MATHEMATICS OF OPERATIONS RESEARCH
Vol. 35, No. 2, May 2010, pp. 347–362
ISSN 0364-765X | EISSN 1526-5471 | 10 | 3502 | 0347

DOI 10.1287/moor.1100.0443

Customer Abandonment in Many-Server Queues

J. G. Dai, Shuangchi He
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
{dai@gatech.edu, heshuangchi@gatech.edu}

We study $G/G/n+GI$ queues in which customer patience times are independent, identically distributed following a general distribution. When a customer's waiting time in queue exceeds his patience time, the customer abandons the system without service. For the performance of such a system, we focus on the abandonment process and the queue length process. We prove that under some conditions, a deterministic relationship between the two stochastic processes holds asymptotically under the diffusion scaling when the number of servers n goes to infinity. These conditions include a minor assumption on the arrival processes that can be time-nonhomogeneous and a key assumption that the sequence of diffusion-scaled queue length processes, indexed by n, is stochastically bounded. We also establish a comparison result that allows one to verify the stochastic boundedness by studying a corresponding sequence of systems without customer abandonment.

Key words: multiserver queues; customer abandonment; many-server heavy traffic; Halfin-Whitt regime; quality- and efficiency-driven regime
MSC2000 subject classification: Primary: 60K25; Secondary: 68M20, 90B22
OR/MS subject classification: Primary: queues–limit theorems, queues–balking and reneging, queue–transient results; secondary: probability–distribution comparisons, probability–stochastic model applications
History: Received April 3, 2009; revised December 31, 2009. Published online in *Articles in Advance* April 30, 2010.

1. Introduction. A $G/G/n$ queue is a classic stochastic system that has been extensively studied in literature (see, for example, Borovkov [4], Iglehart and Whitt [11], Kiefer and Wolfowitz [13], among others). In such a system, there are n identical servers. The customer arrival process to the system is assumed to be general (the first G in the $G/G/n$ notation). Upon his arrival to the system, a customer gets into service immediately if an idle server is available; otherwise, he waits in a buffer with infinite waiting room that holds a first-in-first-out (FIFO) queue. The service times are assumed to be general (the second G), forming an arbitrary sequence of nonnegative random variables. When a server finishes serving a customer, the server takes the leading customer from the waiting buffer; when the queue is empty, the server begins to idle. A $G/G/n$ queue is also referred to as a parallel server queue. Such a queue has been used extensively to model a customer call center (see, for example, survey papers by Aksin et al. [1], Gans et al. [8]).

As pointed out by Garnett et al. [9], customer abandonment is a key factor for call center operations. Also, the customer arrival process to a call center is typically nonhomogeneous in time (see, for example, Brown et al. [5]). This paper studies parallel server queues that allow for both time-nonhomogeneous arrival processes and customer abandonment. In our model, each customer has a patience time; when a customer's waiting time in the queue exceeds his patience time, the customer abandons the system without any service. If the patience times are general, the resulting system is referred to as a $G/G/n+G$ queue. If we further assume that the patience times are independent and identically distributed (iid), it is referred to as a $G/G/n+GI$ queue. The interarrival times and the service times are assumed to be general, without the iid assumptions.

Let $Q(t)$ be the number of customers waiting in queue at time t and $G(t)$ be the cumulative number of customers who have abandoned the system by time t. The purpose of this paper is to establish an asymptotic relationship between the *queue length process* $Q=\{Q(t);\ t\geq 0\}$ and the *abandonment process* $G=\{G(t);\ t\geq 0\}$ in a $G/G/n+GI$ queue when the number of servers n is large.

To motivate such a relationship, consider an $M/M/n+M$ queue in which the sequences of interarrival times, service times, and patience times are all iid and each sequence follows an exponential distribution. Each customer waiting in the queue abandons the system at rate $\alpha\geq 0$. Because of the memoryless property of an exponential distribution, one can argue that, with probability one,

$$G(t)=N\left(\alpha\int_0^t Q(s)\,ds\right)\quad\text{for all } t\geq 0, \tag{1}$$

where $N=\{N(t);\ t\geq 0\}$ is a Poisson process with unity rate.

To further simplify relationship (1), we focus on systems with high arrival rates, following the pioneering work of Halfin and Whitt [10]. Specifically, we consider a sequence of $M/M/n+M$ systems indexed by the number of servers n, each having a homogeneous Poisson arrival process. For the nth system, its arrival rate λ^n depends on n. The arrival rate $\lambda^n\to\infty$ as $n\to\infty$ whereas the service time and the patience time distributions

do not change with n. We use $1/\mu$ to denote the mean service time of each customer and define the traffic intensity of the nth system as $\rho^n = \lambda^n/(n\mu)$. We assume that

$$\lim_{n\to\infty} \sqrt{n}(1-\rho^n) = \beta \quad \text{for some } \beta \in \mathbb{R}. \tag{2}$$

When Condition (2) holds, the sequence of systems is said to be in the *Halfin-Whitt regime* or in the *quality- and efficiency-driven* (*QED*) *regime*.

When the systems are in the Halfin-Whitt regime, one can prove from relationship (1) that for each $T > 0$,

$$\frac{1}{\sqrt{n}} \sup_{0\le t\le T} \left| G^n(t) - \alpha \int_0^t Q^n(s)\, ds \right| \to 0 \quad \text{in probability as } n \to \infty. \tag{3}$$

The main result (Theorem 2.1) of this paper is to prove that the asymptotic relationship (3) holds for a sequence of $G/G/n + GI$ queues with general arrival processes that can be time-nonhomogeneous, assuming that the sequence of diffusion-scaled queue length processes is stochastically bounded.

The heavy traffic condition (2) implies that the sequence of queues is critically loaded in the limit. It is often used to prove stochastic boundedness for the diffusion-scaled queue length processes. However, Condition (2) is not necessary for the stochastic boundedness result. For example, when the sequence of $M/M/n+M$ systems is underloaded (namely, $\lim_{n\to\infty} \rho^n < 1$), the stochastic boundedness still holds. Our main theorem, Theorem 2.1, assumes stochastic boundedness for the sequence of diffusion-scaled queue length processes. The heavy traffic condition (2) is *not* used in the rest of this paper. For a particular sequence of $G/G/n + G$ systems, proving the stochastic boundedness result is by no means easy. The second theorem of this paper, Theorem 2.2, is a comparison result showing that the queue length at any time in a $G/G/n+G$ queue is dominated by the queue length in a corresponding $G/G/n$ queue with longer service times and no abandonment. The comparison result implies that it is sufficient to prove stochastic boundedness for the diffusion-scaled queue length processes in a sequence of $G/G/n$ queues without abandonment.

In Theorem 2.1, α in (3) is replaced by the density (as the right derivative) at zero of the patience time distribution. Under the stochastic boundedness assumption on diffusion-scaled queue length processes, customer waiting times will be proved to converge to zero as $n \to \infty$. Thus, customer abandonment rarely happens when n is large; only those customers who have extremely small patience times can possibly abandon the system. Therefore, the patience time distribution, outside a small neighborhood of zero, barely has any influence on system dynamics. Zeltyn and Mandelbaum [22] observe the same phenomenon and study steady-state quantities for $M/M/n + GI$ queues in the Halfin-Whitt regime. In their results, the patience time distribution affects the limiting quantities only through its density at zero. Recently, Reed and Tezcan [18] refine and generalize the results of Zeltyn and Mandelbaum [22] to the $GI/M/n + GI$ model. By focusing on the patience time distribution on a neighborhood of zero rather than the origin itself, Reed and Tezcan [18] give an improved approximation for the steady-state performance measures.

The asymptotic relationship (3) is the key to proving a many-server heavy traffic limit theorem for a sequence of $G/Ph/n + GI$ queues in Dai et al. [6], where the service times follow a phase-type distribution. The limit theorem in Dai et al. [6] generalizes the result of Puhalskii and Reiman [16] for $G/Ph/n$ queues without abandonment. By using a continuous mapping approach, Dai et al. [6] first proves a heavy traffic limit for $G/Ph/n + M$ queues when either the patience times are exponentially distributed or there is no customer abandonment (corresponding to the case $\alpha = 0$). As a consequence, the stochastic boundedness assumption on the diffusion-scaled queue length processes holds for $G/Ph/n$ queues, and thus Theorem 2.2 in this paper implies that the stochastic boundedness assumption holds for $G/Ph/n+GI$ queues as well.

For the $G/GI/n$ model, Reed [17] proves a many-server limit theorem for one-dimensional customer count processes in the Halfin-Whitt regime. Following the framework of Reed [17], Mandelbaum and Momčilović [14] generalize the limit theorem to the $G/GI/n + GI$ model. Although the current paper and the work of Mandelbaum and Momčilović [14] are contemporary, independent studies, there is a significant overlap between these two papers. For example, Corollary 3 of their paper gives a relationship between the abandonment and the queue length processes; their relationship is similar to our relationship (3). Their Proposition 1, similar to our Theorem 2.2, gives a comparison between queues with and without abandonment. Also, Corollary 1 in their work is similar to Proposition 4.1 in our paper.

