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, Shatin, Hong Kong, 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 has revealed a noticeable impact of delay-related information on phone-customers; for example and somewhat surprisingly, delay announcements can abruptly increase the likelihood to abandon (hang-up). Our starting point is that the latter phenomena can be used to 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). This requires one to go beyond existing theory, which turns out too crude for our needs (e.g., it requires a continuous hazard rate of impatience and hence cannot be applied). We thus develop a refined process and steady-state models, and use them to solve our minimization problem and more. The value and accuracy of our models are demonstrated via extensive numerical experiments.

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, Akşin et al. 2007). Often the goal is a balanced operation that is both Quality- and Efficiency-Driven (QED): customers do not wait too long for available servers and servers do not wait too long for needy customers; for large enough systems, this QED balance translates into waiting and idle times being negligible relative to service time. However, in practice, many call centers are merely Efficiency-Driven (ED) in that they are understaffed, which results in significant delays and consequent abandonment. One
such scenario is depicted in Figures 4 and 5: the first figure reveals extreme understaffing (e.g., 80 agents present at 11:30, while about twice as many are required for 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” (Maister 1985, p. 118). Just as importantly, such information helps customers rationalize 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: there exists empirical support to the fact that this impact is often abrupt or, formally, it is manifested through a discontinuity in the hazard rate of customer patience (underlying the smoothed peaks in Figure 3). Moreover, there is also the practical need and theoretical challenge to accommodate general distributions. It follows that, for tractability, one must resort to fluid or diffusion models with discontinuous primitives, which necessitates refinement of the existing 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). In this case, an estimated duration of delay is announced, which has the following consequences. Some customers choose to balk immediately while others remain online; these will not abandon if served before their patience expires, but they will become irritated once their waiting times reach 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) and Mandelbaum and Zeltyn (2013) have had discussions on such announcements, with the latter observing that, in fact, such announcements, in various call centers, have been found to cause an upward jump in the hazard rate, right after the announcement epoch. Taking a step further, Li et al. (2015) develops a statistical method to estimate the hazard rate as a smooth surface of waiting time and time-of-day. It is shown that peaks in hazard rates, caused by delay announcements, are consistent across different time-of-day.

Both types of announcements share the common feature of giving rise to a non-smooth change in the hazard rate of patience-time distribution. It occurs at a certain “impact point”, which is either the announced waiting time (first type) or the chosen time to make an announcement (second type). 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 possibly 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 a problem for future research, which is here resolved: our refined model offers an 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. (See optimization problem (38) in Section 4.1.1.) This same formulation, of constrained-optimization, is used in Mandelbaum and Zeltyn (2009). It turns out that announcements reduce the staffing level by a magnitude of $O(\sqrt{\lambda})$ (where $\lambda$ 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. (See optimization problem (44) in Section 4.2.1.) It turns out that it is optimal to make an announcement at a time that is approximately the fluid offered waiting time; also here, the announcement reduces staffing by a magnitude of $O(\sqrt{\lambda})$; see Proposition 3 in Section 4.2.

A Queueing Model with a Delay Announcement. Our refined model introduces a general scaling (see (4)) of the patience-time distribution, which precisely captures its fine structures, especially non-smooth changes caused by announcements. As explained in Section 2, hazard-rate scaling and no-scaling of the patience-time distribution are special cases of this general scaling. Our method for analyzing such a refined model is based on the virtual waiting time, which differs from the traditional approach that is based on queue length. The virtual waiting time $V^*(t)$ is roughly the time that an infinitely-patient “virtual” customer would have to wait if arriving at time $t$. The evolution of the virtual waiting time is characterized by (16), which enables us to develop diffusion approximations for systems with patience-time distribution scaling (4). The tractable stationary distribution of the diffusion limit is then used to approximate steady-state performance of its originating queueing systems. 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 distribution and found that a surge is associated with the time of announcement. Akşin-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 queueing 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. Jouini et al. (2011) explored the effect on balking and reneging, of announcing different percentiles of the waiting time distribution. Systems’ performance measures were calculated via an $M/M/s+M$ queue. Through a numerical study, Jouini et al. (2011) explored when informing customers about delays is beneficial and what the percentile should be in these announcements.

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

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

Approximations based on fluid models are surprisingly accurate in the ED regime. Bassamboo and Randhawa (2010) showed that the gap between the steady-state queue length and its fluid approximation is $O(1)$. This enabled the study of optimal capacity sizing for $M/M/n + GI$, based on its fluid approximation, 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 our 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-time distribution at 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:

- A new scaling framework (4) of the patience-time distributions is proposed, which allows the patience-time distribution to change abruptly, as observed in practice.
- Inspired by Liu and Whitt (2014b), we develop a new modeling approach ((9) and (16)–(17)) based on the virtual waiting time. This enables diffusion analysis for systems with scaling (4) of their patience-time distribution. Accurate approximations for performances of the original stochastic system are constructed based on the stationary distribution of the diffusion limit.
• Performance approximations for the fixed-delay model of Armony et al. (2009) are improved by using our approximations. Indeed, quoting the authors of the latter (end of their Section 7), we “better quantify the impact of stochastic fluctuations,” which they left as a problem for future research.

• We jointly optimize (asymptotically) the two problems of optimal-staffing and announcement-timing, while accommodating the empirical phenomenon of the abrupt impact of an announcement.

• Finally, we analyze the performance gap between fluid and diffusion approximations.

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, in Section 3, to 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 5 provides an in-depth discussion of our approximation formulae, and Section 6 offers some concluding remarks. The proofs for theorems and propositions, as well as some complements, are given in the Appendix.

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

2. Model Formulation

Consider a sequence of many-server queueing systems with customer abandonment, indexed by \(n \in \mathbb{N}\). In the \(n\)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 \(\Lambda^n = \{\Lambda^n(t) : t \geq 0\}\). 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. In our model, there are no batch arrivals, namely, $P(\tau_i^n = \tau_{i+1}^n) = 0$, for all $i \in \mathbb{N}$. Assume that there exists a sequence of positive real numbers $\{\lambda_n\}_{n \in \mathbb{N}}$ such that, as $n \to \infty$,

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

(1)

where $\tilde{\Lambda} = \{\tilde{\Lambda}(t) : t \geq 0\}$ is a Brownian motion. Arriving customers are immediately served if any server is idle. Otherwise, they wait in a queue and are served on 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 reaches $u^n_i$, the customer leaves the system immediately without receiving service. Service times are assumed i.i.d., and exponentially distributed with rate $\mu$. Patience times are i.i.d., with a general distribution $F^n(\cdot)$. We also assume that the service times of the initial customers in queue and the remaining service times of the initial customers in service are i.i.d., following the exponential distribution with rate $\mu$. The sequences of service and patience times and the arrival process are mutually independent.

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

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

(2)

$$\frac{\lambda_n F^n_c(\omega^n) - s_n \mu}{\sqrt{\lambda_n}} \to \beta;$$

(3)

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

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

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

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

(4)

where $f_\omega(\cdot)$ is a continuous function. This setting is quite general. A simple special case is where $F^n(x) \equiv F(x)$ and $\omega^n = \omega$ (i.e., without any scaling of the patience-time distribution).
In this case \( f_\omega(x) = f(\omega)x \), where \( f(\cdot) \) is the density function of \( F(\cdot) \). However, the flexibility (4) of allowing the patience-time distribution to vary with \( n \) captures the fine structure that arises from customers’ reaction to delay announcements, as discussed in the introduction (also see Figure 3). This enables us to analyze the impact of such announcements in Section 4, and consequently optimize staffing levels and announcement times, jointly.

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

We now formally model system dynamics. To this end, 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. When \( t \) happens to be the arrival time of an actual customer, the virtual waiting time at \( t \) is regarded as if a virtual customer arrives right after the actual customer. Since there are no batch arrivals, \[
\omega^n_{i} = V^n(\tau^n_{i} - ).
\]

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
\[
A^n(t) = \sum_{i=1}^{\Lambda^n(t)} 1\{u^n_{i} \leq \omega^n_{i}\}. \tag{5}
\]

Any customer who arrives after time 0 cannot receive service before \( V^n(0) \) due to FCFS. For \( t \geq V^n(0) \), let
\[
\kappa^n(t) = \inf\{\tau : \tau + V^n(\tau) > t\}. \tag{6}
\]

All arrivals before \( \kappa^n(t) \) and initial customers are not in queue at time \( t \). Denote by \( Q^n(t) \) the number of customers in the queue at time \( t \). The queue length process, for time \( t \geq V^n(0) \), can be written as
\[
Q^n(t) = \sum_{i=1}^{\Lambda^n(\kappa^n(t))} 1\{u^n_{i} - (t - \tau^n_{i}) > 0\}. \tag{7}
\]

Note that the queue length representation (7) is similar to (6.5) in Liu and Whitt (2014b), and \( \kappa^n(t) \) coincides with their \( t - h^n(t) \) \((h^n(t) \) is the age of the head-of-the-line customer). Let \( B^n(t) \) denote the number of customers who start service during \((0, t] \). Then for \( t \geq V^n(0) \),
\[
B^n(t) - B^n(V^n(0) - ) = \sum_{i=1}^{\Lambda^n(\kappa^n(t))} 1\{u^n_{i} > \omega^n_{i}\}. \tag{8}
\]
Consider the $i$th customer who arrived during $(0,t]$, for some fixed $i$: if $u_i^n \leq \omega_i^n$ then this customer is counted in $A^n(t)$; otherwise, the customer starts service between time $V^n(0)$ and $t + V^n(t)$. There is therefore a simple balance equation regarding the arrival process $\Lambda^n(t)$:

$$\Lambda^n(t) = A^n(t) + B^n(t+V^n(t)) - B^n(V^n(0)-), \quad t \geq 0. \quad (9)$$

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

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

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

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

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

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

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

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

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

$$\widetilde{B}^n(t) = \frac{B^n(t) - s_n \mu t}{\sqrt{\lambda_n}}. \quad (13)$$
From (5), we have
\[
\tilde{\Lambda}^n(t) = \tilde{H}^n(t) + \int_0^t \sqrt{\Lambda_n} \left( F^n(\omega^n + \frac{\tilde{V}^n(x^-)}{\sqrt{\Lambda_n}}) - F^n(\omega^n) \right) d\tilde{\Lambda}^n(x) + F^n(\omega^n)\tilde{\Lambda}^n(t),
\]
where \( \tilde{\Lambda}^n(s) = \Lambda^n(s)/\lambda_n \), which is the fluid scaling of \( \Lambda^n(s) \), and
\[
\tilde{H}^n(t) = \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\Lambda^n(t)} \left( 1_{(u_i^n \leq \omega^n_i)} - F^n(\omega^n_i) \right).
\]

Applying the diffusion scaling to each term in (9), it follows from \( t = tF^n(\omega^n) + tF^n(\omega^n) \) that
\[
\tilde{\Lambda}^n(t) = \tilde{\Lambda}^0(t) + \frac{(s_n \mu - \lambda_n F^n(\omega^n))t}{\sqrt{\lambda_n}} + \tilde{B}^n(t + V^n(t)) - \tilde{B}^n(V^n(0)) + \frac{s_n \mu}{\lambda_n} \left( \tilde{V}^n(t) - \tilde{V}^n(0) \right).
\]

This, with (14), helps us write \( \tilde{V}^n \) as
\[
\tilde{V}^n(t) = \tilde{V}^n(0) - \frac{\lambda_n}{s_n \mu} \int_0^t \sqrt{\lambda_n} \left( F^n(\omega^n + \frac{\tilde{V}^n(x^-)}{\sqrt{\Lambda_n}}) - F^n(\omega^n) \right) d\tilde{\Lambda}^n(x) + \tilde{Y}^n(t),
\]
where
\[
\tilde{Y}^n(t) = \frac{\lambda_n}{s_n \mu} \left[ F^n(\omega^n)\tilde{\Lambda}^n(t) - \tilde{H}^n(t) + \frac{(\lambda_n F^n(\omega^n) - s_n \mu)t}{\sqrt{\lambda_n}} - \tilde{B}^n(t + V^n(t)) + \tilde{B}^n(V^n(0)) \right].
\]

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

We assume that the sequence \( \{\omega^n\}_{n \in \mathbb{N}} \) in (3)–(4) satisfies
\[
\tilde{V}^n(0) \Rightarrow V_0 \quad \text{as } n \to \infty.
\]
Note that this condition is different from customary, and is enough for our analysis due to the same reason mentioned right after Proposition 1. In Appendix EC.5, we provide a verifiable sufficient condition for (18).

**Theorem 1.** In the heavy traffic regime (1)–(4) with the initial condition (18), \( \tilde{V}^n \Rightarrow \tilde{V} \), as \( n \to \infty \), where the limit \( \tilde{V} = \{\tilde{V}(t) : t \geq 0\} \) is the unique solution to
\[
\tilde{V}(t) = \tilde{V}_0 - \rho \int_0^t \left[ f_\omega(\tilde{V}(x)) - \beta \right] dx + \left[ \tilde{\Lambda}(t) - \sqrt{\rho}\mathcal{B}(t) - \sqrt{\rho - \overline{1}} \mathcal{B}_A(t) \right], \quad t \geq 0;
\]
here \( \mathcal{B}_A = \{\mathcal{B}_A(t) : t \geq 0\} \) and \( \mathcal{B} = \{\mathcal{B}(t) : t \geq 0\} \) are two independent standard Brownian motions, which are further independent of \( \tilde{\Lambda} \).

