OPERATIONS RESEARCH
Vol. 65, No. 5, September–October 2017, pp. 1380–1397
ISSN 0030-364X (print), ISSN 1526-5463 (online)
http://pubsonline.informs.org/journal/opre/

Refined Models for Efficiency-Driven Queues with Applications to Delay Announcements and Staffing

Junfei Huang,[a] Avishai Mandelbaum,[b] Hanqin Zhang,[c] Jiheng Zhang[d]

[a] Department of Decision Sciences and Managerial Economics, CUHK Business School, The Chinese University of Hong Kong, Shatin, Hong Kong; [b] Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, 3200003 Haifa, Israel; [c] Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore; [d] Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Contact: junfeih@cuhk.edu.hk, http://orcid.org/0000-0002-3764-354X (JH); avim@ie.technion.ac.il (AM); bizzhq@nus.edu.sg (HZ); j.zhang@ust.hk, http://orcid.org/0000-0003-3025-1495 (JZ)

Received: June 14, 2014
Revised: November 22, 2015; August 3, 2016; November 19, 2016
Accepted: January 25, 2017
Published Online in Articles in Advance: July 17, 2017

Subject Classifications: probability: stochastic model; queues: approximations, limit theorems
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2017.1619

Abstract. Data has revealed a noticeable impact of delay-time-related information on phone-customers; for example and somewhat surprisingly, delay announcements can abruptly increase the likelihood to abandon (hang up). Our starting point is that the latter phenomena can be used to support the control of queue lengths and delays. We do so by timing the announcements appropriately and determining the staffing levels accordingly. To this end, we model a service system as an overloaded $GI/M/s+GI$ queue, in which we seek to minimize the number of servers, s, subject to quality-of-service constraints (e.g., fraction abandoning), while accounting for the instantaneous (hence discontinuous) impact of an announcement on the distribution (hazard rate) of customer patience. For tractability, our analysis is asymptotic as s increases indefinitely, and it is naturally efficiency-driven (namely the servers are highly busy, and hence essentially all customers are delayed in queue prior to service). This requires one to go beyond existing theory, which turns out to be too crude for our needs (e.g., it requires a continuous hazard rate of impatience and hence cannot be applied). We thus develop a refined process and steady-state models, and use them to solve our minimization problem and more. The value and accuracy of our models are demonstrated via extensive numerical experiments.

Funding: The first author's research was supported in part by the Hong Kong Research Grants Council [Project 24500314]. The work of the second author has been partially supported by the U.S.-Israel Binational Science Foundation [Grants 2005175 and 2008480], Israel Science Foundation [Grants 1357/08 and 1955/15] and by the Technion funds for promotion of research and sponsored research. Some of the research of the first and second author was funded by and carried out while they were visiting the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the National Science Foundation—the hospitality of SAMSI is gratefully acknowledged. The fourth author's research was supported in part by the Hong Kong Research Grants Council [Grants GRF-622411 and GRF-624012].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2017.1619.

Keywords: customer abandonment • delay announcement • staffing • call centers • ED+QED • refined approximation

1. Introduction

Motivated by large service centers (mainly call centers), there has been a growing body of research on many-server queues with customer abandonment (Garnett et al. 2002, Gans et al. 2003, Zeltyn and Mandelbaum 2005, Whitt 2006, Akşin et al. 2007). Often the goal is a balanced operation that is both Quality- and Efficiency-Driven (QED): Customers do not wait too long for available servers and servers do not wait too long for needy customers. For large enough systems, this QED balance translates into waiting and idle times being negligible, relative to the service time. However, in practice, many call centers are merely Efficiency-Driven (ED) in that they are understaffed, which results in significant delays and consequent abandonment. One such scenario is depicted in Figures 4 and 5: The first figure reveals extreme understaffing (e.g., 80 agents present at 11:30, while about twice as many are required for good performance); and the second figure demonstrates the severe outcome of such understaffing: 20%–60% abandonment.

Various reasons could lead to an ED operation. For example, call centers could be service-oriented as opposed to revenue-generating (e.g., Whitt 2004); or staffing levels could be inflexible to accommodate temporal peaks or an unexpectedly high demand (e.g., Perry and Whitt 2009). When this happens, and when queues are *invisible* (e.g., call centers) and significant, it makes sense and is hence prevalent to provide customers with delay information. (Such information has no noticeable impact in short-wait conditions; see Hui and Tse 1996.) One reason is to relieve waiting anxiety because "uncertain waits feel longer than predictable finite waits" (Maister 1985, p. 118).

Just as important, such information helps customers decide whether their gain from service is worth the wait. In that case they *abandon* the queue, which, in heavy traffic, could dramatically improve the waiting experience of those opting for service. (For visible queues, and starting with Naor 1969, the analogous option for a customer is to *renege* upon arrival at too long a queue; see Hassin and Haviv 2003.) Delay announcements could thus provide a relatively simple and inexpensive means for improving customer experience and controlling delay—this is the starting point of the present paper.

Specifically, we develop a model for many-server queues in the ED regime, or more precisely, ED+QED refinement. Our model captures the effect of growing impatience on system performance, which is attributed to delay announcements. We then use our model to simultaneously optimize the staffing levels and timing of announcements, subject to service level constraints (e.g., fraction abandoning). However, capturing the impact of announcements on customers' tendency to abandon (hang up) raises a challenge: There exists empirical support to suggest that this impact is often abrupt or, formally, manifested through a discontinuity in the hazard rate of customer patience (underlying the smoothed peaks in Figure 3). Moreover, there is also the practical need and theoretical challenge to accommodate general distributions. It follows that for tractability, one must resort to fluid or diffusion models with discontinuous primitives, which necessitates refinement of the existing models.

Two Types of Delay Announcements. Our refined model is motivated by two types of announcements. The first is to be made upon the arrival of customers who must wait before receiving service; see the "all-exponential model" in Armony et al. (2009). In this case, an estimated duration of delay is announced, which has the following consequences. Some customers choose to balk immediately, while others remain online. Customers will not abandon if served before their patience expires, but they will become irritated once their waiting times reach the announced delay. Such behavior is collectively (statistically) manifested by a sudden increase in the hazard rate of the patience-time distribution at the announced time (e.g., Armony et al. 2009). The second type of announcement is to be made during waiting, e.g., when a customer's waiting time reaches one minute. Here, the announcements provide varying levels of information, ranging from the detailed "your waiting time is expected to be X minutes/seconds," through "you are number X in the queue," to the vague "please hold—an agent will be with you momentarily." Allon and Bassamboo (2011) and Mandelbaum and Zeltyn (2013) have had discussions on such announcements, with the latter observing that such announcements, in various call centers, have been found to be associated with an upward jump in the hazard rate right after the announcement. Taking a step further, Li et al. (2015) developed a statistical method to estimate the hazard rate as a smooth surface of waiting time and time-of-day: It is shown that peaks in the hazard rates, attributed to delay announcements, are consistent across different times-of-day.

In view of Armony et al. (2009), Mandelbaum and Zeltyn (2013), and Li et al. (2015), both types of announcements share the common feature of being associated with a nonsmooth (abrupt) change in the hazard rate of the patience-time distribution. It occurs at a certain "impact point," which is either the announced waiting time (first type) or the chosen time to make an announcement (second type). We make it an assumption that announcements do abruptly increase the likelihood of customer abandonments. Under this assumption, we develop models that quantify the impact of this nonsmooth change on operational performance, which then provides insights and guidelines for the management of congestion. For example, we obtain answers to whether an announcement upon arrival (first type) can reduce staffing costs, and whether an announcement during waiting (second type) should be used (and if so, when).

Refined Models are Needed. With the above motivation, we consider a multiserver queueing system $GI/M/n+GI$, with a possibly nonsmooth patience-time distribution—this distinguishes our model from the existing ones. To elaborate, it has been shown in Whitt (2006), via simulation, that fluid models capture very accurately the performance of ED systems. This was rigorously proved later in Bassamboo and Randhawa (2010), but under some regularity conditions that are not satisfied in the presence of announcements. In concert with that, Armony et al. (2009) demonstrated that fluid models are inaccurate for systems with a delay announcement. For example, in their $M/M/n+GI$ system with $n = 100$ servers—each having a service rate of 1, an arrival rate of 140, and with the hazard rate of patience-time distribution having a jump—the simulated queue length is 17.3 while the fluid approximation is 23.7. The understanding of this gap was left as a problem for future research, which is here resolved: Our refined model offers an improved approximation of 16.4 (see Section 4.1 for details). Another example is approximating the tail probabilities of waiting times in the ED+QED regime. This was studied in Mandelbaum and Zeltyn (2009), and requires the smoothness of the hazard rate of customer patience. For such a system with 100 servers, each having a service rate of 1, and an arrival rate of 120, the simulated tail probability is 0.2574, while the approximation in Mandelbaum and Zeltyn (2009) yields 0.4167. Our refined model produces an accurate approximation of 0.2575 (see Section 5 for extensive numerical experiments). Generally speaking, nonsmooth changes in the hazard rate of patience render the existing models inaccurate, and our refined model closes this accuracy gap successfully.

Control via Announcements, Jointly with Staffing. In addition to improving customer satisfaction psychologically, announcements can also reduce staffing levels while not hurting the service level (as characterized by the tail probabilities of the waiting times, for example). To elaborate, with an announcement upon arrival, we minimize the staffing level, subject to a target bound on the probability that the waiting time exceeds a benchmark (see optimization problem (38) in Section 4.1.1). This same formulation of constrained optimization is used in Mandelbaum and Zeltyn (2009). It turns out that announcements reduce the staffing level by a magnitude of $O(\sqrt{\lambda})$ (where λ is the arrival rate). With announcements during waiting, we simultaneously optimize the timing of an announcement as well as minimize the staffing level, and do so subject to bounds on the tail probability of waiting and on the fraction abandoning (see optimization problem (44) in Section 4.2.1). It turns out that it is optimal to make an announcement at a time that is approximately the fluid offered waiting time; here also, the announcement reduces staffing by a magnitude of $O(\sqrt{\lambda})$ (see Proposition 3 in Section 4.2).

A Queueing Model with a Delay Announcement. Our refined model introduces a general scaling (see (4)) of the patience-time distribution, which precisely captures its fine structures, especially the nonsmooth changes attributed to announcements. As explained in Section 2, hazard-rate scaling and no-scaling of the patience-time distribution are special cases of this general scaling. Our method for analyzing such a refined model is based on the virtual waiting time, which differs from the traditional approach that is based on queue length. The *virtual waiting time* $V^n(t)$ is the time that an infinitely-patient "virtual" customer would have to wait if arriving at time t. The evolution of the virtual waiting time is characterized by (16), which enables us to develop diffusion approximations for systems with patience-time distribution scaling (4). The tractable stationary distribution of the diffusion limit is then used to approximate the steady-state performance of its originating queueing systems. Useful characteristics of the approximation formulae are: (a) closed-form; (b) no need to worry which scaling to choose (hazard-rate scaling vs. none) or how to choose a scaling for the patience-time distribution; and (c) the ability to analyze how the operational performance is affected by a nonsmooth change of the patience-time distribution.

1.1. Literature Review

Announcements. Customers' reaction to announcements within large service systems, in particular call centers, has been studied both empirically and theoretically. Brown et al. (2005) and Mandelbaum and Zeltyn (2013) statistically estimated the hazard rate of patience-time distribution and found that a surge is associated with the time of announcement. Akşin et al. (2016) modeled abandonment decisions endogenously in the presence of delay announcements; they studied how announcements impact customer behavior which, in turn, affects system performance. This led to an empirical approach that combines the estimation of patience parameters, the modeling of abandonment behavior, and a queueing analysis that incorporates that behavior. Yu et al. (2017) explored the impact of delay announcements, using an empirical approach that is based on a medium-sized call center. Their key insights are that delay announcements not only impact customers' perceptions of the system, but also directly impact the waiting costs. Ibrahim et al. (2016) investigated delay announcements in call centers within the framework of an $M/M/n+M$ in the ED regime: The announcement to an arriving customer is the delay of the last customer to enter service. The announcement-dependent customer behavior is then explicitly modeled by letting the joint probability and abandonment rate depend on the announced waiting time. Jouini et al. (2011) explored the effect of announcing different percentiles of the waiting time distribution on balking and reneging. System performance measures were calculated via an $M/M/s+M$ queue. Through a numerical study, Jouini et al. (2011) explored when informing customers about delays is beneficial and what the percentile should be in these announcements.