The two papers, however, differ significantly both philosophically and in terms of assumptions and proof techniques. We believe that our results have laid a framework for a modular approach to proving many-server limit theorems for queues with customer abandonment in the Halfin-Whitt regime: First, prove a limit theorem for queues without customer abandonment using a continuous mapping approach. Then, use our asymptotic

relationship (3) and a modified map to prove a corresponding limit theorem for queues with abandonment. This modular approach is carried out in Dai et al. [6]. Further, we believe that the limit theorem of Mandelbaum and Momčilović [14] for one-dimensional customer count processes can be proved in a simpler approach using our Theorem 2.1 and the limit theorem of Reed [17]. Indeed, using the limit theorem of Reed [17] as well as our comparison result, we can readily see that the stochastic boundedness assumption is satisfied for the $G/GI/n+GI$ queues in the Halfin-Whitt regime. Recently, Kaspi and Ramanan [12] report a measure-valued heavy traffic limit for $G/GI/n$ queues. It is expected that our Theorem 2.1 can be used to generalize their result to the $G/GI/n+GI$ model. Besides these philosophical differences, as the main theorem of this paper, our Theorem 2.1 differs from Corollary 3 of Mandelbaum and Momčilović [14] in the following aspects. First, their corollary is stated as a weak convergence result whereas our asymptotic relationship (3) is a stronger result at a sample path level. Second, their corollary assumes iid service times whereas we assume nothing on service times as long as the stochastic boundedness assumption holds. Third, we impose much weaker assumptions on arrival processes. They assume that each arrival process in the sequence has a time-homogeneous arrival rate and the sequence of arrival processes satisfies a certain functional central limit theorem. In contrast, we assume (10) and (11) for the arrival processes that allow for nonhomogeneous arrival rates and batch arrivals; see §2.2. The latter two features often exclude a functional central limit for the arrival processes. Of course, we need the stochastic boundedness for the diffusion-scaled queue length processes; the assumption implicitly requires that the sequence of queues is not overloaded in the limit.

A key insight in this paper as well as in Mandelbaum and Momčilović [14] is that the exact distribution of patience times is irrelevant in the Halfin-Whitt regime as long as customer abandonment is explicitly built into the model. This phenomenon is in sharp contrast to the one found in Whitt [20] when systems are operated in an overloaded regime known as the *efficiency-driven* (*ED*) *regime*; the system performance there depends crucially upon the patience time distribution and a fluid model is shown to be able to capture that dependency. In particular, Bassamboo and Randhawa [3] demonstrate that for $M/M/n+GI$ queues with certain performance measures and patience time distributions, the optimized staffing levels surprisingly drive the queues to the overloaded regime. In such a case, a fluid model provides accurate approximations for performance measures; the approximation error does not increase with the system size n.

The remainder of the paper is organized as follows. Our main results are presented in §2. Section 3 is dedicated to several preliminary but more general results for $G/G/n+G$ queues. The proof of Theorem 2.2 can also be found in §3. The detailed proof of Theorem 2.1 is given in §4 and focuses on $G/G/n+GI$ queues with iid patience times. We leave the proof of Lemma 3.1 to the appendix.

1.1. Notation. All random variables and processes are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We reserve $\mathbb{E}[\cdot]$ for expectation. The symbols $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}_+$ are used to denote the sets of integers, nonnegative integers, positive integers, real numbers, and nonnegative real numbers, respectively. The space of functions $f\colon \mathbb{R}_+ \to \mathbb{R}$ that are right-continuous on $[0,\infty)$ and have left limits on $(0,\infty)$ is denoted by $\mathbb{D}$, which is endowed with the Skorohod J_1-topology (see, for example, Ethier and Kurtz [7]). For $f \in \mathbb{D}$, $f(t-)$ denotes the left limit of f at $t>0$; given another $f^* \in \mathbb{D}$ that is nondecreasing and takes values in $\mathbb{R}_+$, $f \circ f^*$ denotes the composed function in $\mathbb{D}$ with $(f \circ f^*)(t) = f(f^*(t))$ for $t \geq 0$. For a sequence of random variables (or processes) $\{Z^n;\ n \in \mathbb{N}\}$ taking values in $\mathbb{R}$ (or $\mathbb{D}$), we write $Z^n \Rightarrow Z$ to denote the convergence of Z^n to Z in distribution, where Z is a random variable on $\mathbb{R}$ (or a process on $\mathbb{D}$). For an index set J and a set of random variables $\{Y_j;\ j \in J\}$, $\sigma\{Y_j;\ j \in J\}$ is the σ-field generated by $\{Y_j;\ j \in J\}$. For $x, z \in \mathbb{R}$, $\lfloor x \rfloor = \max\{j \in \mathbb{Z}\colon j \leq x\}$, $x \vee z = \max\{x, z\}$, $x \wedge z = \min\{x, z\}$, and $x^+ = \max\{x, 0\}$. We use 1_S to denote the indicator function of a set $S \subset \Omega$.

2. Heavy traffic setting and main results. In §2.1, we define a $G/G/n+G$ queue for a fixed $n \in \mathbb{N}$ using a sequence of primitive random variables. In §2.2, we introduce a sequence of $G/G/n+GI$ queues and the stochastic boundedness assumption on diffusion-scaled queue length processes. In §2.3, we state two theorems.

2.1. A $G/G/n+G$ queue. To define a $G/G/n+G$ queue, we are given a sequence of primitive random variables $\{\tau_i,\ v_i,\ \gamma_i;\ i \in \mathbb{Z}\}$. For each sample path $\omega \in \Omega$, let

$$X(0,\omega) = \inf\{i \geq 0\colon v_j(\omega) = 0 \text{ for all } j \leq -i\}.$$

We assume that $X(0,\omega) < \infty$ on each sample path ω. The integer $X(0,\omega)$ is interpreted as the number of total customers who are initially in the system. Letting

$$Q(0,\omega) = (X(0,\omega) - n)^+,$$

$Q(0,\omega)$ is interpreted as the number of customers who are waiting in queue at time zero. Thus, customers $i = 1 - X(0,\omega), \ldots, 0$ are in the system initially, with customers $i = 1 - Q(0,\omega), \ldots, 0$ waiting in queue.

We assume $\tau_i(\omega) \le \tau_{i+1}(\omega)$ for each $\omega \in \Omega$ and each $i \in \mathbb{Z}$. One interprets $\tau_i(\omega)$ as the arrival time of the ith customer. We further assume that for each $\omega \in \Omega$, $\tau_1(\omega) > 0$ and $\tau_i(\omega) = 0$ for all $i \le 0$. Thus, by time zero, all customers with indices $i \le 0$ have arrived at the system. $\tau_1(\omega)$ is the arrival time of the first customer after time zero. For $t \ge 0$, let

$$E(t) = \sup\{i \in \mathbb{Z}_+ \colon \tau_i \le t\}. \tag{4}$$

Clearly, $E(t)$ is the number of customers who arrive at the system during $(0, t]$.

For each $i \in \mathbb{Z}$, $v_i(\omega) \ge 0$. One interprets $v_i(\omega)$ as the service time of the ith customer if he has not started his service by time zero or as his remaining service time at time zero if he has started service. For $i \ge 1$, $\gamma_i(\omega) \ge 0$ is interpreted as the patience time of the ith customer. For customer i who is waiting in queue at time zero, $\gamma_i(\omega) > 0$ is interpreted as the remaining patience time of the customer. For customer i who has entered service or abandoned the system by time zero, $\gamma_i(\omega)$ can take any value. By convention, we set $\gamma_i(\omega) = -1$ when $i \le -Q(0,\omega)$. To keep track of the history of the $G/G/n+G$ queue, we define a filtration $\{\mathcal{F}_i;\ i \in \mathbb{Z}_+\}$ by

$$\mathcal{F}_i = \sigma\{\tau_{j+1},\ v_j,\ \gamma_j;\ j \le i\}. \tag{5}$$

Most of this paper studies $G/G/n+GI$ queues. In such a queue, the sequence of patience times $\{\gamma_i;\ i \in \mathbb{N}\}$ is assumed to be iid.