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

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

\[
\tilde{Q}^n(t) = \frac{1}{\sqrt{\lambda_n}} \left( Q^n(t) - \lambda_n \int_{0}^{\omega} F_c^n(x)dx \right), \quad t \geq 0. \tag{20}
\]

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

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

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

where \( F(\cdot) \) is continuous at \( \omega \), and \( F_c(\omega) = 1/\rho \). 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

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

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

\[
\mathbb{E} \left( \mathcal{G}(t_1) \mathcal{G}(t_2) \right) = \int_{t_1 \wedge (t_2 - \omega)}^{t_1} F(t_1 - x) F_c(t_2 - x)dx, \quad \omega < t_1 \leq t_2. \tag{23}
\]

Note that Condition (21) was not needed for Theorem 1, because only those customers who will eventually receive service affect the virtual waiting time. In fact, Theorem 1 required only some “local” property (condition (3)) of the distribution \( F^n \) around \( \omega^n \). However, those customers who will eventually abandon affect the queue length since they will stay in queue until their patience times expire. That is, the queue length process depends on customer patience times. Thus, condition (21) on the “global” property of the distributions \( F^n \) is needed for Theorem 2. Since we impose no refined assumption on the initial state of the system, one can characterize the asymptotic queue process only after the warm-up period \([0, \omega]\). We note that, in analogy to Krichagina and Puhalskii (1997), the Gaussian process \( \mathcal{G} \) can be written as an integral with respect to a Kiefer process; see page 237 of Krichagina and Puhalskii (1997) for details.
3.2. Steady-State Analysis

For the purpose of steady-state analysis, we add the assumption that the arrival process $\Lambda$ is in fact a renewal process; let the inter-arrival time have mean $1/\lambda$ and variance $\theta^2/\lambda^2$. 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

$$\sigma^2 = \theta^2 + 2\rho - 1;$$

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

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

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

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

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

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

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

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

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

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

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

We have established limits for a sequence of systems. These limits will now support an approximation for a single given system, specifically closed-form formulae for its steady-state waiting time and queue length. To this end, and as is often the case (e.g., Reed and Ward (2008) and Reed and Tezcan (2012)), we presume the validity of a limit-interchange, which justifies the steady-state approximation of a queuing 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 \((s, \mu, \lambda, \theta^2, H)\). The heavy traffic assumptions (3) and (4) constitute a mathematical tool that guides on how to capture the structure of the patience-time distribution \( H(\cdot) \) around \( \omega \), 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 (27), set

\[
\rho := \frac{\lambda}{s \mu}, \quad \sigma^2 = \theta^2 + 2\rho - 1, \quad (28)
\]

\[
\beta := \frac{\lambda H_c(\omega) - s \mu}{\sqrt{\lambda}}, \quad (29)
\]

\[
f_\omega(x) := \sqrt{\lambda}\left[H(\omega + \frac{x}{\sqrt{\lambda}}) - H(\omega)\right]. \quad (30)
\]

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

\[
\pi(y) = C \exp\left(-\frac{2\rho}{\sigma^2} \int_0^y [f_\omega(x) - \beta] \, dx\right), \quad (31)
\]

where

\[
C = \left(\int_{-\infty}^{\infty} \exp\left(-\frac{2\rho}{\sigma^2} \int_0^y [f_\omega(x) - \beta] \, dx\right) \, dy\right)^{-1} \quad (32)
\]

is the normalizing constant.

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

\[
\mathbb{P}(W(\infty) > y) = H_c(y)\mathbb{P}\left(\sqrt{\lambda}(V(\infty) - \omega) > \sqrt{\lambda}(y - \omega)\right)
\]

\[
\approx H_c(y) \int_{\sqrt{\lambda}(y - \omega)}^{\infty} C \exp\left(-\frac{2\rho}{\sigma^2} \int_0^v [f_\omega(x) - \beta] \, dx\right) \, dv. \quad (33)
\]

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 (31) and the above formula. For example, the expected steady-state queue length can be calculated as

\[
\mathbb{E}Q(\infty) \approx \lambda \int_0^\omega H_c(x)dx + \frac{1}{\rho} \sqrt{\lambda} \int_{-\infty}^{\infty} x \pi(dx).
\] (34)

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

4. Impact of Delay Announcements

The advantage of our refined approximation is the ability to capture the fine structure of the patience-time distribution. This is necessary for our applications where delay announcements cause a sudden change of the patience-time distribution at a certain “impact point”. As in Mandelbaum and Zeltyn (2009), we now apply our refined approximation to solve two staffing problems for a call center in the ED+QED regime. The first one, arising from the application of a delay announcement upon arrival, is to minimize the staffing level to achieve a service level constraint (see (38)). The second, arising from the application of an announcement during waiting, has the additional option of an announcement epoch, while sharing the same objective of minimizing staffing (see (44)). The patience-time distributions in both applications are not smooth enough to use existing results.

4.1. Delay Announcement Upon Arrival

As described in the Introduction, arriving customers who must wait often receive an announcement upon arrival concerning their anticipated delay. The announced information could include a single number \(\tau\) related to the delay, which is called a fixed delay announcement by 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 times are also affected by the announced information \(\tau\). This effect is modeled by assuming that customers’ patience times follow the conditional distribution \(H(t|\tau)\).

An announced delay \(\omega_e\) is an equilibrium delay if either (i) \(\rho B_e(0) \leq 1\) and \(\omega_e = 0\), or (ii) \(\rho B_e(0) > 1\) and

\[
\rho B_e(\omega_e)H_e(\omega_e|\omega_e) \leq 1 \quad \text{and} \quad \rho B_e(\omega_e)H_e(t|\omega_e) > 1, \quad \text{for } 0 \leq t < \omega_e.
\] (35)
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 $\omega_e$. Our model setting, and specifically the concept of equilibrium delay in the fluid scale, were introduced by Armony et al. (2009). They also proved the uniqueness of equilibrium delay under some regularity conditions (Theorem 4.1 in Armony et al. (2009)), which are satisfied in the following model.

In this subsection, we consider “the all-exponential conditional hazard-rate model” proposed in Armony et al. (2009), where customer arrivals are assumed to be a Poisson process with rate $\lambda$, and their 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

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

(36)

here $h_0$ and $h_1$ are two parameters, which capture abandonment behavior before and after the announcement $\tau$. In order to compare the fluid approximation developed by Armony et al. (2009) with the diffusion approximation given by Proposition 2, $h_0$ and $h_1$ are assumed to be constants independent of $\tau$. It follows from (35) that the unique equilibrium delay is

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

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

Though balking is not formally incorporated in our diffusion analysis, our approximation can easily accommodate it by regarding balking customers as having zero-patience, which is allowed by our assumptions. This gives rise to the patience-time distribution $\hat{H}(\cdot|\omega_e) = B(\omega_e) + B_c(\omega_e)H(\cdot|\omega_e)$. 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 (37), the equilibrium fluid delay is $\omega_e = 0.224$. As pointed out in Armony et al. (2009), their fluid approximation is not nearly close to the simulation when $h_1 = 4 > h_0 = 0.5$, though it agrees closely when $h_1 = h_0 = 0.5$. In Table 1 it is seen that our diffusion analysis yields much improved approximations. The columns “Simulated” and “Fluid” are taken from Table 1 in Armony et al. (2009). The columns “Diffusion” are calculated using formula (31). In particular, $E[Q(\infty)]$ is calculated based on
(34), and $E[W(\infty); B_c]$ based on (33). The calculation of $P(W(\infty) \leq \omega | S)$, the conditional probability that the steady-state waiting time is less than $\omega$, follows from

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

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

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Fluid</th>
<th>$h_1 = 0.5$</th>
<th>Simulated</th>
<th>Diffusion</th>
<th>$h_1 = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E[Q(\infty)]$</td>
<td>23.7</td>
<td>24.3</td>
<td>23.7</td>
<td>17.3</td>
<td>16.4</td>
</tr>
<tr>
<td>$E(W(\infty); B_c)$</td>
<td>.212</td>
<td>.217</td>
<td>.224</td>
<td>.155</td>
<td>.151</td>
</tr>
<tr>
<td>$P(W(\infty) \leq \omega</td>
<td>S)$</td>
<td>1</td>
<td>.512</td>
<td>.512</td>
<td>.764</td>
</tr>
</tbody>
</table>

Table 1  Comparing fluid and diffusion approximations with simulated performance metrics.

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 be restricted 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 $\theta^2/\lambda^2$ and the patience-time distribution $H(\cdot)$ be given. We seek the number of servers $s$ such that staffing cost is minimized while adhering to the given service level agreement $(z, \alpha)$ as follows:

$$\min_s \quad s$$

$$\text{s.t.} \quad P(W(\infty) > z) \leq \alpha.$$  \hspace{1cm} (38)

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

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

$$s = \left\lceil \frac{\lambda}{\mu} H_c(\omega) - \frac{\beta}{\mu} \sqrt{\lambda} \right\rceil.$$  \hspace{1cm} (39)
Note that $\beta$ is the only element needed to determine the number of servers $s$. We calculate it via (33), by solving the optimization problem (38). This gives rise to

$$\max \beta \quad \text{s.t.} \quad H_c(\omega) \int_0^\infty C \exp \left(-\frac{2\rho}{\sigma^2} \int_0^y [f_\omega(x) - \beta] \, dx \right) \, dy \leq \alpha.$$  \hspace{1cm} (40)

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

**Example 1.** Customers arrive according to a Poisson process with rate $\lambda$. Service times are exponentially distributed with rate $\mu$. The patience-time distribution is

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

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

![Staffing level and probability that waiting time exceeds $\omega$.](image)

In this example, set the parameter $\omega = 1/6$ and $\kappa = 5$ in (41). Assume that the individual service rate $\mu = 1$ and the arrival rate $\lambda = 120$. It is clear that $H(\cdot)$ is not differentiable at the fluid offered waiting time $\omega$. Thus, we cannot use existing results such as Mandelbaum and Zeltyn (2009). Applying the staffing rule (39) by numerically solving (40) with $\alpha = 0.3$, we get that the optimal number of servers is 97. To show the accuracy of our approach, we plot the tail probability vs. the number of servers in Figure 1. The vertical axis is the probability that waiting time exceeds 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 accurate: the curve obtained almost overlaps the curve by simulation. In fact, the optimal number of servers required in order to achieve $P(W(\infty) > \omega) \leq \alpha$ 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 distribution is as follows

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

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

In this example, set the parameter \( \omega = \log 1.2 / h_0 \), \( h_0 = 1.0 \) and \( \kappa = 100 \) in (42). Assume that the individual service rate \( \mu = 1 \) and the arrival rate \( \lambda = 120 \). In Figure 2, the number of servers ranges from 91 to 110. The vertical axis is the probability that waiting time exceeds the fluid offered waiting time. Again, we apply the staffing rule (39) by numerically solving (40) with \( \alpha = 0.4 \), to show that the optimal number of servers is 96. To demonstrate the accuracy, and also to compare against the staffing rule given by Mandelbaum and Zeltyn (2009), we plot the tail probability 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 \( \alpha = 0.4 \), gives rise to the optimal number of servers 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.

Remark 1. To apply the staffing method in Section 4.1.1 to the delay announcement model in Section 4.1, one simply replaces \( H_c(\omega) \) by \( \tilde{H}(\omega_c|\omega_e) \) in (40), where \( \omega_e = \tau \) is the equilibrium delay. It is furthermore natural to choose \( z = \tau \), since then the constraint in (38) bounds the fraction of “broken promises”. For the all-exponential model, the optimal staffing level
is determined by (39), with $\beta$ replaced by $\beta^*$—the optimal solution to (40). 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, if properly designed under the appropriate circumstances, an announcement can reduce staffing without hurting service-levels.

4.2. Delay Announcement During Waiting – Asymptotic Optimality

Instead of making announcements upon arrival, one can make an announcement while customers are waiting. Such announcements have been shown 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 a patience-time distribution in an Israeli call center. The 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 unwilling-to-wait customers.)

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).
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 (44) in the sequel).

A Queueing Model. Based on the above-mentioned empirical support 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 distribution is

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

here $h_0$ and $h_1$ are given parameters with $h_0 < h_1$, which captures the increase of hazard rate caused by the announcement made at $\tau$. In this regard, it is not necessary to specify the exact content of the announcement as long as the hazard rate has a jump at $\tau$. We comment here that the form (43) 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 (43) is also relevant and insightful, in that it captures the essence of the data in Figure 3, in which an announcement abruptly and
temporarily increases the hazard rate. Not only is \( h_0 < h_1 \) supported by empirical data, it is also quite intuitive. If \( h_0 = h_1 \), the announcement does not make any difference, thus no need to study it. Suppose \( h_0 > h_1 \), meaning that the announcement “encourages” customers to remain online. Essentially, such encouragement would add a load to the system. If we maintain the same staffing, then service quality worsens, as measured by the probability of waiting time exceeding a certain threshold. To maintain the same service quality, one must staff more servers and, hence, it is better to have no announcement. Finally, as suggested by Mandelbaum and Zeltyn (2013): no matter the purpose of an announcement, the ultimate result is conceivably an encouragement for customers to abandon the system.

4.2.1. An Optimization Model We now investigate whether the management of an overloaded call center can benefit from a delay announcement, with the intention of minimizing 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:

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

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

**Proposition 3.** For the stylized model (43) and (44), let \( (s^\lambda_*, \tau^\lambda_*) \) be an optimal solution to (44). Denote by \( \tau_* \) 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
\[ s^* = \frac{\lambda}{\mu} (1 - \alpha_1) - \sqrt{\lambda} \frac{\beta_*}{\mu} + o(\sqrt{\lambda}), \] (45)
where \( \beta_* \) is the unique solution to
\[ \begin{align*}
\max_{\beta} & \quad \beta \\
\text{s.t.} & \quad (1 - \alpha_1) \int_0^\infty \exp \left( -\frac{2\rho}{\sigma^2} \int_0^y [h_{\tau_*}(x) - \beta] dx \right) dy \\
& \qquad \times \int_{-\infty}^\infty \exp \left( -\frac{2\rho}{\sigma^2} \int_0^y [h_{\tau_*}(x) - \beta] dx \right) dy \leq \alpha_2,
\end{align*} \] (46)
with \( h_{\tau_*}(x) = \lim_{\lambda \to \infty} \sqrt{\lambda} [H(\tau_* + \frac{x}{\sqrt{\lambda}}) - H(\tau_* \mid \tau_*)]. \)

The proof of this proposition is presented in Appendix EC.2. The key message is that, in the asymptotic sense, it is optimal to make the announcement so that the abandonment fraction is exactly \( \alpha_1 \), and then set the optimal staffing level according to (45). The staffing level then depends on \( \beta_* \) only. Comparing this to the case without announcement (with \( f_{\tau_*}(x) \) in (46) replaced by \( e^{-h_{\omega}x} \)), it is easy to verify that \( \beta_* \) in the case with announcement is smaller; hence the staffing level in (45) is reduced by \( O(\sqrt{\lambda}) \).