The ED and ED + QED Regime. The ED regime was introduced in Garnett et al. (2002), which was then followed by ample research on the many-server queues in that regime. Whitt (2004) studied both diffusion approximations and steady-state limits for an ED Markovian model. The ED+QED regime arose in Mandelbaum and Zeltyn (2009), as a refinement to ED that accommodates approximations to the tail probabilities for $M/M/n+G$. Dai et al. (2010) analyzed diffusion models for $G/Ph/n+M$ systems in both the QED and ED+QED regimes.

A closely related paper is Liu and Whitt (2014b): They established process level diffusion approximations for models where the arrival rate and staffing level are time dependent, thus allowing the system to alternate between overloaded and underloaded regimes. Our paper differs from theirs in terms of patience-time distribution scaling and methodology. More specifically, their patience-time distribution is independent of n, and is assumed to have a continuous density function. In our model, being motivated by systems with delay announcements, we allow the patience-time distribution to be more general (e.g., the one in Armony et al. 2009). This requires a general scaling framework (4) of patience-time distributions, which allows the patience-time distributions to have abrupt changes. If the patience-time distributions are independent of n and have a continuous density function, the method in Liu and Whitt (2014b) applies to the stationary model (see their Section 10 for the connection to the ED+QED regime). However, we have not been able to apply the method in Liu and Whitt (2014b) to our general scaling framework (4), thus necessitating the development of a new method. Our base is a conservation law that connects the numbers of customer arrivals, abandonments, and service completions (see (9)). This law is formalized by a system dynamic equation for the virtual waiting time (see (16)), which supports our diffusion approximations.

Fluid Models. Fluid approximations are useful in the ED regime. The pioneering work by Whitt (2006) introduced fluid systems with general service and patience-time distributions, and established the first fluid limit in a discrete-time framework. It gave rise to simple, yet effective, approximations for various performance measures, based on the equilibrium of the fluid model. Whitt's fluid approximation, in continuous time, was formally justified by Kang and Ramanan (2010) and Zhang (2013) using measure-valued processes. The papers by Kang and Pang (2011, 2013) would be helpful for readers to relate Whitt (2006) to the different approach in Kang and Ramanan (2010) and Zhang (2013). The paper by Liu and Whitt (2012a) completed the story started in Whitt (2006) by bringing the model to the time-varying setting. Long and Zhang (2014) proved that the fluid model $G/GI/n{+}GI$ converges to an equilibrium state, following the result for $G/M/n{+}GI$ in Section 5 of Liu and Whitt (2011a). A sequence of works by Liu and Whitt (2011b, 2012b, 2014a) comprehensively analyzed, from theory to algorithms, networks of many-server fluid queues in the time-varying setting.

Approximations based on fluid models are surprisingly accurate in the ED regime. Bassamboo and Randhawa (2010) showed that the gap between the steady-state queue length and its fluid approximation is $O(1)$. This enabled the study of optimal capacity sizing for $M/M/n{+}GI$, based on its fluid approximation, to minimize the sum of the capacity costs and long-term average customer-related costs. An in-depth discussion on the gap between fluid and diffusion approximations is provided in our Section 5.4. This adds to Bassamboo and Randhawa (2010), because their assumptions do not hold in the presence of non-smooth patience-time distributions, which one might find in the presence of delay announcements.

Tail Probabilities of the Waiting Time. In calculating performance measures for $M/M/n{+}GI$, Zeltyn and Mandelbaum (2005) and Mandelbaum and Zeltyn (2009) identified the important role of the derivative of the patience-time distribution at the fluid offered waiting time. In particular, they studied the tail probability of waiting, which is beyond the scope of the fluid models. Their method took advantage of explicit expressions for steady-state performance (which would not have been possible without Poisson arrivals and exponential service times). We must resort to diffusion limits in order to accommodate general arrivals and patience-time distributions.

1.2. Main Contribution

To summarize, the contributions of this paper are as follows:

- A new scaling framework (4) of the patience-time distribution is proposed, which allows it to change abruptly, as observed in practice.
- Inspired by Liu and Whitt (2014b), we develop a new modeling approach ((9) and (16)–(17)) based on the virtual waiting time. This enables diffusion analysis for systems with scaling (4) of their patience-time distribution. Accurate approximations for performances of the original stochastic system are constructed based on the stationary distribution of the diffusion limit.
- Performance approximations for the fixed-delay model of Armony et al. (2009) are improved by using our approximations. Indeed, quoting the authors of the latter (end of their Section 7), we "better quantify the impact of stochastic fluctuations," which they left as a problem for future research.
- We jointly optimize (asymptotically) the two problems of optimal-staffing and announcement-timing, under the assumption (supported by our data and experience) that an announcement causes an abrupt change in the likelihood of abandonment.
- Finally, we analyze the performance gap between fluid and diffusion approximations, for a given patience-time distribution.

1.3. Organization and Notation

The rest of the paper is organized as follows. Section 2 introduces the queueing model and the heavy-traffic regime. We then proceed, in Section 3, to derive the diffusion limits and their stationary distribution, and then use the latter to approximate the steady-state performance of the originating queueing system. Based on the approximation formulae, in Sections 4.1 and 4.2, we investigate the impact of an announcement upon arrival and during waiting, respectively. Section 5 provides an in-depth discussion of our approximation formulae, and Section 6 offers some concluding remarks. The proofs for the theorems and propositions, as well as some complements, are given in the appendix.

We conclude the Introduction with the convention and notations that are used throughout the paper.

All random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, unless otherwise specified; $\mathbb{E}$ is the expectation associated with the probability $\mathbb{P}$. Let $\mathbb{N}$ and $\mathbb{R}$ denote the set of natural numbers and real numbers, respectively. Let $\mathbf{D}([0,\infty),\mathbb{R})$ be the space of right-continuous functions with left-limits, defined on $[0,\infty)$ and taking real values. We equip this space with the Skorohod J_1-topology (see Ethier and Kurtz 1986). For a sequence of random elements $\{X^n\}_{n\in\mathbb{N}}$, taking values in a metric space, we write $X^n \Rightarrow X$ to denote the convergence of X^n to X in distribution. For any $a, b \in \mathbb{R}$, we set $a^+ = \max(a, 0)$ and $a \wedge b = \min(a, b)$. For any probability distribution function $F(\cdot)$, let $F_c(x) = 1 - F(x)$. For any two real-valued nonnegative functions f and g, we write $f(n) = O(g(n))$ if $\limsup_{n\to\infty} f(n)/g(n) < \infty$, and $f(n) = o(g(n))$ if $\limsup_{n\to\infty} f(n)/g(n) = 0$.

2. Model Formulation

Consider a sequence of many-server queueing systems with customer abandonment, indexed by $n \in \mathbb{N}$. In the nth system, there is a single class (queue) of customers who are served by s_n statistically identical servers. Customers arrive according to a counting process $\Lambda^n = \{\Lambda^n(t)\colon t \geqslant 0\}$. For $i \geqslant 1$, let

$$\tau_i^n = \inf\{t \geqslant 0\colon \Lambda^n(t) \geqslant i\}$$

represent the time of the ith arrival to the nth system. In our model, there are no batch arrivals, namely, $\mathbb{P}(\tau_i^n = \tau_{i+1}^n) = 0$, for all $i \in \mathbb{N}$. Assume that there exists a sequence of positive real numbers $\{\lambda_n\}_{n\in\mathbb{N}}$ such that, as $n \to \infty$, we have $\lambda_n \to \infty$ and

$$\tilde{\Lambda}^n \Rightarrow \tilde{\Lambda} \quad \text{with } \tilde{\Lambda}^n(t) = \frac{1}{\sqrt{\lambda_n}}(\Lambda^n(t) - \lambda_n t), \tag{1}$$

where $\tilde{\Lambda} = \{\tilde{\Lambda}(t)\colon t \geqslant 0\}$ is a Brownian motion. Arriving customers are immediately served if any server is idle. Otherwise, they wait in a queue and are served on a first-come, first-served (FCFS) basis. The ith arriving customer requires a service time of v_i^n, and has patience time u_i^n: once waiting time reaches u_i^n, the customer leaves the system immediately without receiving service. Service times are assumed i.i.d., and exponentially distributed with rate μ. Patience times are i.i.d., with a general distribution $F^n(\cdot)$. We also assume that the service times of the initial customers in queue and the remaining service times of the initial customers in service are i.i.d., following the exponential distribution with rate μ. The sequences of service and patience times and the arrival process are mutually independent.

In concert with the assumptions on the ED regime (Whitt 2006) and ED+QED regime (Mandelbaum and Zeltyn 2009), we assume that there exist $\rho > 1$, $\beta \in \mathbb{R}$, and $\{\omega^n\}_{n\in\mathbb{N}}$ converging to ω with $0 < \omega < \infty$, such that, as $n \to \infty$,

$$\lim_{n\to\infty} \frac{\lambda_n}{s_n \mu} = \rho > 1, \tag{2}$$

$$\frac{\lambda_n F_c^n(\omega^n) - s_n \mu}{\sqrt{\lambda_n}} \to \beta; \tag{3}$$

(2)–(3) imply that $F_c^n(\omega^n) \to 1/\rho$. In the special case where $F^n \equiv F$ and $\omega^n \equiv \omega$, (2)–(3) imply (3.6) of Theorem 3.1 in Whitt (2006). Thus ω can be interpreted as the *fluid offered waiting time*. In fact, we show in Proposition 1 that ω serves as the fluid limit of the offered waiting time also when F^n does vary with n.

Note that if $\rho = 1$ and $\omega^n = 0$, then (2)–(3) become the QED regime (see Garnett et al. 2002, Reed and Tezcan 2012). There is a difference, in both analysis and results of the QED and ED+QED regimes. For example, the virtual waiting time in the QED regime is of order $1/\sqrt{\lambda_n}$, while that in the ED+QED regime oscillates around $\omega^n > 0$, in order of $1/\sqrt{\lambda_n}$. Consequently, the diffusion limit of the virtual waiting time is always nonnegative in the QED regime (centered by 0), while it can be both positive and negative in the ED+QED regime (centered by $\omega^n > 0$). In this paper, we focus on the ED+QED regime.

Motivated by data and applications, we assume that for ω^n in (3), the patience-time distributions also satisfy, as $n \to \infty$, that

$$\sqrt{\lambda_n}\left[F^n\left(\omega^n + \frac{x}{\sqrt{\lambda_n}}\right) - F^n(\omega^n)\right] \to f_\omega(x), \tag{4}$$

where $f_\omega(\cdot)$ is a continuous function. This setting is quite general. A simple special case is where $F^n(x) \equiv F(x)$ and $\omega^n = \omega$ (i.e., without any scaling of the patience-time distribution). In this case $f_\omega(x) = f(\omega)x$, where $f(\cdot)$ is the density function of $F(\cdot)$. However, the flexibility (4) of allowing the patience-time distribution to vary with n captures a subtle change around the fluid offered waiting time; we refer to this as a *fine structure*, which has been associated with customers' abrupt reaction to delay announcements, as discussed in the Introduction (see also Figure 3). This enables us to analyze the impact of such announcements in Section 4, and consequently optimize staffing levels and announcement times, jointly.

Readers should note that both β and f_ω in (3)–(4) depend on the sequence $\{\omega^n\}_{n\in\mathbb{N}}$. Nevertheless, given arrival processes Λ^n, number of servers, service rate μ, patience-time distribution F^n, and initial system state, for different choices of the sequence $\{\omega^n\}_{n\in\mathbb{N}}$, as long as they satisfy (3)–(4) and (18), it can be shown that their diffusion limits and stationary distributions are consistent. See Appendix EC.4 for detailed explanations.

