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

Junfei Huang
Department of Decision Sciences and Managerial Economics, CUHK Business School, The Chinese University of Hong Kong, Hong Kong S.A.R., China, junfeih@cuhk.edu.hk

Avishai Mandelbaum
Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, 32000 Haifa, Israel, avim@ie.technion.ac.il

Hanqin Zhang
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore, bizzhq@nus.edu.sg

Jiheng Zhang
Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Hong Kong S.A.R., China, j.zhang@ust.hk

Data reveals a noticeable impact of delay-related information on phone-customers; for example, delay announcements abruptly increase the likelihood to abandon (hang-up). Our starting point is that the latter phenomena can be used to harness queue-lengths and delays, and we do so by timing the announcement appropriately and determining 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. the 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 servers are highly busy, hence essentially all customers are delayed in queue prior to service). However, existing theory turns out too crude for our needs. Moreover, it requires continuous hazard-rates of impatience and hence cannot be applied. We thus develop 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.

Key words: 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, Akšin 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 which, for large enough systems, translates
into waiting- and idle-time that are negligible relative to 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 twice as many are required for satisfactory performance); 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 are inflexible enough to accommodate temporal peaks or unexpectedly high demand (e.g., Perry and Whitt (2009)). When this happens, and when queues are invisible (e.g. call centers), it makes sense and is hence prevalent to provide customers with delay information. Indeed, such information has no significant impact in short-wait conditions; see Hui and Tse (1996). One reason is to relieve waiting anxiety since “uncertain waits feel longer than predictable finite waits” (e.g., Armony et al. (2009)). Just as importantly, such information helps customers determine if their gain from service is not worth its wait. In that case they abandon the queue which, in heavy traffic, improves dramatically 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 to 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 its ED+QED refinement, which captures the effect of delay announcements on system performance. We then use this model to simultaneously optimize staffing levels and timing of announcement, subject to service level constraints (e.g. fraction abandoning). Our analysis must hence capture the impact of announcements on customers’ tendency to abandon (hang-up), which raises a challenge: empirical evidence reveals that this impact is often abrupt or, formally, it is manifested through a discontinuity in the hazard-rate function 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 refinements of the existing ones.

Two types of Delay Announcements. Our refined model is motivated by two types of announcements. The first is to be made upon arrival of customers who must wait before receiving service; see the “all-exponential model” in Armony et al. (2009). Often here, an estimated duration of delay is announced, which has the following consequences. Some customers choose to balk immediately while others remain online; these will not abandon if served before their patience
expires, but they will become irritated once their waiting-time reaches the announced delay. Such a behavior is collectively 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 1 minute. Here 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), Mandelbaum and Zeltyn (2013) have discussions on such announcements, with the latter observing that, in fact, any announcement causes an upward jump in the hazard-rate, right after the announcement epoch.

Both types of announcements share the common feature of giving rise to a non-smooth change in the hazard-rate of patience-time. 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). In this paper, we develop models that quantify the impact of announcements 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 cost, and whether an announcement during waiting (second type) should at all 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 possible non-smooth patience-time distribution—this distinguishes our model from 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 service rate 1, arrival rate being 140 and 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 an open problem, which is here resolved: our refined model offers the much improved approximation 16.4 (see Section 4.1 for details). Another example is approximating the tail probabilities of waiting-times in the ED+QED regime, as studied in Mandelbaum and Zeltyn (2009). For such a system with 100 servers, each having service rate 1, and arrival rate being 120, the simulated tail probability is 0.2574 while the approximation in Mandelbaum and Zeltyn (2009) yields 0.4167. Our refined model produces the accurate approximation of 0.2575 (see Section 5 for extensive numerical experiments). Generally speaking, non-smooth changes in the patience hazard-rates render 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 waiting-time, say). 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. (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 $\lambda$ is the arrival rate). With announcements during waiting, we simultaneously optimize the timing of an announcement as well as minimize the staffing level, and doing so subject to bounds on the tail probability of waiting as well as on the fraction abandoning. It turns out optimal to make an announcement at a time that is approximately the fluid offered waiting-time and, also here, the announcement reduces staffing by a magnitude of $O(\sqrt{\lambda})$; see Proposition 3 below.

A Queueing Model with a Delay Announcement. Our refined model introduces a general scaling (see (3)) of the patience-time distribution, which precisely captures its fine structures, especially non-smooth changes caused by announcements. As explained in Section 2, hazard-rate scaling and hence also no-scaling of the patience-time distribution are special cases of this general scaling. Our method of diffusion analysis, in the ED regime, is also new: it is based on the virtual waiting-time, as opposed to queue-length that has been the basis for most analyses of many-server queues. The virtual waiting-time $V^n(t)$ is roughly the time that an infinitely-patient “virtual” customer would have to wait if arriving at time $t$. Its evolution is characterized in (18) which, notably, is one-dimensional (despite the generality of our patience-time distribution). We are thus able to approximate a complex system by a simple one-dimensional diffusion process; furthermore, that diffusion has a tractable stationary distribution, which is used to approximate steady-state performance of its originating queueing system. Useful features 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; (c) analysis of how operational performance is affected by a non-smooth change of the patience-time distribution (accounting for characteristics of the arrival-process and service-rate).

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 and found that a surge is associated with the time of announcement. Aksin-Karaesmen et al. (2013) modeled customers’ abandonment decisions endogenously in the presence of delay announcements; they
studied how announcements impact customers’ 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 queuing analysis that incorporates that behavior. Yu et al. (2013) 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’ perception about the system but they also directly impact waiting costs. Armony et al. (2012) 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. An announcement-dependent customer behavior is then explicitly modeled by letting the joint probability and abandonment rate depend on the announced waiting-time.

**The ED and ED+QED Regime.** The ED regime was introduced in Garnett et al. (2002), followed by ample research of many-server queues in that regime. The ED+QED regime arose in Mandelbaum and Zeltyn (2009), as a refinement to ED that accommodates approximations to tail probabilities. Whitt (2004) studied both diffusion approximations and steady-state limits for an ED Markovian model. Dai et al. (2010) analyzed diffusion models for $G/Ph/n+M$ systems in both the QED and ED regimes. Liu and Whitt (2014) used two-parameter stochastic processes to prove diffusion approximation in a time-varying environment (alternating between underloaded and overloaded time-intervals), for models with exponential service time and generally distributed patience time (without scaling it). Our method of diffusion analysis differs from that of Liu and Whitt (2014) in that we use a one-dimensional process (the virtual waiting-time) rather a two-parameter process. Moreover, our main focus is a general scaling of the patience-time distribution, rather than allowing the system to be time-varying.

**Fluid Models.** Fluid approximations are useful in the ED regime. Whitt (2006) conjectured a fluid model with general service and patience-time distributions; he then constructed simple, yet effective, approximations for various performance measures, based on the equilibrium of the fluid model. The fluid approximation was formally justified by Kang and Ramanan (2010) and Zhang (2013), using measure-valued processes. Then Long and Zhang (2013) proved that the fluid model does converge to an equilibrium state. Approximations based on the fluid model 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, so as to minimize the sum of capacity costs and long-term average customer-related costs. An in-depth discussion on the gap between fluid and diffusion approximations is provided in Section 5.4. This adds to Bassamboo and Randhawa (2010), since their assumptions do not hold in the presence of 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 distribution, at the time which is the fluid offered waiting-time. In particular, they studied the tail-probability of waiting, which is beyond the scope of 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 the following:

- Developing and analyzing a refined model for many-server queues in the ED+QED regime, based on a diffusion approximation for the underlying stochastic processes. Technically, the refinement goes beyond existing theory (see [3]); practically, the corresponding theory can then accommodate real phenomena and applications, such as jointly deciding on announcements and staffing.

- To be more specific, our approximations allow the patience-time distribution to change abruptly, as observed in practice. (In particular, there is a requirement for neither continuous differentiability, as in Bassamboo and Randhawa (2010), nor even the existence of a derivative, as in Mandelbaum and Zeltyn (2009).)

- Introducing a new method, based on a one-dimensional process (the virtual waiting-time), to obtain the diffusion limit. This circumvents the complexity of measure-valued or two-parameter processes, which have been used so far.

- Analyzing the stationary distribution of the diffusion, and constructing effective approximations for performances of its originating system.

- Using our approximation to study the fixed-delay mode of Armony et al. (2009), and verifying numerically that our approximations are accurate. This resolves a question that was left open in Armony et al. (2009).

- Jointly optimizing (asymptotically) the two problems of optimal-staffing and announcement-timing, while accommodating the empirical phenomenon of an abrupt announcement impact.

- Analyzing the performance gap between fluid and diffusion approximations (Section 5.4).

1.3. Organization and Notation

The paper is organized as follows. Section 2 introduces the queueing model and the heavy-traffic regime. We then proceed to Section 3, where we derive diffusion limits and their stationary distribution, and then use the latter to approximate 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 4.1.1 provides (asymptotically) optimal staffing for problems motivated by announcements. Section 5 gives an in-depth discussion of our approximation formulae, and Section 6 concludes the paper. Proofs of all propositions are given in Appendix EC.1 and several auxiliary lemmas for our diffusion analysis appear in Appendix EC.2.

We conclude the introduction with convention and notation 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 \(F^c(\cdot) = 1 - F(\cdot)\). For any two real-valued nonnegative functions \(f\) and \(g\), we say that (1) \(f(n) = O(g(n))\) if \(\limsup_{n \to \infty} \frac{f(n)}{g(n)} < \infty\); (2) \(f(n) = o(g(n))\) if \(\limsup_{n \to \infty} \frac{f(n)}{g(n)} = 0\).

2. Model Formulation

Consider a sequence of many-server queueing systems with customer abandonment, indexed by \(n \in \mathbb{N}\). In the \(n\)th 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 \(\Lambda^n = \{\Lambda^n(t) : t \geq 0\}\). Assume that there exists a sequence of positive real numbers \(\{\lambda_n\}_{n \in \mathbb{N}}\) such that as \(n \to \infty\),

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

where \(\tilde{\Lambda} = \{\tilde{\Lambda}(t) : t \geq 0\}\) is a Brownian motion. Arriving customers are immediately served if any servers are idle. Otherwise, they wait in a queue and are served on the first-come first-served (FCFS) basis. The \(i\)th arriving customer requires a service time of \(v^n_i\), and has patience-time \(u^n_i\): once waiting-time exceeds \(u^n_i\), the customer leaves the system immediately without receiving service. Service times are assumed to be independent and have the same exponential distribution with mean \(1/\mu\), and

\[
\lim_{n \to \infty} \frac{\lambda_n}{\mu s_n} = \rho > 1.
\]

Patience-times are i.i.d., with a general distribution \(F^n(\cdot)\). We assume that

\[
\sqrt{\lambda_n} \left[ F^n(\omega^n + \frac{x}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right] \to f_\omega(x),
\]
\[
\frac{\lambda_n F^n_c(\omega^n) - s_n \mu}{\sqrt{\lambda_n}} \to \beta, \quad (4)
\]
as \(n \to \infty\); here \(\{\omega^n\}_{n \in \mathbb{N}}\) is some sequence that converges to a limit \(\omega \in (0, \infty)\), which we call the \textit{fluid offered waiting-time}, \(\beta\) is a constant and \(f_\omega(\cdot)\) is a continuous function. We also assume that customer arrivals, service times and patience-times are all independent of each other. Note that \(\text{[2]}\) is an ED assumption, while \(\text{[4]}\) is an ED+QED assumption (Mandelbaum and Zeltyn (2009)).

The above characterizes a sequence of \(G/M/s_n+GI\) queues, \(n = 1, 2, \ldots\). Condition \(\text{[1]}\) about the arrival processes, and Condition \(\text{[2]}\) about the nominal load, are standard. Condition \(\text{[3]}\) is in fact motivated by data and applications. Indeed, our setting for scaling the patience-time distribution \(F^n(\cdot)\) is quite general. A simple special case is without any scaling: \(F^n(x) \equiv F(x)\) and \(\omega^n = \omega\), in which case \(f_\omega(x) = f(\omega)x\), where \(f(\cdot)\) is the density function of \(F(\cdot)\). However, more generally, the flexibility \(\text{[3]}\) of allowing the patience-time distribution to vary with \(n\), captures the fine structure that arises from customers’ reaction to delay announcements, as discussed in the introduction (also see Figure 3). This enables one to analyze the impact of such announcements in Section 4 and consequently optimize staffing levels and announcement times, jointly.

