

# Customer Abandonment in Many-Server Queues

J. G. Dai and Shuangchi He <sup>1</sup>

May 22, 2009

## Abstract

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-count 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  $n \rightarrow \infty$ . The key assumption is 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.

**Keywords:** multi-server queues, customer abandonment, many-server heavy traffic, Halfin-Whitt regime, quality- and efficiency-driven regime.

**AMS Subject Classification (2000):** 90B20, 68M20, 60J70

## 1 Introduction

A  $G/G/n$  queue is a classic stochastic system that has been extensively studied in literature. 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 a service, 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 customer call centers; see, for examples, survey papers Gans et al. (2003) and Aksin et al. (2007).

As pointed out in Garnett et al. (2002), customer abandonment is a key factor for call center operations. Our paper studies parallel-server queues with customer abandonment. In our model, each customer has a patience time; when a customer's waiting time in queue

---

<sup>1</sup>H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, Email: {dai,heshuangchi}@gatech.edu; research supported in part by NSF grants CMMI-0727400 and CMMI-0825840, and by an IBM Faculty Award

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.

Let  $Q(t)$  be the number of customers in queue at time  $t$ , and  $G(t)$  be the cumulative number of customers who have abandoned the system by time  $t$ . This paper is to establish an asymptotic relationship between the *queue-length* process  $Q = \{Q(t); t \geq 0\}$  and the *abandonment-count* process  $G = \{G(t); t \geq 0\}$  in a  $G/G/n + GI$  queue when the number of servers  $n$  is large and the patience times are independent, identically distributed (iid) following a general distribution (the  $+GI$ ).

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

$$G(t) = N \left( \alpha \int_0^t Q(s) ds \right) \quad (1)$$

for all  $t \geq 0$  where  $N$  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 (1981). Specifically, we consider a sequence of  $M/M/n + M$  systems indexed by the number of servers  $n$ . For the  $n$ th system, its arrival rate  $\lambda^n$  depends on  $n$ . The arrival rate  $\lambda^n \rightarrow \infty$  as  $n \rightarrow \infty$ , 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  $n$ th system as  $\rho^n = \lambda^n/(n\mu)$ . We assume that

$$\lim_{n \rightarrow \infty} \sqrt{n}(1 - \rho^n) = \beta \quad (2)$$

for some constant  $\beta$ . 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 system is in the Halfin-Whitt regime, one can prove from relationship (1) that for each  $T > 0$ ,

$$\frac{1}{\sqrt{n}} \sup_{0 \leq t \leq T} \left| G^n(t) - \alpha \int_0^t Q^n(s) ds \right| \rightarrow 0 \text{ in probability} \quad (3)$$

as  $n \rightarrow \infty$ . The main result (Theorem 1) of this paper is to prove that the asymptotic relationship (3) holds for a sequence of  $G/G/n + GI$  queues, assuming that the sequence of diffusion-scaled queue-length processes is stochastically bounded.

The heavy traffic condition (2) implies that the sequence of systems is critically loaded in the limit. It is often used to prove the stochastic boundedness for the sequence of 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, i.e.,  $\lim_{n \rightarrow \infty} \rho^n < 1$ , the stochastic boundedness still holds. Our main theorem, Theorem 1, assumes the stochastic boundedness for the sequence of diffusion-scaled queue-length processes. In particular, 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 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 the corresponding  $G/G/n$  queue without abandonment. The comparison result implies that it is sufficient to prove stochastic boundedness of the queue-length processes in a sequence of  $G/G/n$  queues without abandonment.

In Theorem 1, the  $\alpha$  in (3) is replaced by the right-derivative at 0 of the patience time distribution. Under the stochastic boundedness assumption on the diffusion-scaled queue-length processes, the waiting times will be proved to converge to 0 as  $n \rightarrow \infty$ . Thus, customer abandonment rarely happens; only those customers who have extremely small patience time may abandon the system. Therefore, the patience time distribution, outside a neighborhood of zero, barely has any influence on the system dynamics. Zeltyn and Mandelbaum (2005) observes the same phenomenon and studies the steady-state quantities of  $M/M/n + GI$  queues in the Halfin-Whitt regime.

Asymptotic relationship (3) is critical to proving a many-server heavy traffic limit theorem for a sequence of  $G/Ph/n + GI$  queues in Dai et al. (2009), where the service times follow a phase-type distribution. It is expected that it will again play a critical role in proving a future many-server heavy traffic limit theorem for a sequence of  $G/GI/n + GI$  queues. For  $G/Ph/n + GI$  queues, the stochastic boundedness assumption follows from Puhalskii and Reiman (2000) and for  $G/GI/n + GI$  queues, it follows from Reed (2007) under some mild assumption of the service time distribution. In that paper, the author proved a many-server heavy traffic limit theorem for the one-dimensional customer-count process, without customer abandonment. A key insight from these limit theorems is that the exact distribution of patience times is irrelevant in the Halfin-Whitt regime as long as the customer abandonment is explicitly built into the model. This phenomenon is in sharp contrast to the one found in Whitt (2006) when the systems is operated in an overloaded regime known as the *Efficiency-Driven regime*; the system performance there depends critically to the patience time distribution and a fluid model is shown to be able to capture that dependency.

The remainder of the paper is organized as follows. The main results are presented in Section 2. The proofs of Theorems 1 and 2 are given in Sections 3 and 4, respectively. A few preliminary results on  $G/G/n + G$  queues and a special case of the initial assumption of Theorem 1 are discussed in the appendix.

## Notation

All random variables and processes are assumed to be defined on a common probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The symbols  $\mathbb{Z}$ ,  $\mathbb{Z}_+$ ,  $\mathbb{R}$ , and  $\mathbb{R}_+$  are used to denote the sets of integers, nonnegative integers, real numbers, and nonnegative real numbers, respectively. The space of functions  $f : \mathbb{R}_+ \rightarrow \mathbb{R}$  that are right-continuous on  $[0, \infty)$  and have left limits in  $(0, \infty)$  is denoted by  $\mathbb{D}$ , which is endowed with the Skorohod  $J_1$ -topology. We reserve the symbol  $\Rightarrow$  for convergence in distribution and  $\mathbb{E}[\cdot]$  for expectation. For a set of random variables  $\{Y_j; j \in J\}$ ,  $\sigma\{Y_j; j \in J\}$  is the  $\sigma$ -field generated by the set of random variables. Given  $\sigma$ -fields  $\mathcal{F}_j \subset \mathcal{F}$  with  $j \in J$ ,  $\bigvee_{j \in J} \mathcal{F}_j$  is the smallest  $\sigma$ -field that contains  $\mathcal{F}_j$  for each  $j \in J$ . For  $a, b \in \mathbb{R}$ ,  $\lfloor a \rfloor = \max\{j \in \mathbb{Z} : j \leq a\}$ ,  $a \vee b = \max\{a, b\}$ ,  $a \wedge b = \min\{a, b\}$ , and  $a^+ = \max\{a, 0\}$ . We use  $1_S$  to denote the indicator function of a set  $S \subset \Omega$ , and use  $e : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  to denote the identity function on  $\mathbb{R}_+$ , that is,  $e(t) = t$  for  $t \geq 0$ . For an  $f \in \mathbb{D}$ ,  $f(t-)$  denotes its left limit at  $t > 0$ .

## 2 Heavy traffic setting and the main results

In Section 2.1, we define a  $G/G/n + G$  queue for a fixed positive integer  $n$  using a sequence of primitive random variables. In Section 2.2, we introduce a sequence of  $G/G/n + GI$  queues and the stochastic boundedness assumption on the sequence of diffusion-scaled queue-length processes. In Section 2.3, we state two main results of this paper.