2.2. Asymptotic framework on $G/G/n+GI$ queues. We consider a sequence of $G/G/n+GI$ queues indexed by the number of servers n. We add a superscript n to the primitive random variables of the nth system and use $\{\mathcal{F}_i^n;\ i \in \mathbb{Z}_+\}$ to denote the associated filtration where

$$\mathcal{F}_i^n = \sigma\{\tau_{j+1}^n,\ v_j^n,\ \gamma_j^n;\ j \le i\}. \tag{6}$$

We assume that

$$\gamma_{i+1}^n \text{ is independent of } \mathcal{F}_i^n \quad \text{for each } i \in \mathbb{Z}_+ \tag{7}$$

and that $\{\gamma_i^n;\ i \in \mathbb{N}\}$ is a sequence of iid random variables with distribution function F that does not change with n. Recall that for $i \ge 1$, γ_i^n is the patience time of the ith customer who arrives after time zero at the nth system. The preceding assumption states that the distribution of these patience times does not depend on the number of servers; this assumption seems reasonable in many cases. For $i \le 0$, γ_i^n is the remaining patience time of a customer who is initially waiting in queue in the nth system. This remaining patience time may depend on how long the customer has been waiting by time zero, and this waiting time in turn may depend on the number of servers n. We further assume that the distribution F satisfies

$$F(0) = 0 \tag{8}$$

and is right-differentiable at zero with right derivative

$$\alpha = \lim_{x \downarrow 0} x^{-1} F(x) < \infty. \tag{9}$$

The arrival process of the nth system is $E^n = \{E^n(t);\ t \ge 0\}$, where $E^n(t)$, defined in (4), denotes the number of customer arrivals in $(0, t]$. The fluid-scaled arrival process $\bar{E}^n$ is defined by

$$\bar{E}^n(t) = \frac{1}{n} E^n(t).$$

The following two assumptions are made upon the arrival processes. First, given an arbitrary $T > 0$, there exists a constant $c_T > 0$ such that

$$\lim_{n \to \infty} \mathbb{P}\left[\inf_{0 \le t \le T} \{\bar{E}^n(t+\delta) - \bar{E}^n(t)\} < \delta c_T\right] = 0 \quad \text{for all } \delta > 0. \tag{10}$$

Second, the sequence of fluid-scaled arrival processes is stochastically bounded, that is, for each $T > 0$,

$$\lim_{a \to \infty} \limsup_{n \to \infty} \mathbb{P}[\bar{E}^n(T) > a] = 0. \tag{11}$$

Roughly speaking, Condition (10) says that when n is large, the number of customer arrivals should be at least $n\delta c_T$ during the time interval $(t, t+\delta]$ for any $t \in [0, T]$. Assumptions (10) and (11) impose very mild constraints on the arrival processes. Clearly, they allow each arrival process E^n to have a time-nonhomogeneous arrival rate.

Recall the queue length process Q^n and the abandonment process G^n of the nth system. We define their respective diffusion-scaled versions $\tilde{Q}^n$ and $\tilde{G}^n$ via

$$\tilde{Q}^n(t) = \frac{1}{\sqrt{n}} Q^n(t) \qquad \text{and} \qquad \tilde{G}^n(t) = \frac{1}{\sqrt{n}} G^n(t).$$

For the main result (Theorem 2.1) of this paper, the key assumption is that the sequence of diffusion-scaled queue length processes is stochastically bounded, namely, for each $T > 0$,

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}\left[\sup_{0\le t\le T} \tilde{Q}^n(t) > a \right] = 0. \tag{12}$$

For Theorem 2.1, we also need to make an assumption on the initial condition. Let G_0^n be the number of customers who are waiting in queue at time zero but will eventually abandon the system. Let

$$\tilde{G}_0^n = \frac{1}{\sqrt{n}} G_0^n.$$

We assume that

$$\tilde{G}_0^n \;\Rightarrow\; 0 \quad \text{as } n \to \infty. \tag{13}$$

2.3. Main results. We state two theorems in this section. The first theorem is the main result of this paper. It says that the asymptotic relationship (3) holds for a sequence of $G/G/n+GI$ queues under certain conditions.

THEOREM 2.1. *Consider a sequence of $G/G/n+GI$ queues that satisfies (7)–(11). Assume that the sequence of diffusion-scaled queue length processes is stochastically bounded and the sequence of queues satisfies the initial condition (13). Then, the asymptotic relationship (3) holds for each $T > 0$.*

The proof of Theorem 2.1 will be presented in §4. All assumptions in Theorem 2.1 are standard except for the stochastic boundedness assumption (12). Verifying this assumption can be a significant task.

We now present the second theorem. The theorem, referred to as the comparison result, shows that the queue length at any time in a $G/G/n+G$ queue is dominated by the queue length in a corresponding $G/G/n$ queue with longer service times and no abandonment. This comparison result implies that to verify the stochastic boundedness assumption (12) for a sequence of $G/G/n+GI$ queues, it is sufficient to prove stochastic boundedness for the queue length processes in corresponding $G/G/n$ queues without abandonment.

To state Theorem 2.2, we consider two FIFO queues: a $G/G/n+G$ queue denoted by $\Sigma^{(1)}$ and a $G/G/n$ queue denoted by $\Sigma^{(2)}$. For $l = 1, 2$, we add a superscript (l) to the primitive random variables and performance processes of $\Sigma^{(l)}$. We assume that all servers in both systems are identical, the arrival processes to both queues are identical, and, at time zero, there are $X(0)$ customers in each system indexed by $i = 1 - X(0), \ldots, 0$. Recall that $v_i^{(l)}$ is the service time of the ith customer if he has not started service by time zero or is his remaining service time at time zero if he has started service. We further assume that

$$v_i^{(1)} \le v_i^{(2)} \quad \text{for all } i \ge 1 - X(0). \tag{14}$$

In short, there are two differences between the queues. First, each customer in $\Sigma^{(1)}$ has an equal or shorter service time than the corresponding customer in $\Sigma^{(2)}$. Second, customers in $\Sigma^{(1)}$ can possibly abandon the system whereas those in $\Sigma^{(2)}$ cannot.

THEOREM 2.2. *Let $Q^{(1)}(t)$ and $Q^{(2)}(t)$ be the respective numbers of customers waiting in queue in $\Sigma^{(1)}$ and $\Sigma^{(2)}$ at time $t \ge 0$. Then, on each sample path,*

$$Q^{(1)}(t) \le Q^{(2)}(t) \quad \textit{for all } t \ge 0.$$

The proof of Theorem 2.2 is given in §3.2.

3. Preliminary results on $G/G/n+G$ queues. In this section, we consider $G/G/n+G$ queues where patience times are not assumed to be iid. In such a queue, the interarrival times, the service times, and the patience times are three arbitrary sequences of nonnegative random variables. In §3.1, we first rigorously define offered waiting times and virtual waiting times. The offered waiting time of each customer is shown to be measurable in Lemma 3.1. In Lemma 3.2, these times are shown to be related at the arrival time of each customer. We then define nominal service starting times. These nominal times are shown to be ordered in the FIFO fashion in Lemma 3.3. A relationship among the offered waiting times, the patience times, and the queue length process is presented in Lemma 3.4. These lemmas will be used in the proof of Theorem 2.1 in §4 when the patience times are specialized to be iid.

3.1. Offered and virtual waiting times. First, we introduce two notions: *offered waiting times* and *virtual waiting times* (see Baccelli [2] and Stanford [19] for discussions on them in single-server queues). In a $G/G/n+G$ queue, for each $i\in\mathbb{Z}$, we use w_i to denote the offered waiting time of the ith customer. For $i\geq 1$, w_i is the amount of time he would have to wait in queue until getting into service if his patience were infinite. For $1-Q(0)\leq i\leq 0$, the ith customer is waiting in queue at time zero and w_i is his remaining waiting time if he has infinite patience. To define an offered waiting time mathematically, it is convenient to introduce the *remaining service time process* $r_i=\{r_i(t);\ t\geq 0\}$ for $i\in\mathbb{Z}$, where $r_i(t)$ is the remaining service time for the ith customer at time t. Fix $\omega\in\Omega$. For each $i\leq -X(0,\omega)$, let $w_i(\omega)=0$ and $r_i(t,\omega)=0$ for all $t\geq 0$, and for $i>-X(0,\omega)$, let

$$w_i(\omega)=\inf\left\{t\geq 0\colon \sum_{j\leq i-1} 1_{\{r_j(\tau_i(\omega)+t,\,\omega)>0\}}<n\right\},\tag{15}$$

and

$$r_i^a(t,\omega)=1_{\{t<\tau_i(\omega)+\gamma_i(\omega)\}}v_i(\omega),\tag{16}$$

$$r_i^s(t,\omega)=1_{\{t<\tau_i(\omega)+w_i(\omega)\}}v_i(\omega)+1_{\{t\geq\tau_i(\omega)+w_i(\omega)\}}(v_i(\omega)-t)^+,\tag{17}$$

$$r_i(t,\omega)=1_{\{0\leq\gamma_i(\omega)\leq w_i(\omega)\}}r_i^a(t,\omega)+(1-1_{\{0\leq\gamma_i(\omega)\leq w_i(\omega)\}})r_i^s(t,\omega)\tag{18}$$

for $t\geq 0$. Equation (15) says that if no arrival occurs after the $(i-1)$st customer, w_i is the amount of time beyond τ_i until one of the n servers becomes idle. Equation (18) says that the ith customer will abandon the queue if $0\leq\gamma_j\leq w_j$ and, in this case, his remaining service time at time t is given by $r_i^a(t)$. Otherwise, he either has received or will receive service and his remaining service time at time t is $r_i^s(t)$. Clearly, recursions (15)–(18) define $w_i(\omega)$ for each $\omega\in\Omega$ and $i\in\mathbb{Z}$.

Our first lemma demonstrates the measurability of each offered waiting time. We leave its proof to the appendix.

Lemma 3.1. *For a $G/G/n+G$ queue, w_i is $\mathcal{F}_k$-measurable for $k\in\mathbb{Z}_+$ and $i\leq k+1$, where the filtration $\{\mathcal{F}_k;\ k\in\mathbb{Z}\}$ is defined by (5).*

For the $G/G/n+G$ queue, we use $W(t)$ to denote its virtual waiting time at time $t\geq 0$. One interprets $W(t)$ as the amount of time a hypothetical customer would have to wait in queue had he arrived at time t with infinite patience. Given $X(0)$, the number of total customers initially in the system, the virtual waiting time at time t can be defined by

$$W(t)=\inf\left\{s\geq 0\colon \sum_{i=1-X(0)}^{E(t)} 1_{\{r_i(t+s)>0\}}<n\right\}.\tag{19}$$

We call $W=\{W(t);\ t\geq 0\}$ the *virtual waiting time process*. The following lemma relates offered waiting times to virtual waiting times at corresponding arrival times.