We now model mathematically system dynamics. To this end, we introduce two notions that correspond to waiting-times. The first is the \textit{offered waiting-time} \(\omega^n_i\), which denotes the time that the \(i\)th arriving customer in the \(n\)th system must wait before receiving service, if that customer would have been infinitely patient, for \(i \geq 1\). The second is the \textit{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 \(n\)th system. For \(i \geq 1\), let

\[
\tau^n_i = \inf\{t \geq 0 : \Lambda^n(t) \geq i\}
\]
represent the time of the \(i\)th arrival to the \(n\)th system. Then we have

\[
\omega^n_i = V^n(\tau^n_i -).
\]

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 \(n\)th system. Clearly, \(A^n(0) = 0\) and

\[
A^n(t) = \sum_{i=1}^{\Lambda^n(t)} 1_{\{u^n_i \leq \omega^n_i\}}.
\]

Introducing \(H^n(t, x) = \sum_{i=1}^{\lfloor t \rfloor} 1_{\{u^n_i \leq x\}}\), \(A^n(t)\) can be also written as

\[
A^n(t) = \int_0^t \int_0^{V^n(y)} dH^n(\Lambda^n(y), x). \quad (5)
\]
Any customer who arrives after time 0 cannot receive service before $V_n(0)$ due to FCFS. For $t \geq V_n(0)$, let

$$\kappa_n(t) = \inf\{\tau : \tau + V_n(\tau) > t\}. \quad (6)$$

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 $t \geq V_n(0)$, can be written as

$$Q_n(t) = \sum_{i=\Lambda_n(\kappa_n(t))}^{\Lambda_n(t)} 1_{\{u_n^i-(t-\tau_n^i)>0\}}. \quad (7)$$

Let $B_n(t)$ denote the number of customers who start service during $(0, t]$. Then for $t \geq V_n(0)$,

$$B_n(t) - B_n(V_n(0)-) = \sum_{i=1}^{\Lambda_n(\kappa_n(t))} 1_{\{u_n^i > \omega_n^i\}}. \quad (8)$$

Consider the $i$th customer who arrived during $(0, t]$, for some fixed $i$: if $u_n^i \leq \omega_n^i$ 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 $\Lambda_n(t)$:

$$\Lambda_n(t) = A_n(t) + B_n(t + V_n(t)) - B_n(V_n(0)-), \quad t \geq 0. \quad (9)$$

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 analysis of system dynamics requires a regularity assumption on the initial state, specifically,

$$\sqrt{\lambda_n} \left(V_n(0) - \omega_n\right) \Rightarrow \tilde{V}_0, \quad (10)$$

as $n \to \infty$, where $\{\omega_n\}_{n \in \mathbb{N}}$ is given by (3)–(4). Note that this is weaker than prevalent assumptions (e.g., Zhang [2013]), which require either that the remaining patience-time for the initial customers follows a certain distribution, or that initial customers are infinitely patient. 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)–(4) with the initial condition (10), for any $T \geq 0$, as $n \to \infty$,

$$\sup_{0 \leq t \leq T} \left|V_n(t) - \omega\right| \Rightarrow 0. \quad (11)$$
The proof is provided in Appendix EC.1. The above proposition serves as a first-order fluid approximation. We now pursue a refined approximation of the stochastic deviation from the fluid limit. Introduce the diffusion-scaled queue-size \( \tilde{Q}^n = \{ \tilde{Q}^n(t) : t \geq 0 \} \) by

\[
\tilde{Q}^n(t) = \frac{1}{\sqrt{\lambda_n}} \left( Q^n(t) - \lambda_n \int_0^t F^n_c(x) dx \right), \quad t \geq 0,
\]

and the diffusion-scaled virtual waiting-time process \( \tilde{V}^n = \{ \tilde{V}^n(t) : t \geq 0 \} \) by

\[
\tilde{V}^n(t) = \sqrt{\lambda_n} \left( V^n(t) - \omega^n \right), \quad t \geq 0,
\]

where \( \{ \omega^n \}_{n \in \mathbb{N}} \) is given by (3)–(4). We shall establish the following diffusion approximation for the virtual waiting-time. This mesoscopic level turns out natural for capturing system dynamics that is triggered by customers’ reactions to delay announcements, as it emerges from our call center data.

**Theorem 1.** In the heavy traffic regime (1)–(4) with the initial condition (10), \( \tilde{V}^n \to \tilde{V} \) as \( n \to \infty \), where the limit \( \tilde{V} = \{ \tilde{V}(t) : t \geq 0 \} \) is the unique solution to

\[
\tilde{V}(t) = \tilde{V}_0 - \rho \int_0^t \left[ f_\omega(\tilde{V}(x)) - \beta \right] dx + \left[ \tilde{\Lambda}(t) - \sqrt{\rho} B(t) - \sqrt{\rho - 1} B_A(t) \right], \quad t \geq 0;
\]

here \( B_A = \{ B_A(t) : t \geq 0 \} \) and \( B = \{ B(t) : t \geq 0 \} \) are two independent standard Brownian motions, independent of \( \tilde{\Lambda} \).

**Proof.** First introduce the following diffusion-scaled processes:

\[
\begin{align*}
\tilde{H}^n(t, x) &= \frac{H^n(\lambda_n t, x) - \lambda_n t F^n(x)}{\sqrt{\lambda_n}}, \\
\tilde{A}^n(t) &= \frac{A^n(t) - \lambda_n t F^n(\omega^n)}{\sqrt{\lambda_n}}, \\
\tilde{B}^n(t) &= \frac{B^n(t) - s_n \mu t}{\sqrt{\lambda_n}}.
\end{align*}
\]

By (5), we have

\[
\begin{align*}
\tilde{A}^n(t) &= \int_0^t \int_0^{V^n(y)} d\tilde{H}^n(\tilde{\Lambda}^n(y), x) \\
&\quad + \int_0^t \sqrt{\lambda_n} \left( F^n(\omega^n + \frac{\tilde{V}^n(s)}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right) d\tilde{\Lambda}^n(s) + F^n(\omega^n) \tilde{\Lambda}^n(t),
\end{align*}
\]

where \( \tilde{\Lambda}^n(s) = \Lambda^n(s)/\lambda_n \), which is the fluid scaling of \( \Lambda^n(s) \). Applying the diffusion scaling to each term in (9), it follows from \( t = t F^n(\omega^n) + t F^n_c(\omega^n) \) that

\[
\begin{align*}
\tilde{\Lambda}^n(t) &= \tilde{A}^n(t) + \frac{(s_n \mu - \lambda_n F^n(\omega^n)) t}{\sqrt{\lambda_n}} \\
&\quad + \tilde{B}^n(t + V^n(t)) - \tilde{B}^n(V^n(0)) + \frac{s_n \mu}{\lambda_n} \left( \tilde{V}^n(t) - \tilde{V}^n(0) \right).
\end{align*}
\]
This helps us write $\tilde{V}^n$ as

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

(18)

where

$$
\tilde{Y}^n(t) = \frac{\lambda_n}{s_n \mu} \left[ F^n(\omega^n) \Lambda^n(t) - \int_0^t \int_0^{V^n(y)} d\tilde{H}^n(\Lambda^n(y), x) + \frac{(\lambda_n F^n(\omega^n) - s_n \mu) t}{\sqrt{\lambda_n}} \right]
$$

$$
- \tilde{B}(t + V^n(t)) + \tilde{B}(V^n(0))
$$

(19)

We now characterize the diffusion limit of $\tilde{Y}^n = \{ \tilde{Y}^n(t) : t \geq 0 \}$. Condition (4) implies that, for any $T > 0$,

$$
\sup_{0 \leq t \leq T} \left| \frac{(\lambda_n F^n(\omega^n) - s_n \mu) t}{\sqrt{\lambda_n}} - \beta t \right| \to 0.
$$

(20)

It then follows from (1), (4), (20), Lemmas EC.1 and EC.3 that $\tilde{Y}^n \Rightarrow \tilde{Y}$, where $\tilde{Y} = \{ \tilde{Y}(t) : t \geq 0 \}$ with

$$
\tilde{Y}(t) = \rho \beta t + \tilde{\Lambda}(t) - \sqrt{\rho - \tilde{B}} \tilde{\Lambda}(t) - \sqrt{\rho} \tilde{B}(t).
$$

By Lemma EC.4, any subsequence of $\{ \tilde{V}^n \}_{n \in \mathbb{N}}$ has a further convergent subsequence, written as $\{ \tilde{V}^{n_k} : n_k \in \mathbb{N} \}$. Let $\{ V_*(t) : t \geq 0 \}$ be its corresponding limit. By straightforward analysis, Condition (3) implies that, as $n \to \infty$,

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

(21)

for any sequence $\{ x^n \}_{n \in \mathbb{N}}$ such that $x^n \to x$. In view of (18), by Assumptions (10) and (21), we have

$$
V_*(t) = \tilde{V}_0 - \rho \int_0^t f_\omega(V_*(x)) dx + \tilde{Y}(t), \quad t \geq 0.
$$

(22)

In view of Theorem 5.15 on page 341 of Karatzas and Shreve (1991), we know that when $f_\omega(\cdot)$ is locally integrable, the solution of (22) is unique in the sense of probability law. Hence, from Condition (3), we conclude weak convergence of $\{ \tilde{V}^n \}_{n \in \mathbb{N}}$ and that the corresponding limit satisfies (14).

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

**Theorem 2.** In the heavy traffic regime (1)–(4) with the initial condition (10), if the sequence of patience-time distributions

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

(23)
then $\tilde{Q}^n \Rightarrow \tilde{Q}$ on the time interval $(\omega, \infty)$, as $n \to \infty$; here the limit $\tilde{Q} = \{\tilde{Q}(t): t \geq 0\}$ is given by

$$\tilde{Q}(t) = \int_{t-\omega}^{t} F_c(t-x) d\tilde{\Lambda}(x) - \int_{t-\omega}^{t} \int_{0}^{1}(s+x \leq t) d\tilde{H}(s,F(x)) + \frac{1}{\rho} \tilde{V}(t-\omega), \quad (24)$$

where $\tilde{H} = \{\tilde{H}(s,t) : s \in [0,\infty), t \in [0,1]\}$ is a Kiefer process, $\tilde{V}$ is given by Theorem 1 and $\tilde{\Lambda}$ is given by (11).

Note that Condition (23) is not needed for Theorem 1. 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 $[0,\omega]$.

Proof of Theorem 3. From Theorem 2.8 in [Whitt, 1980], proving the convergence on $(\omega, \infty)$ is equivalent to proving the convergence on $[\omega+\delta, \infty)$, for any $\delta > 0$, which we now proceed to prove.