We now formally model system dynamics. To this end, we introduce two notions that correspond to waiting times. The first is the *offered waiting time* ω_i^n, which denotes the time that the ith arriving customer in the nth system must wait before receiving service, if that customer would have been infinitely patient, for $i \geqslant 1$. The second is the *virtual waiting time* $V^n(t)$, which is the amount of time that a virtual customer with infinite patience would have to wait before receiving service, had that customer arrived at time t in the nth system. When t happens to be the arrival time of an actual customer, the virtual waiting time at t is regarded as if a virtual customer arrives right after the actual customer. Since there are no batch arrivals,

$$\omega_i^n = V^n(\tau_i^n -).$$

Denote by $A^n(t)$ the number of customers who arrive during the time interval $(0, t]$ and either abandoned or will eventually abandon the nth system. Clearly, $A^n(0) = 0$ and

$$A^n(t) = \sum_{i=1}^{\Lambda^n(t)} \mathbf{1}_{\{u_i^n \leqslant \omega_i^n\}}. \tag{5}$$

Any customer who arrives after time 0 cannot receive service before $V^n(0)$ due to FCFS. For $t \geqslant V^n(0)$, let

$$\kappa^n(t) = \inf\{\tau\colon \tau + V^n(\tau) > t\}. \tag{6}$$

All arrivals before $\kappa^n(t)$ and initial customers are not in queue at time t. Denote by $Q^n(t)$ the number of customers in the queue at time t. The queue length process, for time $t \geqslant V^n(0)$, can be written as

$$Q^n(t) = \sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t)} \mathbf{1}_{\{u_i^n - (t-\tau_i^n) > 0\}}. \tag{7}$$

Note that the queue length representation (7) is similar to (6.5) in Liu and Whitt (2014b), and $\kappa^n(t)$ coincides with their $t - h^n(t)$ ($h^n(t)$ is the age of the head-of-the-line customer). Let $B^n(t)$ denote the number of customers who start service during $(0,\ t]$. Then, for $t \geqslant V^n(0)$,

$$B^n(t) - B^n(V^n(0)-) = \sum_{i=1}^{\Lambda^n(\kappa^n(t))} \mathbf{1}_{\{u_i^n > \omega_i^n\}}. \tag{8}$$

Consider the ith customer who arrived during $(0, t]$, for a fixed i: If $u_i^n \leqslant \omega_i^n$, then this customer is counted in $A^n(t)$. Otherwise, the customer starts service between time $V^n(0)$ and $t + V^n(t)$. There is, therefore, a simple balance equation regarding the arrival process $\Lambda^n(t)$:

$$\Lambda^n(t) = A^n(t) + B^n(t + V^n(t)) - B^n(V^n(0)-), \quad t \geqslant 0. \tag{9}$$

3. Diffusion Approximations and Steady-State Analysis

In this section, we derive our main theoretical results. These include diffusion limits for the virtual waiting time and the number of customers in the system, as well as stationary distributions of the diffusions in the heavy-traffic regime (1)–(4).

3.1. Stochastic Process Limits

Our first result is on the fluid scale, claiming that the virtual waiting time process V^n is asymptotically close to the fluid offered waiting time.

Proposition 1. *In the heavy traffic regime* (1)–(3), *if* $V^n(0) \Rightarrow \omega$ *as* $n \to \infty$, *then for any* $T \geqslant 0$, *as* $n \to \infty$,

$$\sup_{0 \leqslant t \leqslant T} |V^n(t) - \omega| \Rightarrow 0. \tag{10}$$

The proof is provided in Appendix EC.2. The above proposition serves as a first-order fluid approximation. Note that the condition $V^n(0) \Rightarrow \omega$, as $n \to \infty$, is different from prevalent assumptions, which require either that the remaining patience times for the initial customers follow a certain distribution (e.g., Zhang 2013), or that initial customers are infinitely patient (e.g., Mandelbaum and Momčilović 2012). This is because the customer service times are assumed exponential. In this case, the virtual waiting time after 0 only depends on the initial state through the virtual waiting time at 0.

We now pursue a refined approximation of the stochastic deviation from the fluid limit. In light of Proposition 1, we introduce the diffusion-scaled virtual waiting time process $\tilde{V}^n = \{\tilde{V}^n(t)\colon t \geqslant 0\}$ by

$$\tilde{V}^n(t) = \sqrt{\lambda_n}(V^n(t) - \omega^n), \quad t \geqslant 0. \tag{11}$$

It is also necessary to introduce diffusion-scaled abandonment and service processes. As ω^n is the fluid offered waiting time, roughly speaking, the abandonment probability for each arriving customer is $F^n(\omega^n)$. Thus, intuitively, the diffusion-scaled abandonment process $\tilde{A}^n = \{\tilde{A}^n(t)\colon t \geqslant 0\}$ can be defined by

$$\tilde{A}^n(t) = \frac{A^n(t) - \lambda_n t F^n(\omega^n)}{\sqrt{\lambda_n}}. \tag{12}$$

As the system is in the ED+QED regime and service rate is μ, each server is almost surely busy at all times. Hence, the diffusion-scaled service process $\tilde{B}^n = \{\tilde{B}^n(t)\colon t \geqslant 0\}$ should be defined by

$$\tilde{B}^n(t) = \frac{B^n(t) - s_n \mu t}{\sqrt{\lambda_n}}. \tag{13}$$

From (5), we have

$$\begin{aligned}\tilde{A}^n(t) ={}& \tilde{H}^n(t) + \int_0^t \sqrt{\lambda_n} \\ &\cdot \left(F^n\left(\omega^n + \frac{\tilde{V}^n(x-)}{\sqrt{\lambda_n}}\right) - F^n(\omega^n)\right) \mathrm{d}\bar{\Lambda}^n(x) \\ &+ F^n(\omega^n)\tilde{\Lambda}^n(t),\end{aligned} \tag{14}$$

where $\bar{\Lambda}^n(s) = \Lambda^n(s)/\lambda_n$, which is the fluid scaling of $\Lambda^n(s)$, and

$$\tilde{H}^n(t) = \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\Lambda^n(t)} \left(\mathbf{1}_{\{u_i^n \leqslant \omega_i^n\}} - F^n(\omega_i^n)\right). \tag{15}$$

Applying the diffusion scaling to each term in (9), it follows from $t = tF^n(\omega^n) + tF^n_c(\omega^n)$ that

$$\tilde\Lambda^n(t) = \tilde A^n(t) + \frac{(s_n\mu - \lambda_n F^n_c(\omega^n))t}{\sqrt{\lambda_n}} + \tilde B^n(t + V^n(t)) - \tilde B^n(V^n(0)) + \frac{s_n\mu}{\lambda_n}(\tilde V^n(t) - \tilde V^n(0)).$$

This, with (14), helps us write $\tilde V^n$ as

$$\tilde V^n(t) = \tilde V^n(0) - \frac{\lambda_n}{s_n\mu}\int_0^t \sqrt{\lambda_n} \cdot \left(F^n\left(\omega^n + \frac{\tilde V^n(x-)}{\sqrt{\lambda_n}}\right) - F^n(\omega^n)\right) \mathrm{d}\bar\Lambda^n(x) + \tilde Y^n(t), \quad (16)$$

where

$$\tilde Y^n(t) = \frac{\lambda_n}{s_n\mu}\left[F^n_c(\omega^n)\tilde\Lambda^n(t) - \tilde H^n(t) + \frac{(\lambda_n F^n_c(\omega^n) - s_n\mu)t}{\sqrt{\lambda_n}} - \tilde B^n(t + V^n(t)) + \tilde B^n(V^n(0))\right]. \quad (17)$$

Starting with (16), we can establish the following diffusion approximation for the virtual waiting time. This mesoscopic level turns out to be natural for capturing system dynamics that are triggered by customer reactions to delay announcements, as revealed by our call center data.

We assume that the sequence $\{\omega^n\}_{n\in\mathbb{N}}$ in (3)–(4) satisfies

$$\tilde V^n(0) \Rightarrow \tilde V_0 \quad \text{as } n \to \infty. \quad (18)$$

Note that this condition is different from customary, and is enough for our analysis due to the same reason mentioned right after Proposition 1. In Appendix EC.5, we provide a verifiable sufficient condition for (18).

Theorem 1. *In the heavy traffic regime (1)–(4) with the initial condition (18), $\tilde V^n \Rightarrow \tilde V$, as $n \to \infty$, where the limit $\tilde V = \{\tilde V(t)\colon t \geqslant 0\}$ is the unique solution to*

$$\tilde V(t) = \tilde V_0 - \rho\int_0^t [f_\omega(\tilde V(x)) - \beta]\,\mathrm{d}x + \left[\tilde\Lambda(t) - \sqrt{\rho}\mathcal{B}(t) - \sqrt{\rho - 1}\mathcal{B}_A(t)\right], \quad t \geqslant 0; \quad (19)$$

here $\mathcal{B}_A = \{\mathcal{B}_A(t)\colon t \geqslant 0\}$ and $\mathcal{B} = \{\mathcal{B}(t)\colon t \geqslant 0\}$ are two independent standard Brownian motions, which are further independent of $\tilde\Lambda$.

Note that the diffusion limit in (19) generalizes (4.9) in Liu and Whitt (2014b), where $F^n = F$ and F has a continuous density function f, which yields $f_\omega(\tilde V(x)) = f(\omega)\tilde V(x)$. Similar to Liu and Whitt (2014b), the three Brownian motions $\tilde\Lambda$, $\mathcal{B}$ and $\mathcal{B}_A$ capture the stochastic variability of the arrival process, service times and patience times, respectively. The drift term of the diffusion limit contains two parts. The first is the state-dependent drift $f_\omega(\tilde V(\cdot))$, which depends on both $\tilde V(\cdot)$ and the function $f_\omega(\cdot)$ characterizing the patience-time distribution in the general framework (4). The second part β is the drift due to the heavy traffic assumption (3).

Now we consider the diffusion approximation for the queue length process. In view of Proposition 1, the queue length at time t approximately includes the customers who have arrived to the system during time interval $(t - \omega^n, t]$ and have not abandoned by time t. The number of these customers roughly equals $\lambda_n\int_{t-\omega^n}^t F^n_c(t - x)\,\mathrm{d}x$ $(= \lambda_n\int_0^{\omega^n} F^n_c(x)\,\mathrm{d}x)$, since the arrival rate is λ_n (see Whitt 2006). One is led to introduce the diffusion-scaled queue-size $\tilde Q^n = \{\tilde Q^n(t)\colon t \geqslant 0\}$ by

$$\tilde Q^n(t) = \frac{1}{\sqrt{\lambda_n}}\left(Q^n(t) - \lambda_n\int_0^{\omega^n} F^n_c(x)\,\mathrm{d}x\right), \quad t \geqslant 0. \quad (20)$$

Building on the diffusion limit for the virtual waiting time, the diffusion limit of the queue length process is characterized in the following theorem.

Theorem 2. *Assume the heavy traffic regime (1)–(4) with the initial condition (18). Assume further that the sequence of patience-time distributions satisfies*

$$\{F^n(\cdot)\}_{n\in\mathbb{N}} \textit{ converges to } F(\cdot) \text{ on } [0,\omega] \textit{ in total variation}, \quad (21)$$

where $F(\cdot)$ is continuous at ω, and $F_c(\omega) = 1/\rho$. Then $\tilde Q^n \Rightarrow \tilde Q$ on the time interval (ω, ∞), as $n \to \infty$. Here the limit $\tilde Q = \{\tilde Q(t)\colon t \geqslant 0\}$ is given by

$$\tilde Q(t) = \int_{t-\omega}^t F_c(t - x)\,\mathrm{d}\tilde\Lambda(x) + \frac{1}{\rho}\tilde V(t - \omega) + \mathcal{G}(t), \quad (22)$$

where $\tilde\Lambda$ is given by (1), and $\tilde V$ is given by Theorem 1; $\mathcal{G} = \{\mathcal{G}(t)\colon t > \omega\}$ is a Gaussian process independent of $\tilde\Lambda$ and $\tilde V$, with zero mean and covariance

$$\mathbb{E}(\mathcal{G}(t_1)\mathcal{G}(t_2)) = \int_{t_1\wedge(t_2-\omega)}^{t_1} F(t_1 - x)F_c(t_2 - x)\,\mathrm{d}x, \quad \omega < t_1 \leqslant t_2. \quad (23)$$

Note that Condition (21) was not needed for Theorem 1, because only those customers who will eventually receive service affect the virtual waiting time. In fact, Theorem 1 required only a "local" property (condition (4)) of the distribution F^n around ω^n. However, those customers who will eventually abandon affect the queue length since they will stay in queue until their patience times expire. That is, the queue length process depends on customer patience times. Thus, condition (21) on the "global" property of the distributions F^n is needed for Theorem 2. Since we impose no refined assumption on the initial state of the system, one can characterize the asymptotic queue process only after the warm-up period $[0, \omega]$. We note that, in analogy to Krichagina and Puhalskii (1997), the Gaussian process $\mathcal{G}$ can be written as an integral with respect to a Kiefer process; see Krichagina and Puhalskii (1997, p. 237) for details.

3.2. Steady-State Analysis

For the purpose of steady-state analysis, we add the assumption that the arrival process Λ^n is, in fact, a renewal process; let the inter-arrival time have mean $1/\lambda_n$ and variance θ^2/λ_n^2. Then the limit $\tilde{\Lambda}$ in (1) can be written as $\theta\mathscr{B}_\Lambda$, where $\mathscr{B}_\Lambda = \{\mathscr{B}_\Lambda(t)\colon t \geqslant 0\}$ is a standard Brownian motion, independent of $\mathscr{B}$ and $\mathscr{B}_A$ given by Theorem 1. Let

$$\sigma^2 = \theta^2 + 2\rho - 1; \tag{24}$$

then $\tilde{\Lambda} - \sqrt{\rho}\mathscr{B} - \sqrt{\rho-1}\mathscr{B}_A$ is equal in distribution to a process $\sigma\mathscr{W}$, with $\mathscr{W} = \{\mathscr{W}(t)\colon t \geqslant 0\}$ being a standard Brownian motion. According to Theorem 1, the diffusion limit of the virtual waiting time satisfies

$$\tilde{V}(t) = \tilde{V}(0) - \rho\int_0^t \bigl[f_\omega(\tilde{V}(x)) - \beta\bigr]\,\mathrm{d}x + \sigma\mathscr{W}(t), \quad t \geqslant 0. \tag{25}$$

We now calculate the stationary distribution of the diffusion limits $\tilde{V}$ and $\tilde{Q}$. These will be used to derive approximations for performance measures of their originating queueing systems; see Section 3.3.