### 2.1 The $G/G/n + G$ 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

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

We assume that  $X(0, \omega) < \infty$  on each sample path  $\omega$ . The integer  $X(0, \omega)$  is interpreted as the total number of customers that 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 0. Thus, customers  $i = 1 - X(0, \omega), \dots, 0$  are in the system initially, with customers  $i = 1 - Q(0, \omega), \dots, 0$  waiting in queue.

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

$$E(t) = \sup\{i \in \mathbb{Z}_+ : \tau_i \leq t\}. \quad (4)$$

Clearly,  $E(t)$  is the number of customers who have arrived at the system in  $(0, t]$ .

For each  $i \in \mathbb{Z}$ ,  $v_i(\omega) \geq 0$ . One interprets  $v_i(\omega)$  as the service time of the  $i$ th customer if he has not started his service by time 0 or his remaining service time at time 0 if he has started service. For  $i > 0$ ,  $\gamma_i(\omega) \geq 0$  is interpreted as the patience time of the  $i$ th customer. For customer  $i$  who is waiting in the queue at time 0,  $\gamma_i(\omega) \geq 0$  is interpreted as the remaining patience time of the customer. For customer  $i$  who has entered service or abandoned the system by time 0,  $\gamma_i(\omega)$  can take any value. By convention, we set  $\gamma_i(\omega) = -1$  when  $i \leq -Q(0, \omega)$ . Most of this paper studies  $G/G/n + GI$  queues. In such a queue, the sequence of patience times  $\{\gamma_i\}_{i=1}^\infty$  is assumed to be iid.

## 2.2 The asymptotic framework

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 in the  $n$ th system. Defining a filtration  $\{\mathcal{F}_i^n; i \in \mathbb{Z}_+\}$  by

$$\mathcal{F}_i^n = \sigma \{ \tau_{j+1}^n, v_j^n, \gamma_j^n; j \leq i \}, \quad (5)$$

we assume that

$$\gamma_{i+1}^n \text{ is independent of } \mathcal{F}_i^n \text{ for each } i \in \mathbb{Z}_+ \quad (6)$$

and that  $\{\gamma_i^n\}_{i=1}^\infty$  is a sequence of iid random variables with distribution function  $F$ . We further assume the distribution  $F$  satisfies

$$F(0) = 0 \quad (7)$$

and is right-differentiable at 0 with right-derivative

$$\alpha = \lim_{x \downarrow 0} x^{-1} F(x) < \infty. \quad (8)$$

The arrival process in the  $n$ th system is  $E^n = \{E^n(t); t \geq 0\}$ , where  $E^n(t)$ , defined in (4), denotes the number of customer arrivals in  $(0, t]$ . Fixing a constant  $\lambda > 0$ , define the diffusion-scaled arrival process  $\tilde{E}^n = \{\tilde{E}^n(t); t \geq 0\}$  via

$$\tilde{E}^n(t) = \frac{1}{\sqrt{n}} \hat{E}^n(t) \quad \text{and} \quad \hat{E}^n(t) = E^n(t) - n\lambda t$$

for  $t \geq 0$ . We assume that the sequence of diffusion-scaled arrival processes  $\tilde{E}^n$  is stochastically bounded for some  $\lambda > 0$ ; namely, there exists a constant  $\lambda > 0$  such that for each  $T > 0$ ,

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq T} |\tilde{E}^n(t)| > a \right] = 0. \quad (9)$$

Condition (9) is a mild assumption. It is satisfied, for example, when  $E^n(t) = E_0(n\lambda t)$  and  $E_0$  is a renewal process; in this case,  $\tilde{E}^n$  converges to a driftless Brownian motion in distribution as  $n \rightarrow \infty$ .

Recall the queue-length process  $Q^n$  and the abandonment-count process  $G^n$  in the  $n$ th system. Define the diffusion-scaled versions of the queue-length process  $\tilde{Q}^n$  and abandonment-count process  $\tilde{G}^n$  via

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

For the main result (Theorem 1) of this paper, the key assumption is that the sequence of diffusion-scaled queue-length processes is stochastically bounded; namely, for each  $T > 0$ ,

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq T} \tilde{Q}^n(t) > a \right] = 0. \quad (10)$$

For Theorem 1, we also need to make an assumption on initial condition. Let  $G_0^n$  be the number of customers who are in the queue at time 0 but will eventually abandon the system. Let

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

We assume that as  $n \rightarrow \infty$

$$\tilde{G}_0^n \Rightarrow 0. \quad (11)$$

As a special case, the initial assumption (11) holds if we assume that  $\{\gamma_i^n; i = 1 - Q^n(0), \dots, 0\}$ , the remaining patience times of the  $Q^n(0)$  initial customers, are iid and follow the distribution  $F$ . See Lemma 15 in the appendix.

### 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 1.** *Consider a sequence of  $G/G/n + GI$  queues that satisfies (6)–(9). Assume the sequence of diffusion-scaled queue-length processes is stochastically bounded and the sequence of queues satisfies the initial condition (11). Then, the asymptotic relationship (3) holds for any  $T > 0$ .*

The proof of Theorem 1 will be presented in Section 3. All assumptions in Theorem 1 are standard, except the stochastic boundedness assumption (10). Verifying the stochastic boundedness assumption can be a significant task.

We now present the second theorem of the paper. The theorem, known as the comparison result in the rest of this paper, shows that the queue length at any time in a  $G/G/n + G$  queue is dominated by the queue length in the corresponding  $G/G/n$  queue without abandonment. The comparison result implies that, to verify the stochastic boundedness assumption (10) for a sequence of  $G/G/n + GI$  queues, it is sufficient to prove

stochastic boundedness for the queue-length processes in the corresponding sequence of  $G/G/n$  queues without abandonment.

To state Theorem 2, we consider two FIFO queues: a  $G/G/n + G$  queue denoted by  $\Sigma_1$ , and the corresponding  $G/G/n$  queue denoted by  $\Sigma_2$ . All servers in both systems are identical. At time 0, the two systems have identical initial conditions: in each system, there are  $X(0)$  customers indexed by  $i = 1 - X(0), \dots, 0$ , and the  $i$ th customers of  $\Sigma_1$  and  $\Sigma_2$  have the same remaining service time. After time 0, the arrival processes to both queues are identical; for  $i = 1, 2, \dots$ , the  $i$ th customers of  $\Sigma_1$  and  $\Sigma_2$  share the same service requirement. In short, the only difference between the two systems is that customers in  $\Sigma_1$  can abandon the system whereas customers in  $\Sigma_2$  cannot.

**Theorem 2.** *Let  $Q_1(t)$  and  $Q_2(t)$  be the respective numbers of customers waiting in queue of  $\Sigma_1$  and  $\Sigma_2$  at time  $t \geq 0$ . Then on each sample path,*

$$Q_1(t) \leq Q_2(t)$$

for all  $t \geq 0$ .

The proof of Theorem 2 is given in Section 4.

### 3 Proof of Theorem 1

Let  $A^n(t)$  denote, among all customers who have arrived at the  $n$ th system by time  $t \geq 0$ , the number of those who will eventually abandon the queue. Our first result is the following proposition showing that  $A^n$  and  $G^n$  are asymptotically close.

**Proposition 1.** *Under the conditions of Theorem 1,*

$$\tilde{A}^n - \tilde{G}^n \Rightarrow 0$$

as  $n \rightarrow \infty$ , where

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

The proof of the proposition will be given in Section 3.4. Given Proposition 1, to prove Theorem 1, it suffices to prove that for each  $T > 0$ ,

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t) - \alpha \int_0^t \tilde{Q}^n(s) ds \right| \rightarrow 0 \text{ in probability} \quad (12)$$

as  $n \rightarrow \infty$ .