By (7), when $t \geq V^n(0)$, the queue-length process can be written as

$$Q^n(t) = \sum_{i=\Lambda^n(t-\omega^n)}^{\Lambda^n(t)} 1\{u^n_i+\tau^n_i > t\} + \sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)} 1\{u^n_i+\tau^n_i > t\}$$

$$= \sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)} 1\{u^n_i+\tau^n_i > t\} + \sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)} (1\{u^n_i+\tau^n_i > t\} - F^n_c(t-\tau^n_i))$$

$$+ \int_{t-\omega^n}^{t} F^n_c(t-x) d(\Lambda^n(x) - \lambda_n x) + \lambda_n \int_{t-\omega^n}^{t} F^n_c(t-x) dx.$$

Recalling the definition of $\tilde{H}^n$ in (15), we have

$$\frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)} (1\{u^n_i+\tau^n_i > t\} - F^n_c(t-\tau^n_i)) = -\int_{t-\omega^n}^{t} \int_{0}^{1} 1\{y+x \leq t\} d\tilde{H}^n(\Lambda^n(x),x).$$

Applying the diffusion scaling (12),

$$\tilde{Q}^n(t) = \tilde{M}^n(t) + \frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t-\omega^n)} 1\{u^n_i+\tau^n_i > t\}, \quad (25)$$

where

$$\tilde{M}^n(t) = -\int_{t-\omega^n}^{t} \int_{0}^{1} 1\{y+x \leq t\} d\tilde{H}^n(\Lambda^n(x),x) + \int_{t-\omega^n}^{t} F^n_c(t-x) d\frac{\Lambda^n(x) - \lambda_n x}{\sqrt{\lambda_n}}. \quad (26)$$

The process $\tilde{M}^n(\cdot)$ can be viewed as the diffusion-scaled queue-length process of an infinite-server queue, with service times $u^n_i - \omega^n$, as given by [Krichagina and Puhalskii, 1997]. A slight modification of the proof of Theorem 3.1 in [Krichagina and Puhalskii, 1997] gives that the first term in (26) weakly converges to the process $\{\int_{t-\omega^n}^{t} 1\{s+x \leq t\} d\tilde{H}(s,F(x)) : t \geq \omega + \delta\}$. For the second term, integrating-by-part, we have

$$\int_{t-\omega^n}^{t} F^n_c(t-x) d\frac{\Lambda^n(x) - \lambda_n x}{\sqrt{\lambda_n}} = \tilde{\Lambda}^n(t) - F^n_c(\omega^n) \tilde{\Lambda}^n(t-\omega^n) - \int_{t-\omega^n}^{t} \frac{\Lambda^n(x) - \lambda_n x}{\sqrt{\lambda_n}} dF^n_c(t-x). \quad (27)$$
Using the Skorohod representation theorem, we embed all the random objects in a common probability space. We maintain the original notation for the mapped random objects. On the new probability space, we have

\[
\sup_{0 \leq t \leq T} \left| \frac{\Lambda^n(t) - \lambda_n t}{\sqrt{\Lambda_n}} - \tilde{\Lambda}(t) \right| \to 0, \quad \text{as } n \to \infty, \quad (28)
\]

on each sample path. Note that

\[
\int_{t-\omega}^{t} \frac{\Lambda^n(x) - \lambda_n x}{\sqrt{\Lambda_n}} \, dF^c_n(t-x) - \int_{t-\omega}^{t} \tilde{\Lambda}(x) \, dF^c(t-x) = \int_{t-\omega}^{t} \left( \frac{\Lambda^n(x) - \lambda_n x}{\sqrt{\Lambda_n}} - \tilde{\Lambda}(x) \right) \, dF^c_n(t-x) + \int_{t-\omega}^{t} \tilde{\Lambda}(x) \, dF^c_n(t-x) + \int_{t-\omega}^{t} \tilde{\Lambda}(x) \, d(F^c_n(t-x) - F^c(t-x)) .
\]

The first two terms on the right-hand side converge to zero in probability, following (28) and the assumption \( \omega^n \to \omega \). Since \( \tilde{\Lambda}(\cdot) \) is a Brownian motion, thus for any fixed \( T > 0 \),

\[
\lim_{\Gamma \to \infty} \mathbb{P}( \sup_{0 \leq t \leq T} |\tilde{\Lambda}(t)| \geq \Gamma ) = 0 .
\]

This and \( \{F^n(\cdot)\}_n \) converges to \( F(\cdot) \) in total variation on \([0, \omega]\) imply that the last term also converges to 0 in probability. Combining the above convergence with (27), we conclude that the second term in (26) weakly converges to \( \{\int_{t-\omega}^{t} F_c(t-x) \, d\tilde{\Lambda}(x) : t \geq \omega + \delta \} \). We now focus on the second term in (25).

For any \( M > 0 \) and \( n \in \mathbb{N} \), define the event \( \Omega^n_M = \{ \sup_{\omega + \delta \leq t \leq T} |\sqrt{\Lambda_n}(t - \omega^n - \kappa^n(t))| \leq M \} \). It is clear that, on the event \( \Omega^n_M \), we have

\[
\sum_{i=\Lambda^{n}(\kappa^{n}(t))+1}^{\Lambda^{n}(t-\omega^n)} 1_{\{u_i^n > \omega^n + \frac{M}{\sqrt{\Lambda_n}}\}} \leq \sum_{i=\Lambda^{n}(\kappa^{n}(t))+1}^{\Lambda^{n}(t-\omega^n)} 1_{\{u_i^n + \tau_i^n > t\}} \leq \sum_{i=\Lambda^{n}(\kappa^{n}(t))+1}^{\Lambda^{n}(t-\omega^n)} 1_{\{u_i^n > \omega^n - \frac{M}{\sqrt{\Lambda_n}}\}} .
\]

Introduce

\[
\tilde{G}_M^-(t) = \frac{1}{\sqrt{\Lambda_n}} \left[ \sum_{i=\Lambda^{n}(\kappa^{n}(t))+1}^{\Lambda^{n}(t-\omega^n)} \left( 1_{\{u_i^n > \omega^n + \frac{M}{\sqrt{\Lambda_n}}\}} - F^n_c(\omega^n - \frac{M}{\sqrt{\Lambda_n}}) \right) \right. + F^n_c(\omega^n - \frac{M}{\sqrt{\Lambda_n}}) (\Lambda^n(t - \omega^n) - \Lambda^n(t - \omega^n - \kappa^n(t))) + \left. \left( F^n_c(\omega^n - \frac{M}{\sqrt{\Lambda_n}}) - F^n_c(\omega^n) \right) \lambda_n(t - \omega^n - \kappa^n(t)) \right] ,
\]

\[
\tilde{G}_M^+(t) = \frac{1}{\sqrt{\Lambda_n}} \left[ \sum_{i=\Lambda^{n}(\kappa^{n}(t))+1}^{\Lambda^{n}(t-\omega^n)} \left( 1_{\{u_i^n > \omega^n + \frac{M}{\sqrt{\Lambda_n}}\}} - F^n_c(\omega^n + \frac{M}{\sqrt{\Lambda_n}}) \right) \right. + F^n_c(\omega^n + \frac{M}{\sqrt{\Lambda_n}}) (\Lambda^n(t - \omega^n) - \Lambda^n(t - \omega^n - \kappa^n(t))) + \left. \left( F^n_c(\omega^n + \frac{M}{\sqrt{\Lambda_n}}) - F^n_c(\omega^n) \right) \lambda_n(t - \omega^n - \kappa^n(t)) \right] .
\]
By Theorem 1, we have
\[ \sup V_n(t - \omega^n - \kappa^n(t)) \cdot \lambda_n(t - \omega^n - \kappa^n(t)). \]
Then
\[ \frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda_n(t-\omega^n)}^{\Lambda_n(t+\omega^n)} 1_{\left(\omega^n - \frac{M}{\sqrt{\lambda_n}} \right)} = \tilde{G}_n^- (t) + F_c^n (\omega^n) \sqrt{\lambda_n (t - \omega^n - \kappa^n(t))}, \]
and
\[ \frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda_n(t+\omega^n)}^{\Lambda_n(t+\omega^n)} 1_{\left(\omega^n + \frac{M}{\sqrt{\lambda_n}} \right)} = \tilde{G}_n^+ (t) + F_c^n (\omega^n) \sqrt{\lambda_n (t - \omega^n - \kappa^n(t))}. \]
As a result of being on the event \( \Omega^n_M \),
\[ \sup_{\omega + \delta \leq t \leq T} \left| \tilde{Q}^n(t) - \tilde{M}^n(t) - F_c^n (\omega^n) \sqrt{\lambda_n (t - \omega^n - \kappa^n(t))} \right| \]
\[ \leq \sup_{\omega + \delta \leq t \leq T} |\tilde{G}_n^- (t)| + \sup_{\omega + \delta \leq t \leq T} |\tilde{G}_n^+ (t)|. \]
It is clear, for any fixed \( M > 0 \),
\[ \sup_{\omega + \delta \leq t \leq T} |\tilde{G}_n^- (t)| + \sup_{\omega + \delta \leq t \leq T} |\tilde{G}_n^+ (t)| \Rightarrow 0. \]
As a result, for any \( \eta > 0 \),
\[ \mathbb{P} \left( \sup_{\omega + \delta \leq t \leq T} \left| \tilde{Q}^n(t) - \tilde{M}^n(t) - F_c^n (\omega^n) \sqrt{\lambda_n (t - \omega^n - \kappa^n(t))} \right| \geq \eta \right) \]
\[ \leq \mathbb{P} ((\Omega^n_M)^c) + \mathbb{P} \left( \left( \sup_{\omega + \delta \leq t \leq T} |\tilde{G}_n^- (t)| + \sup_{\omega + \delta \leq t \leq T} |\tilde{G}_n^+ (t)| \geq \eta \right) \right). \] Note that the definition of \( \kappa^n(\cdot) \) in (6) and Proposition 1 imply that, as \( n \to \infty \),
\[ \sup_{V_n(0) \leq t \leq T} \left| \kappa^n(t) - t + \omega \right| \to 0. \]
By the initial condition (10), the probability that \( V^n(0) > \omega + \delta \) is vanishing with \( n \to \infty \). As a result, we have
\[ \sup_{\omega + \delta \leq t \leq T} \left| \kappa^n(t) - t + \omega \right| \to 0. \]
By Theorem 1, \( \sup_{0 \leq t \leq T} \sqrt{\lambda_n} V^n(t) - V^n(t^-) \Rightarrow 0 \), as \( n \to \infty \). From the definition (6), we know that for \( \omega + \delta \leq t \leq T \), \( t \leq \kappa^n(t) + V^n(\kappa^n(t)) \leq t + \sup_{\omega + \delta \leq t \leq T} \left| V^n(\kappa^n(t)) - V^n(\kappa^n(t^-)) \right| \). This, together with (13), implies that, as \( n \to \infty \),
\[ \sup_{\omega + \delta \leq t \leq T} \left| \sqrt{\lambda_n} (t - \kappa^n(t) - V^n(\kappa^n(t))) \right| \]
\[ = \sup_{\omega + \delta \leq t \leq T} \left| \sqrt{\lambda_n} (t - \kappa^n(t) - \omega^n) - \tilde{V}^n(\kappa^n(t)) \right| \Rightarrow 0. \]
The first implication of (31) is that
\[ \lim_{M \to \infty} \lim_{n \to \infty} \mathbb{P} ((\Omega^n_M)^c) = 0. \]
Combining this with (29) and (30), we know that \( \{ \tilde{Q}^n(t) \} \) and \( \{ \tilde{M}^n(t) + F_c^n (\omega^n) \sqrt{\lambda_n (t - \omega^n - \kappa^n(t))} \} \) have the same weak limit. Since (31) also implies that \( \sqrt{\lambda_n} (t - \kappa^n(t) - \omega^n) \Rightarrow \tilde{V}^n(t - \omega) \), the result of the theorem follows. □
3.2. Steady-State Analysis

For the purpose of steady-state analysis, we add the assumption that the arrival process $\Lambda^n$ is a renewal process, where the inter-arrival time has mean $1/\lambda_n$ and variance $\theta^2/\lambda_n^2$. Then the limit $\tilde{\Lambda}$ in (1) can be written as $\theta B_\Lambda$, where $B_\Lambda = \{B_\Lambda(t) : t \geq 0\}$ is a standard Brownian motion, independent of $B$ and $B_A$ given by Theorem 1. Let

$$\sigma^2 = \theta^2 + 2\rho - 1; \quad (32)$$

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

$$\tilde{V}(t) = \tilde{V}(0) - \rho \int_0^t [f_\omega(\tilde{V}(x)) - \beta] \, dx + \sigma W(t), \quad t \geq 0. \quad (33)$$

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

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

$$\lim_{x \to \infty} f_w(x) > \beta, \quad \text{and} \quad \lim_{x \to -\infty} f_w(x) < \beta, \quad (34)$$

where $\beta$ is given in (4). Then the diffusion limit $\tilde{V}$ has a stationary distribution $\pi(\cdot)$ given by (denote by $\tilde{V}(\infty)$ a random variable with density function $\pi(\cdot)$)

$$\pi(y) = C \exp \left( -\frac{2\rho}{\sigma^2} \int_y^\infty [f_\omega(x) - \beta] \, dx \right), \quad y \in \mathbb{R}, \quad (35)$$

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

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

here $\tilde{N}_1$ and $\tilde{N}_2$ are normal random variables, both with mean zero and, respectively, variances $\theta^2 \int_\omega^\infty (F_\omega(x))^2 \, dx$ and $\int_\omega^\infty F(x) F_\omega(x) \, dx$.