Proposition 2. *Assume that $f_w(\cdot)$ in Theorem 1 satisfies*

$$\lim_{x\to\infty} f_w(x) > \beta \quad \textit{and} \quad \lim_{x\to-\infty} f_w(x) < \beta, \tag{26}$$

where β is given in (3). Then the diffusion limit $\tilde{V}$ has a stationary distribution $\pi(\cdot)$ given by

$$\pi(y) = C\exp\biggl(-\frac{2\rho}{\sigma^2}\int_0^y [f_\omega(x) - \beta]\,\mathrm{d}x\biggr), \quad y \in \mathbb{R}, \tag{27}$$

where C is a normalizing constant. Similarly, the stationary distribution of the queue length diffusion limit $\{\tilde{Q}(t)\colon t > \omega\}$ in Theorem 2 exists as well. Denote by $\tilde{Q}(\infty)$ a random variable with such a distribution; then it can be written, in distribution, as

$$\tilde{Q}(\infty) = \tilde{\mathscr{N}}_1 + \tilde{\mathscr{N}}_2 + \frac{1}{\rho}\tilde{V}(\infty),$$

where $\tilde{\mathscr{N}}_1$, $\tilde{\mathscr{N}}_2$ and $\tilde{V}(\infty)$ are mutually independent, $\tilde{V}(\infty)$ is a random variable with density function $\pi(\cdot)$, $\tilde{\mathscr{N}}_1$ and $\tilde{\mathscr{N}}_2$ are normal random variables, both with mean zero and variances $\theta^2\int_0^\omega (F_c(x))^2\,\mathrm{d}x$ and $\int_0^\omega F(x)F_c(x)\,\mathrm{d}x$, respectively.

We comment here that the density function $\pi(\cdot)$, and in particular the normalizing constant C, depend on the limit $f_\omega(\cdot)$ in (4). We later apply this model flexibility in Section 5, demonstrating that C and $\pi(\cdot)$ can thus have different analytical expressions depending on the application. Condition (26) on f_ω is needed to ensure existence of the stationary distribution $\pi(\cdot)$, and it is not a restrictive requirement in our applications; see the examples in Section 4.

3.3. Approximation of the Originating System

We have established limits for a *sequence* of systems. These limits will now support an approximation for a single given system—specifically closed-form formulae for its steady-state waiting time and queue length. To this end, and as is often the case (e.g., Reed and Ward 2008, Reed and Tezcan 2012), we presume the validity of a limit-interchange, which justifies the steady-state approximation of a queueing system by its diffusion approximation. In practice, one can typically observe/estimate system parameters: (i) the number of servers s and individual service rate μ; (ii) the patience-time distribution $H(\cdot)$; and (iii) mean and variance of the inter-arrival time $1/\lambda$ and θ^2/λ^2, respectively. We denote the system by $(s, \mu, \lambda, \theta^2, H)$. The heavy traffic assumptions (3) and (4) constitute a mathematical tool that guides on how to capture the structure of the patience-time distribution $H(\cdot)$ around ω, and its impact on performances for a single system with s servers. We rely on the stationary distribution π in Proposition 2 to obtain the closed-form approximation formulae for this particular system. The key challenge is to map the stationary distribution of the diffusion limit to the one corresponding to the originating system $(s, \mu, \lambda, \theta^2, H)$. In view of (27), set

$$\rho := \frac{\lambda}{s\mu}, \quad \sigma^2 = \theta^2 + 2\rho - 1, \tag{28}$$

$$\beta := \frac{\lambda H_c(\omega) - s\mu}{\sqrt{\lambda}}, \tag{29}$$

$$f_\omega(x) := \sqrt{\lambda}\biggl[H\biggl(\omega + \frac{x}{\sqrt{\lambda}}\biggr) - H(\omega)\biggr]. \tag{30}$$

The above gives rise to our approximation formulae for the particular system $(s, \mu, \lambda, \theta^2, H)$. Let $V(\infty)$ denote the steady-state of its virtual waiting time. Then by Proposition 2, the density of $\sqrt{\lambda}(V(\infty) - \omega)$ can be approximated by

$$\pi(y) = C\exp\biggl(-\frac{2\rho}{\sigma^2}\int_0^y [f_\omega(x) - \beta]\,\mathrm{d}x\biggr), \tag{31}$$

where

$$C = \biggl(\int_{-\infty}^{\infty} \exp\biggl(-\frac{2\rho}{\sigma^2}\int_0^y [f_\omega(x) - \beta]\,\mathrm{d}x\biggr)\mathrm{d}y\biggr)^{-1} \tag{32}$$

is the normalizing constant.

Waiting Time. For the system $(s, \mu, \lambda, \theta^2, H)$, denote by $W(\infty)$ the steady-state of its waiting time. Then $W(\infty)$ is just the minimum between $V(\infty)$ and the customer patience time. Thus, by (31),

$$\begin{aligned}
&\mathbb{P}(W(\infty) > y)\\
&= H_c(y)\mathbb{P}(\sqrt{\lambda}(V(\infty) - \omega) > \sqrt{\lambda}(y - \omega))\\
&\approx H_c(y)\int_{\sqrt{\lambda}(y-\omega)}^{\infty} C\exp\biggl(-\frac{2\rho}{\sigma^2}\int_0^v [f_\omega(x) - \beta]\,\mathrm{d}x\biggr)\mathrm{d}v.
\end{aligned} \tag{33}$$

A special choice of y is ω, which will be used in the optimal staffing problem that we discuss in Section 4.

Queue Length. For the system $(s, \mu, \lambda, \theta^2, H)$, denote by $Q(\infty)$ the steady-state of its queue length. By Proposition 2, $Q(\infty)$ can be approximated by

$$\lambda \int_0^{\omega} H_c(x)\,\mathrm{d}x + \sqrt{\lambda}\left(\tilde{\mathcal{N}}_1 + \tilde{\mathcal{N}}_2 + \frac{\sqrt{\lambda}}{\rho}(V(\infty) - \omega)\right).$$

Consequently, performance measures related to $Q(\infty)$ can be calculated explicitly using (31) and the above formula. For example, the expected steady-state queue length can be calculated as

$$\mathbb{E}Q(\infty) \approx \lambda \int_0^{\omega} H_c(x)\,\mathrm{d}x + \frac{1}{\rho}\sqrt{\lambda} \int_{-\infty}^{\infty} x\pi(\mathrm{d}x). \qquad (34)$$

It is clear from the above that the patience-time distribution significantly affects system performance. In the next section, we shall use the above to analyze delay announcements and solve related staffing problems. The accuracy of the above approximations is demonstrated in Section 5, using various patience-time distributions.

4. Impact of Delay Announcements

The advantage of our refined approximation is the ability to capture the fine structure of the patience-time distribution. This is necessary for our applications where delay announcements cause a sudden change of the patience-time distribution at a certain "impact point." As in Mandelbaum and Zeltyn (2009), we now apply our refined approximation to solve two staffing problems for a call center in the ED+QED regime. The first one, arising from the application of a delay announcement upon arrival, is to minimize the staffing level so as to achieve a service level constraint (see (38)). The second, arising from the application of an announcement during waiting, has the additional option of an announcement epoch, while sharing the same objective of minimizing staffing (see (44)). The patience-time distributions in both applications are not smooth enough to use existing results, but have the feature characterized by (4); here ω can be differently interpreted, depending on the application, which we elaborate on in the following two sections.

4.1. Delay Announcement Upon Arrival

As described in the Introduction, arriving customers who must wait often receive an announcement upon arrival concerning their anticipated delay. The announced information could include a single number τ related to the delay, which is called a *fixed delay announcement* by Armony et al. (2009). Following their model description, customers respond to the announcement by choosing to balk or not. Given a delay announcement of τ, the probability that an arriving customer chooses to balk is $B(\tau)$. Here $B(\cdot)$ is assumed to be a distribution function. Arriving customers who do not balk (with probability $B_c(\tau)$) join the queue and their patience times are also affected by the announced information τ. This effect is modeled by assuming that customers' patience times follow the conditional distribution $H(t \mid \tau)$.

An announced delay ω_e is an *equilibrium delay* if either (i) $\rho B_c(0) \leqslant 1$ and $\omega_e = 0$, or (ii) $\rho B_c(0) > 1$ and

$$\begin{aligned} &\rho B_c(\omega_e) H_c(\omega_e \mid \omega_e) \leqslant 1 \quad \text{and} \\ &\rho B_c(\omega_e) H_c(t \mid \omega_e) > 1, \quad \text{for } 0 \leqslant t < \omega_e. \end{aligned} \qquad (35)$$

When there is a unique equilibrium delay ω_e, the above formal relations capture the facts that, in equilibrium, and asymptotically in the fluid scale, the announced delay τ is equal to the long-run average delay of served customers, and both are equal to the equilibrium delay ω_e. Our model setting, and specifically the concept of equilibrium delay in the fluid scale, were introduced by Armony et al. (2009). They also proved the uniqueness of equilibrium delay under some regularity conditions (Theorem 4.1 in Armony et al. 2009), which are satisfied in the following model.

In this subsection, we consider "the all-exponential conditional hazard-rate model" proposed in Armony et al. (2009), where customer arrivals are assumed to be a Poisson process with rate λ, and their service times are exponentially distributed with rate μ. For a delay announcement of τ, the balking probability is $B(\tau) = 1 - e^{-b\tau}$ and customers' patience is

$$H(t \mid \tau) = \begin{cases} 1 - e^{-h_0 t}, & 0 \leqslant t \leqslant \tau; \\ 1 - e^{-h_0 \tau} e^{-h_1 (t-\tau)}, & t > \tau; \end{cases} \qquad (36)$$

here h_0 and h_1 are two parameters, which capture abandonment behavior before and after the announcement τ. Here the abrupt change from h_0 to h_1 at τ occurs because the promise for delay duration will be violated once waiting time exceeds τ. In order to compare the fluid approximation developed by Armony et al. (2009) with the diffusion approximation given by Proposition 2, h_0 and h_1 are assumed to be constants independent of τ. It follows from (35) that the unique equilibrium delay is

$$\omega_e = \frac{1}{b + h_0} \ln \rho. \qquad (37)$$

With this setting, customers' patience-time distribution has different left and right derivatives at the announced delay $\tau = \omega_e$. Armony et al. (2009) derived fluid approximations for various performance metrics of such a system (see Table 1 therein). However, as they observed, some performance metrics (such as expected queue length and probabilities related to waiting time) are not well approximated by their fluid model—indeed, a refinement is called for.

Though balking is not formally incorporated in our diffusion analysis, our approximation can easily accommodate it by regarding balking customers as having zero-patience, which is allowed by our assumptions. This gives rise to the patience-time distribution $\bar{H}(\cdot \mid \omega_e) = B(\omega_e) + B_c(\omega_e) H(\cdot \mid \omega_e)$. We now consider the same set

Table 1. Comparing Fluid and Diffusion Approximations with Simulated Performance Metrics

Performance measure		$h_1 = 0.5$		$h_1 = 4$	
$h_0 = 0.5$	Fluid	Simulated	Diffusion	Simulated	Diffusion
$\mathbb{E}[Q(\infty)]$	23.7	24.3	23.7	17.3	16.4
$\mathbb{E}(W(\infty); B_c)$	0.212	0.217	0.224	0.155	0.151
$\mathbb{P}(W(\infty) \leqslant \omega_e \mid S)$	1	0.512	0.512	0.754	0.756

of parameters as in Armony et al. (2009). The number of servers $s = 100$ with individual service rate $\mu = 1$. The arrival rate $\lambda = 140$, balking rate $b = 1$ and the patience-time hazard rate $h_0 = 0.5$ and $h_1 = 4$. By (37), the equilibrium fluid delay is $\omega_e = 0.224$. As pointed out in Armony et al. (2009), their fluid approximation is not nearly close to the simulation when $h_1 = 4 > h_0 = 0.5$, though it agrees closely when $h_1 = h_0 = 0.5$. In Table 1, it is seen that our diffusion analysis yields much improved approximations. The columns labeled "Simulated" and "Fluid" are taken from Armony et al. (2009, Table 1). The columns labeled "Diffusion" are calculated using formula (31). In particular, $\mathbb{E}[Q(\infty)]$ is calculated based on (34), and $\mathbb{E}[W(\infty); B_c]$ based on (33). The calculation of $\mathbb{P}(W(\infty) \leqslant \omega_e \mid S)$, the conditional probability that the steady-state waiting time is less than ω_e, follows from

$$\mathbb{P}(W(\infty) \leqslant \omega_e \mid S) = \frac{\mathbb{P}(W(\infty) \leqslant u \wedge \omega_e)}{\mathbb{P}(W(\infty) \leqslant u)},$$

where u is a generic random variable independent of $W(\infty)$, and following the distribution $\tilde{H}(\cdot \mid \omega_e)$; $\mathbb{P}(W(\infty) \leqslant y)$ can be calculated from (33), for any y.