Lemma 3.2. *For a $G/G/n+G$ queue,*

$$W(\tau_i-)\leq w_i\leq W(\tau_i)\quad \text{for } i\geq 1$$

and

$$w_i\leq W(0)\quad \text{for } i\leq 0.$$

Proof. Let $y(t)=\inf\{s\geq 0\colon \sum_{i=1-X(0)}^{E(t-)} 1_{\{r_i(s)>0\}}<n\}$. Then, for any $t'\in[\tau_{E(t-)},t)$, because $E(t')=E(t-)$, using (19) we have $t'+W(t')=t'\vee y(t)$. Thus, $W(t')=(y(t)-t')^+$ and $W(t-)=(y(t)-t)^+$. Because $E(\tau_i-)<i\leq E(\tau_i)$, it follows from (15) and (19) that $W(\tau_i-)\leq w_i\leq W(\tau_i)$. In particular, $W(\tau_i-)=w_i$ if exactly one customer arrives at time τ_i. Using $E(0)=0$ and $\tau_i=0$ for $i\leq 0$, $w_i\leq W(0)$ also follows from (15) and (19). □

Note that the ith customer would begin his service at time $\tau_i + w_i$ if he would not abandon the queue. We call $\tau_i + w_i$ the ith customer's *nominal service starting time*. It follows from (15) that

$$\tau_i + w_i = \inf\left\{s \geq \tau_i \colon \sum_{j \leq i-1} 1_{\{r_j(s)>0\}} < n\right\}. \tag{20}$$

Similarly, we call $t + W(t)$ the nominal service starting time for a customer arriving at time t. It can be written as

$$t + W(t) = \inf\left\{s \geq t \colon \sum_{i=1-X(0)}^{E(t)} 1_{\{r_i(s)>0\}} < n\right\}. \tag{21}$$

Lemma 3.3 states that although customer abandonment is involved, the nominal service starting times are still ordered in the FIFO fashion as in a $G/G/n$ queue without abandonment.

LEMMA 3.3. *For a $G/G/n+G$ queue,*

$$t_1 + W(t_1) \leq t_2 + W(t_2) \quad \textit{for } 0 \leq t_1 \leq t_2$$

and

$$\tau_i + w_i \leq \tau_j + w_j \quad \textit{for any } i, j \in \mathbb{Z} \textit{ with } i \leq j.$$

PROOF. It follows from (16)–(18) that the process $\{1_{\{r_i(t)>0\}};\ t \geq 0\}$ is right-continuous for all $i \in \mathbb{Z}$. Hence, $\sum_{i=1-X(0)}^{E(t)} 1_{\{r_i(t+W(t))>0\}} < n$. If there exists $0 \leq t_1 \leq t_2$ such that $t_1 + W(t_1) > t_2 + W(t_2)$,

$$n \leq \sum_{i=1-X(0)}^{E(t_1)} 1_{\{r_i(t_2+W(t_2))>0\}} \leq \sum_{i=1-X(0)}^{E(t_2)} 1_{\{r_i(t_2+W(t_2))>0\}} < n$$

by (21), which yields a contradiction. Thus, $t_1 + W(t_1) \leq t_2 + W(t_2)$. Using (20), we can prove $\tau_i + w_i \leq \tau_j + w_j$ for $i \leq j$ by a similar argument. □

Lemma 3.4 establishes a pair of inequalities. The inequalities will later allow us to convert a summation of offered waiting times into an integral of the queue length process.

LEMMA 3.4. *For a $G/G/n+G$ queue,*

$$\int_0^t Q(s)\,ds \leq \sum_{i=1-Q(0)}^{E(t)} (\gamma_i \wedge w_i) \leq \int_0^{t+W(t)} Q(s)\,ds \quad \textit{for all } t \geq 0.$$

PROOF. We first observe that for $i \geq 1 - Q(0)$, the ith customer spends $\gamma_i \wedge w_i$ units of time waiting in queue. For $t \geq 0$, let

$$b_i(t) = \begin{cases} 1 & \text{if the } i\text{th customer is waiting in queue at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Then, $q_i(t) = \int_0^t b_i(s)\,ds$ is the ith customer's cumulative waiting time by t. Note that $q_i(t) \leq \gamma_i \wedge w_i$; if the ith customer has gotten into service or abandoned the queue by t, $q_i(t) = \gamma_i \wedge w_i$ holds. For any $0 \leq s \leq t$, the queue length at time s can be counted by $Q(s) = \sum_{i=1-Q(0)}^{E(t)} b_i(s)$. Then,

$$\int_0^t Q(s)\,ds = \sum_{i=1-Q(0)}^{E(t)} \int_0^t b_i(s)\,ds = \sum_{i=1-Q(0)}^{E(t)} q_i(t) \leq \sum_{i=1-Q(0)}^{E(t)} (\gamma_i \wedge w_i).$$

For $1 - Q(0) \leq i \leq E(t)$, the ith customer should have gotten into service or abandoned the system by time $t + W(t)$ because $\tau_i + w_i \leq t + W(t)$ (see Lemmas 3.2 and 3.3). Then, $q_i(t + W(t)) = \gamma_i \wedge w_i$. It follows that

$$\int_0^{t+W(t)} Q(s)\,ds = \sum_{i=1-Q(0)}^{E(t+W(t))} q_i(t + W(t)) \geq \sum_{i=1-Q(0)}^{E(t)} q_i(t + W(t)) = \sum_{i=1-Q(0)}^{E(t)} (\gamma_i \wedge w_i). \quad \square$$

3.2. Proof of Theorem 2.2. To prove Theorem 2.2, for $l = 1, 2$ and $i \geq 1 - X(0)$, let $r_i^{(l)}(t)$ be the remaining service time of the ith customer in $\Sigma^{(l)}$ at time t. Recall that $E(t)$ is the number of customer arrivals to both queues in $(0, t]$.

Proof of Theorem 2.2. The set of customers being served in $\Sigma^{(l)}$ at time $t \geq 0$ can be represented by

$$\Pi^{(l)}(t) = \left\{ i \in \mathbb{Z}\colon\ 1 - X(0) \leq i \leq E(t),\ r_i^{(l)}(t) > 0,\ \sum_{k=1-X(0)}^{i} 1_{\{r_k^{(l)}(t)>0\}} \leq n \right\}. \tag{22}$$

Set $\xi_0 = 0$ and let $0 < \xi_1 \leq \xi_2 \leq \cdots$ be the service completion times in $\Sigma^{(2)}$. By (14), at time $\xi_0 = 0$, $r_i^{(1)}(\xi_0) \leq r_i^{(2)}(\xi_0)$ for all $i \geq 1 - X(0)$.

Suppose that $r_i^{(1)}(t) \leq r_i^{(2)}(t)$ for $0 \leq t \leq \xi_m$. For any $i \in \Pi^{(2)}(\xi_m)$, (22) implies either $i \in \Pi^{(1)}(\xi_m)$ or $r_i^{(1)}(\xi_m) = 0$. Because for $t \in (\xi_m, \xi_{m+1}]$, $r_i^{(2)}(t) = r_i^{(2)}(\xi_m) - (t - \xi_m) \geq 0$ and $r_i^{(1)}(t) = (r_i^{(1)}(\xi_m) - (t - \xi_m))^+$, then $r_i^{(1)}(t) \leq r_i^{(2)}(t)$. If $i \notin \Pi^{(2)}(\xi_m)$, $r_i^{(1)}(t) \leq r_i^{(2)}(t)$ also holds for $t \in (\xi_m, \xi_{m+1}]$ because $r_i^{(2)}(t) = r_i^{(2)}(\xi_m)$ and $r_i^{(1)}(t) \leq r_i^{(1)}(\xi_m)$. By induction, we get $r_i^{(1)}(t) \leq r_i^{(2)}(t)$ for all $t \geq 0$ and $i \geq 1 - X(0)$.

For $t \geq 0$, let

$$m(t) = \begin{cases} \max\{i \in \Pi^{(2)}(t)\} & \text{if } \Pi^{(2)}(t) \neq \varnothing, \\ E(t) & \text{if } \Pi^{(2)}(t) = \varnothing, \end{cases}$$

which is the index of the last customer being served during $(0, t]$ in $\Sigma^{(2)}$. So, $Q^{(2)}(t) = E(t) - m(t)$ and

$$\sum_{i=1-X(0)}^{m(t)-1} 1_{\{r_i^{(2)}(t)>0\}} < n.$$

Because $r_i^{(1)}(t) \leq r_i^{(2)}(t)$ for each $i \geq 1 - X(0)$, the above inequality leads to

$$\sum_{i=1-X(0)}^{m(t)-1} 1_{\{r_i^{(1)}(t)>0\}} < n,$$

which implies $Q^{(1)}(t) \leq E(t) - m(t)$. Therefore, $Q^{(1)}(t) \leq Q^{(2)}(t)$. □

4. Proof of Theorem 2.1. We present the proof of Theorem 2.1 in this section. The proof is decomposed into three propositions; these propositions will be proved in §§4.2–4.4.