We comment here that the density function $\pi(\cdot)$, and in particular the normalizing constant $C$, depend on the limit $f_\omega(\cdot)$ in (3). 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 (34) on $f_\omega$ 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 (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 often done (e.g., Reed and Ward (2008) and 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 $\mu$; (ii) the patience-time distribution $H(\cdot)$; (iii) mean and variance of the inter-arrival time $1/\lambda$ and $\theta^2/\lambda^2$, respectively. We denote the system by $(s, \mu, \lambda, \theta^2, H)$. It is realistically unreasonable to assume that the patience-time distribution depends on the size of the system. The heavy traffic assumptions (2) and (4) merely constitute a mathematical tool that guides on how to capture the structure of the patience-time distribution $H(\cdot)$ around $\omega$, and its impact on performances for a single system with $s$ servers. We rely on the stationary distribution $\pi$ 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 (35), set

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

(36)

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

(37)

$$f_\omega(x) := \sqrt{\lambda} \left[ H(\omega + \frac{x}{\sqrt{\lambda}}) - H(\omega) \right].$$

(38)

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

$$\pi(y) = C \exp \left( -\frac{2\rho}{\sigma^2} \int_0^y [f_\omega(x) - \beta] \, dx \right),$$

(39)

where

$$C = \left( \int_{-\infty}^{\infty} \exp \left( -\frac{2\rho}{\sigma^2} \int_0^y [f_\omega(x) - \beta] \, dx \right) \, dy \right)^{-1}$$

(40)

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 (39),

$$\Pr(W(\infty) > y) = H_c(y) \Pr \left( \sqrt{\lambda}(V(\infty) - \omega) > \sqrt{\lambda}(y - \omega) \right).$$

(40)
\begin{equation}
\approx H_c(y) \int_{\sqrt{\lambda}(y - \omega)}^{\infty} C \exp \left( -\frac{2\rho}{\sigma^2} \int_{0}^{\omega} [f_\omega(x) - \beta] dx \right) dv.
\end{equation}

A special choice of \( y \) is \( \omega \), 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

\[
\lambda \int_{0}^{\omega} H_c(x) dx + \sqrt{\lambda} \left( \tilde{N}_1 + \tilde{N}_2 + \frac{\sqrt{\lambda}}{\rho} (V(\infty) - \omega) \right).
\]

Consequently, performance measures related to \( Q(\infty) \) can be calculated explicitly using (39) and the above formula. For example, the expected steady-state queue-length can be calculated as

\[
\mathbb{E}Q(\infty) \approx \lambda \int_{0}^{\omega} H_c(x) dx + \frac{1}{\rho} \sqrt{\lambda} \int_{-\infty}^{\infty} x \pi(dx).
\]

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”. We now analyze separately announcements upon customer arrival and announcements during waiting.

#### 4.1. Delay Announcement Upon Arrival

Arriving customers who must wait often receive an announcement upon arrival concerning their anticipated delay. The announced information could include a single number \( \tau \) related to the delay, which is called a *fixed delay announcement* by [Armony et al. (2009)](Armony et al. 2009). Customers respond to the announcement by choosing to balk or not. Given a delay announcement of \( \tau \), 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-time is also affected by the announced information \( \tau \). This effect is modeled by assuming that customers’ patience-time follows the conditional distribution \( H(t|\tau) \).

An announced delay \( \omega_e \) is an *equilibrium delay* if

\[
\rho B_c(\omega_e) H_c(\omega_e|\omega_e) \leq 1 \quad \text{and} \quad \rho B_c(\omega_e) H_c(t|\omega_e) > 1 \quad \text{for} \quad 0 \leq t < \omega_e.
\]
When there is a unique equilibrium delay $\omega_e$, the above formal relations capture the facts that, in equilibrium, and asymptotically in the fluid scale, the announced delay $\tau$ is equal to the long-run average delay of served customers, and both are equal to the equilibrium delay $\omega_e$. (The concept of equilibrium delay in the fluid scale was introduced by [Armony et al. 2009].)

In this subsection, we consider the “all-exponential model” proposed in [Armony et al. 2009]. Customers arrive according to a Poisson process with rate $\lambda$. Service times are exponentially distributed with rate $\mu$. For a delay announcement of $\tau$, the balking probability is $B(\tau) = 1 - e^{-b\tau}$ and customers’ patience is

$$H(t|\tau) = \begin{cases} 1 - e^{-h_0t}, & 0 \leq t \leq \tau; \\ 1 - e^{-h_0}\tau e^{-h_1(t-\tau)}, & t > \tau. \end{cases} \quad (44)$$

It follows from (43) that the equilibrium delay is

$$\omega_e = \frac{1}{b + h_0} \ln \rho. \quad (45)$$

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 there). 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 needed.

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 $\tilde{H}(\cdot|\omega_e) = B(\omega_e) + B_c(\omega_e)H(\cdot|\omega_e)$.

We now consider the same set of parameters as in [Armony et al. 2009]. The number of servers $s = 100$ with individual service rate $\mu = 1$. The arrival $\lambda = 140$, balking rate $b = 1$ and the patience-time hazard-rate $h_0 = 0.5$ and $h_1 = 4$. By (45), 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 “Simulated” and “Fluid” are taken from Table 1 in [Armony et al. 2009]. The columns “Diffusion” are calculated using formula (39). In particular, $E[Q(\infty)]$ is calculated based on (42), and $E[W(\infty); B_c]$ based on (41). The calculation of $P(W(\infty) \leq \omega_e|S)$, the conditional probability that the steady-state waiting-time is less than $\omega_e$, follows from

$$P(W(\infty) \leq \omega_e|S) = \frac{P(W(\infty) \leq u \wedge \omega_e)}{P(W(\infty) \leq u)},$$

where $u$ is a generic independent random variable, following the distribution $\tilde{H}(\cdot|\omega_e)$ and $P(W(\infty) \leq y)$ can be calculated from (41) 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
\( P(W(\infty) > z) \) 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 to the setting of delay announcements,
e.g., to the all-exponential model. Let the individual service rate \( \mu \), arrival rate \( \lambda \), variance of the
interarrival time \( \theta^2 \) and the patience-time distribution \( H(\cdot) \) be given. Suppose that the patience-
time distribution \( H(\cdot) \) exhibits a sharp change around \( \omega \). We seek the number of servers \( s \) such
that staffing cost is minimized while adhering to the given service level agreement \((z, \alpha)\) as follows:

\[
\min s \\
\text{s.t. } P(W(\infty) > z) \leq \alpha.
\]

Remark 1. In the analysis of delay announcements, as above, the distribution \( H(\cdot) \) becomes
\( \tilde{H}(\cdot|\omega_e) \), where \( \omega_e \) is the equilibrium delay, and \( \omega_e = \tau \); it is furthermore natural to choose \( z = \tau \),
since then the constraint in (46) bounds the fraction of “broken promises”.

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 dis-
tribution, 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 (46). It is based on the
steady-state approximations (39)–(41) in Section 3.2. From (37), the number of servers is

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

(47)

Note that \( \beta \) is the only element needed to determine the number of servers \( s \). According to (41),
to solve the optimization problem (46), one must solve

\[
\max \beta \\
\text{s.t. } H_c(\omega) \int_0^\infty C \exp \left( -\frac{2\rho}{\sigma^2} \int_0^y [f_\omega(x) - \beta] \, dx \right) \, dy \leq \alpha.
\]

(48)
To apply this to the delay announcement model, we just need to replace $H_c(\omega)$ by $\tilde{H}(\omega_e|\omega_e)$ in (48). For the all-exponential model, the optimal staffing level is determined by (47), with $\beta$ replaced by $\beta_*$—the optimal solution to (48). We see that $\beta_*$ 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 an announcement can reduce the staffing without hurting service-levels.

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 $\lambda$. Service times are exponentially distributed with rate $\mu$. The patience-time distribution is

$$H(x) = \begin{cases} x, & x \leq \omega, \\ \omega + k(x - \omega), & \omega < x \leq \omega + \frac{1-\omega}{k}, \end{cases}$$

(49)

where $\omega$ is the fluid offered waiting-time.

In this example, set the parameter $\omega = 1/6$ and $k = 5$ in (49). 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 $\omega$. Thus, we cannot use existing results such as Mandelbaum and Zeltyn (2009).

Applying the staffing rule (47) by numerically solving (48) 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 vs. the number of servers in Figure 1. The vertical axis is the probability that waiting-time exceeds $\omega$.

![Figure 1](image-url)  

**Figure 1** Staffing level and probability that waiting-time exceeds $\omega$.  

In this example, set the parameter $\omega = 1/6$ and $k = 5$ in (49). 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 $\omega$. Thus, we cannot use existing results such as Mandelbaum and Zeltyn (2009). Applying the staffing rule (47) by numerically solving (48) 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 vs. the number of servers in Figure 1. The vertical axis is the probability that waiting-time exceeds $\omega$. \n
the fluid offered 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 quite accurate: the curve obtained almost overlaps the curve by simulation. In fact, the optimal number of servers required in order to achieve \( \mathbb{P}(W(\infty) > \omega) \leq \alpha \) is almost identical to the solution found by the numerical simulation, for service level \( \alpha \) ranging from 0.1 to 0.4.

**Example 2.** Customers arrive according to a Poisson process with rate \( \lambda \). Service times are exponentially distributed with rate \( \mu \). The hazard-rate of the patience-time is as follows

\[
\tilde{h}(x) = \begin{cases} 
  h_0, & x \leq \omega, \\
  h_0 + \kappa(x - \omega), & x > \omega,
\end{cases}
\]

where \( \omega \) is the fluid offered waiting-time.

![Figure 2](image.png)

**Figure 2** Staffing level and probability that waiting-time exceeds \( \omega \). \( \kappa = 100 \)

In this example, set the parameter \( \omega = \log(\rho)/h_0 \) with \( \rho = 1.2 \), \( h_0 = 1.0 \) and \( \kappa = 100 \) in (50). 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 (47) by numerically solving (48) 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 vs. 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 $\alpha = 0.4$, suggests that the optimal number of servers is 101, which is off by 5.

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

### 4.2. Delay Announcement During Waiting – Asymptotic Optimality

Instead of making announcements upon arrival, one can make an announcement while customers are waiting. Such announcements are known to have an impact on the patience-time distribution. For example, Figure 3 (taken from Mandelbaum and Zeltyn (2013)) depicts a surge of the hazard rate, in response to an announcement that is made when customers’ waiting-times reach 60 seconds.

![Hazard rate of patience time](image)

**Figure 3**  Hazard-rates of patience-time in an Israeli call center. The local peak at 60 seconds is an outcome of the following 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 customers who are unwilling to wait.)

A Motivating Example: Controlling a Surge in Arrivals by Announcements. In this section, we study joint optimality of the staffing and timing of an announcement. To motivate the study, we give an example of where a call center (US bank) faced a surge of demand for a couple of days (Oct 9–11) following a special promotion. Figure 4 shows the actual number of servers employed on Oct 10 and what would be required (calculated using the Garnett function and 4CC, respectively) if we wanted the call center to operate in the QED regime. During most of the day, this required number is almost double the number actually employed. Consequently, the call center experienced a high abandonment fraction (around 50%, as seen in Figure 5).
Figure 4  Required staffing to operate in QED and actual staffing

Figure 5  Fraction of abandonment.