To prove (12), one needs to further analyze the process  $A^n$ . For that, we introduce two notions: *virtual waiting time* and *offered waiting time*. For the  $G/G/n + G$  queue, we use  $w(t)$  to denote its virtual waiting time at time  $t \geq 0$ . The precise definition of  $w(t)$  is

given in (42) in the appendix. 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. For each  $i \geq 1$ , we use  $w_i$  to denote the offered waiting time of the  $i$ th customer, which 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  $i$ th customer is waiting in queue at time 0, so we use  $w_i$  to denote the remaining waiting time of the  $i$ th customer if he had infinite patience. A more precise definition of the offered waiting time can be found in Section A.1 in the appendix. For a 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 in the  $n$ th system.

In the  $n$ th system of the  $G/G/n + GI$  queues, for each customer  $i \geq 1 - Q^n(0)$ , given his patience time  $\gamma_i^n$  and offered waiting time  $w_i^n$ , we can determine whether the customer will eventually abandon the queue: he will wait  $\gamma_i^n$  units of time and leave the system with no service when  $\gamma_i^n \leq w_i^n$ , or wait  $w_i^n$  units of time and get into a server otherwise. One can check that

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

Clearly, the process  $A^n$  can be decomposed into

$$A^n(t) = G_0^n + A_1^n(t) + A_2^n(t), \quad (13)$$

where

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

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

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

Defining the diffusion-scaled processes

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

we have the following two propositions.

**Proposition 2.** *Under the conditions of Theorem 1,*

$$\tilde{A}_1^n \Rightarrow 0$$

as  $n \rightarrow \infty$ .

**Proposition 3.** *Under the conditions of Theorem 1, for any  $T > 0$ ,*

$$\sup_{0 \leq t \leq T} \left| \tilde{A}_2^n(t) - \alpha \int_0^t \tilde{Q}^n(s) ds \right| \Rightarrow 0$$

as  $n \rightarrow \infty$ .

The proofs of Propositions 2 and 3 are presented in Sections 3.2 and 3.3, respectively. Clearly, the proof of Theorem 1 follows from (11), (13), and Propositions 1–3.

### 3.1 Virtual and offered waiting times

This section is a preparation for proving Propositions 1–3 in Sections 3.2–3.4. The main result of this section is the following proposition and its corollaries. Proposition 4 says that a sequence of properly scaled virtual waiting time processes is stochastically bounded when the sequence of diffusion-scaled queue-length processes is stochastically bounded.

**Proposition 4.** *Assume (7)–(10) hold. Then*

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq T} \sqrt{n} w^n(t) > a \right] = 0. \quad (14)$$

We leave the proof of the proposition to the end of this section. We now state a few corollaries to the proposition. For that, define  $g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  by

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

By (7) and (8),  $g$  is right-continuous at 0 and for all  $x \geq 0$ ,

$$F(x) = xg(x).$$

**Corollary 1.** *Assume (7)–(10) hold. Then for any  $T > 0$ ,*

$$\sup_{0 \leq t \leq T} w^n(t) \Rightarrow 0, \quad (15)$$

$$\sup_{0 \leq t \leq T} F(w^n(t)) \Rightarrow 0, \quad (16)$$

$$\sup_{1 \leq i \leq nT} w_i^n \Rightarrow 0, \quad (17)$$

$$\sup_{1 \leq i \leq nT} F(w_i^n) \Rightarrow 0, \quad (18)$$

$$\sup_{1 \leq i \leq nT} |g(w_i^n) - \alpha| \Rightarrow 0, \quad (19)$$

as  $n \rightarrow \infty$ .

*Proof.* Proposition 4 implies (15). Since  $F$  is nondecreasing and right-continuous at 0, by the continuous mapping theorem,

$$\sup_{0 \leq t \leq T} F(w^n(t)) \leq F\left(\sup_{0 \leq t \leq T} w^n(t)\right) \Rightarrow F(0) = 0,$$

which proves (16). To see (17), first Lemma 14 (in the appendix) implies

$$\sup_{1 \leq i \leq E^n(T)} w_i^n \leq \sup_{0 \leq t \leq T} w^n(t) \Rightarrow 0. \quad (20)$$

Next,

$$\mathbb{P}\left[\sup_{1 \leq i \leq nT} w_i^n > \varepsilon\right] \leq \mathbb{P}\left[\sup_{1 \leq i \leq E^n(2\lambda^{-1}T)} w_i^n > \varepsilon\right] + \mathbb{P}\left[\bar{E}^n(2\lambda^{-1}T) \leq T\right], \quad (21)$$

where  $\bar{E}^n$  is the fluid-scaled arrival process defined by

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

From (9), we see that

$$\bar{E}^n \Rightarrow \lambda e. \quad (22)$$

Therefore, (17) follows from (20)–(22). The convergence in (18) can be proved similarly to the one in (16). For any  $\varepsilon > 0$ , since  $g$  is right-continuous at 0, there exists  $\delta > 0$  such that  $|g(x) - \alpha| \leq \varepsilon$  for all  $0 \leq x \leq \delta$ , and thus (19) follows from

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

□

To prove Proposition 4, we first establish a series of lemmas. For the  $G/G/n + GI$  queue, we define a continuous-time filtration  $\{\mathcal{F}^n(t); t \geq 0\}$  by

$$\mathcal{F}^n(t) = \mathcal{F}_{[nt]}^n,$$

where the filtration  $\{\mathcal{F}_i^n; i \in \mathbb{Z}_+\}$  is defined in (5). For any  $a > 0$ , let

$$L_a^n(t) = \sum_{i=1}^{[nt]} \left(1_{\{\gamma_i^n \leq n^{-1/2}a\}} - F(n^{-1/2}a)\right) \quad \text{and} \quad \tilde{L}_a^n(t) = \frac{1}{\sqrt{n}}L_a^n(t).$$

Our first lemma concerns  $L_a^n$ .

**Lemma 1.** For the  $G/G/n + GI$  queue,  $\{(L_a^n(t), \mathcal{F}^n(t)); t \geq 0\}$  is a martingale for any  $a > 0$ , and its quadratic variation is given by

$$[L_a^n](t) = \sum_{i=1}^{\lfloor nt \rfloor} \left( 1_{\{\gamma_i^n \leq n^{-1/2}a\}} - F(n^{-1/2}a) \right)^2. \quad (23)$$

*Proof.* Clearly,  $L_a^n(t)$  is  $\mathcal{F}_{\lfloor nt \rfloor}^n$ -measurable, and thus  $\mathcal{F}^n(t)$ -measurable. For any  $t \geq 0$ ,  $\mathbb{E}[L_a^n(t)] \leq nt$ . For  $0 \leq s \leq t$ , since  $L_a^n$  has independent increments, then

$$\mathbb{E}[L_a^n(t) - L_a^n(s) | \mathcal{F}^n(s)] = \sum_{i=\lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \mathbb{E} \left[ 1_{\{\gamma_i^n \leq n^{-1/2}a\}} - F(n^{-1/2}a) \right] = 0.$$

So  $\{(L_a^n(t), \mathcal{F}^n(t)); t \geq 0\}$  is a martingale.

Since  $L_a^n$  is piecewise constant and  $\sum_{i=1}^{\lfloor nt \rfloor} |1_{\{\gamma_i^n \leq n^{-1/2}a\}} - F(n^{-1/2}a)| \leq nt$ ,  $L_a^n$  is a finite-variation process, from which (23) follows (see, for example, Theorem 2.26 of Protter (2005)).  $\square$

Our next lemma concerns  $\tilde{L}_a^n$ .

**Lemma 2.** Assume (7) and (8) hold. Then for any  $a > 0$ ,

$$\tilde{L}_a^n \Rightarrow 0$$

as  $n \rightarrow \infty$ .