5. Refined Approximations

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

5.1. Using the Density of Patience-Time Distribution

Assume that the patience-time distribution \( H(\cdot) \) has a density at \( \omega \) and write it as \( H'(\omega) \). Then \( f_{\omega}(x) \), defined in (30), can be “well approximated” by \( H'(\omega)x \), and the density \( \pi \) in (31) can be approximately specialized to
\[ \pi(x) \approx C \exp \left( -\frac{\rho}{\sigma^2} \left[ H'(\omega)x^2 - 2\beta x \right] \right). \] (47)
This implies that \( \sqrt{\lambda}(V(\infty) - \omega) \) asymptotically follows a normal distribution. Consequently, it follows from (33) that
\[ \begin{align*}
P(W(\infty) > y) & \approx H_c(y) \int_{\sqrt{\lambda}(y - \omega)}^\infty \exp \left( -\frac{\rho}{\sigma^2} \left[ H'(\omega)x^2 - 2\beta x \right] \right) dx \\
& = H_c(y)P \left( \frac{\beta}{H'(\omega)} + \frac{\sigma}{\sqrt{2\rho H'(\omega)}} N > \sqrt{\lambda}(y - \omega) \right) \\
& = H_c(y)\Phi_c \left( \frac{-\beta\sqrt{2\rho} + \sqrt{2\rho\lambda H'(\omega)(y - \omega)}}{\sigma\sqrt{H'(\omega)}} \right),
\end{align*} \] (48)
where \( N \) 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 (28)–(29). Furthermore, we have

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

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. If \( \beta = 0 \), by the symmetry of the normal distribution and in view of (34), we have

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

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 (30), for large \( \lambda \),

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

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

\[
\pi(x) \approx \begin{cases} 
C \exp \left( -\frac{1}{2} \frac{(x - \frac{\beta \sqrt{H(\omega-)}}{H'(\omega-)})^2}{\frac{\beta^2}{\rho H'(\omega-)}} \right) \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega-)} \right), & x \leq 0, \\
C \exp \left( -\frac{1}{2} \frac{(x - \frac{\beta \sqrt{H(\omega+)}}{H'(\omega+)})^2}{\frac{\beta^2}{\rho H'(\omega+)}} \right) \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega+)} \right), & x > 0.
\end{cases}
\]

where the normalizing constant satisfies

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

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

\[
\mathbb{P}(W(\infty) > \omega) \approx C \cdot H_c(\omega) \Phi_c \left( \frac{-\sqrt{2 \rho \beta}}{\sigma \sqrt{H'(\omega+)}} \right) \frac{\rho \beta^2}{\sigma^2 H'(\omega+)} \frac{1}{\sqrt{H'(\omega+)}}.
\]
It follows from (34) that the expected queue length is

\[
\mathbb{E}Q(\infty) \approx \lambda \left( \omega - \frac{\omega^2}{2} \right) + C \frac{\sqrt{\lambda}}{\rho} \left\{ \int_0^\infty x \exp \left( - \frac{1}{2} \left( x - \frac{\beta}{H'(\omega^+)} \right)^2 \right) \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega^+)} \right) \, dx \right. \\
\left. - \int_0^\infty x \exp \left( - \frac{1}{2} \left( x + \frac{\beta}{H'(\omega^-)} \right)^2 \right) \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega^-)} \right) \, dx \right\} \\
= \lambda \left( \omega - \frac{\omega^2}{2} \right) + C \frac{\sqrt{\lambda}}{\rho} \left\{ \frac{\lambda^2}{2 \rho} \left( \frac{1}{H'(\omega^+)} - \frac{1}{H'(\omega^-)} \right) \right. \\
+ \frac{\sqrt{2\pi}\sigma}{\sqrt{2\rho}} \left\{ \frac{\beta}{(H'(\omega^+))^{3/2}} \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega^+)} \right) \Phi \left( -\frac{\sqrt{2\rho}\beta}{\sigma \sqrt{H'(\omega^+)}} \right) \right. \\
\left. + \frac{\beta}{(H'(\omega^-))^{3/2}} \exp \left( \frac{\rho \beta^2}{\sigma^2 H'(\omega^-)} \right) \Phi \left( -\frac{\sqrt{2\rho}\beta}{\sigma \sqrt{H'(\omega^-)}} \right) \right\}.
\]

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

\[
\mathbb{P}(W(\infty) > \omega) \approx H_c(\omega) \left( 1 + \sqrt{\frac{H'(\omega^+)}{H'(\omega^-)}} \frac{h_N \left( -\frac{\sqrt{2\rho}\beta}{\sigma \sqrt{H'(\omega^+)}} \right)}{h_N \left( \frac{\sqrt{2\rho}\beta}{\sigma \sqrt{H'(\omega^-)}} \right)} \right)^{-1} = H_c(\omega) \times \text{Garnett}_p(x),
\]

where \( \text{Garnett}_p(x) = [1 + \sqrt{\pi} h_N(x)/h_N(-x)]^{-1}, \) \( y = H'(\omega^+)/H'(\omega^-), \) and \( x = -\sqrt{2\rho}\beta/\sigma \sqrt{H'(\omega^-)}. \)

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

(a) \( \mathbb{E}Q(\infty) \)

<table>
<thead>
<tr>
<th>servers</th>
<th>( k = 1 )</th>
<th>Simulated</th>
<th>Approx.</th>
<th>( k = 3 )</th>
<th>Simulated</th>
<th>Approx.</th>
<th>( k = 5 )</th>
<th>Simulated</th>
<th>Approx.</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>.4188 ± .0015</td>
<td>.4167</td>
<td>.3079 ± .0012</td>
<td>.3050</td>
<td>.2619 ± .0013</td>
<td>.2575</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>.4172 ± .0024</td>
<td>.4167</td>
<td>.3053 ± .0018</td>
<td>.3050</td>
<td>.2578 ± .0013</td>
<td>.2575</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.4168 ± .0020</td>
<td>.4167</td>
<td>.3051 ± .0014</td>
<td>.3050</td>
<td>.2574 ± .0012</td>
<td>.2575</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>.4163 ± .0018</td>
<td>.4167</td>
<td>.3048 ± .0014</td>
<td>.3050</td>
<td>.2574 ± .0012</td>
<td>.2575</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>.4173 ± .0013</td>
<td>.4167</td>
<td>.3054 ± .0009</td>
<td>.3050</td>
<td>.2579 ± .0008</td>
<td>.2575</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) \( \mathbb{P}(W(\infty) > \omega) \)

Table 2 Comparison of the approximation and the simulation of Example 1.

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

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

5.3. Using the hazard rate of the Patience-Time Distribution

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

\[
f_\omega(x) = \sqrt{\lambda} \left[ H(\omega + \frac{x}{\sqrt{\lambda}}) - H(\omega) \right]
= \exp \left( - \int_{\omega}^{\omega + \frac{x}{\sqrt{\lambda}}} h(y) dy \right) \sqrt{\lambda} \left[ 1 - \exp \left( - \int_{\omega}^{\omega + \frac{x}{\sqrt{\lambda}}} h(y) dy \right) \right]
= H_c(\omega) \sqrt{\lambda} \left[ 1 - \exp \left( - \int_{0}^{\frac{x}{\sqrt{\lambda}}} h(\omega + y) dy \right) \right]
\approx H_c(\omega) \sqrt{\lambda} \left[ \frac{1}{\sqrt{\lambda}} \int_{0}^{x} h(\omega + \frac{y}{\sqrt{\lambda}}) dy \right] \quad \text{(for large } \lambda \text{)}
= H_c(\omega) \int_{0}^{x} h(\omega + \frac{y}{\sqrt{\lambda}}) dy.
\]

(53)

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

\[
\pi(x) \approx C \exp \left( \frac{2\rho \beta x}{\sigma^2} \right) \exp \left( - \frac{2\rho h_c(\omega)}{\sigma^2} \int_{0}^{\omega + \frac{x}{\sqrt{\lambda}}} h(\omega + \frac{y}{\sqrt{\lambda}}) dy \right),
\]

(54)

with the appropriate normalizing constant \(C\). Based on (54), the probability \(P(W(\infty) > y)\) and expected queue length \(E Q(\infty)\) can be approximated by replacing \(f_\omega(x)\) in (33)–(34) 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 \(\beta = 0\), with other \(\beta\)'s behaving similarly. Table 3 summarizes the comparison for Example 2 for different system sizes with the number of servers ranging from 20 to 400, and \(\kappa = 20, 100\). The column “Appr. G” is obtained via our approximation formulae (33)–(34) with \(f_\omega(x)\) from (30), while column “Appr. H” is calculated by replacing (30) with
Table 3 Comparison of the approximation and the simulation of Example 2.

(a) $E(Q_\infty)$

<table>
<thead>
<tr>
<th>Servers</th>
<th>$\kappa = 20$</th>
<th>$\kappa = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>4</td>
<td>3.34 ± 0.02</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>8.65 ± 0.04</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>18.04 ± 0.05</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>37.56 ± 0.06</td>
</tr>
<tr>
<td>400</td>
<td>80</td>
<td>77.20 ± 0.08</td>
</tr>
</tbody>
</table>

(b) $P(W_\infty > \omega)$

<table>
<thead>
<tr>
<th>Servers</th>
<th>$\kappa = 20$</th>
<th>$\kappa = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>.35785 ± .00143</td>
<td>.3576</td>
</tr>
<tr>
<td>50</td>
<td>.36396 ± .00210</td>
<td>.3641</td>
</tr>
<tr>
<td>100</td>
<td>.37122 ± .00188</td>
<td>.3712</td>
</tr>
<tr>
<td>200</td>
<td>.37858 ± .00175</td>
<td>.3787</td>
</tr>
<tr>
<td>400</td>
<td>.38652 ± .00124</td>
<td>.3859</td>
</tr>
</tbody>
</table>

(53). As Table 3 (a) shows, the larger the parameter $\kappa$ becomes (meaning a steeper change of the hazard rate), the larger the error that the fluid approximation yields in approximating the expected queue length. Since, in this case, the patience-time distribution is differentiable, we can also use the method by Mandelbaum and Zeltyn (2009), which leads to 0.4167 in approximating $P(W_\infty > \omega)$ for all systems in Table 3 (b). This is not nearly as close as either “Appr. G” or “Appr. H”. We also observe that “Appr. G” seems better for the tail-probability of waiting times, and similar or slightly worse for queue length, when compared against “Appr. H”.

We now relate our general setting of scaling the patience-time distribution to the hazard-rate scaling in Reed and Tezcan (2012). Note that our study is in the ED+QED regime ($\omega > 0$), which is different from the QED regime ($\omega = 0$) studied in Reed and Tezcan (2012). From the application point of view, the ED+QED regime is more suitable for the analysis of delay announcements. From the technical viewpoint, our derivation of the corresponding diffusion limit is quite different from that in the QED regime. Specifically, when analyzing the virtual waiting time at time $t$ for the QED regime, the customers in queue at time $t$ who would eventually abandon are negligible. However, under the ED+QED regime, these customers must be accounted for, which makes the analysis more challenging. At the same time, it is worth pointing out that the diffusion limits in the QED and ED+QED regimes share a somewhat similar structure (e.g. the variance is constant and the drift is a continuous function of the state). This similarity helps one to apply the same procedure for calculating the stationary distribution of both diffusion limits.

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

$$h^n(x) = \begin{cases} h(x), & \text{for } x \in [0, \omega], \\ h(\omega + \sqrt{n}(x - \omega)), & \text{for } x \in (\omega, \infty). \end{cases}$$
This implies that $F_n(\omega) = H_c(\omega)$. In view of (53),

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

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

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

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 observed in Table 2 (a) that the fluid approximation could result in a large error, when the left and right derivatives of the patience-time distribution do not agree. We now further study Example 2, where the patience-time distribution is differentiable but not that smooth. Similarly to Table 3, we simulate Example 2 with the same set of parameters ($\mu = 1$, $h_0 = 1$, $\rho = 1.2$, $\omega = \log(\rho)/h_0$ 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>1999929.32 ± 86.96</td>
</tr>
</tbody>
</table>

Table 4 On the accuracy of the fluid approximation in the ED regime as system size becomes large.
Also, from (34) and (56), we deduce the approximation for the expected queue length

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

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 system size becomes large. Our finding on the gap, between the approximations based on fluid and diffusion, concurs with this result. The two gaps are similar under the premise that the diffusion approximation is close to the originating system. Nevertheless, what we offer here is an alternative view 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 established 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 reflect in the diffusion limit, and consequently in the approximation formulae for performance measures.

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

To illustrate the value and generality of our approximations, we connected 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 new method for obtaining diffusion limits of many-server queues, by focusing on the virtual waiting time. Following our method, He (2015) recently developed diffusion approximations for overloaded queues in the nondegenerate slowdown (NDS) scaling. Our method can be also applied in the QED regime, but we leave this for future research.

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

Acknowledgments

The authors are grateful to the thorough review team for providing numerous thoughtful and constructive comments that significantly improved both the mathematical rigor and exposition of this paper. Junfei Huang’s research was supported in part by the Hong Kong Research Grants Council, Early Career Scheme, and Project No. 24500314. Jiheng Zhang’s research was supported in part by the Hong Kong Research Grants Council [Grants GRF-622411 and GRF-624012].

References


Technical Proofs

EC.1. Proof of Theorems 1 and 2

Proof of Theorem 1. We first characterize the diffusion limit of $\bar{Y}^n = \{\bar{Y}^n(t): t \geq 0\}$ given by (17). Condition (3) implies that, for any $T > 0$,

$$\sup_{0 \leq t \leq T} \left| \frac{\lambda_n F_n^n(\omega^n) - s_n \mu}{\sqrt{\lambda_n}} - \beta \right| t = \left| \frac{\lambda_n F_n^n(\omega^n) - s_n \mu}{\sqrt{\lambda_n}} - \beta \right| T \to 0. \quad (\text{EC.1})$$

It then follows from (1), (3), (EC.1), Lemmas EC.2 and EC.4 that $\bar{Y}^n \Rightarrow \bar{Y}$, where $\bar{Y} = \{\bar{Y}(t): t \geq 0\}$ with

$$\bar{Y}(t) = \rho \beta t + \bar{\Lambda}(t) - \sqrt{\rho - \bar{1}B_A(t)} - \sqrt{\bar{\rho}B(t)}.$$ 

By Lemma EC.5 in Section EC.3, any subsequence of $\{\bar{Y}^n\}_{n \in \mathbb{N}}$ has a further convergent subsequence, written as $\{\bar{V}^{nk}\}_{k \in \mathbb{N}}$ such that

$$\bar{V}^{nk} \Rightarrow V_* \quad \text{as } k \to \infty, \quad (\text{EC.2})$$

for some limit $\{V_*(t): t \geq 0\}$. The objective is to characterize the limit as the solution to (19).