4.1.1. Implications to Optimal Staffing.

As proposed by Mandelbaum and Zeltyn (2009), the ED+QED regime is useful for staffing under the constraint satisfaction $\mathbb{P}(W(\infty) > z)$, for a benchmark z. Here we revisit the staffing problem, using our refined approximation, for those applications where the patience-time distribution is not smooth enough to apply the result in Mandelbaum and Zeltyn (2009).

We describe a general approach rather than restrict ourselves to the setting of delay announcements, e.g., to the all-exponential model. Let the individual service rate μ, arrival rate λ, variance of the interarrival time θ^2/λ^2, and the patience-time distribution $H(\cdot)$ be given. We seek the number of servers s such that staffing cost is minimized while adhering to the given service level agreement (z, α) as follows:

$$\begin{aligned} &\min \ s \\ &\text{s.t.} \ \ \mathbb{P}(W(\infty) > z) \leqslant \alpha. \end{aligned} \tag{38}$$

For a general H, we set $z = \omega$. Thus, we often face a situation where one must account for the fine structure of the patience-time distribution around the benchmark z. (For example, the hazard rate has a jump in the above all-exponential model, implying that the left and right derivatives of the patience-time distribution are not equal.) We demonstrate in this subsection that our diffusion analysis, which is general enough to accommodate such a fine structure of the patience-time distribution, can help not only in performance evaluation but also with optimal staffing subject to constraints on the tail probability.

We now propose an asymptotically optimal staffing rule, which solves (38). It is based on the steady-state approximations (31)–(33) in Section 3.2. From (29), the number of servers is

$$s = \left\lceil \frac{\lambda}{\mu} H_c(\omega) - \frac{\beta}{\mu}\sqrt{\lambda} \right\rceil. \tag{39}$$

Note that β is the only element needed to determine the number of servers s. We calculate it via (33), by solving the optimization problem (38). This gives rise to

$$\begin{aligned} &\max \ \beta \\ &\text{s.t.} \ \ H_c(\omega) \int_0^\infty C \exp\left(-\frac{2\rho}{\sigma^2} \int_0^y [f_\omega(x) - \beta]\,\mathrm{d}x\right) \mathrm{d}y \\ &\quad \leqslant \alpha. \end{aligned} \tag{40}$$

To demonstrate the applicability and accuracy of our approximation for staffing, we performed numerical studies for the following two examples, without a delay announcement for clarity. We shall revisit these two examples in Section 5, for an in-depth discussion on performance evaluation.

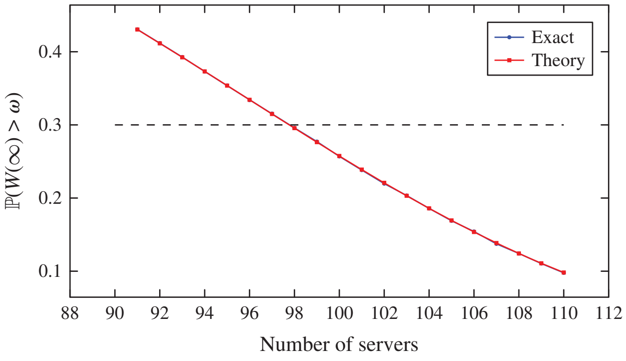
Example 1. Customers arrive according to a Poisson process with rate λ. Service times are exponentially distributed with rate μ. The patience-time distribution is

$$H(x) = \begin{cases} x, & x \leqslant \omega, \\ \omega + \kappa(x - \omega), & \omega < x \leqslant \omega + \dfrac{1-\omega}{\kappa}, \end{cases} \tag{41}$$

where ω is the fluid offered waiting time.

In this example, set the parameter $\omega = 1/6$ and $\kappa = 5$ in (41). Assume that the individual service rate $\mu = 1$ and the arrival rate $\lambda = 120$. It is clear that $H(\cdot)$ is not differentiable at the fluid offered waiting time ω. Thus, we cannot use existing results such as Mandelbaum and Zeltyn (2009). Applying the staffing rule (39) by numerically solving (40) with $\alpha = 0.3$, we get that the optimal number of servers is 97. To show the accuracy of our approach, we plot the tail probability versus the number of servers in Figure 1. The vertical axis is the probability that waiting time exceeds the fluid offered

Figure 1. (Color online) Staffing Level and Probability That Waiting Time Exceeds ω



waiting time. The number of servers ranges from 91 to 110 on the horizontal axis. Figure 1 demonstrates that the approximation, based on our theory, is accurate: The curve obtained almost overlaps the curve by simulation. In fact, the optimal number of servers required in order to achieve $\mathbb{P}(W(\infty) > \omega) \leqslant \alpha$ is almost identical to the solution found by the numerical simulation, for service level α ranging from 0.1 to 0.4.

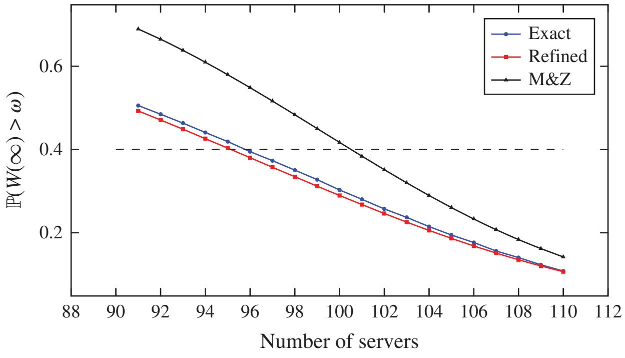
Example 2. Customers arrive according to a Poisson process with rate λ. Service times are exponentially distributed with rate μ. The hazard rate of the patience-time distribution is as follows:

$$\tilde{h}(x) = \begin{cases} h_0, & x \leqslant \omega, \\ h_0 + \kappa(x - \omega), & x > \omega, \end{cases} \tag{42}$$

where ω is the fluid offered waiting time.

In this example, set the parameter $\omega = \ln(1.2)/h_0$, $h_0 = 1.0$ and $\kappa = 100$ in (42). Assume that the individual service rate $\mu = 1$ and the arrival rate $\lambda = 120$. In Figure 2, the number of servers ranges from 91 to 110. The vertical axis is the probability that waiting time exceeds the fluid offered waiting time. Again, we apply the staffing rule (39) by numerically solving (40) with $\alpha = 0.4$, to show that the optimal number of servers is 96. To demonstrate the accuracy, and also to compare against the staffing rule given by Mandelbaum and Zeltyn (2009), we plot the tail probability versus the number of servers in Figure 2. The figure shows that our approximation is much closer to the simulated results. In fact, our refined approximation gives a "near" optimal solution, which is off at most by 1, for any service level ranging from 0.1 to 0.7. Without using it, the error in the staffing level could range from 2 to 8 or even higher. For instance, the staffing level based on the approximation by Mandelbaum and Zeltyn (2009), with $\alpha = 0.4$, gives rise to the optimal number of servers 101, which is off by 5.

Figure 2. (Color online) Staffing Level and Probability That Waiting Time Exceeds ω



Note. Here $\kappa = 100$.

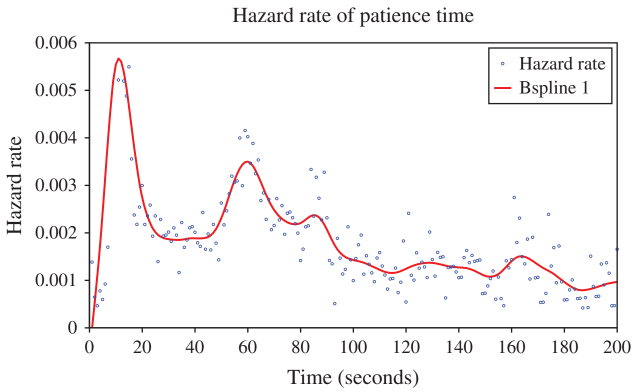
The above two examples demonstrate that our approach can handle more general settings and results in more accurate staffing, when compared to existing methods.

Remark 1. To apply the staffing method in Section 4.1.1 to the delay announcement model in Section 4.1, one simply replaces $H_c(\omega)$ by $\tilde{H}(\omega_e|\omega_e)$ in (40), where $\omega_e = \tau$ is the equilibrium delay. It is, furthermore, natural to choose $z = \tau$, since then the constraint in (38) bounds the fraction of "broken promises." For the all-exponential model, the optimal staffing level is determined by (39), with β replaced by β_*—the optimal solution to (40). We see that β_* essentially depends on $f_\omega(\cdot)$. In comparison to the case without announcement, one can verify that the staffing level in the case with announcement is at least $O(\sqrt{\lambda})$ smaller than the case without announcement; hence, if properly designed under the appropriate circumstances, an announcement can reduce staffing without hurting service levels.

4.2. Delay Announcement During Waiting: Asymptotic Optimality

Instead of making announcements upon arrival, one can make an announcement while customers are

Figure 3. (Color online) Hazard Rates of a Patience-Time Distribution in an Israeli Call Center



Notes. The peak at 60 seconds can be attributed to an announcement at that time: "You are number X in queue, and the first one has been waiting Y seconds." (The peak at 10 seconds arises from unwilling-to-wait customers.)

waiting. As already indicated, such announcements have been often associated with an abrupt change in the patience-time distribution. For example, Figure 3 (taken from Mandelbaum and Zeltyn 2013) depicts a surge of the hazard rate when waiting time reaches 60 seconds: The latter is precisely the time when an announcement is made to customers.

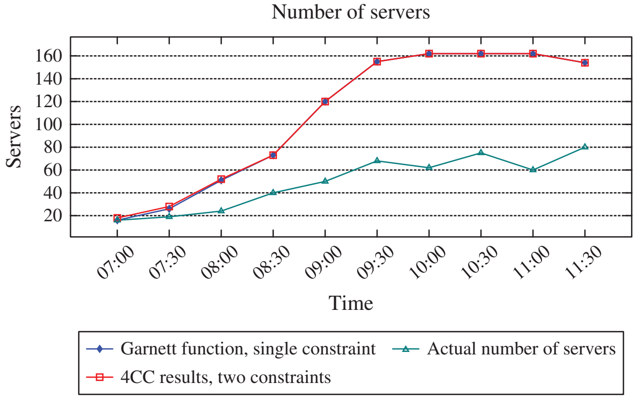
A Queueing Model. Based on the previously-mentioned empirical support that an announcement is associated with a sudden jump, we propose the following stylized (yet insightful) model. The announcement is made when customers' waiting time reaches τ, and the patience-time distribution is accordingly assumed to follow

$$H(x \mid \tau) = \begin{cases} 1 - e^{-h_0 x}, & x \leqslant \tau, \\ 1 - e^{-h_0\tau - h_1(x-\tau)}, & x > \tau; \end{cases} \tag{43}$$

Here h_0 and h_1 are given parameters with $h_0 < h_1$, which captures the increase of the hazard rate caused by the announcement made at τ. The nonsmooth change of the hazard rate, from h_0 to h_1 at τ, can be interpreted as customers' reactions to hearing the announcement. (Note that hazard rate models are, necessarily, not models about individual response, but rather models of aggregate behavior of a customer population.) We comment here that the form (43) was chosen merely for concreteness and simplicity. Indeed, the results in this subsection hold for general patience-time distributions; for example, the hazard rate can also drop back to h_0 from h_1 after a while. The example (43) is also relevant and insightful, in that it captures the essence of the data in Figure 3, in which an announcement abruptly and temporarily increases the hazard rate. Not only is $h_0 < h_1$ supported by empirical data, it is also quite intuitive. If $h_0 = h_1$, the announcement does not make any difference, thus there is no need to study it. Suppose $h_0 > h_1$, meaning that the announcement "encourages" customers to remain online. Essentially, such encouragement would add a load to the system. If we maintain the same staffing, then service quality worsens, as measured by the probability of waiting time exceeding a certain threshold. To maintain the same service quality, one must staff more servers and, hence, it is better to have no announcement. Finally, as suggested by Mandelbaum and Zeltyn (2013), regardless of the purpose of an announcement, the ultimate result is conceivably an encouragement for customers to abandon the system.