Before proving Lemma 2, we introduce the next lemma, which is a special case of the martingale functional central limit theorem. The proof of Lemma 3 can be found, for example, in Whitt (2007). Lemma 3 is used in the proof of Lemma 2. It is also used in the proof of Proposition 2.

**Lemma 3.** Let  $\{(M^n(t), \mathcal{G}^n(t)); t \geq 0\}$  be a local martingale with  $M^n(0) = 0$  for each  $n \geq 1$ . Assume that for any  $T > 0$ ,

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

as  $n \rightarrow \infty$ . Then  $M^n \Rightarrow 0$  in  $\mathbb{D}$  as  $n \rightarrow \infty$ .

*Proof of Lemma 2.* It follows from Lemma 1 that  $\{(\tilde{L}_a^n(t), \mathcal{F}^n(t)); t \geq 0\}$  is a martingale with quadratic variation

$$[\tilde{L}_a^n](t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \left( 1_{\{\gamma_i^n \leq n^{-1/2}a\}} - F(n^{-1/2}a) \right)^2.$$

Since  $F$  is right-continuous at 0, then for any  $T > 0$ , (7) implies

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[ [\tilde{L}_a^n](T) \right] = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^{\lfloor nT \rfloor} \mathbb{E} \left[ \left( \mathbf{1}_{\{\gamma_i^n \leq n^{-1/2}a\}} - F(n^{-1/2}a) \right)^2 \right] \leq \lim_{n \rightarrow \infty} TF(n^{-1/2}a) = 0,$$

which leads to  $[\tilde{L}_a^n](T) \Rightarrow 0$  as  $n \rightarrow \infty$ . Since  $\sup_{0 < t \leq T} |\tilde{L}_a^n(t) - \tilde{L}_a^n(t-)| \leq n^{-1/2}$ , we have

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

Then it follows from Lemma 3 that  $\tilde{L}_a^n \Rightarrow 0$  as  $n \rightarrow \infty$ .  $\square$

For each  $a > 0$ , we next define

$$G_a^n(t) = \sum_{i=E^n(t)+1}^{E^n(t+n^{-1/2}a)} \mathbf{1}_{\{\gamma_i^n \leq n^{-1/2}a\}}$$

that counts the number of customers that have arrived at the  $n$ th system during the time interval  $(t, t + n^{-1/2}a]$  and whose patience times are no more than  $n^{-1/2}a$ . In diffusion scaling, let

$$\tilde{G}_a^n(t) = \frac{1}{\sqrt{n}} G_a^n(t).$$

**Lemma 4.** *Assume (7)–(9) hold. Then for any  $a > 0$ ,*

$$\tilde{G}_a^n \Rightarrow 0$$

as  $n \rightarrow \infty$ .

*Proof.* Note that for each  $n \geq 1$  and  $t \geq 0$ ,

$$G_a^n(t) = F(n^{-1/2}a)(E^n(t + n^{-1/2}a) - E^n(t)) + L_a^n(\bar{E}^n(t + n^{-1/2}a)) - L_a^n(\bar{E}^n(t)).$$

Thus, in diffusion scaling,

$$\tilde{G}_a^n(t) = F(n^{-1/2}a)(\tilde{E}^n(t + n^{-1/2}a) - \tilde{E}^n(t) + a\lambda) + \tilde{L}_a^n(\bar{E}^n(t + n^{-1/2}a)) - \tilde{L}_a^n(\bar{E}^n(t)).$$

As a sequence of processes indexed by  $n$ , the third term on the right side  $\tilde{L}_a^n \circ \bar{E}^n \Rightarrow 0$  as  $n \rightarrow \infty$  due to (22), Lemma 2, and the random-time-change theorem. Similarly, the second term  $\tilde{L}_a^n \circ \bar{E}^n(\cdot + n^{-1/2}a) \Rightarrow 0$ . It follows from assumption (9) that

$$\lim_{c \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq T} \left| \tilde{E}^n(t + n^{-1/2}a) - \tilde{E}^n(t) \right| > c \right] = 0,$$

from which the first term (as a process) converges to zero in distribution by using the fact that  $F(n^{-1/2}a) \rightarrow 0$  as  $n \rightarrow \infty$ . It follows that  $\tilde{G}_a^n \Rightarrow 0$  as  $n \rightarrow \infty$ .  $\square$

*Proof of Proposition 4.* We first claim that

$$E^n(t + \delta) - E^n(t) \leq Q^n(t + \delta) + G_{n^{1/2}\delta}^n(t) \quad (24)$$

when  $0 < \delta < w^n(t)$ . To see (24), fix  $t \geq 0$ . For  $\delta \in (0, w^n(t))$  and  $\tau_i^n \in (t, t + \delta]$ , since  $t + \delta < t + w^n(t) \leq \tau_i^n + w_i^n$  (see Lemmas 13 and 14 in the appendix), the  $i$ th customer will not get into service by time  $t + \delta$ , so he will either be waiting in queue or have abandoned the system by then—the latter case implies  $\gamma_i^n < \delta$ . This proves (24).

Next, for any  $a > 0$ , (24) implies

$$\begin{aligned} & \mathbb{P} \left[ \sup_{0 \leq t \leq T} w^n(t) > n^{-1/2}a \right] \\ & \leq \mathbb{P} \left[ \inf_{0 \leq t \leq T} \left( E^n(t + n^{-1/2}a) - E^n(t) - Q^n(t + n^{-1/2}a) - G_a^n(t) \right) \leq 0 \right] \\ & = \mathbb{P} \left[ \sup_{0 \leq t \leq T} \left( \tilde{E}^n(t) - \tilde{E}^n(t + n^{-1/2}a) + \tilde{Q}^n(t + n^{-1/2}a) + \tilde{G}_a^n(t) \right) \geq \lambda a \right] \\ & \leq \mathbb{P} \left[ \sup_{0 \leq t \leq T} \left| \tilde{E}^n(t) - \tilde{E}^n(t + n^{-1/2}a) \right| > \frac{\lambda a}{3} \right] + \mathbb{P} \left[ \sup_{0 \leq t \leq T} \tilde{G}_a^n(t) > \frac{\lambda a}{3} \right] \\ & + \mathbb{P} \left[ \sup_{0 \leq t \leq 2T} \tilde{Q}^n(t) > \frac{\lambda a}{3} \right] + \mathbb{P} \left[ n^{-1/2}a > T \right]. \end{aligned}$$

By Lemma 4, we get

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq T} \sqrt{n} w^n(t) > a \right] \\ & \leq \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq T} \left| \tilde{E}^n(t) - \tilde{E}^n(t + n^{-1/2}a) \right| > \frac{\lambda a}{3} \right] + \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq 2T} \tilde{Q}^n(t) > \frac{\lambda a}{3} \right]. \end{aligned}$$

Therefore, (14) follows from (9) and (10).  $\square$

### 3.2 Proof of Proposition 2

This section is dedicated to the proof of Proposition 2. Again, we will use the martingale convergence theorem (Lemma 3). First, for  $i \geq 1$  and  $t \geq 0$  let

$$L_i^n = \sum_{j=1}^i \left( 1_{\{\gamma_j^n \leq w_j^n\}} - F(w_j^n) \right), \quad L^n(t) = L_{[nt]}^n, \quad \text{and} \quad \tilde{L}^n(t) = \frac{1}{\sqrt{n}} L^n(t).$$