Our attention is now focused on a sequence of $G/G/n + GI$ queues that satisfy Conditions (7)–(13). Let $A^n(t)$ denote, among all customers who have arrived at the nth system by time $t \geq 0$, the number of those who will eventually abandon the queue. The process A^n has a diffusion-scaled version given by

$$\tilde{A}^n(t) = \frac{1}{\sqrt{n}} A^n(t).$$

Our first result is the following proposition showing that A^n and G^n are asymptotically close under diffusion scaling.

Proposition 4.1. *Under the conditions of Theorem 2.1,*

$$\tilde{A}^n - \tilde{G}^n \ \Rightarrow\ 0 \quad \textit{as } n \to \infty.$$

The proof of Proposition 4.1 is presented in §4.4. Given Proposition 4.1, to prove Theorem 2.1 it suffices to show that, for each $T > 0$,

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t) - \alpha \int_0^t \tilde{Q}^n(s)\, ds \right| \to 0 \quad \text{in probability as } n \to \infty. \tag{23}$$

To prove (23), one needs to further analyze the process A^n. For the nth system of the sequence of $G/G/n+GI$ queues, we use $W^n(t)$ and w_i^n to denote the corresponding virtual and offered waiting times. For each customer $i \geq 1 - Q^n(0)$ given his patience time γ_i^n and offered waiting time w_i^n, one can determine whether the customer

will eventually abandon the queue: He will wait γ_i^n units of time and leave the system with no service when $\gamma_i^n \le w_i^n$ or wait w_i^n units of time and get into a server otherwise. This implies the following expression:

$$A^n(t) = \sum_{i=1-Q^n(0)}^{E^n(t)} 1_{\{\gamma_i^n \le w_i^n\}}.$$

Clearly, the process A^n can be decomposed into

$$A^n(t) = G_0^n + A_1^n(t) + A_2^n(t), \tag{24}$$

where

$$G_0^n = \sum_{i=1-Q^n(0)}^{0} 1_{\{\gamma_i^n \le w_i^n\}}$$

is the number of customers who are initially waiting in queue but will eventually abandon the queue,

$$A_1^n(t) = \sum_{i=1}^{E^n(t)} (1_{\{\gamma_i^n \le w_i^n\}} - F(w_i^n)) \qquad \text{and} \qquad A_2^n(t) = \sum_{i=1}^{E^n(t)} F(w_i^n).$$

Defining the diffusion-scaled versions

$$\tilde{A}_1^n(t) = \frac{1}{\sqrt{n}} A_1^n(t) \qquad \text{and} \qquad \tilde{A}_2^n(t) = \frac{1}{\sqrt{n}} A_2^n(t),$$

we have the following two propositions.

PROPOSITION 4.2. *Under the conditions of Theorem 2.1,*

$$\tilde{A}_1^n \;\Rightarrow\; 0 \quad \text{as } n \to \infty.$$

PROPOSITION 4.3. *Under the conditions of Theorem 2.1, for any $T > 0$,*

$$\sup_{0 \le t \le T} \left| \tilde{A}_2^n(t) - \alpha \int_0^t \tilde{Q}^n(s)\, ds \right| \;\Rightarrow\; 0 \quad \text{as } n \to \infty.$$

The proofs of Propositions 4.2 and 4.3 are presented in §§4.2 and 4.3, respectively. Clearly, the proof of Theorem 2.1 follows from (13), (24), and Propositions 4.1–4.3.

4.1. Virtual waiting time processes for $G/G/n + GI$ queues. This section is a preparation for proving Propositions 4.1–4.3 in §§4.2–4.4. The main result here is Proposition 4.4, which says that for the sequence of $G/G/n + GI$ queues, the virtual waiting time processes converge to zero in probability.

PROPOSITION 4.4. *Assume that (8) and (10)–(12) hold. Then,*

$$W^n \;\Rightarrow\; 0 \quad \text{as } n \to \infty.$$

Intuitively, this convergence follows from the following observation. Fix $t > 0$. By assumption (10), for any $\delta > 0$ small enough, there exists a constant $c > 0$ such that when n is large, there are at least $n\delta c$ customer arrivals during the time interval $(t, t+\delta]$. When n is large, all customers who arrived before time t must have entered service or abandoned the system by time $t+\delta$. Otherwise, except for a small abandoned portion, those who arrived during $(t, t+\delta]$ must reside in queue at time $t+\delta$ because of the FIFO discipline, contradicting the stochastic boundedness assumption on the diffusion-scaled queue length processes. Therefore, the virtual waiting time $W^n(t)$ should be no more than δ. Because $\delta > 0$ can be arbitrary, this implies that $W^n(t)$ goes to zero as n goes to infinity.

Before presenting the proof of Proposition 4.4, we give a few corollaries that will be used in later proofs. Define $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ by

$$g(x) = \begin{cases} \alpha & \text{for } x = 0, \\ x^{-1}F(x) & \text{for } x > 0. \end{cases}$$

Under assumptions (8) and (9), g is right-continuous at zero and

$$F(x) = xg(x) \quad \text{for all } x \ge 0.$$

COROLLARY 4.1. *Assume that (8)–(12) hold. Then, for each $T > 0$,*

$$\sup_{1\le i\le E^n(T)} w_i^n \;\Rightarrow\; 0, \tag{25}$$

$$\sup_{1\le i\le E^n(T)} F(w_i^n) \;\Rightarrow\; 0, \tag{26}$$

$$\sup_{1\le i\le E^n(T)} |g(w_i^n)-\alpha| \;\Rightarrow\; 0, \tag{27}$$

$$\sup_{1\le i\le nT} F(w_i^n) \;\Rightarrow\; 0, \tag{28}$$

as $n\to\infty$.

PROOF. First, (25) follows from Lemma 3.2 and Proposition 4.4. Because F is nondecreasing and right-continuous at zero, by the continuous mapping theorem,

$$\sup_{1\le i\le E^n(T)} F(w_i^n) \le F\left(\sup_{1\le i\le E^n(T)} w_i^n\right) \;\Rightarrow\; F(0)=0,$$

which proves (26). For any $\varepsilon>0$, because g is right-continuous at zero, there exists $\delta>0$ such that $|g(x)-\alpha|\le\varepsilon$ for all $0\le x\le\delta$ and thus (27) follows from

$$\limsup_{n\to\infty}\mathbb{P}\left[\sup_{1\le i\le E^n(T)}|g(w_i^n)-\alpha|>\varepsilon\right]\le\limsup_{n\to\infty}\mathbb{P}\left[\sup_{1\le i\le E^n(T)} w_i^n>\delta\right]=0.$$

Also, for any $\varepsilon>0$,

$$\mathbb{P}\left[\sup_{1\le i\le nT}F(w_i^n)>\varepsilon\right]\le\mathbb{P}\left[\sup_{1\le i\le E^n(c_T^{-1}T)}F(w_i^n)>\varepsilon\right]+\mathbb{P}[\bar{E}^n(c_T^{-1}T)<T],$$

where $c_T>0$ is the constant given in (10). Then, (28) follows from (10) and (26). □

To prove Proposition 4.4, we introduce the following processes. For each $\delta>0$, let

$$L_\delta^n(t)=\sum_{i=1}^{\lfloor nt\rfloor}(1_{\{\gamma_i^n\le\delta\}}-F(\delta)) \qquad\text{and}\qquad \bar{L}_\delta^n(t)=\frac{1}{n}L_\delta^n(t). \tag{29}$$

Because the patience times $\{\gamma_i^n;\ i\in\mathbb{N}\}$ are iid, the functional law of large numbers suggests

$$\bar{L}_\delta^n \;\Rightarrow\; 0 \quad\text{as } n\to\infty. \tag{30}$$

For each $\delta>0$, we also define

$$G_\delta^n(t)=\sum_{i=E^n(t)+1}^{E^n(t+\delta)}1_{\{\gamma_i^n\le\delta\}}, \tag{31}$$

which counts the number of customers who arrive at the nth system during $(t,t+\delta]$ but whose patience times are no more than δ. It has a fluid-scaled version given by

$$\bar{G}_\delta^n(t)=\frac{1}{n}G_\delta^n(t).$$

We further introduce the fluid-scaled queue length process $\bar{Q}^n$, given by

$$\bar{Q}^n(t)=\frac{1}{n}Q^n(t),$$

which, by the stochastic boundedness assumption (12), satisfies

$$\bar{Q}^n \;\Rightarrow\; 0 \quad\text{as } n\to\infty. \tag{32}$$

PROOF OF PROPOSITION 4.4. We first claim that when $0 < \delta < W^n(t)$,

$$E^n(t+\delta) - E^n(t) \le Q^n(t+\delta) + G^n_\delta(t). \tag{33}$$

To see (33), fix $t \ge 0$. For $\delta \in (0, W^n(t))$ and $\tau^n_i \in (t, t+\delta]$, because $t+\delta < t+W^n(t) \le \tau^n_i + w^n_i$ (see Lemmas 3.2 and 3.3), the ith customer will not get into service by time $t+\delta$, so he will either be waiting in queue or will have abandoned the system by then—the latter case implies $\gamma^n_i < \delta$. This proves (33).