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 waiting tail-probability (see (53) in the sequel).

A Queueing Model. Based on empirical findings that a sudden jump is associated with an announcement, we propose the following stylized (yet insightful) model. If the announcement is made when customers’ waiting-time reaches $\tau$, then the patience-time is

$$H(x|\tau) = \begin{cases} 
1 - e^{-h_0 x}, & x \leq \tau, \\
1 - e^{-h_0 \tau - h_1(x-\tau)}, & x > \tau; 
\end{cases} \quad (51)$$

here $h_0$ and $h_1$ are given parameters,

$$h_0 < h_1, \quad (52)$$

that capture the increase of hazard-rate caused by the announcement made at $\tau$. We comment here that the form [51] was chosen for both correctness 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 (51) 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 Assumption 52 supported by empirical data, it is also quite intuitive. If \( h_0 = h_1 \), the announcement does not make any difference, thus 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 pointed out by Mandelbaum and Zeltyn (2013): no matter the purpose of an announcement, the ultimate result is an encouragement for customers to abandon the system.

**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 staffing level while subject to a pre-specified service level. The latter is characterized by two constraints: fraction of abandonment less than \( \alpha_1 \) and fraction of those waiting above a threshold less than \( \alpha_2 \). 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 is closely related to revenue generation, as abandonment typically means revenue loss. The second constraint caters mainly to customer satisfaction. The optimization problem is hence formulated as follows:

\[
\begin{align*}
\min_{s, \tau} & \quad s \\
\text{s.t} & \quad \mathbb{P}(\text{Ab}) \leq \alpha_1, \\
& \quad \mathbb{P}(W(\infty) > \omega_\tau \wedge \tau) \leq \alpha_2,
\end{align*}
\]

(53)

where \( \omega_\tau \) is determined by the equation \( H(\omega_\tau | \tau) = \alpha_1 \). It should be pointed out that both constraints depend on the announcement time \( \tau \). 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 (53) similarly to (46), by using the approximation formula (41) from our diffusion analysis, with some technical adjustment. To gain insight, we analyze a large-scale limit where the arrival rate \( \lambda \) increases indefinitely. Let \( (s^\lambda, \tau^\lambda) \) be an optimal solution to (53).

**Proposition 3.** Denote by \( \tau_* \) the unique solution to \( H(\tau_* | \tau) = \alpha_1 \). Then any optimal announcement epoch \( \tau^\lambda_* \) satisfies

\[
\lim_{\lambda \to \infty} \sqrt{\lambda}(\tau_* - \tau^\lambda_*) = 0.
\]
Moreover, the optimal number of servers is
\[ s^\lambda = \frac{\lambda}{\mu} (1 - \alpha_1) - \sqrt{\lambda} \frac{\beta_\star}{\mu} + o(\sqrt{\lambda}), \tag{54} \]
where \( \beta_\star \) is the unique solution to
\[ \max_{\beta} \beta \text{ s.t. } (1 - \alpha_1) \int_{-\infty}^{\infty} \exp\left( -\frac{2\beta}{\sigma^2} \int_{0}^{y} [f_{\tau_\star}(x) - \beta] \, dx \right) \, dy < \alpha_2, \tag{55} \]
with \( f_{\tau_\star}(x) = \lim_{\lambda \to \infty} \sqrt{\lambda} \left[ H(\tau_\star + \frac{x}{\sqrt{\lambda}}) - H(\tau_\star) \right] \).

The proof of this proposition is presented in Appendix [EC.1]. The key message is that, in the asymptotic sense, it is optimal to make the announcement so that the abandonment fraction is exactly \( \alpha_1 \), and then set the optimal staffing level according to (54). The staffing level then depends on \( \beta_\star \) only. Comparing this to the case without announcement (with \( f_{\tau_\star}(x) \) in (55) replaced by \( e^{-h_{\omega}x} \)), it is easy to verify that \( \beta_\star \) in the case with announcement is smaller; hence the staffing level in (54) is reduced by \( O(\sqrt{\lambda}) \).

5. Refined Approximations

The approximation (39), together with the derived formulae (41)–(42) for the steady-state, is general, and does not depend on special properties of the patience-time distribution. This generality is its strength and weakness: it provides an accurate recipe that is free of details. This freedom is at the cost of additional insight and closed-form expressions. Such advantages are consequences of a special structure, which we exploit in the present section. 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

Assume that the patience-time distribution \( H(\cdot) \) has a density at \( \omega \) and write it as \( H'(\omega) \). Then \( f_\omega(x) \), defined in (38), can be “well approximated” by \( H'(\omega)x \), and the density \( \pi \) in (39) can be approximately specialized to
\[ \pi(x) \approx C \exp \left( -\frac{\rho}{\sigma^2} \left[ H'(\omega)x^2 - 2\beta x \right] \right). \tag{56} \]
This implies that \( \sqrt{\lambda}(V(\infty) - \omega) \) asymptotically follows a normal distribution. Consequently, it follows from (41) that
\[
\mathbb{P}(W(\infty) > y) \approx H_{\epsilon}(y) \int_{\sqrt{\lambda}(y-\omega)}^{\infty} C \exp \left( -\frac{\rho}{\sigma^2} \left[ H'(\omega)x^2 - 2\beta x \right] \right) \, dx \\
= H_{\epsilon}(y) \mathbb{P}\left( \frac{\beta}{H'(\omega)} + \frac{\sigma}{\sqrt{2\rho H'(\omega)}} N > \sqrt{\lambda}(y - \omega) \right)
\]
\[ = H_c(y) \Phi_c \left( \frac{-\beta \sqrt{2 \rho} + \sqrt{2 \rho \lambda H'(\omega)(y - \omega)}}{\sigma \sqrt{H'(\omega)}} \right), \tag{57} \]

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 
\[ s = H_c(\omega) \frac{\lambda}{\mu} + \delta \sqrt{\frac{\lambda}{\mu}}, \] with \( \delta \) any finite constant, and the arrival process is a Poisson process, then we have \( \sigma^2 = 2 \rho \) and \( \beta = -\delta \sqrt{\mu} \) by (36)–(37). Furthermore, we have
\[ P(W(\infty) > \omega) \approx H_c(\omega) \Phi_c \left( \frac{\delta \sqrt{\mu}}{\sqrt{H'(\omega)}} \right). \]

This is consistent with Theorem 4.3 in [Mandelbaum and Zeltyn (2009)]. There is also a connection to [Bassamboo and Randhawa (2010)], where accuracy of the fluid approximation to the expected queue-length is studied. By the symmetry of the normal distribution, in view of (42), we have
\[ \mathbb{E}Q(\infty) \approx \lambda \int_0^\omega H_c(x) \, dx. \tag{58} \]

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 \( \omega, H'(\omega+) \) and \( H'(\omega-) \), are not equal. Following (38), for large \( \lambda \),
\[ f_\omega(x) = \sqrt{\lambda} \left[ H(\omega + \frac{x}{\sqrt{\lambda}}) - H(\omega) \right] \]
\[ \approx \begin{cases} H'(\omega-) x, & x \leq 0, \\ H'(\omega+) x, & x > 0. \end{cases} \tag{59} \]

The density \( \pi \) in (39) can be then approximately specialized to
\[ \pi(x) \approx \begin{cases} C \exp \left( -\frac{1}{2} \frac{(x - \frac{\beta}{H'(\omega-)})^2}{\frac{1}{2 \rho H'(\omega-)}} \right) \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega-)} \right), & x \leq 0, \\ C \exp \left( -\frac{1}{2} \frac{(x - \frac{\beta}{H'(\omega+)})^2}{\frac{1}{2 \rho H'(\omega+)}} \right) \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega+)} \right), & x > 0, \end{cases} \tag{60} \]

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 (41) by
\[ P(W(∞) > ω) \approx CH_c(ω)Φ_c\left(\frac{-√2ρβ}{σH'(ω^+)}\right)\exp\left(\frac{ρβ^2}{σ^2H'(ω^+)}\right) - \frac{1}{√H'(ω^+)} . \quad (61) \]

It follows from (42) that the expected queue-length is
\[ EQ(∞) \approx λ\left(ω - \frac{ω^2}{2}\right) + C\sqrt{λ} \left\{ \int_0^∞ x \exp\left(-\frac{1}{2}\frac{(x - \frac{H'(ω^+)β}{ρ})^2}{\frac{1}{2}ρH'(ω^+)}\right) \exp\left(\frac{ρβ^2}{σ^2H'(ω^-)}\right) dx \right\} \]
\[ = λ\left(ω - \frac{ω^2}{2}\right) + C\sqrt{λ} \left\{ \frac{σ^2}{2ρ}\left[H'(ω^+) - \frac{1}{H'(ω^-)}\right] + \sqrt{2πσ}\frac{β}{(H'(ω^+))^{3/2}} \exp\left(\frac{ρβ^2}{σ^2H'(ω^+)}\right)Φ_c\left(-\frac{√2ρβ}{σH'(ω^+)}\right) \right\} + \frac{β}{(H'(ω^-))^{3/2}} \exp\left(\frac{ρβ^2}{σ^2H'(ω^-)}\right)Φ\left(-\frac{√2ρβ}{σH'(ω^-)}\right) \} . \]

The last derivation is in fact a generalization of the Garnett function, introduced in Garnett et al. (2002). To make the connection, let \( h_N(·) \) be the hazard-rate function for the standard Normal distribution. Then (61) can be written as
\[ P(W(∞) > ω) \approx H_c(ω) \left(1 + \sqrt{H'(ω^+)}/h_N\left(\frac{-√2ρβ}{σH'(ω^+)}\right)\right)^{-1} = H_c(ω) × Garnett_y(x) \]
where Garnett_y(x) = \( [1 + \sqrt{y}h_N(x)/h_N(-x)]^{-1}, y = H'(ω^+)/H'(ω^-), \) and \( x = -√2ρβ/(σ√H'(ω^-)) \).

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

Table 1 summarizes the comparison for Example 1 with a different right-derivative \( k = 1, 3, 5 \). The column “Approx.” is obtained via our approximation formulae (41)–(42). 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., \( H(·) \) 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 1 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 (38), we compute

\[
 f_\omega(x) = \sqrt{\lambda} \left[ H(\omega + \frac{x}{\sqrt{\lambda}}) - H(\omega) \right] \\
= \exp \left( - \int_0^\omega h(y)dy \right) \sqrt{\lambda} \left[ 1 - \exp \left( - \int_\omega^{\omega + \frac{x}{\sqrt{\lambda}}} h(y)dy \right) \right] \\
= H_c(\omega) \sqrt{\lambda} \left[ 1 - \exp \left( - \int_0^{\omega + \frac{x}{\sqrt{\lambda}}} h(\omega + y)dy \right) \right] \\
\approx H_c(\omega) \sqrt{\lambda} \left[ \frac{1}{\sqrt{\lambda}} \int_0^x h(\omega + \frac{y}{\sqrt{\lambda}})dy \right] \quad \text{(for large } \lambda) \\
= H_c(\omega) \int_0^x h(\omega + \frac{y}{\sqrt{\lambda}})dy. 
\]  

From (39), the density \( \pi \) of \( \sqrt{\lambda}(V(\infty) - \omega) \) can be approximated by

\[
\pi(x) \approx C \exp \left( \frac{2\rho \beta x}{\sigma^2} \right) \exp \left( -\frac{2\rho}{\sigma^2} H_c(\omega) \int_0^x \int_0^v h(\omega + \frac{y}{\sqrt{\lambda}})dydv \right), 
\]  

with the appropriate normalizing constant \( C \). Based on (63), the probability \( \mathbb{P}(W(\infty) > y) \) and expected queue-length \( \mathbb{E}Q(\infty) \) can be approximated by replacing \( f_\omega(x) \) in (41)–(42) with \( H_c(\omega) \int_0^x h(\omega + \frac{y}{\sqrt{\lambda}})dy \).