**Lemma 5.** *Assume (6) holds. Then  $\{(L_i^n, \mathcal{F}_i^n); i \in \mathbb{Z}_+\}$  is a martingale.*

*Proof.* By Lemma 12 (in the appendix),  $w_j^n$  is  $\mathcal{F}_i^n$ -measurable for  $1 \leq j \leq i+1$ ; then  $L_i^n$  is  $\mathcal{F}_i^n$ -measurable. Since  $w_i^n$  is  $\mathcal{F}_{i-1}^n$ -measurable whereas  $\gamma_i^n$  is independent of  $\mathcal{F}_{i-1}^n$ ,

$$\mathbb{E} [L_i^n - L_{i-1}^n | \mathcal{F}_{i-1}^n] = \mathbb{E} \left[ 1_{\{\gamma_i^n \leq w_i^n\}} | \mathcal{F}_{i-1}^n \right] - F(w_i^n) = 0.$$

Also, we have  $\mathbb{E}[|L_i^n|] \leq i$ . So  $\{(L_i^n, \mathcal{F}_i^n); i \in \mathbb{Z}_+\}$  is a martingale.  $\square$

**Lemma 6.** *Assume (6) holds. Then  $\{(L^n(t), \mathcal{F}^n(t)); t \geq 0\}$  is a martingale with quadratic variation*

$$[L^n](t) = \sum_{i=1}^{\lfloor nt \rfloor} \left( 1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n) \right)^2. \quad (25)$$

*Proof.* It follows from Lemma 5 that  $L^n$  is adapted to  $\{\mathcal{F}^n(t); t \geq 0\}$ . It is a martingale because for  $0 \leq s \leq t$ ,  $\mathbb{E}[|L^n(t)|] = \mathbb{E}[|L^n_{\lfloor nt \rfloor}|] < \infty$  and  $\mathbb{E}[L^n(t) | \mathcal{F}^n(s)] = \mathbb{E}[L^n_{\lfloor nt \rfloor} | \mathcal{F}_{\lfloor ns \rfloor}^n] = L^n_{\lfloor ns \rfloor} = L^n(s)$ . Since  $L^n$  is piecewise constant and  $\sum_{i=1}^{\lfloor nt \rfloor} |1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n)| \leq nt$ ,  $L^n$  is a finite-variation process and (25) follows.  $\square$

**Lemma 7.** *Assume (6)–(10) hold. Then*

$$\tilde{L}^n \Rightarrow 0$$

as  $n \rightarrow \infty$ .

*Proof.* Using Lemma 6,  $\{(\tilde{L}^n(t), \mathcal{F}^n(t)); t \geq 0\}$  is a martingale with quadratic variation

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

By (18) and the fact  $F(w_i^n) \leq 1$ , we obtain

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

As  $\gamma_i^n$  is independent of  $\mathcal{F}_{i-1}^n$  but  $w_i^n$  is  $\mathcal{F}_{i-1}^n$ -measurable (see Lemma 12 in the appendix),

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

Then for any  $T > 0$ ,

$$\begin{aligned} \mathbb{E} \left[ [\tilde{L}^n](T) \right] &= \frac{1}{n} \sum_{i=1}^{\lfloor nT \rfloor} \mathbb{E} \left[ \left( 1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n) \right)^2 \right] \leq \frac{1}{n} \sum_{i=1}^{\lfloor nT \rfloor} \mathbb{E} [F(w_i^n)] \\ &\leq T \mathbb{E} \left[ \sup_{1 \leq i \leq nT} F(w_i^n) \right]. \end{aligned}$$

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

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

Then  $\tilde{L}^n \Rightarrow 0$  as  $n \rightarrow \infty$  follows from Lemma 3.  $\square$

*Proof of Proposition 2.* Since  $\tilde{A}_1^n(t) = \tilde{L}^n(\bar{E}^n(t))$ , then  $\tilde{A}_1^n \Rightarrow 0$  as  $n \rightarrow \infty$  follows from (22), Lemma 7, and the random-time-change theorem.  $\square$

### 3.3 Proof of Proposition 3

This section is devoted to the proof of Proposition 3. The key to the proof is to establish the following lemma.

**Lemma 8.** *Assume (7)–(10) hold. Then for any  $T > 0$ ,*

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

as  $n \rightarrow \infty$ .

Assuming Lemma 8, we now provide the proof of Proposition 3.

*Proof of Proposition 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

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

Lemma 8 leads to  $\tilde{A}_{21}^n \Rightarrow 0$  as  $n \rightarrow \infty$ . Also by (10) and Lemma 8, for any  $t \geq 0$ ,

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

Using (19) and (22),  $\sup_{1 \leq i \leq E^n(T)} |g(w_i^n) - \alpha| \Rightarrow 0$  as  $n \rightarrow \infty$ , so that

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

which concludes the proof.  $\square$

It remains to prove Lemma 8. For that, we need to establish two lemmas. In the first lemma, we demonstrate the stochastic boundedness of the process  $\tilde{A}_2$ .

**Lemma 9.** *Assume (7)–(10) hold. Then*

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \tilde{A}_2^n(t) > a \right] = 0 \quad (26)$$

for any  $t \geq 0$ .

*Proof.* Since

$$\tilde{A}_2^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} F(w_i^n) \leq \bar{E}^n(t) \sup_{1 \leq i \leq E^n(t)} \sqrt{n} w_i^n \sup_{1 \leq i \leq E^n(t)} g(w_i^n),$$

then

$$\begin{aligned} \mathbb{P} \left[ \tilde{A}_2^n(t) > a \right] &\leq \mathbb{P} \left[ \bar{E}^n(t) > 2\lambda t \right] + \mathbb{P} \left[ \sup_{1 \leq i \leq E^n(t)} g(w_i^n) > \alpha + \varepsilon \right] \\ &\quad + \mathbb{P} \left[ \sup_{1 \leq i \leq E^n(t)} \sqrt{n} w_i^n > \frac{a}{2\lambda(\alpha + \varepsilon)t} \right] \end{aligned}$$

for any  $t > 0$  and  $\varepsilon > 0$ . By Proposition 4 in Section 3.1 and Lemma 14 in the appendix,

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{1 \leq i \leq E^n(t)} \sqrt{n} w_i^n > a \right] = 0.$$

Using (19), (22), and the random-time-change theorem,  $\sup_{1 \leq i \leq E^n(t)} |g(w_i^n) - \alpha| \Rightarrow 0$  for any  $t > 0$ . Also by (22) and the fact  $\tilde{A}_2^n(0) = 0$ , we see (26) hold.  $\square$

The next lemma establishes a pair of inequalities. The inequalities allow us to convert the summation of offered waiting times into an integral of the queue-length process, and therefore allow us to prove Lemma 8.

**Lemma 10.** *For any  $t \geq 0$ ,*

$$\int_0^t Q^n(s) ds \leq \sum_{i=1-Q^n(0)}^{E^n(t)} (\gamma_i^n \wedge w_i^n) \leq \int_0^{t+w^n(t)} Q^n(s) ds. \quad (27)$$

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

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

Then  $q_i^n(t) = \int_0^t b_i^n(s) ds$  is the  $i$ th customer's cumulative waiting time by  $t$ . Note that  $q_i^n(t) \leq \gamma_i^n \wedge w_i^n$ , and if the  $i$ th customer has got service or abandoned the queue by  $t$ ,  $q_i^n(t) = \gamma_i^n \wedge w_i^n$  holds. For any  $0 \leq s \leq t$ , the queue length at time  $s$  can be counted by

$$Q^n(s) = \sum_{i=1-Q^n(0)}^{E^n(t)} b_i^n(s).$$

Then

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

For  $1 - Q^n(0) \leq i \leq E^n(t)$ , the  $i$ th customer should have got into service or abandoned the system by time  $t + w^n(t)$ , because  $\tau_i^n + w_i^n \leq t + w^n(t)$  (see Lemmas 13 and 14 in the appendix). Then  $q_i^n(t + w^n(t)) = \gamma_i^n \wedge w_i^n$ . It follows that