Assume $\delta > 0$ is small enough so that $F(\delta) < 1/2$. For each $T > 0$, (33) implies

$$\mathbb{P}\left[\sup_{0\le t\le T} W^n(t) > \delta\right] \le \mathbb{P}\left[\inf_{0\le t\le T}\{E^n(t+\delta) - E^n(t) - G^n_\delta(t) - Q^n(t+\delta)\} \le 0\right].$$

By (29) and (31),

$$G^n_\delta(t) = F(\delta)(E^n(t+\delta) - E^n(t)) + L^n_\delta(\bar{E}^n(t+\delta)) - L^n_\delta(\bar{E}^n(t))$$

and thus

$$E^n(t+\delta) - E^n(t) - G^n_\delta(t) \ge \tfrac{1}{2}(E^n(t+\delta) - E^n(t)) - L^n_\delta(\bar{E}^n(t+\delta)) + L^n_\delta(\bar{E}^n(t)).$$

Then, we have

$$\begin{aligned}
&\mathbb{P}\left[\sup_{0\le t\le T} W^n(t) > \delta\right] \\
&\le \mathbb{P}\left[\inf_{0\le t\le T}\left\{\frac{1}{2}(\bar{E}^n(t+\delta) - \bar{E}^n(t)) - \bar{L}^n_\delta(\bar{E}^n(t+\delta)) + \bar{L}^n_\delta(\bar{E}^n(t)) - \bar{Q}^n(t+\delta)\right\} \le 0\right] \\
&\le \mathbb{P}\left[\inf_{0\le t\le T}\{\bar{E}^n(t+\delta) - \bar{E}^n(t)\} \le \frac{\delta c_T}{2}\right] + \mathbb{P}\left[\sup_{0\le t\le T} |\bar{L}^n_\delta(\bar{E}^n(t+\delta))| \ge \frac{\delta c_T}{12}\right] \\
&\quad + \mathbb{P}\left[\sup_{0\le t\le T} |\bar{L}^n_\delta(\bar{E}^n(t))| \ge \frac{\delta c_T}{12}\right] + \mathbb{P}\left[\sup_{0\le t\le T} \bar{Q}^n(t+\delta) \ge \frac{\delta c_T}{12}\right].
\end{aligned}$$

By (11) and (30), we see $\bar{L}^n_\delta \circ \bar{E}^n \Rightarrow 0$ as $n \to \infty$. This, together with (10) and (32), yields $W^n \Rightarrow 0$. □

4.2. Proof of Proposition 4.2. This section is dedicated to the proof of Proposition 4.2. First, we define a continuous-time filtration $\{\mathcal{F}^n(t);\ t \ge 0\}$ by

$$\mathcal{F}^n(t) = \mathcal{F}^n_{\lfloor nt\rfloor},$$

where the filtration $\{\mathcal{F}^n_i;\ i \in \mathbb{Z}_+\}$ is defined by (6). Next, let

$$H^n_i = \sum_{j=1}^{i} (1_{\{\gamma^n_j \le w^n_j\}} - F(w^n_j))h(w^n_j) \quad \text{for each } i \in \mathbb{Z}_+,$$

where $h\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a Borel measurable function such that $0 \le h(x) \le 1$ for all $x \in \mathbb{R}_+$. We further let

$$H^n(t) = H^n_{\lfloor nt\rfloor} \qquad \text{and} \qquad \tilde{H}^n(t) = \frac{1}{\sqrt{n}} H^n(t).$$

Now, we introduce a series of results on the process H^n.

LEMMA 4.1. *Assume that (7) holds. Then, $\{(H^n_i, \mathcal{F}^n_i);\ i \in \mathbb{Z}_+\}$ is a martingale.*

PROOF. Lemma 3.1 assures that w^n_j is $\mathcal{F}^n_i$-measurable for $1 \le j \le i+1$. Then, H^n_i is $\mathcal{F}^n_i$-measurable. Because w^n_i is $\mathcal{F}^n_{i-1}$-measurable whereas γ^n_i is independent of $\mathcal{F}^n_{i-1}$,

$$\mathbb{E}[H^n_i - H^n_{i-1} \mid \mathcal{F}^n_{i-1}] = \mathbb{E}[1_{\{\gamma^n_i \le w^n_i\}} \mid \mathcal{F}^n_{i-1}]h(w^n_i) - F(w^n_i)h(w^n_i) = 0.$$

Also, we have $\mathbb{E}[|H^n_i|] \le i$. Thus, $\{(H^n_i, \mathcal{F}^n_i);\ i \in \mathbb{Z}_+\}$ is a martingale. □

LEMMA 4.2. *Assume that (7) holds. Then, $\{(H^n(t), \mathcal{F}^n(t));\ t \ge 0\}$ is a martingale with quadratic variation*

$$[H^n](t) = \sum_{i=1}^{\lfloor nt\rfloor} (1_{\{\gamma^n_i \le w^n_i\}} - F(w^n_i))^2 h(w^n_i)^2. \tag{34}$$

PROOF. By Lemma 4.1, H^n is adapted to $\{\mathcal{F}^n(t);\ t \geq 0\}$. It is a martingale because for $0 \leq s \leq t$, $\mathbb{E}[|H^n(t)|] = \mathbb{E}[|H^n_{\lfloor nt \rfloor}|] < \infty$ and $\mathbb{E}[H^n(t) \mid \mathcal{F}^n(s)] = \mathbb{E}[H^n_{\lfloor nt \rfloor} \mid \mathcal{F}^n_{\lfloor ns \rfloor}] = H^n_{\lfloor ns \rfloor} = H^n(s)$. Because H^n is piecewise constant and $\sum_{i=1}^{\lfloor nt \rfloor} |1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n)|h(w_i^n) \leq nt$, H^n is a finite-variation process from which (34) follows (see Theorem 2.26 of Protter [15]). □

LEMMA 4.3. *Assume that* (7), (8), *and* (10)–(12) *hold. Then,*

$$\tilde{H}^n \ \Rightarrow\ 0 \quad \text{as } n \to \infty.$$

Before proving Lemma 4.3, we introduce a martingale convergence lemma, which is a degenerate case of the martingale functional central limit theorem. Its proof can be found in Whitt [21].

LEMMA 4.4. *Let* $\{(M^n(t), \mathcal{G}^n(t));\ t \geq 0\}$ *be a local martingale with* $M^n(0) = 0$ *for each* $n \in \mathbb{N}$. *Assume that, for any* $T > 0$,

$$\mathbb{E}\left[\sup_{0<t\leq T} |M^n(t) - M^n(t-)|\right] \to 0 \qquad \text{and} \qquad [M^n](T) \ \Rightarrow\ 0 \quad \text{as } n \to \infty.$$

Then, $M^n \Rightarrow 0$ *as* $n \to \infty$.

PROOF OF LEMMA 4.3. Using Lemma 4.2, $\{(\tilde{H}^n(t), \mathcal{F}^n(t));\ t \geq 0\}$ is a martingale with quadratic variation

$$[\tilde{H}^n](t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} (1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n))^2 h(w_i^n)^2.$$

Fix $T > 0$. By (28) and the fact that $F(w_i^n) \leq 1$, we obtain

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{1\leq i\leq nT} F(w_i^n)\right] = 0.$$

Because γ_i^n is independent of $\mathcal{F}^n_{i-1}$ but w_i^n is $\mathcal{F}^n_{i-1}$-measurable (see Lemma 3.1),

$$\mathbb{E}[(1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n))^2 h(w_i^n)^2 \mid \mathcal{F}^n_{i-1}] = (1 - F(w_i^n))F(w_i^n)h(w_i^n)^2 \leq F(w_i^n).$$

Then,

$$\mathbb{E}[[\tilde{H}^n](T)] \leq \frac{1}{n}\sum_{i=1}^{\lfloor nT \rfloor} \mathbb{E}[F(w_i^n)] \leq T\mathbb{E}\left[\sup_{1\leq i\leq nT} F(w_i^n)\right].$$

It follows that $\mathbb{E}[[\tilde{H}^n](T)] \to 0$ and hence $[\tilde{H}^n](T) \Rightarrow 0$ as $n \to \infty$. Because $\sup_{0<t\leq T} |\tilde{H}^n(t) - \tilde{H}^n(t-)| \leq n^{-1/2}$, we have

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{0<t\leq T} |\tilde{H}^n(t) - \tilde{H}^n(t-)|\right] = 0.$$

Then, it follows from Lemma 4.4 that $\tilde{H}^n \Rightarrow 0$ as $n \to \infty$. □

For the proof of Proposition 4.2, we let

$$L^n(t) = \sum_{j=1}^{\lfloor nt \rfloor} (1_{\{\gamma_j^n \leq w_j^n\}} - F(w_j^n)) \qquad \text{and} \qquad \tilde{L}^n(t) = \frac{1}{\sqrt{n}} L^n(t).$$

PROOF OF PROPOSITION 4.2. Note that $L^n(t) = H^n(t)$ when $h = 1$. Lemma 4.3 implies that $\tilde{L}^n \Rightarrow 0$ as $n \to \infty$. Because $\tilde{A}_1^n(t) = \tilde{L}^n(\bar{E}^n(t))$, it follows from (11) that $\tilde{A}_1^n \Rightarrow 0$ as $n \to \infty$. □

4.3. Proof of Proposition 4.3. We present the proof of Proposition 4.3 in this section. The crucial step is using Lemma 3.4 to establish the following lemma that converts a summation of offered waiting times to an integral of the queue length process.