To this end, write the second term on the right-hand side of (16) as

$$\int_0^t \sqrt{\lambda_n} \left( F^n(\omega^n + \frac{\bar{V}^n(x-)}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right) d\bar{\Lambda}^n(x) \quad (\text{EC.3})$$

$$= \int_0^t f_\omega(\bar{V}^n(x-))d\bar{\Lambda}^n(x) + \int_0^t \left( \sqrt{\lambda_n} \left( F^n(\omega^n + \frac{\bar{V}^n(x-)}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right) - f_\omega(\bar{V}^n(x-)) \right) d\bar{\Lambda}^n(x).$$

According to Lemma 8.3 of Dai and Dai (1999) and (EC.2), the first term in (EC.3) converges to $\int_0^t f_\omega(V_*(x)) dx$ along the subsequence $\{n_k\}_{k \in \mathbb{N}}$ as $k \to \infty$. By Condition (1) on the arrival process, for any $\varepsilon > 0$ there exists an $N_1$ such that when $n \geq N_1$,

$$\mathbb{P} \left( \sup_{0 \leq t \leq T} |\bar{\Lambda}^n(t)| \geq 2T \right) \leq \frac{\varepsilon}{2}.$$ 

By the tightness proved in Lemma EC.5 of Section EC.3, for the above $\varepsilon$, there also exist $M > 0$ and $N_2$ such that for all $n \geq N_2$,

$$\mathbb{P} \left( \sup_{0 \leq t \leq T} |\bar{V}^n(t)| \geq M \right) \leq \frac{\varepsilon}{2}.$$ 

With the help of Lemma 4.1 of Dai (1995), by Condition (4), $\sqrt{\lambda_n} \left( F^n(\omega^n + \frac{x}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right)$ converges to $f_\omega(x)$ uniformly on compact sets. Thus, for any given $\delta > 0$, we can find an $N_3$ such that for $n \geq N_3$ and $x \in [-M, M]$,

$$\left| \sqrt{\lambda_n} \left( F^n(\omega^n + \frac{x}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right) - f_\omega(x) \right| \leq \frac{\delta}{2T}.$$
So we can conclude that for all \( n \geq \max(N_1, N_2, N_3) \),
\[
\mathbb{P} \left( \sup_{0 \leq t \leq T} \left| \int_0^t \left( \sqrt{\lambda_n} (F^n(\omega^n + \frac{\tilde{V}^n(x-)}{\sqrt{\lambda_n}}) - F^n(\omega^n)) - f_\omega(\tilde{V}^n(x-)) \right) d\tilde{\Lambda}^n(x) \right| \geq \delta \right) \\
\leq \mathbb{P} \left( \sup_{0 \leq t \leq T} |\tilde{V}^n(t)| \geq M \right) + \mathbb{P} \left( \sup_{0 \leq t \leq T} |\tilde{\Lambda}^n(t)| \geq 2T \right) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\]
This proves that the last term in (EC.3) converges to 0. As a result, (EC.3) converges to \( \int_0^t f_\omega(V^\ast_n(x))dx \). It follows from (16), Assumptions (18) and the above proved convergence of (EC.3) that the limit process \( \{ V^\ast_n(t) : t \geq 0 \} \) is a solution to (19).

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 (19) is unique in the sense of probability law. Hence, from Condition (4), we conclude weak convergence of \( \{ \tilde{V}^n \}_{n \in \mathbb{N}} \) and that the corresponding limit satisfies (19).

**Proof of Theorem 2.** 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
\[
Q^n(t) = \sum_{i=\Lambda^n(\kappa^n(t))}^{\Lambda^n(t)} 1_{\{u^n_i + \tau^n_i > t\}} + \sum_{i=\Lambda^n(t-\omega^n) + 1}^{\Lambda^n(t-\omega^n)} 1_{\{u^n_i + \tau^n_i > t\}}
\]
\[
= \sum_{i=\Lambda^n(\kappa^n(t))}^{\Lambda^n(t)} 1_{\{u^n_i + \tau^n_i > t\}} + \sum_{i=\Lambda^n(t-\omega^n) + 1}^{\Lambda^n(t)} (1_{\{u^n_i + \tau^n_i > t\}} - F^n_c(t - \tau^n_i))
\]
\[
+ \int_{t-\omega^n}^t F^n_c(t-x) d(\Lambda^n(x) - \lambda_n x) + \lambda_n \int_{t-\omega^n}^t F^n_c(t-x) dx.
\]
Applying the diffusion scaling (20),
\[
\tilde{Q}^n(t) = \tilde{M}^n(t) + \frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(\kappa^n(t)) + 1}^{\Lambda^n(t-\omega^n)} 1_{\{u^n_i + \tau^n_i > t\}}, \tag{EC.4}
\]
where
\[
\tilde{M}^n(t) = \frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(t-\omega^n) + 1}^{\Lambda^n(t)} (1_{\{u^n_i + \tau^n_i > t\}} - F^n_c(t - \tau^n_i)) + \int_{t-\omega^n}^t F^n_c(t-x) d\tilde{\Lambda}^n(x). \tag{EC.5}
\]
Following the idea given by Liu and Whitt (2014b), the process \( \tilde{M}^n(\cdot) \) can be viewed as the diffusion-scaled queue length process of an infinite-server queue, with service times \( u^n_i \wedge \omega^n \).

With a modification to the proof for Theorem 3.1 (more specifically, Lemma 5.3) in Krichagina and Puhalskii (1997), we obtain that the first term in (EC.5) weakly converges to the process \( \{ G(t) : t \geq \omega + \delta \} \). See Appendix EC.6 for details on the modifications. For the second term, integrating-by-part, we have
\[
\int_{t-\omega^n}^t F^n_c(t-x) d\tilde{\Lambda}^n(x) = \tilde{\Lambda}^n(t) - F^n_c(\omega^n)\tilde{\Lambda}^n(t-\omega^n) - \int_{t-\omega^n}^t \tilde{\Lambda}^n(x) dF^n_c(t-x). \tag{EC.6}
\]
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

\[
\sup_{0 \leq t \leq T} |\tilde{\Lambda}^n(t) - \Lambda(t)| \to 0, \quad n \to \infty, \tag{EC.7}
\]
on each sample path. Note that

\[
\begin{align*}
\int_{t-\omega}^{t} \tilde{\Lambda}^n(t) dF^n_c(t-x) - \int_{t-\omega}^{t} \tilde{\Lambda}(x) dF_c(t-x) \\
= \int_{t-\omega}^{t} \left( \tilde{\Lambda}^n(t) - \tilde{\Lambda}(x) \right) dF^n_c(t-x) + \int_{t-\omega}^{t} \tilde{\Lambda}(x) dF^n_c(t-x) \\
+ \int_{t-\omega}^{t} \tilde{\Lambda}(x) (F^n_c(t-x) - F_c(t-x)).
\end{align*}
\]

The first two terms on the right-hand side converge to zero in probability, following (EC.7) and $|F^n(\omega^n) - F^n(\omega)| \leq |F^n(\omega) - F(\omega)| + |F^n(\omega^n) - F(\omega)| \to 0$. Since $\tilde{\Lambda}$ is a Brownian motion, thus for any fixed $T > 0$,

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

This and $\{F^n\}_{n \in \mathbb{N}}$ converges to $F$ in total variation on $[0, \omega]$ imply that the last term also converges to zero in probability. Combining the above convergence with (EC.6), we conclude that the second term in (EC.5) weakly converges to $\{\int_{t-\omega}^{t} F_c(t-x)d\tilde{\Lambda}(x) : t \geq \omega + \delta\}$. So we have

\[
\tilde{\Lambda}^n(t) \Rightarrow \mathcal{G}(t) + \int_{t-\omega}^{t} F_c(t-x)d\tilde{\Lambda}(x) \quad \text{on} \quad [\omega + \delta, \infty). \tag{EC.8}
\]

Now we consider the second term on the left-hand side in (EC.4). 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

\[
\sum_{i=\Lambda^n(\kappa^n(t)) + 1}^{\Lambda^n(t-\omega^n)} 1_{\{u^n_i > \omega^n + \frac{M}{\sqrt{\lambda_n}}\}} \leq \sum_{i=\Lambda^n(\kappa^n(t)) + 1}^{\Lambda^n(t-\omega^n)} 1_{\{u^n_i + \tau^n_i > t\}} \leq \sum_{i=\Lambda^n(\kappa^n(t)) + 1}^{\Lambda^n(t-\omega^n)} 1_{\{u^n_i > \omega^n - \frac{M}{\sqrt{\lambda_n}}\}}.
\]

Introduce

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

Then
\[
\frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(\kappa^n(t))}^{\Lambda^n(t-\omega^n)} 1_{\{\omega^n > \omega\} - \frac{M}{\sqrt{\lambda_n}}} = \tilde{G}^n_{M-}(t) + F^c_n(\omega^n) \sqrt{\lambda_n(t-\omega^n - \kappa^n(t))},
\]
\[
\frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(\kappa^n(t))}^{\Lambda^n(t+\omega^n)} 1_{\{\omega^n > \omega\} + \frac{M}{\sqrt{\lambda_n}}} = \tilde{G}^n_{M+}(t) + F^c_n(\omega^n) \sqrt{\lambda_n(t-\omega^n - \kappa^n(t))}.
\]

Due to being on the event \( \Omega^*_M \),
\[
\sup_{\omega+\delta \leq t \leq T} \left| \tilde{Q}^n(t) - \tilde{M}^n(t) - F^c_n(\omega^n) \sqrt{\lambda_n(t-\omega^n - \kappa^n(t))} \right|
\]
\[
\leq \sup_{\omega+\delta \leq t \leq T} |\tilde{G}^n_{M-}(t)| + \sup_{\omega+\delta \leq t \leq T} |\tilde{G}^n_{M+}(t)|.
\]

As a result, for any \( \eta > 0 \),
\[
\mathbb{P} \left( \sup_{\omega+\delta \leq t \leq T} \left| \tilde{Q}^n(t) - \tilde{M}^n(t) - F^c_n(\omega^n) \sqrt{\lambda_n(t-\omega^n - \kappa^n(t))} \right| \geq \eta \right) \leq \mathbb{P}(\Omega^*_M^c) + \mathbb{P} \left( \sup_{\omega+\delta \leq t \leq T} |\tilde{G}^n_{M-}(t)| + \sup_{\omega+\delta \leq t \leq T} |\tilde{G}^n_{M+}(t)| \geq \eta \right). \tag{EC.9}
\]

Note that the definition of \( \kappa^n(\cdot) \) in (6) and Proposition 1 imply that, as \( n \to \infty \),
\[
\sup_{V^n(0) \leq t \leq T} \left| \kappa^n(t) - t + \omega \right| \Rightarrow 0.
\]

By the initial condition (18), the probability that \( V^n(0) > \omega + \delta \) is vanishing with \( n \to \infty \). As a result, we have
\[
\sup_{\omega+\delta \leq t \leq T} \left| \kappa^n(t) - t + \omega \right| \Rightarrow 0.
\]

Then it is clear, for any fixed \( M > 0 \),
\[
\sup_{\omega+\delta \leq t \leq T} |\tilde{G}^n_{M-}(t)| + \sup_{\omega+\delta \leq t \leq T} |\tilde{G}^n_{M+}(t)| \Rightarrow 0. \tag{EC.10}
\]

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 \( \omega + \delta \leq t \leq T \), \( t \leq \kappa^n(t) + V^n(\kappa^n(t)) \leq t + \sup_{\omega+\delta \leq t \leq T} |V^n(\kappa^n(t)) - V^n(\kappa^n(t^-))| \). This, together with (11), implies that, as \( n \to \infty \),
\[
\sup_{\omega+\delta \leq t \leq T} \left| \sqrt{\lambda_n}(t - \kappa^n(t) - V^n(\kappa^n(t))) \right|
\]
\[
= \sup_{\omega+\delta \leq t \leq T} \left| \sqrt{\lambda_n}(t - \kappa^n(t) - \omega^n) - \tilde{V}^n(\kappa^n(t)) \right| \Rightarrow 0. \tag{EC.11}
\]
The first implication of (EC.11) is that
\[
\lim_{M \to \infty} \lim_{n \to \infty} \mathbb{P}((\Omega_M^n)^c) = 0.
\]
Combining this with (EC.10) and (EC.9), we know that \( \{\widetilde{Q}^n(t)\} \) and \( \{\widetilde{M}^n(t) + F^n_c(\omega^n)\sqrt{n}(t - \omega^n - \kappa^n(t))\} \) have the same weak limit. Since (EC.11) also implies that \( \sqrt{n}(t - \kappa^n(t) - \omega^n) \Rightarrow \widetilde{V}^n(t - \omega) \), the result of the theorem follows from (EC.8).

\section*{EC.2. Proofs of Propositions 1–3}

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

\textit{Proof of Proposition 1.} It suffices to show, in view of the convergence \( \omega^n \to \omega \), as \( n \to \infty \), that for any \( T > 0 \) and \( \delta \in (0, \omega/2) \),
\[
\mathbb{P}\left( \sup_{0 \leq t \leq T} |\widetilde{V}^n(t) - \omega^n| \geq \delta \right) \to 0 \text{ as } n \to \infty.
\]
(EC.12)

Define \( \widetilde{V}^n(t) = V^n(t) - \omega^n \), \( \eta^*_1 = \inf \{ t \geq 0 : \widetilde{V}^n(t) \geq \delta \} \) and \( \eta^*_2 = \inf \{ t \geq 0 : \widetilde{V}^n(t) \leq -\delta \} \). Let \( \Omega^n_1(\delta, T) = \{ \eta^*_1 \leq \eta^*_2, \eta^*_1 \leq T \} \), \( \Omega^n_2(\delta, T) = \{ \eta^*_1 > \eta^*_2, \eta^*_2 \leq T \} \), and \( \Omega^n_0(\delta) = \{ \eta^*_1(0) \leq \delta/4 \} \). In view of (18), to get (EC.12), 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 \( \Omega^n_2(\delta, T) \cap \Omega^n_0(\delta) \) is similar. On the set \( \Omega^n_1(\delta, T) \cap \Omega^n_0(\delta) \), define \( \eta^*_1 = \sup \{ 0 \leq t \leq \eta^*_1 : \widetilde{V}^n(t) \leq \delta/3 \} \). By the definitions of \( \eta^*_1 \) and \( \eta^*_2 \), we clearly have that \( \widetilde{V}^n(\eta^*_1) \geq \delta \), and \( \widetilde{V}^n(\eta^*_2) \leq \delta/3 \). In view of \( F^n(\omega^n + x) \geq F^n(\omega^n) \) for any \( x \geq 0 \), by (16)–(17), we have that
\[
\widetilde{V}^n(\eta^*_1) - \widetilde{V}^n(\eta^*_2) \leq \frac{\lambda_n}{s_n \mu} \left[ F^n_c(\omega^n) \frac{\tilde{\Lambda}^n(\eta^*_1) - \tilde{\Lambda}^n(\eta^*_2)}{\sqrt{\lambda_n}} - \frac{1}{\lambda_n} \sum_{i=1}^{\Lambda^n(\eta^*_1)} \mathbb{1}(u^n_i \leq \omega^n) - F^n(\omega^n) \right]
\]  
+ \( \frac{(\lambda_n F^n_c(\omega) - s_n \mu)(\eta^*_1 - \eta^*_2)}{\lambda_n} \) \( - \frac{\tilde{B}^n(\eta^*_1) + V^n(\eta^*_1)) - \tilde{B}^n(\eta^*_2) + V^n(\eta^*_2))}{\sqrt{\lambda_n}} \].
(EC.13)