Consider now Example 2 where the density function of the patience-time distribution exists but the hazard-rate has a very steep change around \( \omega \). The individual service rate \( \mu = 1 \). Assume that \( \rho = 1.2 \) and \( h_0 = 1 \); hence the offered waiting-time \( \omega = \log(1.2) \). As in the previous example, we
report only the study for $\beta = 0$. 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 (41)–(42) with $f_\omega(x)$ from (38), while column “Appr. H” is calculated by replacing (38) with (62). As Table 3 (a) shows, the larger the parameter $\kappa$ becomes (meaning a steeper change of the hazard-rate function), 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 $P(W_\infty > \omega)$ 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 relate our general setting of scaling the patience-time distribution to the hazard-rate scaling in Reed and Tezcan (2012). We point out 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 the diffusion limits in the QED and ED+QED regimes share some 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 $F^n_c(\omega) = H_c(\omega)$. In view of (62),

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

Therefore, by (63), the approximation of the density of $\sqrt{\lambda}(V(\infty) - \omega)$ can be written as

$$\pi(x) \approx \begin{cases} C \exp\left(\frac{2\rho \beta x}{\sigma^2}\right) \exp\left(-\frac{\rho}{4\sigma^2}H_c(\omega)h(\omega)x^2\right), & \text{if } x \leq 0, \\ C \exp\left(\frac{2\rho \beta x}{\sigma^2}\right) \exp\left(-\frac{2\rho}{\sigma^2}H_c(\omega)\int_0^x h(\omega + y)dy\right), & \text{if } x > 0, \end{cases}$$

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

5.4. On the Gap between Fluid and Diffusion Models

We have seen 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 ($\mu = 1$, $h_0 = 1$, $\rho = 1.2$, $\omega = \log(\rho)/h_0$ 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 $\{10^i, i = 1, \ldots, 6\}$.

<table>
<thead>
<tr>
<th>Servers</th>
<th>$\kappa = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulated</td>
</tr>
<tr>
<td>10</td>
<td>1.30 ± .00</td>
</tr>
<tr>
<td>100</td>
<td>16.34 ± .04</td>
</tr>
<tr>
<td>1000</td>
<td>191.58 ± .22</td>
</tr>
<tr>
<td>100000</td>
<td>19980.10 ± 22.18</td>
</tr>
<tr>
<td>1000000</td>
<td>199929.32 ± 86.96</td>
</tr>
</tbody>
</table>

Table 4 On the accuracy of the fluid approximation in the ED regime as system size becomes large.

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
Figure 6  Gap between the fluid and diffusion approximations.

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.

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 (50) into (62), we get the steady-state density,

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

where

$$
C\kappa = \left[ \int_{-\infty}^{\infty} \exp \left( -\frac{h_0 x^2}{2} - \frac{\kappa x^3}{6\sqrt{\lambda}} \right) dx + \frac{\sqrt[4]{2\pi}}{2\sqrt{h_0}} \right]^{-1} \to \frac{\sqrt{h_0}}{\sqrt{2\pi}}, \quad \text{as } \lambda \to \infty.
$$

Also, from (42) and (65), we deduce the approximation for the expected queue-length

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

where

$$
\text{gap}(\lambda) = C\kappa \int_0^\infty x \exp\left(-\frac{x^2}{2}\right) \left( \exp\left(-\frac{\kappa x^3}{6\sqrt{\lambda}}\right) - 1 \right) \sqrt{\lambda} dx.
$$

It is easy to verify the bound $\left| \left( \exp\left(-\frac{\kappa x^3}{6\sqrt{\lambda}}\right) - 1 \right) \sqrt{\lambda} \right| \leq \frac{\kappa x^3}{6}$, 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 $x \geq 0$. Consequently,

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

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 size of system 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 on the gap of 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 is continuously differentiable.

6. Conclusion

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

The fine structure of patience-time naturally arises from delay announcements. Applying our approximation formulae, we thus investigate 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 also prescribe the optimal staffing rule in 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 distributions. 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 existing literature.

From the technical point of view, we offer a novel method for obtaining diffusion limits of many-server queues, by focusing on the virtual waiting-time. This method could be also applied in the QED regime, but we leave this for future research. Also, one may consider a combination of the two types of announcements, or even repeated announcements in future studies.

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 could be convenient to call them back.

Acknowledgments

Jiheng Zhang’s research is supported in part by the Hong Kong Research Grants Council [Grants GRF-622411 and GRF-624012].
References


Technical Proofs

EC.1. Proofs of Propositions 1–3

In this section, we provide the proofs for Propositions 1–3.

Proof of Proposition 1. It suffices to show that, for any $T > 0$ and $\delta \in (0, \omega/2)$,

$$
\mathbb{P}\left( \sup_{0 \leq t \leq T} |V^n(t) - \omega^n| \geq \delta \right) \to 0 \text{ as } n \to \infty. \quad (EC.1)
$$

Define $\hat{V}^n(t) = V^n(t) - \omega^n$, $\eta^n_1 = \inf\{t \geq 0 : \hat{V}^n(t) \geq \delta\}$ and $\eta^n_2 = \inf\{t \geq 0 : \hat{V}^n(t) \leq -\delta\}$. Let $\Omega^n_1(\delta, T) = \{\eta^n_1 \leq \eta^n_2, \eta^n_1 \leq T\}$, $\Omega^n_2(\delta, T) = \{\eta^n_1 > \eta^n_2, \eta^n_1 \leq T\}$, and $\Omega^n_0(\delta) = \{\hat{V}^n(0) \leq \delta/4\}$. In view of (10), to get (EC.1), it is sufficient to prove that the probabilities of the events $\Omega^n_1(\delta, T) \cap \Omega^n_2(\delta)$ and $\Omega^n_2(\delta, T) \cap \Omega^n_0(\delta)$ vanish as $n$ converges to infinity. We will only consider the event $\Omega^n_1(\delta, T) \cap \Omega^n_0(\delta)$, since the analysis of $\Omega^n_1(\delta, T) \cap \Omega^n_2(\delta)$ is similar. On the set $\Omega^n_1(\delta, T) \cap \Omega^n_0(\delta)$, define $\eta^n_{12} = \sup\{0 \leq t \leq \eta^n_1 : \hat{V}^n(t) \leq \delta/3\}$ for any $\delta$. By the definitions of $\eta^n_1$ and $\eta^n_{12}$, we clearly have that

$$
\hat{V}^n(\eta^n_1) \geq \delta, \text{ and } \hat{V}^n(\eta^n_{12} -) \leq \frac{\delta}{3}.
$$

In view of $F^n(\omega^n + x) \geq F^n(\omega^n)$ for any $x \geq 0$, by (18)–(19), we have that

$$
\hat{V}^n(\eta^n_1) - \hat{V}^n(\eta^n_{12} -) \leq \frac{\lambda_n}{s_n \mu} \left[ F^n_c(\omega^n) \frac{\bar{\Lambda}^n(\eta^n_1)}{\sqrt{\lambda_n}} - \frac{1}{\lambda_n} \sum_{i=\Lambda(\eta^n_{12})}^{\Lambda^n(\eta^n_1)} (1_{\{\omega^n_{1i} \leq \omega^n\}} - F^n(\omega^n)) \right]
$$

$$
- \frac{(\lambda_n F^n_c(\omega) - s_n \mu)(\eta^n_1 - \eta^n_{12})}{\lambda_n} - \frac{\bar{V}^n(\eta^n_1 + V^n(\eta^n_1)) - \bar{V}^n(\eta^n_{12} + V^n(\eta^n_{12} -))}{\sqrt{\lambda_n}}. \quad (EC.2)
$$

By (1), (2) and (4), we know that, as $n \to \infty$,

$$
\mathbb{P}\left( \left| \frac{\lambda_n F^n_c(\omega)}{s_n \mu} \cdot \frac{\bar{\Lambda}^n(\eta^n_1) - \bar{\Lambda}^n(\eta^n_{12} -)}{\sqrt{\lambda_n}} \right| > \frac{\delta}{6} \right) \to 0, \quad (EC.3)
$$

$$
\mathbb{P}\left( \left| \frac{1}{s_n \mu} \sum_{i=\Lambda(\eta^n_{12})}^{\Lambda^n(\eta^n_1)} (1_{\{\omega^n_{1i} \leq \omega^n\}} - F^n(\omega^n)) \right| > \frac{\delta}{6} \right) \to 0, \quad (EC.4)
$$

$$
\mathbb{P}\left( \left| \frac{1}{s_n \mu} (\lambda_n F^n_c(\omega^n) - s_n \mu)(\eta^n_1 - \eta^n_{12}) \right| > \frac{\delta}{6} \right) \to 0. \quad (EC.5)
$$

To get that the last term in (EC.2) also vanishes, let $S^n(t)$ denote the number of customers in service at time $t$, and $D^n(t)$ the number of departures through service completion by time $t$. We can relate these two processes with $B^n(t)$ by

$$
B^n(t) = D^n(t) + S^n(t) - S^n(0), \quad (EC.6)
$$
which implies that

\[
B^n(\eta_1^n + V^n(\eta_1^n)) - B^n(\eta_1^n + V^n(\eta_1^n)) = D^n(\eta_1^n + V^n(\eta_1^n)) - D^n(\eta_1^n + V^n(\eta_1^n)) \\
+ S^n(\eta_1^n + V^n(\eta_1^n)) - S^n(\eta_1^n + V^n(\eta_1^n)).
\]

As \(V^n(\cdot)\) is always positive on \([\eta_1^n + V^n(\eta_1^n), \eta_1^n + V^n(\eta_1^n)]\), all the servers are busy; hence \(S^n(\eta_1^n + V^n(\eta_1^n)) = S^n(\eta_1^n + V^n(\eta_1^n)) = s_n\). As a result, noticing that the service time is exponential with rate \(\mu\),

\[
B^n(\eta_1^n + V^n(\eta_1^n)) - B^n(\eta_1^n + V^n(\eta_1^n)) = D^n(\eta_1^n + V^n(\eta_1^n)) - D^n(\eta_1^n + V^n(\eta_1^n)) \\
= S(s_n(\eta_1^n + V^n(\eta_1^n))) - S(s_n(\eta_1^n + V^n(\eta_1^n))).
\]

where \(\{S(t) : t \geq 0\}\) is a Poisson process with rate \(\mu\). Hence, we have that, as \(n \to \infty\),

\[
p\left(\frac{\lambda_n}{s_n \mu} \left| \frac{B^n(\eta_1^n + V^n(\eta_1^n)) - B^n(\eta_1^n + V^n(\eta_1^n))}{\sqrt{\lambda_n}} \right| > \frac{\delta}{6}\right) \to 0.
\]

Combining (EC.2)–(EC.5) and (EC.7), the probability of the event \(\Omega_3(\delta, T) \cap \Omega_3(\delta)\) will vanish as \(n \to \infty\).

Proof of Proposition 2. First consider the stationary distribution of the diffusion limit for the virtual waiting-time process. Introduce \(g(x) = \rho f(x) - \beta\). Note that, in view of (34), \(\lim_{x \to \infty} g(x) > 0\) and \(\lim_{x \to \infty} g(x) < 0\). Now let \(\mathcal{X} = \{\mathcal{X}(t) : t \geq 0\}\) be the solution to the following stochastic differential equation:

\[
d\mathcal{X}(t) = -g(\mathcal{X}(t))dt + \sigma d\mathcal{W}(t), \quad t \geq 0.
\]

It is enough to prove that the stationary distribution of \(\mathcal{X}\) has the density

\[
\pi(y) = C \exp \left( -\frac{2}{\sigma^2} \int_0^y g(x)dx \right), \tag{EC.8}
\]

where \(C\) is a normalizing constant. Noting that the generator of \(\mathcal{X}\) is

\[
\mathcal{A} = \frac{\sigma^2}{2} \frac{d^2}{dx^2} - g(x) \frac{d}{dx},
\]

it is enough to prove that the function \(\pi\) in (EC.8) satisfies

\[
\int_{\mathbb{R}} \mathcal{A} f(x) \pi(x)dx = 0, \tag{EC.9}
\]

for all \(f(\cdot)\) in the class of bounded, twice continuously differentiable functions (see Ethier and Kurtz (1986), page 248). However, with \(\lim_{x \to \infty} g(x) > 0\), it can be easily verified that, with \(\pi\) given in (EC.8), we have

\[
\int_{\mathbb{R}_+} \mathcal{A} f(x) \pi(x)dx = C \sigma^2 \int_{\mathbb{R}_+} d \left[ \exp \left( -\frac{2}{\sigma^2} \int_0^y g(x)dx \right) f'(y) \right] = -C \frac{\sigma^2}{2} f'(0). \tag{EC.10}
\]
Similarly,
\[
\int_{\mathbb{R}_-} Af(x)\pi(x)\,dx = \frac{C\sigma^2}{2} \int_{\mathbb{R}_-} d\left[ \exp\left( -\frac{2}{\sigma^2} \int_0^y g(x)\,dx \right) f'(y) \right] = \frac{C\sigma^2}{2} f'(0).
\] (EC.11)

We now conclude (EC.9) by summing up (EC.10) and (EC.11). This implies (EC.8), and hence (35).