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

□

*Proof of Lemma 8.* Since  $w_i^n \leq w^n(0)$  for  $1 - Q^n(0) \leq i \leq 0$  (see Lemma 14 in the appendix),

$$\frac{1}{\sqrt{n}} \sum_{i=1-Q^n(0)}^0 (\gamma_i^n \wedge w_i^n) \leq \tilde{Q}^n(0) w^n(0) \Rightarrow 0$$

as  $n \rightarrow \infty$ , by (10) and (15). Since  $\sup_{0 \leq t \leq T} (t + w^n(t)) = T + w^n(T)$  (see Lemma 13 in the appendix), using (10) and (15) again,

$$\sup_{0 \leq t \leq T} \int_t^{t+w^n(t)} \tilde{Q}^n(s) ds \leq \sup_{0 \leq t \leq T+w^n(T)} \tilde{Q}^n(t) \sup_{0 \leq t \leq T} w^n(t) \Rightarrow 0.$$

So Lemma 10 implies

$$\sup_{0 \leq t \leq T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (\gamma_i^n \wedge w_i^n) - \int_0^t \tilde{Q}^n(s) ds \right| \Rightarrow 0. \quad (28)$$

By (15), Proposition 2, Lemma 9, and Lemma 14 in the appendix,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} 1_{\{\gamma_i^n \leq w_i^n\}} w_i^n \leq (\tilde{A}_1^n(T) + \tilde{A}_2^n(T)) \sup_{1 \leq i \leq E^n(T)} w_i^n \Rightarrow 0.$$

Since

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

then

$$\sup_{0 \leq t \leq T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} w_i^n - \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (\gamma_i^n \wedge w_i^n) \right| \Rightarrow 0. \quad (29)$$

The proof of the lemma follows from (28) and (29).  $\square$

### 3.4 Proof of Proposition 1

Recall that  $G^n(t)$  is the number of customers that have abandoned the system during  $(0, t]$ , and  $A^n(t)$  is the number of customers who have arrived during  $(0, t]$  but will eventually abandon the system. Clearly, for each  $n$  and  $t \geq 0$ ,

$$G^n(t) \leq A^n(t).$$

We now establish a lower-bound for  $G^n$ . For  $t \geq 0$ , define

$$\zeta^n(t) = \max \{i \in \mathbb{Z} : \tau_i^n + w_i^n \leq t\}.$$

By Lemma 13 in the appendix,  $\tau_i^n + w_i^n \leq t$  for every  $i \leq \zeta^n(t)$ . Therefore, for each customer  $i \leq \zeta^n(t)$ , the customer should have entered service or abandoned the queue by time  $t$ . Since

$$A^n(\tau_{\zeta^n(t)}^n) = \sum_{i=1-Q^n(0)}^{\zeta^n(t)} 1_{\{\gamma_i^n \leq w_i^n\}},$$

one has  $A^n(\tau_{\zeta^n(t)}^n) \leq G^n(t)$ . Therefore, we have

$$A^n(\tau_{\zeta^n(t)}^n) \leq G^n(t) \leq A^n(t). \quad (30)$$

We will prove Proposition 1 by showing that the upper- and lower-bounds in (30) of  $G^n$  are asymptotically close. Before we present the proof, we establish the following lemma that is used in the proof of Proposition 1.

**Lemma 11.** *Assume (6)–(10) hold. Then for any  $T > 0$ ,*

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t + w^n(t)) - \tilde{A}^n(t) \right| \Rightarrow 0 \quad (31)$$

as  $n \rightarrow \infty$ .

*Proof.* Recall the decomposition of  $A^n$  in (13). It follows that

$$\tilde{A}^n(t + w^n(t)) - \tilde{A}^n(t) = \tilde{A}_1^n(t + w^n(t)) - \tilde{A}_1^n(t) + \tilde{A}_2^n(t + w^n(t)) - \tilde{A}_2^n(t).$$

Therefore,

$$\begin{aligned} \sup_{0 \leq t \leq T} \left| \tilde{A}^n(t + w^n(t)) - \tilde{A}^n(t) \right| &\leq \sup_{0 \leq t \leq T} \left| \tilde{A}_1^n(t + w^n(t)) - \tilde{A}_1^n(t) \right| + \sup_{0 \leq t \leq T} \left| \tilde{A}_1^n(t) \right| \\ &\quad + \sup_{0 \leq t \leq T} \left| \tilde{A}_2^n(t + w^n(t)) - \tilde{A}_2^n(t) \right|. \end{aligned}$$

Note that

$$\tilde{A}_2^n(t + w^n(t)) - \tilde{A}_2^n(t) \leq \left( \tilde{E}^n(t + w^n(t)) - \tilde{E}^n(t) + \sqrt{n}\lambda w^n(t) \right) \sup_{0 \leq t \leq T + w^n(T)} F(w^n(t)).$$

By (9), (14), and (16), we obtain

$$\sup_{0 \leq t \leq T} \left| \tilde{A}_2^n(t + w^n(t)) - \tilde{A}_2^n(t) \right| \Rightarrow 0,$$

which, together with (15) and Proposition 2, leads to (31).  $\square$

*Proof of Proposition 1.* By definition, for each  $t \geq 0$ ,  $\tau_{\zeta^n(t)}^n \leq t$ . Since  $A^n$  is nondecreasing,

$$A^n(t) - A^n(\tau_{\zeta^n(t)}^n) \geq 0.$$

When  $t \leq \tau_{\zeta^n(t)+1}^n$ ,

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

Since  $\tau_{\zeta^n(t)+1}^n + w_{\zeta^n(t)+1}^n > t$ , we have

$$A^n(t) - A^n(\tau_{\zeta^n(t)}^n) \leq A^n(t) - A^n(\tau_{\zeta^n(t)+1}^n) + 1 \leq A^n(\tau_{\zeta^n(t)+1}^n + w_{\zeta^n(t)+1}^n) - A^n(\tau_{\zeta^n(t)+1}^n) + 1.$$

Also, when  $t > \tau_{\zeta^n(t)+1}^n$ , by Lemmas 13 and 14 in the appendix,  $\tau_{\zeta^n(t)+1}^n + w_{\zeta^n(t)+1}^n \leq t + w^n(t)$ . It follows that for any  $T > 0$ ,

$$\sup_{0 \leq t \leq T} \left( A^n(t) - A^n(\tau_{\zeta^n(t)}^n) \right) \leq \sup_{0 \leq t \leq T} \left( A^n(t + w^n(t)) - A^n(t) \right) + 1.$$

Then Lemma 11 yields

$$\sup_{0 \leq t \leq T} \left( \tilde{A}^n(t) - \tilde{A}^n(\tau_{\zeta^n(t)}^n) \right) \Rightarrow 0$$

as  $n \rightarrow \infty$ , from which and (30) the proposition follows.  $\square$

## 4 Proof of Theorem 2

*Proof of Theorem 2.* For  $j = 1, 2$ , let  $r_{j,i}(t) \geq 0$  be the remaining service time of the  $i$ th customer of  $\Sigma_j$  at time  $t \geq 0$ . (See (36) in Section A.1 for the definition of remaining service time for a customer.) For  $0 \leq s \leq t$ , we have  $r_{j,i}(s) \geq r_{j,i}(t)$ . Let  $X_j(t)$ , with  $X_j(0) = X(0)$ , denote the number of customers in  $\Sigma_j$  at time  $t$ . Recall that  $E(t)$  is the number of customer arrivals to both queues during the time interval  $(0, t]$ . The set of customers being served in  $\Sigma_j$  at time  $t$  can be represented by

$$\Pi_j(t) = \left\{ i \in \mathbb{Z} : 1 - X(0) \leq i \leq E(t), r_{j,i}(t) > 0, \sum_{k=1-X(0)}^i 1_{\{r_{j,k}(t) > 0\}} \leq n \right\}. \quad (32)$$