A Motivating Example: Controlling a Surge in Arrivals by an Announcement. To motivate our joint optimization of staffing and timing of the announcement, we give an example where a call center (U.S. bank) faced a surge of demand over three days (October 9–11), following a special promotion. Figure 4 shows the actual number of servers employed on October 10 and what would be required (calculated using the Garnett function in Garnett et al. 2002 and 4CC in http://ie.technion.ac.il/serveng/4CallCenters, respectively) if we wanted the call center to operate in the QED regime. During most of the day, this required a number almost double the number actually employed. Consequently, the call center experienced a high abandonment fraction (around 50%, as seen in Figure 5).

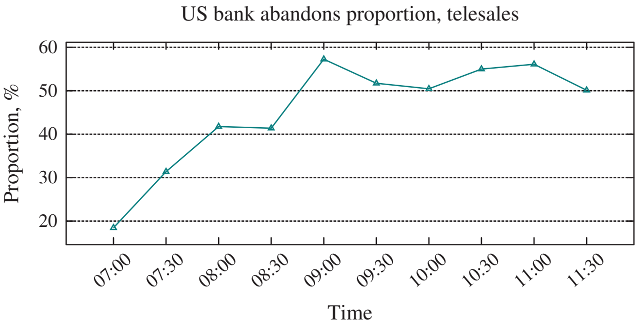
Figure 4. (Color online) Required Staffing to Operate in the QED Regime (Red) vs. Actual Staffing (Blue)



We thus propose a joint staffing and announcement solution, in order to minimize staffing cost with the help of an announcement, while satisfying a service-quality level specified by both the abandonment probability and the tail probability of waiting time (see (44) in the sequel).

4.2.1. An Optimization Model. We now investigate whether the management of an overloaded call center can benefit from a delay announcement, with the intention of minimizing the staffing level while subject to a prespecified service level. The latter is characterized by two constraints: fraction of abandonment less than α_1 and fraction of those waiting above a threshold less than α_2. We require $\alpha_2 < 1 - \alpha_1$ since otherwise the ED+QED regime is not suitable for the staffing (see Remark 4.6 in Mandelbaum and Zeltyn 2009 for explanation). The first constraint on abandonment probability $\mathbb{P}(Ab)$ is closely related to revenue generation, as abandonment typically means revenue

Figure 5. (Color online) Fraction of Abandonment Under Actual Staffing



loss. The second constraint caters mainly to customer satisfaction. The optimization problem is hence formulated as follows:

$$\begin{aligned} \min_{s,\tau}\ & s \\ \text{s.t}\ & \mathbb{P}(\mathrm{Ab}) \leqslant \alpha_1, \\ & \mathbb{P}(W(\infty) > \omega_\tau \wedge \tau) \leqslant \alpha_2, \end{aligned} \tag{44}$$

where ω_τ is determined by the equation $H(\omega_\tau \mid \tau) = \alpha_1$. It should be pointed out that both constraints depend on the announcement time τ. There is an interesting trade-off here: Making an early announcement helps satisfy the second constraint, but may cause too much abandonment (recall that $h_1 > h_0$), thus violating the first constraint. On the other hand, a late announcement has no impact on the system.

We study the optimization problem (44) similarly to (38), by using the approximation formula (33) from our diffusion analysis, with some technical adjustments. To gain insight, we analyze a large-scale limit where the arrival rate λ increases indefinitely. To highlight the dependence on the arrival rate λ for the optimal solution to (44), a superscript λ is added to each of its components.

Proposition 3. *For the stylized model (43) and (44), let $(s_*^\lambda, \tau_*^\lambda)$ be an optimal solution to (44). Denote by τ_* the unique solution to $H(\tau_* \mid \tau_*) = \alpha_1$. Then any optimal announcement epoch τ_*^λ satisfies*

$$\lim_{\lambda\to\infty} \sqrt{\lambda}(\tau_* - \tau_*^\lambda) = 0.$$

Moreover, the optimal number of servers is

$$s_*^\lambda = \frac{\lambda}{\mu}(1-\alpha_1) - \sqrt{\lambda}\frac{\beta_*}{\mu} + o(\sqrt{\lambda}), \tag{45}$$

where β_* is the unique solution to

$$\begin{aligned} \max_{\beta}\ & \beta \\ \text{s.t.}\ & (1-\alpha_1)\frac{\int_0^\infty \exp(-(2\rho/\sigma^2)\int_0^y [h_{\tau_*}(x)-\beta]\,\mathrm{d}x)\,\mathrm{d}y}{\int_{-\infty}^\infty \exp(-(2\rho/\sigma^2)\int_0^y [h_{\tau_*}(x)-\beta]\,\mathrm{d}x)\,\mathrm{d}y} \\ & \leqslant \alpha_2, \end{aligned} \tag{46}$$

with $h_{\tau_*}(x) = \lim_{\lambda\to\infty} \sqrt{\lambda}[H(\tau_* + x/\sqrt{\lambda} \mid \tau_*) - H(\tau_* \mid \tau_*)]$.

The proof of this proposition is presented in Appendix EC.2. The key message is that, in the asymptotic sense, it is optimal to make the announcement so that the abandonment fraction is exactly α_1, and then set the optimal staffing level according to (45). The staffing level then depends on β_* only. Comparing this to the case without announcement (with $f_{\tau_*}(x)$ in (46) replaced by $e^{-h_0\omega}x$), it is easy to verify that β_* in the case with announcement is smaller; hence, the staffing level in (45) is reduced by $O(\sqrt{\lambda})$.

Remark 2. The abrupt change in the patience hazard rate, characterized by h_0 and h_1 in (43), is assumed to be independent of τ. We note, however, that this change may depend on the announcement time τ in practice, but this is left for future research.

5. Refined Approximations

The approximation (31), together with the derived formulae (33)–(34) for the steady-state, does not depend on a particular characteristic (such as derivative or hazard rate) of the patience-time distribution. This generality provides an accurate recipe that is free of details. With additional structure information about the patience-time distribution (e.g., exact values of the left and right derivatives), it leads to insights beyond what has been previously known. At the end of the section, we use such a structure to shed light on the asymptotic gap between fluid and diffusion approximation.

5.1. Using the Density of Patience-Time Distribution

Assume that the patience-time distribution $H(\cdot)$ has a density at ω and write it as $H'(\omega)$. Then $f_\omega(x)$, defined in (30), can be "well approximated" by $H'(\omega)x$, and the density π in (31) can be approximately specialized to

$$\pi(x) \approx C\exp\left(-\frac{\rho}{\sigma^2}[H'(\omega)x^2 - 2\beta x]\right). \tag{47}$$

This implies that $\sqrt{\lambda}(V(\infty) - \omega)$ asymptotically follows a normal distribution. Consequently, it follows from (33) that

$$\begin{aligned} &\mathbb{P}(W(\infty) > y) \\ &\quad\approx H_c(y)\int_{\sqrt{\lambda}(y-\omega)}^{\infty} C\exp\left(-\frac{\rho}{\sigma^2}[H'(\omega)x^2 - 2\beta x]\right)\mathrm{d}x \\ &\quad= H_c(y)\mathbb{P}\left(\frac{\beta}{H'(\omega)} + \frac{\sigma}{\sqrt{2\rho H'(\omega)}}\mathcal{N} > \sqrt{\lambda}(y-\omega)\right) \\ &\quad= H_c(y)\Phi_c\left(\frac{-\beta\sqrt{2\rho} + \sqrt{2\rho\lambda}H'(\omega)(y-\omega)}{\sigma\sqrt{H'(\omega)}}\right), \end{aligned} \tag{48}$$

where $\mathcal{N}$ is a standard normal random variable and $\Phi(\cdot)$ is its distribution function.

This case is related to Mandelbaum and Zeltyn (2009). Specifically, if the number of servers is $s = H_c(\omega)\cdot(\lambda/\mu) + \delta\sqrt{\lambda/\mu}$, with any finite constant δ, and the arrival process is a Poisson process, then we have $\sigma^2 = 2\rho$ and $\beta = -\delta\sqrt{\mu}$ by (28)–(29). Furthermore, we have

$$\mathbb{P}(W(\infty) > \omega) \approx H_c(\omega)\Phi_c\left(\frac{\delta\sqrt{\mu}}{\sqrt{H'(\omega)}}\right).$$

This is consistent with Mandelbaum and Zeltyn (2009, Theorem 4.3). There is also a connection to Bassamboo and Randhawa (2010), where accuracy of the fluid approximation to the expected queue length is studied. If $\beta = 0$, by the symmetry of the normal distribution and in view of (34), we have

$$\mathbb{E}Q(\infty) \approx \lambda\int_0^\omega H_c(x)\,\mathrm{d}x. \tag{49}$$

The expected queue length, derived from the diffusion approximation, has thus been reduced to the one given by the fluid approximation. This provides an alternative support for why the fluid model in itself gives an accurate approximation to queue length, a phenomenon discussed in Bassamboo and Randhawa (2010).

5.2. Using the Left and Right Derivatives of the Patience-Time Distribution

Assume now that the left and right derivatives of the patience-time distribution $H(\cdot)$ at ω, $H'(\omega+)$ and $H'(\omega-)$, are not equal. Following (30), for large λ,

$$f_\omega(x) = \sqrt{\lambda}\left[H\left(\omega + \frac{x}{\sqrt{\lambda}}\right) - H(\omega)\right] \approx \begin{cases} H'(\omega-)x, & x \leqslant 0, \\ H'(\omega+)x, & x > 0. \end{cases} \tag{50}$$

The density π in (31) can be then approximately specialized to

$$\pi(x) \approx \begin{cases} C\exp\left(-\dfrac{1}{2}\dfrac{(x-\beta/(H'(\omega-)))^2}{(\sigma^2/2)(1/(\rho H'(\omega-)))}\right) \\ \quad \cdot\exp\left(\dfrac{\rho\beta^2}{\sigma^2 H'(\omega-)}\right), \quad x \leqslant 0, \\ C\exp\left(-\dfrac{1}{2}\dfrac{(x-\beta/(H'(\omega+)))^2}{(\sigma^2/2)(1/(\rho H'(\omega+)))}\right) \\ \quad \cdot\exp\left(\dfrac{\rho\beta^2}{\sigma^2 H'(\omega+)}\right), \quad x > 0, \end{cases} \tag{51}$$

where the normalizing constant satisfies

$$C^{-1} = \Phi_c\left(\frac{-\sqrt{2\rho}\beta}{\sigma\sqrt{H'(\omega+)}}\right)\exp\left(\frac{\rho\beta^2}{\sigma^2 H'(\omega+)}\right)\frac{1}{\sqrt{H'(\omega+)}} + \Phi\left(\frac{-\sqrt{2\rho}\beta}{\sigma\sqrt{H'(\omega-)}}\right)\exp\left(\frac{\rho\beta^2}{\sigma^2 H'(\omega-)}\right)\frac{1}{\sqrt{H'(\omega-)}}.$$

The steady-state probability, that waiting time exceeds ω, can be approximated via (33) by

$$\mathbb{P}(W(\infty) > \omega) \approx C\cdot H_c(\omega)\Phi_c\left(\frac{-\sqrt{2\rho}\beta}{\sigma\sqrt{H'(\omega+)}}\right) \cdot\exp\left(\frac{\rho\beta^2}{\sigma^2 H'(\omega+)}\right)\frac{1}{\sqrt{H'(\omega+)}}. \tag{52}$$