By (1)–(3), we know that, as \( n \to \infty \),
\[
\mathbb{P}\left( \left| \frac{\lambda_n F^n_c(\omega)}{s_n \mu} \cdot \frac{\tilde{\Lambda}^n(\eta^*_1) - \tilde{\Lambda}^n(\eta^*_2)}{\sqrt{\lambda_n}} \right| > \delta \right) \to 0, \quad \text{(EC.14)}
\]
\[
\mathbb{P}\left( \left| \frac{1}{s_n \mu} (\lambda_n F^n_c(\omega) - s_n \mu)(\eta^*_1 - \eta^*_2) \right| > \delta \right) \to 0. \quad \text{(EC.15)}
\]

Using Lemma EC.1, we have
\[
\mathbb{P}\left( \left| \frac{1}{s_n \mu} \sum_{i=1}^{\Lambda^n(\eta^*_1)} \mathbb{1}(u^n_i \leq \omega^n) - F^n(\omega^n) \right| > \delta \right) \to 0. \quad \text{(EC.16)}
\]
To get that the last term in (EC.13) 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

\[
B^n(t) = D^n(t) + S^n(t) - S^n(0),
\]

which implies that

\[
B^n(\eta^n_1 + V^n(\eta^n_1)) - B^n(\eta^n_{12} + V^n(\eta^n_{12}^-)) = D^n(\eta^n_1 + V^n(\eta^n_1)) - D^n(\eta^n_{12} + V^n(\eta^n_{12}^-))
\]

\[
+ S^n(\eta^n_1 + V^n(\eta^n_1)) - S^n(\eta^n_{12} + V^n(\eta^n_{12}^-)).
\]

As \( V^n(\cdot) \) is always positive on \([\eta^n_{12} + V^n(\eta^n_{12}^-), \eta^n_1 + V^n(\eta^n_1)]\), all the servers are busy; hence \( S^n(\eta^n_1 + V^n(\eta^n_1)) = S^n(\eta^n_{12} + V^n(\eta^n_{12}^-)) = s_n \). As a result, noticing that the service time is exponential with rate \( \mu \),

\[
B^n(\eta^n_1 + V^n(\eta^n_1)) - B^n(\eta^n_{12} + V^n(\eta^n_{12}^-)) = D^n(\eta^n_1 + V^n(\eta^n_1)) - D^n(\eta^n_{12} + V^n(\eta^n_{12}^-))
\]

\[
= S(s_n(\eta^n_1 + V^n(\eta^n_1)) - S(s_n(\eta^n_{12} + V^n(\eta^n_{12}^-)))
\]

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

\[
P\left( \frac{\lambda_n}{s_n \mu} \frac{\tilde{B}^n(\eta^n_1 + V^n(\eta^n_1)) - \tilde{B}^n(\eta^n_{12} + V^n(\eta^n_{12}^-))}{\sqrt{\lambda_n}} \bigg| > \frac{\delta}{\tilde{\eta}_n} \right) \to 0. \tag{EC.18}
\]

Combining (EC.13)–(EC.16) and (EC.18), the probability of the event \( \bar{\Omega}_1^n(\delta, T) \cap \bar{\Omega}_0^n(\delta) \) 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 \( g(x) = \rho(f_x(x) - \beta) \). Note that, in view of (26), \( \lim_{x \to \infty} g(x) > 0 \) 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

\[
\pi(y) = C \exp \left( -\frac{2}{\sigma^2} \int_0^y g(x)dx \right), \tag{EC.19}
\]

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

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

it is enough to prove that the function \( \pi \) in (EC.19) satisfies

\[
\int_{\mathbb{R}} Af(x)\pi(x)dx = 0, \tag{EC.20}
\]
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.19), we have

\[
\int_{\mathbb{R}_+} A f(x) \pi(x) dx = \frac{C \sigma^2}{2} \int_{\mathbb{R}_+} d \left[ \exp \left( -\frac{2}{\sigma^2} \int_0^y g(x) dx \right) f'(y) \right] = -\frac{C \sigma^2}{2} f'(0). \tag{EC.21}
\]

Similarly,

\[
\int_{\mathbb{R}_-} A f(x) \pi(x) dx = \frac{C \sigma^2}{2} \int_{\mathbb{R}_-} d \left[ \exp \left( -\frac{2}{\sigma^2} \int_0^y g(x) dx \right) f'(y) \right] = \frac{C \sigma^2}{2} f'(0). \tag{EC.22}
\]

We now conclude (EC.20) by summing up (EC.21) and (EC.22). This implies (EC.19), and hence (27).

For the stationary distribution of the diffusion limit of the queue length, note that, for \( t > \omega \), \( G(t) \) is normally distributed with zero mean and variance \( \int_0^x F(x) F_c(x) dx \) due to (23). Similarly, \( \int_{t-\omega}^t F_c(t-x) d\tilde{\Lambda}(x) \) follows a zero-mean normal distribution with variance \( \theta^2 \int_{-\infty}^\infty (F_c(x))^2 dx \). 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:

\[
\begin{align*}
(i) \quad \limsup_{\lambda \to \infty} \sqrt{\lambda} (\tau_\lambda - \tau^*_\lambda) > 0 \quad \text{and} \quad (ii) \quad \liminf_{\lambda \to \infty} \sqrt{\lambda} (\tau_\lambda - \tau^*_\lambda) < 0.
\end{align*}
\]

To that end, we first note that for any nondecreasing function \( f(\cdot) \), the function defined by

\[
\int_0^\infty \frac{\exp \left( -\frac{\rho}{\sigma^2} \int_0^y [f(x) - \beta] dx \right) dy}{\int_{-\infty}^\infty \exp \left( -\frac{\rho}{\sigma^2} \int_0^y [f(x) - \beta] dx \right) dy}
\]

is continuous and strictly increasing in \( \beta \). In the remainder of the proof, we first propose a feasible solution, and then compare it with any optimal solution that satisfies either 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 \( \tau_\lambda \). Then by the definition of \( \tau_\lambda \), \( \omega_{\tau_\lambda} = \tau_\lambda \). Hence, the first constraint on the fraction of abandonment in (44) holds. Thus we only consider the second constraint on waiting time in (44), which further becomes \( \mathbb{P}(W^\lambda(\infty) > \tau_\lambda) \leq \alpha_2 \). Let \( \hat{\tau}^\lambda \) be its optimal solution. (We append the superscript \( \lambda \) to emphasize the dependency on the arrival rate.) It follows from (29) and (32) that the optimal number of servers for announcement time \( \tau_\lambda \) is given by (see also problem (38), and (39)–(40))

\[
\hat{\tau}^\lambda = \frac{\lambda}{\mu} H_c(\tau_\lambda | \tau_\lambda) - \frac{\beta_\lambda}{\mu} \sqrt{\lambda} + o(\sqrt{\lambda}), \tag{EC.24}
\]

where \( \beta_\lambda \) solves (46). Then by the continuity and monotonicity of the function given by (EC.23) with \( f(\cdot) = h_{\tau_\lambda}(\cdot) \),

\[
H_c(\tau_\lambda | \tau_\lambda) \cdot \frac{\int_0^\infty \exp \left( -\frac{\rho}{\sigma^2} \int_0^y [h_{\tau_\lambda}(x) - \beta_\lambda] dx \right) dy}{\int_{-\infty}^\infty \exp \left( -\frac{\rho}{\sigma^2} \int_0^y [h_{\tau_\lambda}(x) - \beta_\lambda] dx \right) dy} = \alpha_2. \tag{EC.25}
\]
This is equivalent to
\[
(1 - \alpha_1 - \alpha_2) \cdot \int_0^\infty \exp \left( -\frac{\rho}{2\sigma^2} \int_0^y [h_{\tau_*}(x) - \beta_*]dx \right) dy
= \alpha_2 \int_{-\infty}^0 \exp \left( -\frac{\rho}{2\sigma^2} \int_0^y [h_{\tau_*}(x) - \beta_*]dx \right) dy.
\] (EC.26)

By the definition of \( f_{\tau_*}(\cdot) \),
\[
h_{\tau_*}(x) = \begin{cases} 
  e^{-h_0 \tau_*} h_1 x, & \text{if } x \geq 0, \\
  e^{-h_0 \tau_*} h_0 x, & \text{if } x < 0.
\end{cases}
\] (EC.27)

Obviously, \((\delta^\lambda_\tau, \tau_*^\lambda)\) is a feasible solution to our original problem (44). (Indeed it is the staffing level (45).)

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. \( \lim_{\lambda \to \infty} \sqrt{\lambda}(\tau_* - \tau_*^\lambda) > 0 \). We will first focus on the subcase that
\[
0 < \lim_{\lambda \to \infty} \sqrt{\lambda}(\tau_* - \tau_*^\lambda) < \infty.
\] (EC.28)

Note that in this case \( \omega_{\tau_*}^\lambda > \tau_*^\lambda \) (because \( H(\tau_*^\lambda | \tau_*^\lambda) < \alpha_1 \)), then the constraint on waiting time in (44) becomes \( \mathbb{P}(W^\lambda(\infty) > \tau_*^\lambda) \). Similar to (EC.24)–(EC.25) (noticing that the constraint \( \mathbb{P}(\text{Ab}) \) can be achieved from the first order), the optimal number of servers is
\[
s_*^\lambda = \frac{\lambda}{\mu} H_c(\tau_*^\lambda | \tau_*^\lambda) - \frac{\beta_*^\lambda}{\mu} \sqrt{\lambda} + o(\sqrt{\lambda}),
\] (EC.29)

where \( \beta_*^\lambda \) solves
\[
\max_{\beta} \beta \quad \text{s.t.} \quad H_c(\tau_*^\lambda | \tau_*^\lambda) \cdot \frac{\int_0^\infty \exp \left( -\frac{\rho}{2\sigma^2} \int_0^y [f_{\tau_*^\lambda}(x) - \beta]dx \right) dy}{\int_{-\infty}^0 \exp \left( -\frac{\rho}{2\sigma^2} \int_0^y [f_{\tau_*^\lambda}(x) - \beta]dx \right) dy} \leq \alpha_2
\] (EC.30)

with \( f_{\tau_*^\lambda}(x) = \sqrt{\lambda}[H(\tau_*^\lambda + \frac{x}{\sqrt{\lambda}} | \tau_*^\lambda) - H(\tau_*^\lambda | \tau_*^\lambda)] \). From the definitions of \( f_{\tau_*^\lambda}(x) \) and \( H_c \),
\[
f_{\tau_*^\lambda}(x) = \begin{cases} 
  e^{-h_0 \tau_*^\lambda} \sqrt{\lambda} \left( 1 - \exp(-h_1 \frac{x}{\sqrt{\lambda}}) \right), & \text{if } x \geq 0, \\
  e^{-h_0 \tau_*^\lambda} \sqrt{\lambda} \left( 1 - \exp(-h_0 \frac{x}{\sqrt{\lambda}}) \right), & \text{if } -\tau_*^\lambda \sqrt{\lambda} \leq x < 0.
\end{cases}
\] (EC.31)

This, by the continuity and monotonicity of the function given by (EC.23), similar to (EC.26), implies that
\[
\left( H_c(\tau_*^\lambda | \tau_*^\lambda) - \alpha_2 \right) \cdot \int_0^\infty \exp \left( -\frac{\rho}{2\sigma^2} \int_0^y [f_{\tau_*^\lambda}(x) - \beta_*^\lambda]dx \right) dy
= \alpha_2 \cdot \int_{-\infty}^0 \exp \left( -\frac{\rho}{2\sigma^2} \int_0^y [f_{\tau_*^\lambda}(x) - \beta_*^\lambda]dx \right) dy.
\] (EC.32)

In view of (EC.28) and (EC.31),
\[
\lim_{\lambda \to \infty} \tau_*^\lambda = \tau_* \quad \text{and} \quad \lim_{\lambda \to \infty} f_{\tau_*^\lambda}(x) = h_{\tau_*}(x) = \begin{cases} 
  e^{-h_0 \tau_*} h_1 x, & \text{if } x \geq 0, \\
  e^{-h_0 \tau_*} h_0 x, & \text{if } x < 0.
\end{cases}
\] (EC.33)
This together with (EC.32) implies that

\[
(1 - \alpha_1 - \alpha_2) \cdot \int_0^\infty \exp \left( - \frac{\rho}{2\sigma^2} \int_0^y [h_{\tau_*}(x) - \lim_{\lambda \to \infty} \beta_\lambda^*] dx \right) dy
= \alpha_2 \int_{-\infty}^0 \exp \left( - \frac{\rho}{2\sigma^2} \int_0^y [h_{\tau_*}(x) - \lim_{\lambda \to \infty} \beta_\lambda^*] dx \right) dy.
\]  

Using the continuity and monotonicity of the function given by (EC.23), therefore, we have

\[
\beta_* = \liminf_{\lambda \to \infty} \beta_\lambda^*.
\]  

Similarly, taking the upper limit on both sides of (EC.32), we have

\[
\beta_* = \limsup_{\lambda \to \infty} \beta_\lambda^*.
\]  

On the other hand, by (EC.28)–(EC.29) and the definition of \( H \), we have

\[
0 < \lim_{\lambda \to \infty} \sqrt{\lambda} \left( H_e(\tau_*^\lambda | \tau_*^\lambda) - H_e(\tau_*^\lambda | \tau_*^\lambda) \right) = h_0 e^{-h_0 x} \times \lim_{\lambda \to \infty} \sqrt{\lambda} (\tau_*^\lambda - \tau_*^\lambda) < \infty.
\]  

It then follows from (EC.24), (EC.29) and (EC.35)–(EC.37) that \( \lim_{\lambda \to \infty} (s_*^\lambda - \hat{s}_*^\lambda) / \sqrt{\lambda} > 0 \), which is a contradiction with the optimality of \( s_*^\lambda \).

Now consider the subcase

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

The above argument still works if we replace \( \tau_*^\lambda \) by \( \tau_*^\lambda + M/\sqrt{\lambda} \) for any \( M > 0 \). Then we again obtain a contradiction with the optimality of \( s_*^\lambda \). Hence the proof of Case (i) is complete.