LEMMA 4.5. *Assume that* (7)–(12) *hold. Then, for each* $T > 0$,

$$\sup_{0\leq t\leq T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} w_i^n - \int_0^t \tilde{Q}^n(s)\, ds \right| \ \Rightarrow\ 0 \quad \text{as } n \to \infty.$$

Assuming Lemma 4.5, we now provide the proof of Proposition 4.3.

PROOF OF PROPOSITION 4.3. We decompose $\tilde{A}_2^n(t)$ into

$$\tilde{A}_2^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} F(w_i^n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} w_i^n g(w_i^n) = \tilde{A}_{21}^n(t) + \tilde{A}_{22}^n(t) + \alpha \int_0^t \tilde{Q}^n(s)\, ds,$$

where

$$\tilde{A}_{21}^n(t) = \frac{\alpha}{\sqrt{n}} \sum_{i=1}^{E^n(t)} w_i^n - \alpha \int_0^t \tilde{Q}^n(s)\, ds \qquad \text{and} \qquad \tilde{A}_{22}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (g(w_i^n) - \alpha) w_i^n.$$

Lemma 4.5 leads to $\tilde{A}_{21}^n \Rightarrow 0$. Also, by (12) and Lemma 4.5,

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} w_i^n > a \right] = 0.$$

Then, it follows from (27) that

$$\sup_{0\le t\le T} |\tilde{A}_{22}(t)| \le \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} w_i^n \sup_{1\le i\le E^n(T)} |g(w_i^n) - \alpha| \;\Rightarrow\; 0,$$

which concludes the proof. □

It remains to prove Lemma 4.5. Let

$$H_F^n(t) = \sum_{i=1}^{\lfloor nt \rfloor} (1_{\{\gamma_i^n \le w_i^n\}} - F(w_i^n)) F(w_i^n) \qquad \text{and} \qquad \tilde{H}_F^n(t) = \frac{1}{\sqrt{n}} H_F^n(t).$$

Because $H_F^n(t) = H^n(t)$ if $h = F$, Lemma 4.3 leads to

$$\tilde{H}_F^n \;\Rightarrow\; 0 \quad \text{as } n \to \infty. \tag{35}$$

In the next lemma, we demonstrate the stochastic boundedness of the process $\tilde{A}_2$.

LEMMA 4.6. *Assume that* (7)–(12) *hold. Then, for any* $T > 0$,

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}[\tilde{A}_2^n(T) > a] = 0. \tag{36}$$

PROOF. Fix $T > 0$. For $t \ge 0$, we decompose $\tilde{A}_2^n(t)$ into

$$\tilde{A}_2^n(t) = \tilde{A}_{23}^n(t) + \tilde{A}_{24}^n(t),$$

where

$$\tilde{A}_{23}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (1 - F(w_i^n)) F(w_i^n) \qquad \text{and} \qquad \tilde{A}_{24}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} F(w_i^n)^2.$$

Then, (36) holds if we have

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}[\tilde{A}_{23}^n(T) > a] = 0, \tag{37}$$

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}[\tilde{A}_{24}^n(T) > a] = 0. \tag{38}$$

First, using Lemma 3.4, we get

$$\sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n \le \sum_{i=1-Q^n(0)}^{E^n(t)} (\gamma_i^n \wedge w_i^n) \le \int_0^{t+W^n(t)} Q^n(s)\, ds.$$

It follows from (12) and Proposition 4.4 that

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n > a \right] = 0. \tag{39}$$

Note that

$$\tilde{A}^n_{23}(t) - \frac{\alpha}{\sqrt{n}} \sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n = \tilde{H}^n_F(\bar{E}^n(t)) + \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n (g(w_i^n) - \alpha). \tag{40}$$

The first term on the right-hand side of (40) satisfies $\tilde{H}^n_F \circ \bar{E}^n \Rightarrow 0$ by (11) and (35). The second term satisfies

$$\sup_{0\le t\le T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n (g(w_i^n) - \alpha) \right| \le \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n \sup_{1\le i\le E^n(T)} |g(w_i^n) - \alpha| \;\Rightarrow\; 0$$

by (27) and (39). Now, (37) follows from (39) and (40).

If $\tilde{A}^n_{23}(T) < \tilde{A}^n_{24}(T)$ holds, there must exist $1 \le i \le E^n(T)$ such that $F(w_i^n) > 1 - F(w_i^n)$, namely, $\sup_{1\le i\le E^n(T)} F(w_i^n) > 1/2$. Then, (38) follows from (26) and (37). □

Proof of Lemma 4.5. Because $w_i^n \le W^n(0)$ for $1 - Q^n(0) \le i \le 0$ (see Lemma 3.2), by (12) and Proposition 4.4 we have

$$\frac{1}{\sqrt{n}} \sum_{i=1-Q^n(0)}^{0} (\gamma_i^n \wedge w_i^n) \le \tilde{Q}^n(0) W^n(0) \;\Rightarrow\; 0.$$

Because $\sup_{0\le t\le T}(t + W^n(t)) = T + W^n(T)$ (see Lemma 3.3), by (12) and Proposition 4.4 again we have

$$\sup_{0\le t\le T} \int_t^{t+W^n(t)} \tilde{Q}^n(s)\, ds \le \sup_{0\le t\le T+W^n(T)} \tilde{Q}^n(t) \sup_{0\le t\le T} W^n(t) \;\Rightarrow\; 0.$$

Thus, Lemma 3.4 implies

$$\sup_{0\le t\le T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (\gamma_i^n \wedge w_i^n) - \int_0^t \tilde{Q}^n(s)\, ds \right| \;\Rightarrow\; 0. \tag{41}$$

By Proposition 4.2, Lemma 4.6, and (25),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} 1_{\{\gamma_i^n \le w_i^n\}} w_i^n \le (\tilde{A}^n_1(T) + \tilde{A}^n_2(T)) \sup_{1\le i\le E^n(T)} w_i^n \;\Rightarrow\; 0.$$

Because

$$\sum_{i=1}^{E^n(t)} w_i^n - \sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n \le w_i^n\}} w_i^n \le \sum_{i=1}^{E^n(t)} (\gamma_i^n \wedge w_i^n) \le \sum_{i=1}^{E^n(t)} w_i^n,$$

we have

$$\sup_{0\le t\le T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} w_i^n - \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (\gamma_i^n \wedge w_i^n) \right| \;\Rightarrow\; 0 \quad \text{as } n \to \infty. \tag{42}$$

The assertion of Lemma 4.5 follows from (41) and (42). □

4.4. Proof of Proposition 4.1. Now, we are ready to prove Proposition 4.1. Recall that $G^n(t)$ is the number of customers who abandon the system during $(0, t]$, and $A^n(t)$ is the number of customers who have arrived by time t but will eventually abandon the system. Clearly, $G^n(t) \le A^n(t)$ for $t \ge 0$. We now establish a lower bound for G^n. For $t \ge 0$, define

$$\zeta^n(t) = \inf\{s \ge 0\colon\ s + W^n(s) > t\}.$$

By Lemma 3.2, $\tau_i^n + w_i^n \le \tau_i^n + W^n(\tau_i^n) \le t$ for all $\tau_i^n < \zeta^n(t)$ so that each customer arriving before time $\zeta^n(t)$ should have entered service or abandoned the queue by time t. This implies $A^n((\zeta^n(t) - \delta)^+) \le G^n(t)$ for all $\delta > 0$ and $t \ge 0$. Setting $\delta = n^{-1}$, we have

$$A^n((\zeta^n(t) - n^{-1})^+) \le G^n(t) \le A^n(t). \tag{43}$$

We will prove Proposition 4.1 by showing that the upper and lower bounds of G^n in (43) are asymptotically close.

