

Durrande-Moreau, A. 1999. Waiting for service: Ten years of empirical research. *Internat. J. Service Indust. Management* **10** 171–189.

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

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

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

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

Hui, M., 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., W. Whitt. 2008. Real-time delay estimation based on delay history in the $GI/M/s$ queue. *Manufacturing Service Oper. Management*. Forthcoming.

Larson, R. C. 1987. Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* **35** 895–905.

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

Mandelbaum, A., A. Sakov, S. Zeltyn. 2000. Empirical analysis of a call center. Technical report, Faculty of Industrial Engineering and Management, The Technion, Haifa, Israel.

Munichor, N., A. Rafaeli. 2006. Numbers or apologies? Customer reactions to tele-waiting time fillers. *J. Appl. Psych.* **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. *Oper. Res.* **52**(2) 213–228.

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

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

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

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

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

Whitt, W. 2006. Fluid models for multi-server queues with abandonments. *Oper. Res.* **54** 37–54.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems* **51** 361–402.

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