It follows from (34) that the expected queue length is

$$\begin{aligned} &\mathbb{E}Q(\infty) \\ &\approx \lambda\left(\omega - \frac{\omega^2}{2}\right) + C\frac{\sqrt{\lambda}}{\rho}\left\{\int_0^\infty x\exp\left(-\frac{1}{2}\frac{(x-\beta/(H'(\omega+)))^2}{(\sigma^2/2)(1/(\rho H'(\omega+)))}\right)\right. \\ &\quad \cdot\exp\left(\frac{\rho\beta^2}{\sigma^2 H'(\omega+)}\right)\mathrm{d}x \\ &\quad -\int_0^\infty x\exp\left(-\frac{1}{2}\frac{(x+\beta/(H'(\omega-)))^2}{(\sigma^2/2)(1/(\rho H'(\omega-)))}\right) \\ &\quad \left.\cdot\exp\left(\frac{\rho\beta^2}{\sigma^2 H'(\omega-)}\right)\mathrm{d}x\right\} \\ &= \lambda\left(\omega - \frac{\omega^2}{2}\right) + C\frac{\sqrt{\lambda}}{\rho}\left\{\frac{\sigma^2}{2\rho}\left[\frac{1}{H'(\omega+)} - \frac{1}{H'(\omega-)}\right]\right. \\ &\quad + \frac{\sqrt{2\pi}\sigma}{\sqrt{2\rho}}\left[\frac{\beta}{(H'(\omega+))^{3/2}}\exp\left(\frac{\rho\beta^2}{\sigma^2 H'(\omega+)}\right)\Phi_c\left(-\frac{\sqrt{2\rho}\beta}{\sigma\sqrt{H'(\omega+)}}\right)\right. \\ &\quad \left.\left. + \frac{\beta}{(H'(\omega-))^{3/2}}\exp\left(\frac{\rho\beta^2}{\sigma^2 H'(\omega-)}\right)\Phi\left(-\frac{\sqrt{2\rho}\beta}{\sigma\sqrt{H'(\omega-)}}\right)\right]\right\}. \end{aligned}$$

The last derivation is, in fact, a generalization of the Garnett function, introduced in Garnett et al. (2002). To make the connection, let $h_{\mathcal{N}}(\cdot)$ be the hazard rate for the standard Normal distribution. Then (52) can be written as

$$\begin{aligned} &\mathbb{P}(W(\infty) > \omega) \\ &\approx H_c(\omega)\left(1 + \sqrt{\frac{H'(\omega+)}{H'(\omega-)}}\frac{h_{\mathcal{N}}(-\sqrt{2\rho}\beta/(\sigma\sqrt{H'(\omega+)}))}{h_{\mathcal{N}}(\sqrt{2\rho}\beta/(\sigma\sqrt{H'(\omega-)}))}\right)^{-1} \\ &= H_c(\omega)\times\text{Garnett}_y(x), \end{aligned}$$

where we set $\text{Garnett}_y(x) = [1+\sqrt{y}h_{\mathcal{N}}(x/\sqrt{y})/h_{\mathcal{N}}(-x)]^{-1}$, $y = H'(\omega+)/H'(\omega-)$, and $x = -\sqrt{2\rho}\beta/(\sigma\sqrt{H'(\omega-)})$.

Now consider Example 1 for systems with the number of servers ranging in $\{20, 50, 100, 200, 400\}$. The individual service rate $\mu = 1$. Consider the overloaded case where $\rho = 1.2$, thus the offered waiting time $\omega = 1/6$. We have tested extensively the accuracy of our approximation formulae by experimenting with $\beta \in \{0, -1, 1\}$. To save space, we only report the case $\beta = 0$ (the corresponding arrival rates are $\{24, 60, 120, 240, 480\}$). Other values of β give rise to similar accuracy. (The role of β is emphasized when we discuss the related staffing problem in Section 4.1.1.)

Table 2 summarizes the comparison for Example 1 with a different right derivative $k = 1, 3, 5$. The column "Approx." is obtained via our approximation formulae (33)–(34). The column "Simulation" is obtained by simulating such a system with the given parameters. The number after "±" indicates the half-width 95% confidence interval. Note that when $k = 1$, the left and right derivatives are the same, i.e., $H(\cdot)$ is differentiable at the fluid offered waiting time ω. In this case, our approximation for the expected queue length coincides with the fluid approximation. As Table 2 shows, the larger the difference between the right and left derivatives becomes (k becomes larger), the larger is the error from the fluid approximation in estimating the expected queue length.

5.3. Using the Hazard Rate of the Patience-Time Distribution

Assume that the hazard rate of the patience-time distribution $H(\cdot)$ exists, and denote it by $h(\cdot)$. Following (30),

Table 2. Comparison of the Approximation and the Simulation of Example 1

(a) $\mathbb{E}(Q_\infty)$

		$k=1$		$k=3$		$k=5$	
Servers	Fluid	Simulated	Approx.	Simulated	Approx.	Simulated	Approx.
20	3.67	3.82 ± 0.02	3.67	2.61± 0.01	2.30	2.27 ± 0.01	1.87
50	9.17	9.05 ± 0.04	9.17	6.98 ± 0.03	6.99	6.36 ± 0.03	6.32
100	18.33	17.98 ± 0.06	18.33	14.94 ± 0.04	15.25	14.01 ± 0.03	14.31
200	36.67	36.19 ± 0.07	36.67	31.84 ± 0.05	32.31	30.51 ± 0.05	30.97
400	73.33	72.88 ± 0.09	73.33	66.71 ± 0.06	67.18	64.81 ± 0.06	65.28

(b) $\mathbb{P}(W_\infty > \omega)$

	$k=1$		$k=3$		$k=5$	
Servers	Simulated	Approx.	Simulated	Approx.	Simulated	Approx.
20	0.4188 ± 0.0015	0.4167	0.3079 ± 0.0012	0.3050	0.2619 ±0.0013	0.2575
50	0.4172 ± 0.0024	0.4167	0.3053 ± 0.0018	0.3050	0.2578 ± 0.0013	0.2575
100	0.4168 ± 0.0020	0.4167	0.3051 ± 0.0014	0.3050	0.2574 ± 0.0012	0.2575
200	0.4163 ± 0.0018	0.4167	0.3048 ± 0.0014	0.3050	0.2574 ± 0.0012	0.2575
400	0.4173 ± 0.0013	0.4167	0.3054 ± 0.0009	0.3050	0.2579 ± 0.0008	0.2575

we compute

$$\begin{aligned}
f_\omega(x) &= \sqrt{\lambda}\left[H\left(\omega+\frac{x}{\sqrt{\lambda}}\right)-H(\omega)\right]\\
&= \exp\left(-\int_0^\omega h(y)\,\mathrm{d}y\right)\sqrt{\lambda}\\
&\quad\cdot\left[1-\exp\left(-\int_\omega^{\omega+x/\sqrt{\lambda}} h(y)\,\mathrm{d}y\right)\right]\\
&= H_c(\omega)\sqrt{\lambda}\left[1-\exp\left(-\int_0^{x/\sqrt{\lambda}} h(\omega+y)\,\mathrm{d}y\right)\right]\\
&\approx H_c(\omega)\sqrt{\lambda}\left[\frac{1}{\sqrt{\lambda}}\int_0^x h\left(\omega+\frac{y}{\sqrt{\lambda}}\right)\mathrm{d}y\right] \quad \text{(for large } \lambda)\\
&= H_c(\omega)\int_0^x h\left(\omega+\frac{y}{\sqrt{\lambda}}\right)\mathrm{d}y. \qquad (53)
\end{aligned}$$

From (31), the density π of $\sqrt{\lambda}(V(\infty)-\omega)$ can be approximated by

$$\begin{aligned}
\pi(x) &\approx C\exp\left(\frac{2\rho\beta x}{\sigma^2}\right)\\
&\quad\cdot\exp\left(-\frac{2\rho}{\sigma^2}H_c(\omega)\int_0^x\int_0^v h\left(\omega+\frac{y}{\sqrt{\lambda}}\right)\mathrm{d}y\,\mathrm{d}v\right), \qquad (54)
\end{aligned}$$

with the appropriate normalizing constant C. Based on (54), the probability $\mathbb{P}(W(\infty)>y)$ and expected queue length $\mathbb{E}Q(\infty)$ can be approximated by replacing $f_\omega(x)$ in (33)–(34) with $H_c(\omega)\int_0^x h(\omega+y/\sqrt{\lambda})\,\mathrm{d}y$.

Consider now Example 2, where the density function of the patience-time distribution exists, but the hazard rate has a very steep change around ω. The individual service rate $\mu=1$. Assume that $\rho=1.2$ and $h_0=1$; hence, the offered waiting time $\omega=\ln(1.2)$. As in the previous example, we report only the study for $\beta=0$, with other β's behaving similarly. Table 3 summarizes the comparison for Example 2 for different system sizes with the number of servers ranging from 20 to 400, and $\kappa=20, 100$. The column "Appr. G" is obtained via our approximation formulae (33)–(34) with $f_\omega(x)$ from (30), while column "Appr. H" is calculated by replacing (30) with (53). As Table 3(a) shows, the larger the parameter κ becomes (meaning a steeper change of the hazard rate), the larger the error that the fluid approximation yields in approximating the expected queue length. Since, in this case, the patience-time distribution is differentiable, we can also use the method by Mandelbaum and Zeltyn (2009), which leads to 0.4167 in approximating $\mathbb{P}(W_\infty>\omega)$ for all systems in Table 3(b). This is not nearly as close as either "Appr. G" or "Appr. H." We also observe that "Appr. G" seems better for the tail probability of waiting times, and similar or slightly worse for queue length, when compared against "Appr. H."

We now relate our general setting of scaling the patience-time distribution to the hazard-rate scaling in Reed and Tezcan (2012). Note that our study is in the ED+QED regime ($\omega>0$), which is different from the QED regime ($\omega=0$) studied in Reed and Tezcan (2012). From the application point of view, the ED+QED regime is more suitable for the analysis of delay announcements. From the technical viewpoint, our derivation of the corresponding diffusion limit is quite different from that in the QED regime. Specifically, when analyzing the virtual waiting time at time t for the QED regime, the customers in queue at time t who would eventually abandon are negligible. However, under the ED+QED regime, these customers must be accounted for, which makes the analysis more challenging. At the same time, it is worth pointing out that

Table 3. Comparison of the Approximation and the Simulation of Example 2

(a) $\mathbb{E}(Q_\infty)$

		$\kappa = 20$			$\kappa = 100$		
Servers	Fluid	Simulated	Appr. G	Appr. H	Simulated	Appr. G	Appr. H
20	4	3.34 ± 0.02	3.1599	2.6690	2.70 ± 0.01	2.4354	1.9113
50	10	8.65 ± 0.04	8.7328	8.2553	7.48 ± 0.04	7.025	7.0144
100	20	18.04 ± 0.05	18.3797	17.9050	16.30 ± 0.04	16.6151	16.1387
200	40	37.56 ± 0.06	38.0092	37.5344	35.11 ± 0.06	35.5413	35.0705
400	80	77.20 ± 0.08	77.9364	77.1579	73.84 ± 0.07	74.2668	73.7983

(b) $\mathbb{P}(W_\infty > \omega)$

	$\kappa = 20$			$\kappa = 100$		
Servers	Simulated	Appr. G	Appr. H	Simulated	Appr. G	Appr. H
20	0.35785 ± 0.00143	0.3576	0.3271	0.28828 ± 0.00122	0.2879	0.2493
50	0.36396 ± 0.00210	0.3641	0.3461	0.29371 ± 0.00172	0.2938	0.2723
100	0.37122 ± 0.00188	0.3712	0.3589	0.30348 ± 0.00142	0.3037	0.2895
200	0.37858 ± 0.00175	0.3787	0.3703	0.31552 ± 0.00142	0.3157	0.3062
400	0.38652 ± 0.00124	0.3859	0.3801	0.32922 ± 0.00102	0.3286	0.3221

the diffusion limits in the QED and ED+QED regimes share a somewhat similar structure (e.g., the variance is constant and the drift is a continuous function of the state). This similarity helps one to apply the same procedure for calculating the stationary distribution of both diffusion limits.

Consider a sequence of many-server queues indexed by n. The patience-time distribution $F^n(\cdot)$ has hazard rate $h^n(\cdot)$ given by

$$h^n(x) = \begin{cases} h(x), & \text{for } x \in [0, \omega], \\ h(\omega + \sqrt{\lambda_n}(x - \omega)), & \text{for } x \in (\omega, \infty). \end{cases}$$

This implies that $F_c^n(\omega) = H_c(\omega)$. In view of (53),

$$f_\omega(x) \approx \begin{cases} H_c(\omega) h(\omega) x, & x \leqslant 0, \\ H_c(\omega) \int_0^x h(\omega + y)\,\mathrm{d}y, & x > 0. \end{cases}$$

Therefore, by (54), the approximation of the density of $\sqrt{\lambda}(V(\infty) - \omega)$ can be written as

$$\pi(x) \approx \begin{cases} C\exp\left(\dfrac{2\rho\beta x}{\sigma^2}\right)\exp\left(-\dfrac{\rho}{\sigma^2}H_c(\omega)h(\omega)x^2\right), \\ \quad \text{if } x \leqslant 0, \\ C\exp\left(\dfrac{2\rho\beta x}{\sigma^2}\right) \\ \quad \cdot\exp\left(-\dfrac{2\rho}{\sigma^2}H_c(\omega)\displaystyle\int_0^x\int_0^v h(\omega+y)\,\mathrm{d}y\,\mathrm{d}v\right), \\ \quad \text{if } x > 0, \end{cases} \tag{55}$$

with an appropriately normalizing constant C. This same structure also arises for the QED diffusion in Reed and Tezcan (2012, Proposition 3.2).