Case (ii) Similar to case (i) we assume \( \lim_{\lambda \to \infty} \sqrt{\lambda} (\tau_*^\lambda - \tau_*^\lambda) < 0 \). Now \( \omega_*^\lambda = \tau_*^\lambda \) as \( \tau_*^\lambda \) is larger than \( \tau_* \). Note that the constraint on waiting time in (44) becomes \( P(W^\lambda(\infty) > \omega_*^\lambda) \). So the optimal number of servers is

\[
s_*^\lambda = \frac{\lambda}{\mu} H_e(\omega_*^\lambda | \tau_*^\lambda) - \frac{\hat{\beta}_*^\lambda}{\mu} \sqrt{\lambda} + o(\sqrt{\lambda}),
\]  

where \( \hat{\beta}_*^\lambda \) is the optimal solution to problem given by (EC.30) with replacing \( f_{\tau_*^\lambda}(\cdot) \) by \( f_{\omega_*^\lambda, \tau_*^\lambda}(\cdot) \), where \( f_{\omega_*^\lambda, \tau_*^\lambda} (x) = \sqrt{\lambda}[H(\omega_*^\lambda + \tau_*^\lambda | \tau_*^\lambda) - H(\omega_*^\lambda | \tau_*^\lambda)] \). Note that \( H_e(\omega_*^\lambda | \omega_*^\lambda) = H_e(\omega_*^\lambda | \tau_*^\lambda) \), so the difference between \( \hat{s}_*^\lambda \) (see (EC.24)) and \( s_*^\lambda \) lies in the difference between \( \beta_* \) and \( \hat{\beta}_*^\lambda \). If

\[
\liminf_{\lambda \to \infty} (\beta_* - \hat{\beta}_*^\lambda) > 0,
\]  

then \( \lim_{\lambda \to \infty} (s_*^\lambda - \hat{s}_*^\lambda) / \sqrt{\lambda} > 0 \), which is again a contradiction with the optimality of \( s_*^\lambda \). Thus to complete the proof for Case (ii), it is sufficient to show (EC.39).

To prove (EC.39), again by the continuity and monotonicity of the function given by (EC.23), similar to (EC.26), we have

\[
(1 - \alpha_1 - \alpha_2) \cdot \int_0^\infty \exp \left( - \frac{\rho}{2\sigma^2} \int_0^y [f_{\omega_*^\lambda, \tau_*^\lambda}(x) - \hat{\beta}_*^\lambda] dx \right) dy
\]
= α_2 \int_{-\infty}^{0} \exp \left( -\frac{\rho}{2\sigma^2} \int_{0}^{y} [f_{\omega, n^*}(x) - \hat{\beta}_n^*]dx \right) dy. \quad \text{(EC.40)}

From the definitions of \( f_{\omega, n^*}(x) \) and \( H(x|\tau) \), and \( \omega_{n^*} = \tau_n \),

\[
f_{\omega, n^*}(x) = \begin{cases} e^{-h_0 \tau_n \sqrt{\lambda}(1 - \exp(-h_0 x))}, & \text{if } -\sqrt{\lambda} \tau_n \leq x \leq \sqrt{\lambda}(\tau_n - \tau_*) , \\ e^{-h_0 \tau_n \sqrt{\lambda}(1 - \exp(-h_1 (x - \sqrt{\lambda}(\tau_n - \tau_*)/\sqrt{\lambda}))) - h_0 (\sqrt{\lambda}(\tau_n - \tau_*)/\sqrt{\lambda}) }, & \text{if } x > \sqrt{\lambda}(\tau_n - \tau_*) . \end{cases}
\]

For notation simplicity, assume \( \lim_{\lambda \to \infty} \sqrt{\lambda}(\tau_n - \tau_*) \) exists and denote it by \( \bar{\tau} \). Then the above equation yields that

\[
\lim_{\lambda \to \infty} f_{\omega, n^*}(x) = \begin{cases} e^{-h_0 \tau_n} h_0 x, & \text{if } x \leq \bar{\tau} , \\ e^{-h_0 \tau_n} (h_1(x - \bar{\tau}) + h_0 \bar{\tau}), & \text{if } x > \bar{\tau} . \end{cases} \quad \text{(EC.41)}
\]

Combining (EC.40) and (EC.41) yields that

\[
(1 - \alpha_1 - \alpha_2) \cdot \left[ \int_{0}^{\bar{\tau}} \exp \left( -\frac{\rho}{2\sigma^2} \int_{0}^{y} \left[ e^{-h_0 \tau_n} h_0 x - \limsup_{\lambda \to \infty} \hat{\beta}_n^* \right]dx \right) dy \\
+ \int_{\bar{\tau}}^{\infty} \exp \left( -\frac{\rho}{2\sigma^2} \int_{0}^{\bar{\tau}} \left[ e^{-h_0 \tau_n} h_0 x - \limsup_{\lambda \to \infty} \hat{\beta}_n^* \right]dx - \int_{\bar{\tau}}^{y} \left[ e^{-h_0 \tau_n} (h_1(x - \bar{\tau}) + h_0 \bar{\tau}) - \limsup_{\lambda \to \infty} \hat{\beta}_n^* \right]dx \right) dy \right] \\
= \alpha_2 \int_{-\infty}^{0} \exp \left( -\frac{\rho}{2\sigma^2} \int_{0}^{y} \left[ e^{-h_0 \tau_n} h_0 x - \limsup_{\lambda \to \infty} \hat{\beta}_n^* \right]dx \right) dy.
\]

Notice that by (EC.27) and (EC.41), \( \lim_{\lambda \to \infty} f_{\omega, n^*}(x) = h_{\tau_n}(x) \) for \( x \leq 0 \), and \( \lim_{\lambda \to \infty} f_{\omega, n^*}(x) < h_{\tau_n}(x) \) for \( x > 0 \). Thus using the fact \( 1 - \alpha_1 - \alpha_2 > 0 \) and \( h_0 < h_1 \), and the definition of \( \beta_n \) (see (EC.26)), we have that \( \limsup_{\lambda \to \infty} \hat{\beta}_n^* < \beta_n \) by the monotonicity of the function given by (EC.23), which is equivalent to (EC.39).

In summary, Cases (i) and (ii) do not hold. Thus, we have \( \lim_{\lambda \to \infty} \sqrt{\lambda}(\tau_n - \tau_*) = 0 \). Also from this proof, we see that (45) is the minimal number of servers. Hence, the proof of the proposition is complete.

**EC.3. Several Auxiliary Lemmas**

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

Define the associated filtration with the \( n \)th system by \( \{ \mathcal{F}_n^k : k \geq 1 \} \) by

\[
\mathcal{F}_n^k = \sigma \{ \tau_{n}^{k+1}, \ell \leq k \} .
\]

Then we have

**Lemma EC.1.** \( \{ \sum_{i=1}^{[\lambda_n t]}(1_{[\omega_{n,i}^n \leq \omega_{n-1}^n]} - F_n(\omega_{n,i}^n)) : t \geq 0 \} \) is a Martingale with respect to the filtration \( \{ \mathcal{F}_n^{[\lambda_n t]} : t \geq 0 \} \). Furthermore,

\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} (1_{[\omega_{n,i}^n \leq \omega_{n-1}^n]} - F_n(\omega_{n,i}^n)) \Rightarrow 0 .
\]
Proof. First note that for each $i$, $u^n_{i+1}$ is independent of $\mathcal{F}^n_i$. By Lemma 3.1 in Dai and He (2010), $\omega^n_i$ and $\omega^n_j$ are $\mathcal{F}^n_{i-1}$-measurable for $j < i$. Also $u^n_i$ is measurable w.r.t. $\mathcal{F}^n_{i-1}$. Since the conditional probability
\[
\mathbb{E}[(1\{u^n_i \leq \omega^n_i\} - F^n(\omega^n_i)) | \mathcal{F}^n_{i-1}] = 0,
\]
so $\{\sum_{i=1}^{\lfloor \lambda n t \rfloor} (1\{u^n_i \leq \omega^n_i\} - F^n(\omega^n_i)) : t \geq 0\}$ is a Martingale with respect to the filtration $\{\mathcal{F}^n_{\lfloor \lambda n t \rfloor} : t \geq 0\}$. Therefore, we have the first part of the lemma.

Now prove the second part of the lemma. Clearly, $\{1/\lambda_n \sum_{i=1}^{\lfloor \lambda n t \rfloor} (1\{u^n_i \leq \omega^n_i\} - F^n(\omega^n_i)) : t \geq 0\}$ is also a Martingale. Its quadratic variation is given by
\[
\frac{1}{\lambda^2_n} \sum_{i=1}^{\lfloor \lambda n t \rfloor} (1\{u^n_i \leq \omega^n_i\} - F^n(\omega^n_i))^2 \leq \frac{1}{\lambda^2_n} [\lambda_n t] \to 0.
\]
Therefore, the second part is proved. \qed

LEMA EC.2. Under the same assumptions as Theorem 1, as $n \to \infty$
\[
\tilde{B}^n \Rightarrow 1/\sqrt{\rho} \mathcal{B},
\]
where $\mathcal{B} = \{\mathcal{B}(t) : t \geq 0\}$ is a standard Brownian motion.

Proof. In view of (EC.17), 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
\[
\tilde{B}^n(t) = \tilde{D}^n(t) + \tilde{S}^n(t) - \tilde{S}^n(0).
\]
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 $S(\int_0^t (X^n(x) \wedge s_n) dx)$, where $\{S(t) : t \geq 0\}$ is a Poisson process with rate $\mu$. By (10),
\[
\sup_{0 \leq t \leq T} \frac{(s_n - X^n(t))^+}{\lambda_n} \Rightarrow 0.
\]
This, together with (2), implies
\[
\tilde{D}^n \Rightarrow 1/\sqrt{\rho} \mathcal{B}.
\]
By the initial condition (18) and (EC.46), we have that the last two terms in (EC.45) will converge to zero. Hence, (EC.44) directly follows from (EC.47). \qed

LEMA EC.3. Under the same assumptions as Theorem 1, the sequence of stochastic processes $\{V^n\}_{n \in \mathbb{N}}$ 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 \( M \):

\[
\mathbb{P} \left\{ \sup_{0 \leq t \leq T} |\tilde{V}^n(t)| \geq M \right\} \leq 4\varepsilon.
\]

To this end, define

\[
\varsigma^n_i = \inf\{ t \geq 0 : \tilde{V}^n(t) \geq 1 \}, \quad \varsigma^n_i = \inf\{ t \geq 0 : \tilde{V}^n(t) \leq -M \},
\]

\[
\Omega^n_1(M, T) = \{ \varsigma^n_1 \leq \varsigma^n_2, \varsigma^n_1 \leq T \}, \quad \Omega^n_2(M, T) = \{ \varsigma^n_2 > \varsigma^n_1, \varsigma^n_1 \leq T \}.
\]

Hence we only need to show that, for all large enough \( n \) and \( M \),

\[
\mathbb{P} (\Omega^n_1(M, T)) \leq 2\varepsilon \quad \text{and} \quad \mathbb{P} (\Omega^n_2(M, T)) \leq 2\varepsilon. \tag{EC.48}
\]

We will first consider the event \( \Omega^n_1(M, T) \). By the definition of \( \varsigma^n_1 \), we must have that \( \tilde{V}^n(\varsigma^n_1) \geq \tilde{V}^n(\varsigma^n_1) \). In other words, if \( \tilde{V}^n \) has a jump at \( \varsigma^n_1 \), then it must be an upward jump. Since \( \tilde{V}^n(t) \in [-M, M] \) on \((0, \varsigma^n_1]\), for any \( t \in (0, \varsigma^n_1] \) and small positive \( \delta \in (0, t) \), by (16),

\[
\tilde{V}^n(t) - \tilde{V}^n(t - \delta) = -\frac{\lambda_n}{s_n \mu} \int_{t-\delta}^{t} \sqrt{\lambda_n} \left( F^n(\omega^n + \frac{	ilde{V}^n(x)}{\sqrt{\lambda_n}}) - F^n(\omega^n) \right) d\bar{\Lambda}^n(x) + \tilde{Y}^n(\varsigma^n_1) - \tilde{Y}^n(\varsigma^n_1 - \delta). \tag{EC.49}
\]

Since \( \tilde{V}^n(0) \) is stochastically bounded, we can choose \( M \) large enough such that

\[
\mathbb{P} (\Omega^n_0(M)) = \mathbb{P} \left( \tilde{V}^n(0) \leq \frac{M}{4} \right) \geq 1 - \varepsilon,
\]

where \( \Omega^n_0(M) \) is defined in the proof of Proposition 1. Define \( \varsigma^n_{12} = \sup\{ 0 \leq t \leq \varsigma^n_1 : \tilde{V}^n(t) \leq M/2 \} \cup \varnothing \). We know that on the event \( \Omega^n_0(M) \), \( \varsigma^n_{12} > 0 \). By the definition of \( \varsigma^n_1 \) and \( \varsigma^n_{12} \), we clearly have that

\[
\tilde{V}^n(\varsigma^n_1) \geq M, \quad \text{and} \quad \tilde{V}^n(\varsigma^n_{12} -) \leq \frac{M}{2}. \tag{EC.50}
\]

Note that the process \( \tilde{V}^n(\cdot) \) is larger than \( M/2 \) (thus larger than \( 0 \)) on the interval \([\varsigma^n_{12}, \varsigma^n_1]\). By (EC.49) and the fact that \( F^n(\omega^n + x) \geq F^n(\omega^n) \) for any \( x \geq 0 \),

\[
\tilde{V}^n(\varsigma^n_1) - \tilde{V}^n(\varsigma^n_{12} -) \leq \tilde{Y}^n(\varsigma^n_1) - \tilde{Y}^n(\varsigma^n_{12} -). \tag{EC.51}
\]

By (EC.50) and (EC.51),

\[
\mathbb{P} \left( \Omega^n_0(M) \cap \Omega^n_1(M, T) \right) \leq \mathbb{P} \left( \sup_{t \in [0, T]} |\tilde{Y}^n(t)| \geq \frac{M}{4} \right). \tag{EC.52}
\]

We now prove the stochastic boundedness of \( \tilde{Y}^n \). Recall the definition of \( \tilde{Y}^n \) in (17). The first and the third term on the right side of (17) is stochastically bounded by Conditions (1), (2) and (3). The last two terms are stochastically bounded by Lemma EC.2. It now remains to show the stochastic boundedness of the third term, which can be written as \( \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\Lambda^n(t)} (1 \{ \omega_i^n \leq \omega^n \} - F^n(\omega^n)) \). According to Condition (1), it is enough to show the stochastic boundedness of
\[ \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{[\lambda_n t]} (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i)) \]. From Doob’s inequality for martingale (see Lemma EC.1), for any \( M \geq 0, \)

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

\[
= \frac{1}{M^2} \sum_{i=1}^{[\lambda_n T]} \mathbb{E}(1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))^2 \leq M^{-2}T.
\]

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.52) 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.4.** Under the same assumptions as Theorem 1, as \( n \to \infty \)

\[ \tilde{H}^n(\cdot) \Rightarrow (1/\rho)\sqrt{\rho-1}B_A(\cdot). \] (EC.53)

Here \( \tilde{H}^n(\cdot) \) is given by (15) and \( B_A = \{B_A(t) : t \geq 0\} \) is a standard Brownian motion which is independent of \( \{B(t) : t \geq 0\} \).