PROOF OF PROPOSITION 4.1. First, we see that (23) holds by (13), (24), and Propositions 4.2–4.3. Because W^n is right-continuous, $\zeta^n(t)+W^n(\zeta^n(t))\geq t$ for all $t\geq 0$. It follows that

$$A^n(t)-A^n((\zeta^n(t)-n^{-1})^+)\leq A^n(\zeta^n(t)+W^n(\zeta^n(t)))-A^n((\zeta^n(t)-n^{-1})^+). \tag{44}$$

In diffusion scaling, we have

$$\begin{aligned}&\tilde{A}^n(\zeta^n(t)+W^n(\zeta^n(t)))-\tilde{A}^n((\zeta^n(t)-n^{-1})^+)\\&\leq\left|\tilde{A}^n(\zeta^n(t)+W^n(\zeta^n(t)))-\alpha\int_0^{\zeta^n(t)+W^n(\zeta^n(t))}\tilde{Q}^n(s)\,ds\right|\\&\quad+\left|\tilde{A}^n((\zeta^n(t)-n^{-1})^+)-\alpha\int_0^{(\zeta^n(t)-n^{-1})^+}\tilde{Q}^n(s)\,ds\right|+\alpha\int_{(\zeta^n(t)-n^{-1})^+}^{\zeta^n(t)+W^n(\zeta^n(t))}\tilde{Q}^n(s)\,ds.\end{aligned} \tag{45}$$

By (23), Proposition 4.4, and the fact that $\zeta^n(t)\leq t$, the first term on the right-hand side of (45) satisfies

$$\begin{aligned}&\sup_{0\leq t\leq T}\left|\tilde{A}^n(\zeta^n(t)+W^n(\zeta^n(t)))-\alpha\int_0^{\zeta^n(t)+W^n(\zeta^n(t))}\tilde{Q}^n(s)\,ds\right|\\&\leq\sup_{0\leq t\leq T}\left|\tilde{A}^n(t+W^n(t))-\alpha\int_0^{t+W^n(t)}\tilde{Q}^n(s)\,ds\right|\ \Rightarrow\ 0.\end{aligned}$$

Similarly, the second term satisfies

$$\sup_{0\leq t\leq T}\left|\tilde{A}^n((\zeta^n(t)-n^{-1})^+)-\alpha\int_0^{(\zeta^n(t)-n^{-1})^+}\tilde{Q}^n(s)\,ds\right|\leq\sup_{0\leq t\leq T}\left|\tilde{A}^n(t)-\alpha\int_0^t\tilde{Q}^n(s)\,ds\right|\ \Rightarrow\ 0.$$

Using (12), Proposition 4.4, and the fact that $\zeta^n(t)\leq t$, the third term satisfies

$$\sup_{0\leq t\leq T}\int_{(\zeta^n(t)-n^{-1})^+}^{\zeta^n(t)+W^n(\zeta^n(t))}\tilde{Q}^n(s)\,ds\leq\sup_{0\leq t\leq T}W^n(t)\sup_{0\leq t\leq T}\tilde{Q}^n(t)+\frac{1}{n}\sup_{0\leq t\leq T}\tilde{Q}^n(t)\ \Rightarrow\ 0.$$

Therefore,

$$\sup_{0\leq t\leq T}\{\tilde{A}^n(\zeta^n(t)+W^n(\zeta^n(t)))-\tilde{A}^n((\zeta^n(t)-n^{-1})^+)\}\ \Rightarrow\ 0\quad\text{as } n\to\infty. \tag{46}$$

Finally, Proposition 4.1 follows from (43), (44), and (46). □

Appendix. Proof of Lemma 3.1.

PROOF OF LEMMA 3.1. For each $m\in\mathbb{Z}_+$, let $v_{i,m}=1_{\{i>-m\}}v_i$, $\gamma_{i,m}=1_{\{i>-m\}}\gamma_i-1_{\{i\leq -m\}}$, and $\mathcal{F}_{k,m}=\sigma\{\tau_{i+1},v_{i,m},\gamma_{i,m};\ i\leq k\}$. Because $v_i=\lim_{m\to\infty}v_{i,m}$ and $\gamma_i=\lim_{m\to\infty}\gamma_{i,m}$ for all $i\in\mathbb{Z}$, we have $\mathcal{F}_k=\bigvee_{m=0}^{\infty}\mathcal{F}_{k,m}$, where $\bigvee_{m=0}^{\infty}\mathcal{F}_{k,m}$ is the smallest σ-field that contains each $\mathcal{F}_{k,m}$ for $m\in\mathbb{Z}_+$. Given $k\geq 0$ and $m\geq 0$, we define $w_{i,m}$ and $r_{i,m}(t)$ recursively via a similar procedure to that in (15)–(18) for w_i and $r_i(t)$: For $i\leq -m$, let $w_{i,m}=0$ and $r_{i,m}(t)=0$ for $t\geq 0$; for $i\geq -m+1$, let

$$w_{i,m}=\inf\left\{t\geq 0:\ \sum_{j\leq i-1}1_{\{r_{j,m}(\tau_i+t)>0\}}<n\right\}, \tag{47}$$

and

$$r^a_{i,m}(t)=1_{\{t<\tau_i+\gamma_{i,m}\}}v_{i,m}, \tag{48}$$

$$r^s_{i,m}(t)=1_{\{t<\tau_i+w_{i,m}\}}v_{i,m}+1_{\{t\geq\tau_i+w_{i,m}\}}(v_{i,m}-t)^+, \tag{49}$$

$$r_{i,m}(t)=1_{\{0\leq\gamma_{i,m}\leq w_{i,m}\}}r^a_{i,m}(t)+(1-1_{\{0\leq\gamma_{i,m}\leq w_{i,m}\}})r^s_{i,m}(t) \tag{50}$$

for $t\geq 0$. By (47), we get $w_{i,m}=0$ for $i\leq -m+n$.

Fix integers $k\geq 0$ and $m\geq 0$. We would like to show that

$$w_{i,m}\text{ is }\mathcal{F}_{k,m}\text{-measurable}\quad\text{for each } i\leq k+1. \tag{51}$$

Assume that there exists an integer $j\leq k$ such that $w_{i,m}$ is $\mathcal{F}_{k,m}$-measurable for all $i\leq j$. Clearly, $j=(-m+n)\wedge k$ is such a choice. To prove (51), by induction on j it remains to show that $w_{j+1,m}$ is also $\mathcal{F}_{k,m}$-measurable. To see this, for any $t\geq 0$, $r^a_{i,m}(t)$, $r^s_{i,m}(t)$ and $r_{i,m}(t)$ are $\mathcal{F}_{k,m}$-measurable. By (48)–(50), the process $r_{i,m}$ is right-continuous and thus $r_{i,m}(\tau_i+t)$ is $\mathcal{F}_{k,m}$-measurable for $i\leq j$ and $t\geq 0$ because τ_i is $\mathcal{F}_{k,m}$-measurable. Because $\{w_{j+1,m}\leq t\}=\{\sum_{i\leq j}1_{\{r_{i,m}(\tau_{j+1}+t)>0\}}<n\}$, we conclude that $w_{j+1,m}$ is $\mathcal{F}_{k,m}$-measurable, thus proving (51).

Given $\omega\in\Omega$, we have $v_i(\omega)=v_{i,m}(\omega)$ and $\gamma_i(\omega)=\gamma_{i,m}(\omega)$ for all $m\geq X(0,\omega)$ and $i\in\mathbb{Z}$. One can check that $w_i(\omega)=w_{i,m}(\omega)$ for $m\geq X(0,\omega)$ and thus $w_i=\lim_{m\to\infty}w_{i,m}$. Therefore, w_i is $\mathcal{F}_k$-measurable for $i\leq k+1$.

Acknowledgments. This research is supported in part by NSF Grants CMMI-0727400 and CMMI-0825840 and by an IBM Faculty Award. The authors thank Guodong Pang for helpful discussions on time-nonhomogeneous arrival processes and the anonymous referees for their thoughtful comments.

References

[1] Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16** 665–688.
[2] Baccelli, F., P. Boyer, G. Hebuterne. 1984. Single-server queues with impatient customers. *Adv. Appl. Probab.* **16** 887–905.
[3] Bassamboo, A., R. S. Randhawa. 2010. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* Forthcoming.
[4] Borovkov, A. 1967. On limit laws for service processes in multi-channel systems. *Siberian Math. J.* **8** 746–763.
[5] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.
[6] Dai, J. G., S. He, T. Tezcan. 2010. Many-server diffusion limits for $G/Ph/n+GI$ queues. *Ann. Appl. Probab.* Forthcoming.
[7] Ethier, S. N., T. G. Kurtz. 1986. *Markov Processes: Characterization and Convergence*. Wiley, New York.
[8] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.
[9] Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 208–227.
[10] Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.
[11] Iglehart, D. L., W. Whitt. 1970. Multiple channel queues in heavy traffic. I. *Adv. Appl. Probab.* **2** 150–177.
[12] Kaspi, H., K. Ramanan. 2009. SPDE limits of many server queues. Technical report, Division of Applied Mathematics, Brown University, Providence, RI.
[13] Kiefer, J., J. Wolfowitz. 1955. On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78** 1–18.
[14] Mandelbaum, A., P. Momčilović. 2009. Queues with many servers and impatient customers. http://iew3.technion.ac.il/serveng/References/MM0309.pdf. Preprint.
[15] Protter, P. 2005. *Stochastic Integration and Differential Equations*, 2nd ed. Springer, New York.
[16] Puhalskii, A. A., M. I. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32** 564–595. Correction: **36**, 971 (2004).
[17] Reed, J. 2009. The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* **19** 2211–2269.
[18] Reed, J., T. Tezcan. 2009. Hazard rate scaling for the $GI/M/n+GI$ queue. http://pages.stern.nyu.edu/~jreed/Papers/ReedTezcan121009.pdf. Preprint.
[19] Stanford, R. E. 1979. Reneging phenomena in single channel queues. *Math. Oper. Res.* **4** 162–178.
[20] Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54** 37–54.
[21] Whitt, W. 2007. Proofs of the martingale FCLT. *Probab. Surveys* **4** 268–302.
[22] Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* **51** 361–402.