5.4. On the Gap Between Fluid and Diffusion Models

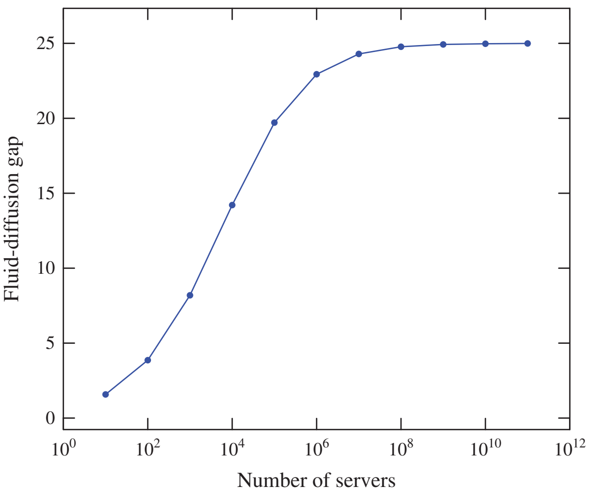
We observed in Table 2(a) that the fluid approximation could result in a large error, when the left and right derivatives of the patience-time distribution do not agree. We now further study Example 2, where the patience-time distribution is differentiable but not that smooth. Similarly to Table 3, we simulate Example 2 with the same set of parameters ($\mu = 1$, $h_0 = 1$, $\rho = 1.2$, $\omega = \ln(\rho)/h_0$ and $\kappa = 100$), but consider a wide range of system size. In Table 4, we compare the simulated queue length and the one obtained by diffusion approximation, where the number of servers ranges in $\{10^i, i = 1, \ldots, 6\}$.

To give a graphical view of how the gap relates to system size, we plot the difference between "Fluid" and "Appr. H" in Figure 6. One can observe that the gap stabilizes around 25, as the system size becomes fairly large. However, for practical purposes, system size is normally in the hundreds, not millions. Thus, the diffusion correction term does play an important role.

Table 4. On the Accuracy of the Fluid Approximation in the ED Regime as System Size Becomes Large

	$\kappa = 100$			
Servers	Simulated	Fluid	Appr. G	Appr. H
10	1.30 ± 0.00	2	0.9983	0.4238
100	16.34 ± 0.04	20	16.6151	16.1387
1,000	191.58 ± 0.22	200	192.2776	191.8092
10,000	1,986.39 ± 1.58	2,000	1,986.265	1,985.7893
100,000	19,980.10 ± 22.18	20,000	19,980.7746	19,980.2881
1,000,000	199,929.32 ± 86.96	200,000	199,977.555	199,977.0616

Figure 6. (Color online) Gap Between the Fluid and Diffusion Approximations



In fact, using the hazard-rate approximation, we now demonstrate for Example 2, that the gap between the fluid and diffusion approximation can be calculated explicitly, and it is indeed $O(1)$. Plugging (42) into (53), we get the steady-state density,

$$\tilde{\pi}(x) = \begin{cases} C_\kappa \exp\left(-\dfrac{1}{\sigma^2}(h_0 x^2 - 2\beta\rho x)\right), & x \leqslant 0, \\ C_\kappa \exp\left(-\dfrac{1}{\sigma^2}\left(h_0 x^2 + \dfrac{\kappa x^3}{3\sqrt{\lambda}} - 2\beta\rho x\right)\right), & x > 0, \end{cases} \tag{56}$$

where

$$C_\kappa = \left[\int_{-\infty}^{\infty} \exp\left(-\frac{h_0 x^2}{2} - \frac{\kappa x^3}{6\sqrt{\lambda}}\right) \mathrm{d}x + \frac{\sqrt{2\pi}}{2\sqrt{h_0}}\right]^{-1} \to \frac{\sqrt{h_0}}{\sqrt{2\pi}}, \quad \text{as } \lambda \to \infty.$$

Also, from (34) and (56), we deduce the approximation for the expected queue length

$$\mathbb{E}[Q(\infty)] \approx \lambda \int_0^\omega H_c(x)\,\mathrm{d}x + \mathrm{gap}(\lambda),$$

where

$$\mathrm{gap}(\lambda) = C_\kappa \int_0^\infty x \exp\left(-\frac{x^2}{2}\right)\left(\exp\left(-\frac{\kappa x^3}{6\sqrt{\lambda}}\right) - 1\right)\sqrt{\lambda}\,\mathrm{d}x.$$

It is easy to verify the bound

$$\left|\left(\exp\left(\frac{-\kappa x^3}{6\sqrt{\lambda}}\right) - 1\right)\sqrt{\lambda}\right| \leqslant \frac{\kappa x^3}{6},$$

and that

$$\left(\exp\left(\frac{-\kappa x^3}{6\sqrt{\lambda}}\right) - 1\right)\sqrt{\lambda} \to \frac{\kappa x^3}{6},$$

as $\lambda \to \infty$ for all $x \geqslant 0$. Consequently,

$$\lim_{\lambda\to\infty} \mathrm{gap}(\lambda) = -\frac{\kappa}{6}\frac{1}{\sqrt{2\pi}}\int_0^\infty x^4 \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x = -\frac{\kappa}{6} \times \frac{3}{2} = -\frac{\kappa}{4},$$

where the integral was evaluated via integration by parts. When $\kappa = 100$, the limit is -25, which is consistent with Figure 6.

Our study of the gap relates to Bassamboo and Randhawa (2010), who studied the gap between the fluid-approximation and the steady-state of the originating system. It is proved in Bassamboo and Randhawa (2010) that the latter gap is $O(1)$, as the system size becomes large. Our finding on the gap between the approximations based on fluid and diffusion concurs with this result. The two gaps are similar under the premise that the diffusion approximation is close to the originating system. Nevertheless, what we offer here is an alternative view of the gap using fluid approximation, under more general conditions than Bassamboo and Randhawa (2010). Indeed, they require, in their Assumption 1, that the density function of the patience-time distribution be continuously differentiable.

6. Conclusion

In this paper, we have established diffusion limits for many-server queues with abandonment in a fairly general setting of scaling the patience-time distribution. Such a generality allows the fine structure of the patience-time distribution to be manifested in the diffusion limit, and consequently in the approximation formulae for the performance measures.

The fine structure of the patience-time distribution can be naturally attributed to delay announcements. Applying our approximation formulae, we have thus investigated the impact of delay announcements in two settings—first when the announcement is made upon arrival, and next when it is made once customers' waiting time exceeds a threshold. We have also prescribed the optimal staffing rule in the presence of a delay announcement.

To illustrate the value and generality of our approximations, we connect them to existing approaches of scaling the patience-time distribution. Moreover, the application of our general formulae does not require the choice of a scaling method, and it applies to more general settings than those in the literature.

From the technical point of view, we offer a new method of obtaining the diffusion limits for many-server queues, by focusing on the virtual waiting time. Following our method, He (2015) recently developed diffusion approximations for overloaded queues in the nondegenerate slowdown (NDS) scaling. We believe that our method can be also applied in the QED regime, and we leave this for future research.

Another worthy direction to pursue is the study of multiple announcements, first upon arrival and subsequent ones during waiting, with the latter possibly interacting with customers: For example, encouraging an abandonment but simultaneously obtaining information about when it would be convenient to call them back. As a final point, deeper statistical validation, individual psychological modeling, and the effect of announcements on model primitives, all naturally call for further study.

Acknowledgments

The authors are grateful to the thorough review team for providing numerous thoughtful and constructive comments that significantly improved both the mathematical rigor and exposition of this paper.

References

Akşin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

Akşin Z, Ata B, Emadi SM, Su C-L (2016) Impact of delay announcements in call centers: An empirical approach. *Oper. Res.* 65(1):242–265.

Allon G, Bassamboo A (2011) The impact of delaying the delay announcements. *Oper. Res.* 59(5):1198–1210.

Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57(1):66–81.

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* 58(5):1398–1413.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.

Dai JG, He S, Tezcan T (2010) Many-server diffusion limits for $G/Ph/n+GI$ queues. *Ann. Appl. Probab.* 20(5):1854–1890.

Ethier SN, Kurtz TG (1986) *Markov Processes*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics (John Wiley & Sons, New York).

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Springer, New York).

He S (2015) Diffusion approximation for efficiency-driven queues: A space-time scaling approach. Working paper, http://arxiv.org/abs/1506.06309v3.

Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: An integrative model of service evaluation. *J. Marketing* 60(2):81–90.

Ibrahim R, Armony M, Bassamboo A (2016) Does the past predict the future? The case of delay announcements in service systems. *Management Sci.* 63(6):1762–1780.

Jouini O, Akşin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing Service Oper. Management* 13(4):534–548.

Kang W, Pang G (2011) Fluid limit of a many-server queueing network with abandonment. Working paper.

Kang W, Pang G (2013) Equivalence of fluid models for $G_t/GI/N+GI$ queues. Working paper.

Kang W, Ramanan K (2010) Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* 20(6):2204–2260.

Krichagina EV, Puhalskii AA (1997) A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Syst.* 25(1–4):235–280.

Li G, Huang JZ, Shen H (2015) Nonparametric two-way hazards model with application to call center waiting times. Working paper, University of North Carolina.

Liu Y, Whitt W (2011a) Large-time asymptotics for the $G_t/M_t/s_t+GI_t$ many-server fluid queue with abandonment. *Queueing Syst.* 67(2):145–182.

Liu Y, Whitt W (2011b) A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.* 59(4):835–846.

Liu Y, Whitt W (2012a) The $G_t/GI/s_t+GI$ many-server fluid queue. *Queueing Syst.* 71(4):405–444.

Liu Y, Whitt W (2012b) A many-server fluid limit for the queueing model experiencing periods of overloading. *Oper. Res. Lett.* 40(5):307–312.

Liu Y, Whitt W (2014a) Algorithms for time-varying networks of many-server fluid queues. *INFORMS J. Comput.* 26(1):59–73.

Liu Y, Whitt W (2014b) Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab.* 24(1):378–421.

Long Z, Zhang J (2014) Convergence to equilibrium states for fluid models of many-server queues with abandonment. *Oper. Res. Lett.* 42(6–7):388–393.

Maister D (1985) The psychology of waiting lines. Czepiel J, Solomon M, Suprenant C, eds. *The Service Encounter: Managing Employee/Customer Interaction in Service Businesses* (Lexington Books, Lexington, MA), 113–123.

Mandelbaum A, Momčilović P (2012) Queues with many servers and impatient customers. *Math. Oper. Res.* 37(1):41–65.

Mandelbaum A, Zeltyn S (2009) Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* 57(5):1189–1205.

Mandelbaum A, Zeltyn S (2013) Data-stories about (im)patient customers in tele-queues. *Queueing Syst.* 75(2–4):115–146.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Perry O, Whitt W (2009) Responding to unexpected overloads in large-scale service systems. *Management Sci.* 55(8):1353–1367.

Reed JE, Tezcan T (2012) Hazard rate scaling of the abandonment distribution for the $GI/M/n+GI$ queue in heavy traffic. *Oper. Res.* 60(4):981–995.

Reed JE, Ward AR (2008) Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Math. Oper. Res.* 33(3):606–644.

Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* 50(10):1449–1461.

Whitt W (2006) Fluid models for multiserver queues with abandonments. *Oper. Res.* 54(1):37–54.

Yu Q, Allon G, Bassamboo A (2017) How do delay announcements shape customer behavior? An empirical study. *Management Sci.* 63(1):1–20.

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

Zhang J (2013) Fluid models of many-server queues with abandonment. *Queueing Syst.* 73(2):147–193.

Junfei Huang is an assistant professor in the Department of Decision Sciences and Managerial Economics at the Chinese University of Hong Kong. His research interests are in asymptotic analysis and optimal control of queueing systems and their applications in manufacturing and services.

Avishai Mandelbaum is a professor and the serving dean of the Faculty of Industrial Engineering and Management, Technion, Israel. He holds the Benjamin and Florence Free chair in Operations Research, Statistics and Service Engineering, and he is an INFORMS fellow. His research has covered analysis, asymptotics and control of stochastic models, with a present focus on data-based research and applications to queueing theory/science and service systems (e.g., teleservices, hospitals).

Hanqin Zhang is a professor in the Department of Decision Sciences, School of Business, National University of Singapore. His research interests include queueing networks, stochastic optimal control, and supply chain management.

Jiheng Zhang is an associate professor in the Department of Industrial Engineering and Logistics Management at the Hong Kong University of Science and Technology. His research interests are in performance evaluation and optimal control via asymptotic analysis of queueing systems arising from applications in manufacturing and services.