**Proof.** We first prove a convergence result for the sequence of processes given by \( \{\frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{[\lambda_n t]} (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i)) : t \geq 0\} \). By Lemma EC.1, the quadratic variation of martingale \( \{\frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{[\lambda_n t]} (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i)) : t \geq 0\} \) is

\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))^2.
\]

We calculate it in the following:

\[
\mathbb{E}\left[ \frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} \left( (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i)) + F^n(\omega_n^i) - F^n(\omega_n^i) \right) \right]^2
\]

\[
= \mathbb{E}\left[ \frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} \left( (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(F^n(\omega_n^i) - F^n(\omega_n^i)) \right) \right]^2
\]

\[
= \frac{2}{(\lambda_n)^2} \sum_{1 \leq j < i \leq [\lambda_n t]} \mathbb{E}\left[ (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(F^n(\omega_n^i) - F^n(\omega_n^i)) \right]\cdot (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(F^n(\omega_n^i) - F^n(\omega_n^i)) \right]
\]

\[
+ \frac{1}{(\lambda_n)^2} \sum_{i=1}^{[\lambda_n t]} \mathbb{E}\left[ (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(F^n(\omega_n^i) - F^n(\omega_n^i)) \right]^2.
\]

Then by conditioning on \( F^n_{i-1} \), we have

\[
\mathbb{E}\left[ (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(F^n(\omega_n^i) - F^n(\omega_n^i)) \right]\cdot (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(\sum_{j < i} (F^n(\omega_n^j) - F^n(\omega_n^j))) \right]
\]

\[
= \mathbb{E}\left[ (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(F^n(\omega_n^i) - F^n(\omega_n^i)) \right]\cdot \mathbb{E}\left[ (F^n(\omega_n^i) - F^n(\omega_n^i))(\sum_{j < i} (F^n(\omega_n^j) - F^n(\omega_n^j))) \right]
\]

\[
= \mathbb{E}\left[ (1_{\{u_n^i \leq \omega_n^i\}} - F^n(\omega_n^i))(F^n(\omega_n^i) - F^n(\omega_n^i)) \right]\cdot \mathbb{E}\left[ (F^n(\omega_n^i) - F^n(\omega_n^i))(\sum_{j < i} (F^n(\omega_n^j) - F^n(\omega_n^j))) \right].
\]
By (EC.43) the first term on the right-hand side of (EC.54) is 0. Note that the second term on the right-hand side of (EC.54) converges to 0 as \( n \to \infty \), so the expectation on the left-hand side of (EC.54) must converge to 0. As a result,

\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} \left( (1_{[u^n_i \leq \omega^n_i]} - F^n(\omega^n_i))^2 - F^n(\omega^n_i) F^n(\omega^n_i) \right) \Rightarrow 0.
\]

On the other hand, on the event given by \( \{ \Lambda^n(t+1) \geq \lambda_n t \} \)

\[
\frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} (F^n(\omega^n_i) F^n(\omega^n_i) - F^n(\omega^n_i) F^n(\omega^n_i))
\]

\[
= \frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} \left( [F^n(\omega^n_i) - F^n(\omega^n_i)] F^n(\omega^n_i) + F^n(\omega^n_i) [F^n(\omega^n_i) - F^n(\omega^n_i)] \right)
\]

\[
\leq \sup_{0 \leq s \leq t+1} 2 |F^n(V^n(s)) - F^n(\omega^n)| t
\]

\[
= \frac{2}{\sqrt{\lambda_n}} \sup_{0 \leq s \leq t+1} \left| \sqrt{\lambda_n} \left( F^n(\omega^n) + \frac{1}{\sqrt{\lambda_n}} \tilde{V}^n(s) - F^n(\omega^n) \right) \right| t,
\]

which vanishes to 0 following from Condition (4) and the stochastic boundedness of \( \{ \tilde{V}^n \}_{n \in \mathbb{N}} \) given by Lemma EC.3. Note that (2) and (3) imply that for any \( T > 0 \), as \( n \to \infty \),

\[
F^n(\omega^n) \to 1 - \frac{1}{\rho} \quad \text{and} \quad P \left( \inf_{0 \leq s \leq T} (\Lambda^n(t+1) - \lambda_n t) \geq 0 \right) \to 1.
\]

As a result,

\[
\lim_{n \to \infty} \frac{1}{\lambda_n} \sum_{i=1}^{[\lambda_n t]} (1_{[u^n_i \leq \omega^n_i]} - F^n(\omega^n_i))^2 \Rightarrow \lim_{n \to \infty} F^n(\omega^n_i) F^n(\omega^n_i) t = \frac{\rho - 1}{\rho^2} t.
\]

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 \( \{ \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{[\lambda_n t]} (1_{[u^n_i \leq \omega^n_i]} - F^n(\omega^n_i)) : t \geq 0 \} \) weakly converges to the process \( \frac{\sqrt{\rho}}{\rho} B_A \). The result of this lemma then follows from the random-time-change theorem. 

**Lemma EC.5.** 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.3, it suffices to study the modulus of continuity for \( \{ \tilde{V}^n \}_{n \in \mathbb{N}} \).

By (1), for any \( \varepsilon > 0 \),

\[
\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P} \left( \sup_{s \in [0,T], |s-t| < \delta} |\tilde{\Lambda}^n(s) - \tilde{\Lambda}^n(t)| > \varepsilon \right) = 0. \quad (EC.55)
\]

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

\[
\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P} \left( \sup_{s \in [0,T], |s-t| < \delta} |\tilde{Y}^n(s) - \tilde{Y}^n(t)| > \varepsilon \right) = 0. \quad (EC.56)
\]

Let \( \Omega^n_{\text{c}}(M, T) \) be the complement of \( \Omega^n_{\text{c}}(M, T) \cup \Omega^n_{\text{c}}(M, T) \) which are given in the proof of Lemma EC.3, then

\[
\lim_{M \to \infty} \liminf_{n \to \infty} \mathbb{P} (\Omega^n_{\text{c}}(M, T)) = 1. \quad (EC.57)
\]
On the event $\Omega^n_c(M, T)$, it follows from (EC.49) that

$$\left| \bar{V}^n(t) - \bar{V}^n(t - \delta) \right| \leq C^n_M \cdot \left( \bar{\Lambda}^n(t) - \bar{\Lambda}^n(t - \delta) \right) + \left| \bar{Y}^n(t) - \bar{Y}^n(t - \delta) \right|,$$

where $C^n_M = \max \left\{ \sqrt{\lambda_n} \left( F_n(\omega^n + \frac{M}{\sqrt{n}}) - F_n(\omega^n) \right), \sqrt{\lambda_n} \left( F_n(\omega^n) - F_n(\omega^n - \frac{M}{\sqrt{n}}) \right) \right\}$. By Condition (4), $C^n_M$ is bounded by some finite number $C_M$ which may depend on $M$. So for any $M > 0$,

$$\mathbb{P} \left( \sup_{s,t \in [0,T]} |\bar{V}^n(s) - \bar{V}^n(t)| > \varepsilon \right) \leq (1 - \mathbb{P} (\Omega^n_c(M, T))) + \mathbb{P} \left( \sup_{s,t \in [0,T]} |\bar{\Lambda}^n(s) - \bar{\Lambda}^n(t)| > \frac{\varepsilon}{2C_M} \right) + \mathbb{P} \left( \sup_{s,t \in [0,T]} |\bar{Y}^n(s) - \bar{Y}^n(t)| > \frac{\varepsilon}{2} \right).$$

By first letting $n$ go to infinite, then $\delta$ to zero and finally $M$ go to infinite, we can show that

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

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

**EC.4. Discussion on the Sequence of $\{\omega^n\}$**

As discussed after condition (4) that both $\beta$ and $f_\omega$ depend on the sequence $\{\omega^n\}_{n \in \mathbb{N}}$, one may wonder whether different sequences of $\{\omega^n\}_{n \in \mathbb{N}}$ satisfying conditions (3)–(4) and assumption (18) will give us inconsistent results. (Inconsistence means that arguments based on different sequences of $\{\omega^n\}_{n \in \mathbb{N}}$ may give contradictions.) In this section, we argue that this inconsistence is impossible.

We are given two sequences $\{\omega^n_{(1)}\}_{n \in \mathbb{N}}$ and $\{\omega^n_{(2)}\}_{n \in \mathbb{N}}$ with $\lim_{n \to \infty} \omega^n_{(1)} = \lim_{n \to \infty} \omega^n_{(2)} = \omega$ such that $i = 1, 2$,

$$\frac{\lambda_n F^n(\omega^n_{(i)}) - s_n \mu \sqrt{\lambda_n}}{\sqrt{\lambda_n}} \to \beta^{(i)}, \quad (EC.58)$$

$$\sqrt{\lambda_n} \left( F^n(\omega^n_{(i)}) + \frac{x}{\sqrt{\lambda_n}} - F^n(\omega^n_{(i)}) \right) \to f^{(i)}(x), \quad (EC.59)$$

$$\sqrt{\lambda_n} \left( V^n(0) - \omega^n_{(i)} \right) \Rightarrow \bar{V}_0^{(i)}. \quad (EC.60)$$

It directly follows from (EC.60) that

$$\lim_{n \to \infty} \sqrt{\lambda_n}(\omega^n_{(2)} - \omega^n_{(1)}) = x_0 \in \mathbb{R}. \quad (EC.61)$$

We first look at the case of the virtual waiting time. Define $\tilde{V}_1^n(t) = \sqrt{\lambda_n}(V^n(t) - \omega^n_{(1)})$ and $\tilde{V}_2^n(t) = \sqrt{\lambda_n}(V^n(t) - \omega^n_{(2)})$. Then

$$\tilde{V}_2^n(t) = \tilde{V}_1^n(t) + \sqrt{\lambda_n}(\omega^n_{(1)} - \omega^n_{(2)}). \quad (EC.62)$$
It directly follows (EC.61)–(EC.62) that
\[
\tilde{V}_i^n \Rightarrow \tilde{V}^{(i)}, \quad i = 1, 2.
\] (EC.63)

To prove the results for two sequences \(\{\omega_n^{(1)}\}_{n \in \mathbb{N}}\) and \(\{\omega_n^{(2)}\}_{n \in \mathbb{N}}\) to be consistent, it is sufficient to show that the diffusion approximations for \(\omega_n^{(1)}\) and \(\omega_n^{(2)}\) given by Theorem 1 also satisfy (EC.63). To this end, note that (EC.61) together with (EC.58)–(EC.59) implies that
\[
\beta^{(2)} = -f^{(1)}(x_0) + \beta^{(1)}, \quad f^{(2)}(x) = f^{(1)}(x_0 + x) - f^{(1)}(x_0).
\] (EC.64)

By (EC.58)–(EC.60) and Theorem 1, we have, as \(n \to \infty\)
\[
\tilde{V}_i^n \Rightarrow \tilde{V}^{(i)},
\] (EC.65)
where
\[
\tilde{V}^{(i)}(t) = \tilde{V}^{(i)}(0) - \rho \int_0^t \left[ f^{(i)}(\tilde{V}^{(i)}(x)) - \beta^{(i)} \right] dx + \left[ \tilde{\Lambda}(t) - \sqrt{\rho B(t)} - \sqrt{\rho - 1}B_A(t) \right].
\] (EC.66)

In view of (EC.64), it follows from (EC.66) that
\[
\tilde{V}^{(2)} = \tilde{V}^{(1)} - x_0.
\] (EC.67)

Therefore, different choices of \(\{\omega_n\}_{n \in \mathbb{N}}\) do not give us inconsistent results based on Theorem 1 when conditions (1)–(4) and assumption (18) hold!

Similarly, from Theorem 2, we can prove that different choices of \(\{\omega_n\}_{n \in \mathbb{N}}\) also do not give us inconsistent results for the diffusion approximations for the queue length process.

**EC.5. Discussion on the Initial State**

We discuss Assumption (18) on the initial state. Usually, the initial state is given by the queue length and patience times; see Liu and Whitt (2014b), Mandelbaum and Momčilović (2012), and Reed and Tezcan (2012). In the following, we hence provide a sufficient condition for (18) in terms of queue length and patience times. Analysis for what general initial conditions imply (18) is left for future research. To the best of our knowledge, the first work focusing on the initial state is Aras et al. (2014), which studied the impact of initial content (e.g., initial age process) on the system performances.

**Lemma EC.6.** Denote by \(Q^n(0)\) the number of customers who are initially in queue. Assume those initial customers in queue are infinitely patient. If \(\frac{Q^n(0) - s_n \mu^n}{\sqrt{x_n}} \Rightarrow \tilde{Q}_0\) for some random variable \(\tilde{Q}_0\) and \(\omega^n \to \omega\), then as \(n \to \infty\),
\[
\tilde{V}^{(0)}(0) \Rightarrow \tilde{V}_0,
\] (EC.68)
where \(\tilde{V}_0 = \rho(\tilde{Q}_0 - \sqrt{\frac{\mu}{\rho}}N)\) with \(N\) being a standard normal random variable independent of \(\tilde{Q}_0\).
Proof. Recall that $D^n(t)$ is the number of departures through service completion by time $t$. As all customers initially in queue will eventually receive service, the virtual waiting time $V^n(0)$ satisfies

$$D^n(V^n(0))^+ \leq Q^n(0) < D^n(V^n(0)).$$

As $D^n$ is a Poisson process with rate $s_n \mu$ on $[0,V^n(0)]$, with probability one, there is only one departure at time $V^n(0)$. This gives

$$D^n(V^n(0)) = Q^n(0) + 1. \quad \text{(EC.69)}$$

Dividing both sides by $\lambda_n$, then one can see that

$$V^n(0) \Rightarrow \omega, \quad \text{as} \quad n \to \infty.$$ 

From (EC.69), we get

$$\tilde{V}^n(0) = \frac{\lambda_n}{s_n \mu} \left( \frac{Q^n(0) - s_n \mu \omega^n}{\sqrt{\lambda_n}} - \frac{D^n(V^n(0)) - s_n \mu V^n(0)}{\sqrt{\lambda_n}} + 1 \right).$$

With $V^n(0) \Rightarrow \omega$, we have $\frac{D^n(V^n(0)) - s_n \mu V^n(0)}{\sqrt{\lambda_n}} \Rightarrow \sqrt{\pi} \mathcal{N}$; here $\mathcal{N}$ follows the standard normal distribution. Together with the assumption that $\frac{Q^n(0) - s_n \mu \omega^n}{\sqrt{\lambda_n}} \Rightarrow \tilde{Q}_0$, we get the convergence of $\tilde{V}^n(0)$. This completes the proof. \hfill \square

**EC.6. Proving the Convergence of the First Term in (EC.5)**

To analyze the first term on the right-hand side of (EC.5), we need a modification of the proof of Lemma 5.3 in Krichagina and Puhalskii (1997). Such a modification is needed because we allow the distribution $F^n(\cdot)$ to vary with $n$ while Lemma 5.3 in Krichagina and Puhalskii (1997) only deals with a fixed $F(\cdot)$ (i.e., $F^n(\cdot) \equiv F(\cdot)$ for all $n$). We now demonstrate how to modify their proof to allow the distribution $F^n(\cdot)$ to vary with $n$. For a function $g(\cdot)$, let $g(x-)$ denote its left-hand limit at $x$.