For the stationary distribution of the diffusion limit of the queue-length, note that, for \( t > \omega \),
\[
\int_{t-\omega}^t \int_0^t 1_{\{s+x \leq t\}} \,d\tilde{H}(s,F(x)) \text{ is normally distributed with zero mean and variance } \int_0^\omega F(x)F_c(x)\,dx
\]
(see Krichagina and Puhalskii (1997)). Similarly, \( \int_{t-\omega}^t F_c(t-x)\,d\tilde{\lambda}(x) \) follows a zero-mean normal distribution with variance \( \theta^2 \int_0^\omega (F_c(x))^2\,dx \). Hence the second result is implied by Theorem 2. This completes the proof. \( \square \)

Proof of Proposition 3. We will prove the statement by contradiction, via considering two cases:

(i) \( \limsup_{\lambda \to \infty} \sqrt{\lambda}(\tau_\omega - \tau_\omega^\lambda) > 0 \) and (ii) \( \liminf_{\lambda \to \infty} \sqrt{\lambda}(\tau_\omega - \tau_\omega^\lambda) < 0 \).

We first propose a feasible solution, and then compare it with any optimal solution that satisfies any one of the above two cases, to get a contradiction to the optimality of an optimal solution.

A feasible solution. This solution is constructed as follows. Suppose an announcement is made exactly at time \( \tau_\omega \). Then by the definition of \( \tau_\omega, \omega_{\tau_\omega} = \tau_\omega \). Hence, the first constraint on the fraction of abandonment in (53) holds. Thus we only consider the second constraint on waiting-time in (53), which further becomes \( \mathbb{P}(W^\lambda(\infty) > \tau_\omega) \leq \alpha_2 \). Let \( \tilde{s}_\omega^\lambda \) be its optimal solution. (We append the superscript \( \lambda \) to emphasize the dependency on the arrival rate.) It follows from the discussion in problem (46) (see (48) and the definition of C given by (40)) that
\[
\tilde{s}_\omega^\lambda = \frac{\lambda}{\mu} H_c(\tau_\omega | \tau_\omega) - \sqrt{\lambda} \frac{\beta}{\mu} + o(\sqrt{\lambda}),
\] (EC.12)
where \( \beta_\omega \) solves (55). Obviously, \((\tilde{s}_\omega^\lambda, \tau_\omega)\) is a feasible solution to our original problem (53). (Indeed it is the staffing level (54).)

Case (i). There is a subsequence along which the limit will be positive. To simplify the notation, we still use \( \lambda \) to index the subsequence, i.e. \( \lim_{\lambda \to \infty} \sqrt{\lambda}(\tau_\omega - \tau_\omega^\lambda) > 0 \). Note that in this case \( \omega_{\tau_\omega^\lambda} > \tau_\omega^\lambda \) (because \( H(\tau_\omega^\lambda | \tau_\omega^\lambda) < \alpha_1 \)), then the constraint on waiting-time in (53) becomes \( \mathbb{P}(W(\infty) > \tau_\omega^\lambda) \).

Similar to the relationship between (46) and (47) (noticing that the constraint \( \mathbb{P}(Ab) \) can be achieved from the first order), the optimal number of servers is
\[
s_\omega^\lambda = \frac{\lambda}{\mu} H_c(\tau_\omega^\lambda | \tau_\omega^\lambda) - \beta_\omega / \mu \sqrt{\lambda} + o(\sqrt{\lambda}),
\] (EC.13)
where \( \beta_\omega \) solves
\[
\max_{\beta} \beta \quad \text{s.t.} \quad H_c(\tau_\omega^\lambda | \tau_\omega^\lambda) \cdot \int_{-\infty}^\infty \exp\left( -\frac{\rho^2}{2\sigma^2} \int_0^y [f_{\tau_\omega^\lambda}(x) - \beta]\,dx \right)\,dy \leq \alpha_2
\] (EC.14)
with \( f_{\tau_n^*}(x) = \sqrt{N}[H(\tau_n^* + \frac{x}{\lambda}) - H(\tau_n^*)] \).

We now try to compare \( s_n^\lambda \) and \( \tilde{s}_n^\lambda \) (see (EC.12)). From our assumption on the patience-time distributions, it is easy to see \( f_{\tau_n^*} \approx f_{\tau_n^*} \). Hence \( \beta_n \approx \tilde{\beta}_n \). While since \( \tau_n^\lambda \) is strictly less than \( \tau_n \), \( H_c(\tau_n^\lambda | \tau_n) > H_c(\tau_n | \tau_n) \). Moreover, since \( \lim_{\lambda \to \infty} \sqrt{N}(\tau_n - \tau_n^\lambda) > 0 \), \( \lim_{\lambda \to \infty} \sqrt{N}[H_c(\tau_n^\lambda | \tau_n^\lambda) - H_c(\tau_n | \tau_n)] > 0 \). It then follows from (EC.12) and (EC.13) that \( \lim_{\lambda \to \infty} (s_n^\lambda - \tilde{s}_n^\lambda)/\sqrt{N} > 0 \), which is a contradiction with the optimality of \( s_n^\lambda \).

Case (ii) Similar to case (i) we assume \( \lim_{\lambda \to \infty} \sqrt{N}(\tau_n - \tau_n^\lambda) < 0 \). Now \( \omega_n = \tau_n \) as \( \tau_n^\lambda \) is larger than \( \tau_n \). Note that the constraint on waiting-time in (53) becomes \( \mathbb{P}(W^\lambda(\infty) > \omega_n) \). So the optimal number of servers is

\[
s_n^\lambda = \frac{\lambda}{\mu} H_c(\omega_n^\lambda | \tau_n^\lambda) - \frac{\tilde{\beta}_n}{\mu} \sqrt{N} + o(\sqrt{N}), \tag{EC.15}\]

where \( \tilde{\beta}_n \) is the optimal solution to

\[
\max \quad \beta \\
\text{s.t.} \quad H_c(\omega_n^\lambda | \tau_n^\lambda) \times \frac{\int_{0}^{\infty} \exp \left( -\frac{2\rho}{\lambda^2} \int_{0}^{y} [f_{\omega_n^\lambda, \tau_n^\lambda}(x) - \beta] dx \right) dy}{\int_{-\infty}^{\infty} \exp \left( -\frac{2\rho}{\lambda^2} \int_{0}^{y} [f_{\omega_n^\lambda, \tau_n^\lambda}(x) - \beta] dx \right) dy} \leq \alpha_2, \tag{EC.16}\]

with \( f_{\omega_n^\lambda, \tau_n^\lambda}(x) = \sqrt{N}[H(\omega_n^\lambda + \frac{x}{\lambda}) - H(\omega_n^\lambda)] \). Note that \( H_c(\omega_n^\lambda | \omega_n^\lambda) = H_c(\omega_n^\lambda | \tau_n^\lambda) \), so the difference between \( \tilde{s}_n^\lambda \) (see (EC.12)) and \( s_n^\lambda \) lies in the difference between \( \beta_n \) and \( \tilde{\beta}_n \). As \( \int_{0}^{t} f_{\omega_n^\lambda, \tau_n^\lambda}(x) dx < \int_{0}^{t} f_{\tau_n}(x) dx \), we have \( \lim_{\lambda \to \infty} (\beta_n - \tilde{\beta}_n) > 0 \). As a result, \( \lim_{\lambda \to \infty} (s_n^\lambda - \tilde{s}_n^\lambda)/\sqrt{N} > 0 \), which is again a contradiction with the optimality of \( s_n^\lambda \).

In summary, Cases (i) and (ii) do not hold. Thus, we have \( \lim_{\lambda \to \infty} \sqrt{N}(\tau_n - \tau_n^\lambda) = 0 \). Also from this proof, we see that (54) is the minimal number of servers. Hence, the proof of the proposition is complete.

\[\square\]

**EC.2. Several Auxiliary Lemmas**

In the following, we establish three technical lemmas which support the proofs of Theorems 1 and 2.

**Lemma EC.1.** Under the same assumptions as Theorem 1 as \( n \to \infty \)

\[
\tilde{B}^n \Rightarrow \frac{1}{\sqrt{p}} B, \tag{EC.17}\]

where \( B = \{B(t) : t \geq 0\} \) is a standard Brownian motion.

**Proof.** In view of (EC.6), we first look at the departure process \( \{D_n(t) : t \geq 0\} \). We introduce the following two diffusion scalings:

\[
\tilde{D}^n(t) = \frac{D^n(t) - s_n \mu t}{\sqrt{\lambda_n}}, \quad \tilde{S}^n(t) = \frac{S^n(t) - s_n}{\sqrt{\lambda_n}}.
\]
Then
\[ \tilde{B}^n(t) = \tilde{D}^n(t) + \tilde{S}^n(t) - \tilde{S}^n(0). \] (EC.18)

Let \( X^n(t) \) denote the total number of customers at time \( t \) in the \( n \)th system. Then the departure process \( D^n(t) \) can be represented as \( S(\int_0^t X^n(x) \land s_n) \, dx) \), where \( \{S(t) : t \geq 0\} \) is a Poisson process with rate \( \mu \). By (11),
\[ \sup_{0 \leq t \leq T} \left( s_n - X^n(t) \right) \Rightarrow 0. \] (EC.19)

This, together with (2), implies \( \tilde{D}^n \Rightarrow 1 / \sqrt{\rho} \tilde{B} \). (EC.20)

By the initial condition (10) and (EC.19), we have that the last two terms in (EC.18) will converge to zero. Hence, (EC.17) directly follows from (EC.20). \( \square \)

**Lemma EC.2.** Under the same assumptions as Theorem 1, the sequence of stochastic processes \( \{\tilde{V}^n\}_{n \in \mathbb{N}} \) is stochastically bounded.