Set  $\xi_0 = 0$  and let  $0 < \xi_1 \leq \xi_2 \leq \dots$  be the service completion times in  $\Sigma_2$ . At time  $\xi_0 = 0$ ,  $r_{1,i}(\xi_0) = r_{2,i}(\xi_0)$  for all  $i \geq 1 - X(0)$  by assumption. Suppose that  $r_{1,i}(t) \leq r_{2,i}(t)$  for  $0 \leq t \leq \xi_m$ . Then for any  $i \in \Pi_2(\xi_m)$ , (32) implies either  $i \in \Pi_1(\xi_m)$  or  $r_{1,i}(\xi_m) = 0$ ; since for any  $t \in (\xi_m, \xi_{m+1}]$ ,  $r_{2,i}(t) = r_{2,i}(\xi_m) - (t - \xi_m) \geq 0$  and  $r_{1,i}(t) = (r_{1,i}(\xi_m) - (t - \xi_m))^+$ , then  $r_{1,i}(t) \leq r_{2,i}(t)$ . If  $i \notin \Pi_2(\xi_m)$ ,  $r_{1,i}(t) \leq r_{2,i}(t)$  also holds for  $t \in (\xi_m, \xi_{m+1}]$  because  $r_{2,i}(t) = r_{2,i}(\xi_m)$  and  $r_{1,i}(t) \leq r_{1,i}(\xi_m)$ . By induction, we get  $r_{1,i}(t) \leq r_{2,i}(t)$  for all  $t \geq 0$ . Let  $m(t) = E(t)$  if  $\Pi_2(t) = \emptyset$ , and  $m(t) = \max \Pi_2(t)$  otherwise. Then  $Q_2(t) = E(t) - m(t)$  and  $\sum_{i=1-X(0)}^{m(t)-1} 1_{\{r_{2,i}(t) > 0\}} < n$ . Since  $r_{1,i}(t) \leq r_{2,i}(t)$  for each  $i$ ,  $\sum_{i=1-X(0)}^{m(t)-1} 1_{\{r_{1,i}(t) > 0\}} < n$ , which implies  $Q_1(t) \leq E(t) - m(t)$ . Therefore, we obtain  $Q_1(t) \leq Q_2(t)$ .  $\square$

## A On $G/G/n + G$ queues

In this section, we present some results for a general  $G/G/n + G$  queue, where the inter-arrival times, the service times, and the patience times are three arbitrary sequences of random variables.

### A.1 Measurability of offered waiting times

Recall the offered waiting time  $w_i$  is the waiting time in queue of the  $i$ th customer if he were infinitely patient. We first show that the offered waiting times can be recursively defined. We then prove that the offered waiting time sequence  $\{w_i; i \in \mathbb{Z}\}$  is adapted to a certain filtration.

To define the offered waiting times mathematically, it is convenient to define the remaining service time process  $\{r_i(t); t \geq 0\}$  for each customer  $i \in \mathbb{Z}$ . For each time  $t \geq 0$ ,  $r_i(t)$  is the remaining service time for the  $i$ th customer at time  $t$ . Fix a  $\omega \in \Omega$ . For each  $i \leq -X(0, \omega)$ , let  $w_i(\omega) = 0$  and  $r_i(t, \omega) = 0$  for  $t \geq 0$ . For  $i > -X(0, \omega)$ , let

$$w_i(\omega) = \inf \left\{ t \geq 0 : \sum_{j \leq i-1} 1_{\{r_j(\tau_i(\omega) + t, \omega) > 0\}} < n \right\} \quad (33)$$

and

$$r_i^a(t, \omega) = 1_{\{t < \tau_i(\omega) + \gamma_i(\omega)\}} v_i(\omega), \quad (34)$$

$$r_i^s(t, \omega) = 1_{\{t < \tau_i(\omega) + w_i(\omega)\}} v_i(\omega) + 1_{\{t \geq \tau_i(\omega) + w_i(\omega)\}} (v_i(\omega) - t)^+, \quad (35)$$

$$r_i(t, \omega) = 1_{\{0 \leq \gamma_i(\omega) \leq w_i(\omega)\}} r_i^a(t, \omega) + (1 - 1_{\{0 \leq \gamma_i(\omega) \leq w_i(\omega)\}}) r_i^s(t, \omega) \quad (36)$$

for  $t \geq 0$ . Equation (33) says that if no arrival occurs after the  $(i-1)$ st customer,  $w_i$  is the amount of time beyond  $\tau_i$  until one of the  $n$  servers becomes idle. Equation (36) says that the  $i$ th customer will abandon the queue if  $0 \leq \gamma_j \leq w_j$ , and in this case his remaining service time is given by  $r_i^a(t)$ ; otherwise, he either has received or will receive service, and his remaining service time is  $r_i^s(t)$ . Clearly, recursions (33)–(36) define  $w_i(\omega)$  for each  $\omega \in \Omega$  and  $i \in \mathbb{Z}$ .

For each  $k \in \mathbb{Z}_+$ , let

$$\mathcal{F}_k = \sigma \{ \tau_{i+1}, v_i, \gamma_i; i \leq k \}.$$

The main result of this section is the following measurability result for the offered waiting times.

**Lemma 12.** *For a  $G/G/n + G$  queue,  $w_i$  is  $\mathcal{F}_k$ -measurable for  $k \in \mathbb{Z}_+$  and  $i \leq k+1$ .*

*Proof.* 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  $\mathcal{F}_{k,m} = \sigma \{ \tau_{i+1}, v_{i,m}, \gamma_{i,m}; i \leq k \}$ . We have  $\mathcal{F}_k = \bigvee_{m=0}^{\infty} \mathcal{F}_{k,m}$  because  $v_i = \lim_{m \rightarrow \infty} v_{i,m}$  and  $\gamma_i = \lim_{m \rightarrow \infty} \gamma_{i,m}$  for all  $i \in \mathbb{Z}$ . Given  $k \geq 0$  and  $m \geq 0$ , we define  $w_{i,m}$  and  $r_{i,m}(t)$  recursively via a similar procedure as in (33)–(36) 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

$$w_{i,m} = \inf \left\{ t \geq 0 : \sum_{j \leq i-1} 1_{\{r_{j,m}(\tau_i+t) > 0\}} < n \right\} \quad (37)$$

and

$$r_{i,m}^a(t) = 1_{\{t < \tau_i + \gamma_{i,m}\}} v_{i,m}, \quad (38)$$

$$r_{i,m}^s(t) = 1_{\{t < \tau_i + w_{i,m}\}} v_{i,m} + 1_{\{t \geq \tau_i + w_{i,m}\}} (v_{i,m} - t)^+, \quad (39)$$

$$r_{i,m}(t) = 1_{\{0 \leq \gamma_{i,m} \leq w_{i,m}\}} r_{i,m}^a(t) + (1 - 1_{\{0 \leq \gamma_{i,m} \leq w_{i,m}\}}) r_{i,m}^s(t) \quad (40)$$

for  $t \geq 0$ . By (37), 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

$$w_{i,m} \text{ is } \mathcal{F}_{k,m}\text{-measurable for each } i \leq k+1. \quad (41)$$

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 (41), 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  $t \geq 0$ ,  $r_{i,m}^a(t)$ ,  $r_{i,m}^s(t)$ , and  $r_{i,m}(t)$  are  $\mathcal{F}_{k,m}$ -measurable. By (38)–(40), 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  $\tau_i$  is  $\mathcal{F}_{k,m}$ -measurable. Since  $\{w_{j+1,m} \leq t\} = \{\sum_{i \leq j} 1_{\{r_{i,m}(\tau_{j+1}+t) > 0\}} < n\}$ , we conclude that  $w_{j+1,m}$  is  $\mathcal{F}_{k,m}$ -measurable, thus proving (41).