To make the connection easy, we adopt the same notation as theirs without conflicting with the notation already used in the above. We denote the first term on the right-hand side of our (EC.5) by

$$M^n_2(t) := \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\Lambda^n(t)} \left( 1_{\{u^n_i + \tau^n_i > t\}} - F^n_{\Lambda}(t - \tau^n_i) \right).$$

Define $v^n_i = u^n_i \land \omega^n$ and let $F^n_{\omega}(\cdot)$ be the distribution of $v^n_i$. Let $u_1$ be a random variable with distribution $F(\cdot)$ and $F_{\omega}(\cdot)$ be the distribution of $u_1 \land \omega$. Define $U^n$ as in their (2.23) but change $n$ to $\lambda_n$. Change $V^n$ in their (3.24) to $V^n(t,x) = U^n(\tilde{\Lambda}^n(t), F^n_{\omega}(x)), t \geq 0, x \geq 0$, and $L^n(t,x)$ in their (3.18) to

$$L^n(t,x) = \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\Lambda^n(t)} \left[ 1_{\{v^n_i \leq x\}} - \int_0^{x \land v^n_i} \frac{dF^n_{\omega}(y)}{1 - F^n_{\omega}(y-)} \right].$$
Then following the same argument leading to their (3.19), we have
\[ V^\alpha(t, x) = -\int_0^t \frac{V^n(t, y-)}{1 - F^\alpha_n(y-)} dF^\alpha_n(y) + L^\alpha(t, x), \quad t \geq 0, x \geq 0. \]

Also, following the same argument leading to their (3.20), we have
\[ M^\alpha_n(t) = G^\alpha(t) - H^\alpha(t), \]
where
\[ G^\alpha(t) = \int_0^t \frac{V^n(t-x, x-)}{1 - F^\alpha_n(x-)} dF^\alpha_n(x), \]
\[ H^\alpha(t) = \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\Lambda^\alpha(t)} \left( 1_{\{s^n_i \leq t\}} - \int_0^{\tau_n^\alpha(t-t^n_i)} \frac{dF^\alpha_n(u)}{1 - F^\alpha_n(u-)} \right). \]

In Krichagina and Puhalskii (1997), the proof of Lemma 5.3, which shows the convergence of \( \{M^\alpha_2(t)\} \), needs (i) the tightness of \( \{M^\alpha_2\}_{n \in \mathbb{N}} \) with \( M^\alpha_2 = \{M^\alpha_2(t) : t \geq 0\} \) (Lemma 3.8) and (ii) finite dimensional distribution convergence of \( \{M^\alpha_2\}_{n \in \mathbb{N}} \). In the following we show (i) and (ii) separately.

(i) **Tightness.** Lemma 3.8 is essentially a summary of Lemmas 3.4 (tightness of \( \{G^\alpha\}_{n \in \mathbb{N}} \) with \( G^\alpha = \{G^\alpha(t) : t \geq 0\} \)) and 3.7 (tightness of \( \{H^\alpha\}_{n \in \mathbb{N}} \) with \( H^\alpha = \{H^\alpha(t) : t \geq 0\} \)), both of which require Lemma 3.1 and 3.2 and some additional analysis. We can directly use their Lemma 3.1 because it is for uniform distribution and does not involve general distributions. In our model, \( \Lambda^\alpha \) plays the role of \( \alpha^\alpha \) in their Lemma 3.2. Our assumption (1) enables us to use their Lemma 3.2.

We now show that the conclusion of their Lemma 3.4 (tightness of \( \{G^\alpha\}_{n \in \mathbb{N}} \)) holds for our model. Let \( T_0 = \sup_{n} \omega^n + 1 \). Let \( T_0 = \sup_{n} \omega^n + 1 \), and define piecewise-linear function \( \psi^n(\cdot) \) such that its graph linearly connects \((0, 0), (\omega^n_1, \omega^n_1), (\omega, \omega^n)\) and \((T_0, T_0)\) (that is, \( \psi^n(0) = 0, \psi^n(\omega^n) = \omega^n, \psi^n(\omega) = \omega^n \) and \( \psi^n(T_0) = T_0 \)), where
\[ \omega^n_1 = \begin{cases} \omega^n - d_n, & \text{if } \omega^n \leq \omega, \\ \omega - d_n, & \text{if } \omega^n > \omega \end{cases} \]
with
\[ \frac{F^\alpha_n(\omega^n -) - F^\alpha_n(\omega^n_1)}{\lambda_n} \leq |\omega^n - \omega|, \]
and \( d_n \in (0, 1/\lambda_n^{1/2}) \). Assumptions (3)–(4) and assumption \( F(\omega-) = 1 - 1/\rho \) imply that \( F^\alpha_n(\omega^n -) - F^\alpha_n(\omega^n -) \rightarrow 0 \) as \( n \rightarrow \infty \). Note that \( \psi^n(\omega) = \omega^n \), then \( \Delta F^\alpha_n(\psi^n(\omega)) \rightarrow \Delta F^\alpha_n(\omega) \).

(For a function \( f \), \( \Delta f(t) := f(t) - f(t-) \)). Then, \( \sup_{0 \leq t \leq \tau_0} |\psi^n(t) - t| \leq |\omega^n - \omega| \rightarrow 0 \) and \( \sup_{0 \leq t \leq \tau_0} |F^\alpha_n(\psi^n(t)) - F^\alpha_n(t)| \leq \frac{1}{\lambda_n} + |F^\alpha_n(\omega^n -) - F^\alpha_n(\omega^n -)| + \sup_{0 \leq t \leq \omega^n} |F^\alpha_n(t) - F^\alpha_n(t)| \rightarrow 0 \) as \( n \rightarrow \infty \). One can also verify that the probability distribution \( F^\alpha_n(\psi^n(\cdot)) \) converges to \( F(\cdot) \) in total variation. Indeed, for any measurable set \( A \),
\[ |F^\alpha_n(\psi^n(A)) - F(\cdot)| \leq |F^\alpha_n(\psi^n(A, x \leq \omega^n)) - F^\alpha_n(A, x \leq \omega^n)| \]
\[ + |F^\alpha_n(\psi^n(A, \omega^n < x < \omega)) - F^\alpha_n(A, \omega^n < x < \omega)| \]
\[ + |F^\alpha_n(\psi^n(A, x \geq \omega)) - F^\alpha_n(A, x \geq \omega)| \]
\[ \leq |F^\alpha_n(A, x \leq \omega^n) - F^\alpha_n(A, x \leq \omega^n)| \]
\[ + \frac{1}{\lambda_n^{1/2}} + |F^\alpha_n(\omega^n -) - F^\alpha_n(\omega^n -)| + |F^\alpha_n(\omega^n -) - F^\alpha_n(\omega^n)| \]
\[ + |\Delta F^\alpha_n(\psi^n(\omega)) - \Delta F^\alpha_n(\omega)|. \]
Then the convergence in total variation of $F^n_{\lambda}(\psi^n(\cdot))$ to $F_{\lambda}$ follows from the convergence in total variation of $F^n_{\lambda}$ to $F_{\lambda}$ on $[0, \omega)$ (from the condition (21)), $F^n(\omega^n) \to F(\omega)$ and the fact that $|\Delta F^n_{\lambda}(\psi^n(\omega)) - \Delta F_{\lambda}(\omega)| \to 0$.

Let $D([0, \infty), D[0, \infty])$ denote the space of all $D[0, \infty]$-valued right continuous function with left limits defined on $[0, \infty)$; see Talreja and Whitt (2009) for detailed analysis on this space. Introduce a sequence of mappings $\Psi^n : D([0, \infty), D[0, \infty]) \to D[0, \infty)$ and $\Psi : D([0, \infty), D[0, \infty)) \to D[0, \infty)$ by

$$\Psi^n(z)(t) = \int_0^t \frac{z(t)}{1 - F^n_{\lambda}(\psi^n(x))} dF^n_{\lambda}(\psi^n(x)), \ t \geq 0,$$

$$\Psi(z)(t) = \int_0^t \frac{z(t)}{1 - F_{\lambda}(x)} dF_{\lambda}(x), \ t \geq 0,$$

for any $z \in D([0, \infty), D[0, \infty))$. Then for any $T \geq 0$,

$$\sup_{0 \leq t \leq T} |\Psi^n(z^n)(t) - \Psi(z)(t)|$$

$$\leq \sup_{0 \leq t \leq T} \left| \int_0^t \frac{z^n(t)}{1 - F^n_{\lambda}(\psi^n(x))} d\left( F^n_{\lambda}(\psi^n(x)) - F_{\lambda}(x) \right) \right|$$

$$+ \sup_{0 \leq t \leq T} \left| \int_0^t \left( \frac{z^n(t)}{1 - F^n_{\lambda}(\psi^n(x))} - \frac{z(t)}{1 - F_{\lambda}(x)} \right) dF_{\lambda}(x) \right|.$$

The first term on the right-hand side in the above converges to zero because $F^n_{\lambda}(\psi^n(\cdot))$ converges to $F_{\lambda}(\cdot)$ in total variation. If $z^n$ converges to $z$ which is continuous in both variables, then the second term converges to zero because

$$\sup_{0 \leq t \leq T} |\psi^n(t) - t| \to 0 \text{ and } \sup_{0 \leq t \leq T} |F^n_{\lambda}(\psi^n(t)) - F_{\lambda}(t)| \to 0 \text{ as } n \to \infty.$$ 

Note that

$$G^n(\psi^n(t)) = \int_0^{\psi^n(t)} \frac{V^n(\psi^n(x) - x, -x)}{1 - F^n_{\lambda}(x)} dF^n_{\lambda}(x) = \int_0^t \frac{V^n(\psi^n(t) - \psi^n(x), \psi^n(x) - x)}{1 - F^n_{\lambda}(\psi^n(x))} dF^n_{\lambda}(\psi^n(x))$$

can be represented as $G^n(\psi^n(t)) = \Psi^n(U^n(\hat{\Lambda}^n(\cdot), \cdot))(t)$. From the fact that $U^n(\hat{\Lambda}^n(\cdot), \cdot)$ converges to a function which is continuous in both variables, then similar to the argument below (3.26) in Krichagina and Puhalskii (1997) we know that $\{G^n\}_{n \in N}$ is $C$-tight, i.e., the conclusion in their Lemma 3.4 holds for our model.

We next show that the conclusion of their Lemma 3.7 (tightness of $\{H^n\}_{n \in N}$) holds for our model. Define $H^n_k$ the same as their (3.28) with changing their $F(\cdot)$ to $F^n_{\lambda}(\cdot)$. The conclusion of their Lemma 3.5, which shows that $H^n_k$ is a square-integrable martingale, still holds because it is for a specific system and no limit is involved. Based on Lemma 3.5, the arguments for proving their Lemma 3.7 still work after we change their $F(\cdot)$ to $F^n_{\lambda}(\cdot)$ (i.e., for our model). This is because the arguments for their (3.60) and thereafter still hold for $F^n_{\lambda}(\cdot)$, except that we need to use the weak law of large numbers for the triangular array $\frac{1}{\lambda_n} \sum_{i=1}^{\lfloor \lambda_n t \rfloor} \int_0^{\psi^n(u)} \frac{dF^n_{\lambda}(u)}{1 - F^n_{\lambda}(u-)}$, where $\lambda_n$ is the average arrival rate.
The weak law of large numbers for triangular array can be applied because we can bound its second moment:

\[
\mathbb{E} \left[ \left( \int_0^{i/n} \frac{dF^n_\Lambda(x)}{1 - F^n_\Lambda(x-)} \right)^2 \right] = \int_0^{\infty} \int_0^{x} \frac{dF^n_\Lambda(x_1)}{1 - F^n_\Lambda(x_1-)} \frac{dF^n_\Lambda(x_2)}{1 - F^n_\Lambda(x_2-)} dF^n_\Lambda(x)
\]

Combining (i) and (ii), we have shown that their Lemma 5.3 also holds for our model.

(ii) Finite dimensional distribution convergence. Introduce

\[
M^n_{2k}(t) = \sum_{i=1}^{k} \square U^n((\tilde{\Lambda}^n(s_{i-1}^k), 0), (\tilde{\Lambda}^n(s_i^k), F^n_\Lambda(t-s_i^k))),
\]

with the operator \( \square \) defined in the same way as the equation below their (2.19). Essentially, we change their \( F(\cdot) \) in their (5.18) to \( F^n_\Lambda(\cdot) \). From the definitions, \( F^n_\Lambda(x) = F_\Lambda(x) = 1 \) if \( x \geq \omega^n \vee \omega \).

As a result, \( F^n_\Lambda(t) \to F_\Lambda(t) \) for \( t > \omega \). From (21) and \( \omega^n \to \omega \), \( F^n_\Lambda(t) \to F_\Lambda(t) \) for \( t < \omega \). As a result, for all \( t \neq \omega \), \( F^n_\Lambda(t) \to F_\Lambda(t) \). In the proof of (a) and (b) on their page 270, it is enough to consider \( \{s_i^k\} \) such that \( t - s_i^k \neq \omega \) for all \( i \geq 1 \) and \( k \). Then the convergence of their (5.17) holds by also using the convergence of \( F^n_\Lambda(t - s_i^k) \) to \( F_\Lambda(t - s_i^k) \) and \( \tilde{\Lambda}^n(s_{i-1}^{k-1}) \) to \( \tilde{\Lambda}(s_{i-1}^{k-1}) \). As a result, their (a) on page 270 still holds. For (b) on their page 270, their (5.20) and the argument on the top of page 273 still works. Thus, we have the convergence of finite dimensional distribution.

Combining (i) and (ii), we have shown that their Lemma 5.3 also holds for our model.