**Proof.** It suffices to show that, for any \( T > 0 \) and \( \varepsilon > 0 \), the following holds for all large enough \( n \) and \( M \):
\[ \Pr \left\{ \sup_{0 \leq t \leq T} |\tilde{V}^n(t)| \geq M \right\} \leq 4\varepsilon. \]

To this end, define
\[ \varsigma_1^n = \inf \{t \geq 0 : \tilde{V}^n(t) \geq M\}, \quad \varsigma_2^n = \inf \{t \geq 0 : \tilde{V}^n(t) \leq -M\}, \]
\[ \Omega_1^n(M, T) = \{\varsigma_1^n \leq \varsigma_2^n, \varsigma_1^n \leq T\}, \quad \Omega_2^n(M, T) = \{\varsigma_1^n > \varsigma_2^n, \varsigma_2^n \leq T\}. \]

Hence we only need to show that, for all large enough \( n \) and \( M \),
\[ \Pr (\Omega_1^n(M, T)) \leq 2\varepsilon \quad \text{and} \quad \Pr (\Omega_2^n(M, T)) \leq 2\varepsilon. \] (EC.21)

We will first consider the event \( \Omega_1^n(M, T) \). By the definition of \( \varsigma_1^n \), we must have that \( \tilde{V}^n(\varsigma_1^n) \geq \tilde{V}^n(\varsigma_2^n -) \). In other words, if \( \tilde{V}^n \) has a jump at \( \varsigma_1^n \), then it must be an upward jump. Since \( \tilde{V}^n(t) \in [-M, M] \) on \([0, \varsigma_1^n]\), for any \( t \in (0, \varsigma_1^n] \) and small positive \( \delta \in (0, t) \), by (18),
\[ \tilde{V}^n(t) - \tilde{V}^n(t - \delta) = -\frac{\lambda_n}{s_n M} \int_{t-\delta}^t \sqrt{\lambda_n} \left( F^n(\omega^n + \frac{\tilde{V}^n(x)}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right) d\Lambda^n(x) \]
\[ + \tilde{Y}^n(\varsigma_1^n) - \tilde{Y}^n(\varsigma_1^n - \delta). \] (EC.22)

Since \( \tilde{V}^n(0) \) is stochastically bounded, we can choose \( M \) large enough such that
\[ \Pr (\Omega_0^n(M)) = \Pr \left( \tilde{V}^n(0) \leq \frac{M}{4} \right) \geq 1 - \varepsilon, \]
We know that on the event $\Omega$, $\tilde{V}^n(\xi^n_1) \geq M$, and $\tilde{V}^n(\xi^n_{12}) \leq \frac{M}{2}$. (EC.23)

Note that the process $\tilde{V}^n(\cdot)$ is larger than $M/2$ (thus larger than 0) on the interval $[\xi^n_{12}, \xi^n_1]$. By (EC.22) and the fact that $F^n(\omega^n + x) \geq F^n(\omega^n)$ for any $x \geq 0$,

$$\tilde{V}^n(\xi^n_1) - \tilde{V}^n(\xi^n_{12}) \leq \tilde{V}^n(\xi^n_1) - \tilde{V}^n(\xi^n_{12}).$$

By (EC.23) and (EC.24),

$$\mathbb{P}\left( \Omega^n_0(M) \cap \Omega^n_1(M,T) \right) \leq \mathbb{P}\left( \sup_{t \in [0,T]} |\tilde{V}^n(t)| \geq \frac{M}{4} \right).$$

We now prove the stochastic boundness of $\tilde{V}^n$. Recall the definition of $\tilde{Y}^n$ in (19). The first and the third term on the right side of (19) is stochastically bounded by Conditions (1), (2) and (4). The last two terms are stochastically bounded by Lemma EC.1. It now remains to show the stochastic boundness of the third term, which can be written as $\frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\lambda_n} (1_{\{u^n_i \leq \omega^n_i\}} - F^n(\omega^n))$. According to Condition (1), it is enough to show the stochastic boundness of $\frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\lambda_n} (1_{\{u^n_i \leq \omega^n_i\}} - F^n(\omega^n))$.

From Doob’s inequality or Kolmogorov’s inequality for martingale, for any $M \geq 0$,

$$\mathbb{P}\left( \sup_{0 \leq t \leq T} \left| \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\lambda_n} (1_{\{u^n_i \leq \omega^n_i\}} - F^n(\omega^n)) \right| \geq M \right) \leq \frac{1}{M^2} \mathbb{E}\left[ \left| \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\lambda_n} (1_{\{u^n_i \leq \omega^n_i\}} - F^n(\omega^n)) \right|^2 \right] = \frac{1}{M^2} \frac{\lambda_n}{\sqrt{\lambda_n}} \sum_{i=1}^{\lambda_n} \mathbb{E}(1_{\{u^n_i \leq \omega^n_i\}} - F^n(\omega^n))^2 \leq M^{-2} T.$$

Using the stochastic boundedness of $\{\tilde{Y}^n, n \geq 1\}$, we can choose $M$ large enough such that the probability on the right-hand side of (EC.25) is less than $\varepsilon$. So we have that $\mathbb{P}(\Omega^n_1(M,T)) \leq 2\varepsilon$ for large enough $M$. A symmetric argument shows that $\mathbb{P}(\Omega^n_2(M,T)) \leq 2\varepsilon$ for large enough $M$. So we have proved stochastic boundedness.

**Lemma EC.3.** Under the same assumptions as Theorem 1 as $n \to \infty$

$$\int_0^t \int_0^{\tilde{Y}^n(y)} d\tilde{H}^n(\tilde{\Lambda}^n(y),x) \Rightarrow (1/\rho)\sqrt{\rho - 1} \mathcal{B}_A(\cdot).$$

(EC.26)

Here $\mathcal{B}_A = \{\mathcal{B}(t) : t \geq 0\}$ is a standard Brownian motion which is independent of $\{\mathcal{B}(t) : t \geq 0\}$.

**Proof.** Note that $\int_0^t \int_0^{\tilde{Y}^n(y)} d\tilde{H}^n(\tilde{\Lambda}^n(y),x) = \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\lambda_n} (1_{\{u^n_i \leq \omega^n_i\}} - F^n(\omega^n_i))$. We first prove a convergence result for the sequence of processes given by $\{\frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\lambda_n} (1_{\{u^n_i \leq \omega^n_i\}} - F^n(\omega^n_i)) : t \geq 0\}$,
which is the same as $\tilde{H}^n(\cdot)$ in Section 4.2 of [Dai and He (2010)] with $h = 1$. Their Lemmas 4.1 and 4.2 still hold. Thus, we have a martingale whose quadratic variation is
\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} (1_{\{u_i^n \leq \omega_i^n\}} - F^n(\omega_i^n))^2.
\]
We calculate the following:
\[
\mathbb{E}\left[\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} \left((1_{\{u_i^n \leq \omega_i^n\}} - F^n(\omega_i^n))^2 - F^n(\omega_i^n)F_c^n(\omega_i^n)\right)\right]^2
\]
\[
= \mathbb{E}\left[\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} \left((1_{\{u_i^n \leq \omega_i^n\}} - F^n(\omega_i^n))(F_c^n(\omega_i^n) - F^n(\omega_i^n))\right)\right]^2
\]
\[
= \frac{1}{(\lambda_n)^2} \sum_{i=1}^{[\lambda_n t]} \mathbb{E} \left((1_{\{u_i^n \leq \omega_i^n\}} - F^n(\omega_i^n))(F_c^n(\omega_i^n) - F^n(\omega_i^n))\right)^2 \to 0.
\]
(In the last equality, we use martingale to argue that the cross terms have expectation 0.) As a result,
\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} \left((1_{\{u_i^n \leq \omega_i^n\}} - F^n(\omega_i^n))^2 - F^n(\omega_i^n)F_c^n(\omega_i^n)\right) \Rightarrow 0.
\]
On the other hand,
\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} (F^n(\omega_i^n)F_c^n(\omega_i^n) - F^n(\omega^n)F_c^n(\omega^n))
\]
\[
= \frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} [F^n(\omega_i^n) - F^n(\omega^n)] F_c^n(\omega_i^n) + F^n(\omega^n) [F_c^n(\omega_i^n) - F^n(\omega_i^n)]
\]
\[
\leq \sup_{0 \leq s \leq t} 2|F^n(V^n(s)) - F^n(\omega^n)| t
\]
\[
= \frac{2}{\sqrt{\lambda_n}} \sup_{0 \leq s \leq t} \left| \sqrt{\lambda_n} \left( F^n(\omega^n) + \frac{1}{\sqrt{\lambda_n}} \tilde{V}^n(s) \right) - F^n(\omega^n) \right| t,
\]
which vanishes to 0 following Condition [3] and the stochastic boundedness of $\{\tilde{V}^n\}_{n \in \mathbb{N}}$. Note that [2] and [4] imply that as $n \to \infty$,
\[
F^n(\omega^n) \to 1 - \frac{1}{\rho},
\]
As a result,
\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} (1_{\{u_i^n \leq \omega_i^n\}} - F^n(\omega_i^n))^2 \Rightarrow \lim_{n \to \infty} F^n(\omega^n)F_c^n(\omega^n)t = \frac{\rho - 1}{\rho^2} t.
\]
Then from the martingale convergence theorem (Theorem 8.1 (ii) of [Pang et al. (2007)]), we know that the sequence of the processes given by $\{ \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{[\lambda_n t]} (1_{\{u_i^n \leq \omega_i^n\}} - F^n(\omega_i^n)) : t \geq 0 \}$ weakly converges to the process $\frac{\sqrt{\rho - 1}}{\rho} B_A$. The result of this lemma then follows from the random-time-change-theorem.
Lemma EC.4. Under the same assumptions as Theorem 1 the sequence of stochastic processes \( \{ \tilde{V}_n \}_{n \in \mathbb{N}} \) is tight.

Proof. In view of Lemma EC.2 it suffices to study the modulus of continuity for \( \{ \tilde{V}_n \}_{n \in \mathbb{N}} \). By (1), for any \( \epsilon > 0 \),

\[
\limsup_{\delta \to 0} \lim_{n \to \infty} \mathbb{P} \left( \sup_{s,t \in [0,T] \atop |s-t| < \delta} |\tilde{A}^n(s) - \tilde{A}^n(t)| > \epsilon \right) = 0. \tag{EC.27}
\]

By Conditions (1), (2) and (4), and Lemma EC.3, for any \( \epsilon > 0 \),

\[
\limsup_{\delta \to 0} \lim_{n \to \infty} \mathbb{P} \left( \sup_{s,t \in [0,T] \atop |s-t| < \delta} |\tilde{Y}^n(s) - \tilde{Y}^n(t)| > \epsilon \right) = 0. \tag{EC.28}
\]

Let \( \Omega_c^n(M,T) \) be the complement of \( \Omega_1^n(M,T) \cup \Omega_2^n(M,T) \), then

\[
\lim_{M \to \infty} \liminf_{n \to \infty} \mathbb{P}(\Omega_c^n(M,T)) = 1. \tag{EC.29}
\]

On the event \( \Omega_c^n(M,T) \), it follows from (EC.22) that

\[
|\tilde{V}^n(t) - \tilde{V}^n(t-\delta)| \leq C_M^n \cdot \left( |\tilde{A}^n(t) - \tilde{A}^n(t-\delta)| + |\tilde{Y}^n(t) - \tilde{Y}^n(t-\delta)| \right),
\]

where \( C_M^n = \max \left\{ \sqrt{\lambda_n(F^n(\omega^n) + \frac{M}{\sqrt{n}})} - F^n(\omega^n), \sqrt{\lambda_n(F^n(\omega^n) - F^n(\omega^n - \frac{M}{\sqrt{n}}))} \right\} \). By Condition (3), \( C_M^n \) is bounded by some finite number \( C_M \) which may depend on \( M \). So for any \( M > 0 \),

\[
\mathbb{P}\left( \sup_{s,t \in [0,T] \atop |s-t| < \delta} |\tilde{V}^n(s) - \tilde{V}^n(t)| > \epsilon \right) \leq (1 - \mathbb{P}(\Omega_c^n(M,T))) + \mathbb{P} \left( \sup_{s,t \in [0,T] \atop |s-t| < \delta} |\tilde{A}^n(s) - \tilde{A}^n(t)| > \frac{\epsilon}{2C_M^n} \right)
\]

\[
+ \mathbb{P}\left( \sup_{s,t \in [0,T] \atop |s-t| < \delta} |\tilde{Y}^n(s) - \tilde{Y}^n(t)| > \frac{\epsilon}{2} \right).
\]

By first choosing \( M \) and \( n \) large enough and then choosing \( \delta \) small enough, we can show that

\[
\limsup_{\delta \to 0} \lim_{n \to \infty} \mathbb{P}\left( \sup_{s,t \in [0,T] \atop |s-t| < \delta} |\tilde{V}^n(s) - \tilde{V}^n(t)| > \epsilon \right) = 0.
\]

This shows that the modulus of continuity for \( \{ \tilde{V}_n \}_{n \in \mathbb{N}} \) will vanish as \( n \to \infty \). Hence we have the lemma. \( \square \)