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  $w_i = \lim_{m \rightarrow \infty} w_{i,m}$ . Therefore,  $w_i$  is  $\mathcal{F}_k$ -measurable for  $i \leq k + 1$ .  $\square$

## A.2 Virtual and offered waiting times

In this section, we present two frequently used lemmas on virtual and offered waiting times in the  $G/G/n + G$  queue. Similar to the offered waiting time defined in (33), one can define virtual waiting time  $w(t)$  via

$$w(t) = \inf \left\{ s \geq 0 : \sum_{i=1-X(0)}^{E(t)} 1_{\{r_i(t+s) > 0\}} < n \right\}. \quad (42)$$

The  $i$ th 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  $i$ th customer's *nominal service-starting time*. It follows from (33) that

$$\tau_i + w_i = \inf \left\{ s \geq \tau_i : \sum_{j \leq i-1} 1_{\{r_j(s) > 0\}} < n \right\}. \quad (43)$$

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

$$t + w(t) = \inf \left\{ s \geq t : \sum_{i=1-X(0)}^{E(t)} 1_{\{r_i(s) > 0\}} < n \right\} \quad (44)$$

where  $E(t)$  is again the number of arrivals to the queue in  $(0, t]$ . The next lemma states that although there is customer abandonment in the  $G/G/n + G$  queue, the nominal service-starting times are still ordered in the FIFO fashion as in a  $G/G/n$  queue without abandonment.

**Lemma 13.** *For a  $G/G/n + G$  queue,*

$$t_1 + w(t_1) \leq t_2 + w(t_2)$$

for  $0 \leq t_1 \leq t_2$ . Similarly,

$$\tau_i + w_i \leq \tau_j + w_j$$

for any  $i, j \in \mathbb{Z}$  with  $i \leq j$ .

*Proof.* By (34)–(36), the process  $1_{\{r_i(\cdot) > 0\}}$  is right-continuous. Hence, for  $s = t + w(t)$ ,  $\sum_{i=1-X(0)}^{E(t)} 1_{\{r_i(s) > 0\}} < n$ . If there are  $0 \leq t_1 \leq t_2$  such that  $t_1 + w(t_1) > t_2 + w(t_2)$ , then

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

which leads to a contradiction. Thus,  $t_1 + w(t_1) \leq t_2 + w(t_2)$ . Using (43) we can prove  $\tau_i + w_i \leq \tau_j + w_j$  for  $i \leq j$  via a similar argument.  $\square$

Our last lemma relates the offered waiting times to the virtual waiting times at the corresponding arrival times.

**Lemma 14.** *For a  $G/G/n + G$  queue,*

$$w(\tau_i-) \leq w_i \leq w(\tau_i)$$

for  $i \geq 1$ ;

$$w_i \leq w(0)$$

for  $i \leq 0$ .

*Proof.* Let  $y(t) = \inf\{s \geq 0 : \sum_{i=1-X(0)}^{E(t-)} 1_{\{r_i(s) > 0\}} < n\}$ . Then for any  $t' \in [\tau_{E(t-)}, t)$ , since  $E(t') = E(t-)$ , using (44) we have  $t' + w(t') = t' \vee y(t)$ ; thus,  $w(t') = (y(t) - t')^+$  and  $w(t-) = (y(t) - t)^+$ . Since  $E(\tau_i-) < i \leq E(\tau_i)$ , it follows from (43) and (44) 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  $\tau_i$ .

Using  $E(0) = 0$  and  $\tau_i = 0$  for  $i \leq 0$ ,  $w_i \leq w(0)$  also follows from (43) and (44).  $\square$

## B On the initial assumption (11)

**Lemma 15.** *Consider a sequence of  $G/G/n + GI$  queues that satisfies (7)–(10). Assume that the patience times of the customers who are waiting in queue at time 0 are iid and follow the distribution  $F$ . Then the initial assumption (11) holds.*

*Proof.* For each  $n$ , let  $\{z_i^n\}_{i=-\infty}^0$  be a sequence of iid random variables having the distribution  $F$ , where  $z_i^n = \gamma_i^n$  for  $i = 1 - Q^n(0), \dots, 0$ . By Lemma 14,

$$\tilde{G}_0^n = \frac{1}{\sqrt{n}} \sum_{i=1-Q^n(0)}^0 1_{\{z_i^n \leq w_i^n\}} \leq \tilde{H}_0^n,$$

where

$$\tilde{H}_0^n = \frac{1}{\sqrt{n}} \sum_{i=1-Q^n(0)}^0 1_{\{z_i^n \leq w^n(0)\}}.$$

Then (11) follows once we prove  $\tilde{H}_0^n \Rightarrow 0$  as  $n \rightarrow \infty$ .

For  $a > 0$  and  $b > 0$ , let

$$\tilde{H}_{a,b}^n = \frac{1}{\sqrt{n}} \sum_{1-b\sqrt{n} \leq i \leq 0} 1_{\{z_i^n \leq n^{-1/2}a\}},$$

which satisfies  $\mathbb{E}[\tilde{H}_{a,b}^n] \leq bF(n^{-1/2}a) \rightarrow 0$  as  $n \rightarrow \infty$  by (7). Hence,  $\tilde{H}_{a,b}^n \Rightarrow 0$  as  $n \rightarrow \infty$ . Given  $\varepsilon > 0$ , we have

$$\mathbb{P}[\tilde{H}_0^n > \varepsilon] \leq \mathbb{P}[\tilde{H}_{a,b}^n > \varepsilon] + \mathbb{P}[w^n(0) > n^{-1/2}a] + \mathbb{P}[\tilde{Q}^n(0) > b],$$

so that

$$\limsup_{n \rightarrow \infty} \mathbb{P}[\tilde{H}_0^n > \varepsilon] \leq \limsup_{n \rightarrow \infty} \mathbb{P}[w^n(0) > n^{-1/2}a] + \limsup_{n \rightarrow \infty} \mathbb{P}[\tilde{Q}^n(0) > b].$$

Letting  $a \rightarrow \infty$  and  $b \rightarrow \infty$ , it follows from Proposition 4 and (10) that  $\tilde{H}_0^n \Rightarrow 0$  as  $n \rightarrow \infty$ .  $\square$

## References

- AKSIN, Z., ARMONY, M. and MEHROTRA, V. (2007). The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management*, **16** 665–688.
- DAI, J. G., HE, S. and TEZCAN, T. (2009). Many-server diffusion limits for  $G/Ph/n + GI$  queues. School of Industrial and Systems Engineering, Georgia Institute of Technology.
- GANS, N., KOOLE, G. and MANDELBAUM, A. (2003). Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, **5** 79–141.
- GARNETT, O., MANDELBAUM, A. and REIMAN, M. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management*, **4** 208–227.
- HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, **29** 567–588.
- PROTTER, P. (2005). *Stochastic Integration and Differential Equations*. 2nd ed. Springer, New York, NY.
- PUHALSKII, A. and REIMAN, M. (2000). The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, **32** 564–595. Correction: **36**, 971 (2004).

- REED, J. E. (2007). The  $G/GI/N$  queue in the Halfin-Whitt regime I: infinite server queue system equations. Submitted for publication.
- WHITT, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research*, **54** 37–54.
- WHITT, W. (2007). Proofs of the martingale FCLT. *Probability Surveys*, **4** 268–302.
- ZELTYN, S. and MANDELBAUM, A. (2005). Call centers with impatient customers: many-server asymptotics of the  $M/M/n + G$  queue. *Queueing Systems*, **51** 361–402.