MATHEMATICS OF OPERATIONS RESEARCH
Vol. 37, No. 1, February 2012, pp. 41–65
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.1110.0530

Queues with Many Servers and Impatient Customers

Avishai Mandelbaum
Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 3200, Israel,
avim@tx.technion.ac.il

Petar Momčilović
Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida 32611,
petar@ise.ufl.edu

The asymptotic many-server queue with abandonments, $G/GI/N+GI$, is considered in the quality- and efficiency-driven (QED) regime. Here the number of servers and the offered load are related via the square-root rule, as the number of servers increases indefinitely. QED performance entails short waiting times and scarce abandonments (high quality) *jointly* with high servers' utilization (high efficiency), which is feasible when many servers cater to a single queue. For the $G/GI/N+GI$ queue, we derive diffusion approximations for both its queue-length and virtual-waiting-time processes. Special cases, for which closed-form analysis is provided, are the $G/M/N+GI$ and $G/D/N+GI$ queues, thus expanding and generalizing existing results.

Key words: multiserver queue; abandonment; QED regime; Halfin–Whitt regime; diffusion approximation
MSC2000 subject classification: Primary: 60K25; secondary: 90B22
OR/MS subject classification: Primary: queues-balking and reneging, queues-multichannel; secondary: queues-limit theorems, queues-diffusion models
History: Received March 15, 2009; revised February 26, 2010. Published online in *Articles in Advance* January 13, 2012.

1. Introduction. The quality- and efficiency-driven (QED) regime achieves, jointly, high levels of system's *efficiency*, as manifested by servers' high utilization, and service *quality*, namely, customers' short waiting times and hence scarce abandonments. QED performance is achievable in carefully balanced queueing systems with many servers—indeed, with few servers, efficiency and quality must be traded off against each other. Within the $G/GI/N$ framework, or more precisely $G/GI/N+GI$, the QED regime with abandonments is characterized by the relation $N=R+\beta\sqrt{R}+o(\sqrt{R})$, for some scalar β; here N is the number of servers and R is the offered load, namely, the arrival rate multiplied by average service time. (This *square-root staffing* relation also characterizes the QED regime without abandonments, but β must then be taken positive to ensure stability.)

1.1. Relevance. Recent interest in multiserver queues with impatient customers is due to their applicability in modeling medium-to-large-scale customer call/contact centers. In such service operations, abandonments arise naturally and, in fact, *must* be accounted for in models (see §2 in Garnett et al. [12] for an elaboration). Additionally, well-run call centers are QED (Gans et al. [11]) or some relatives of it (e.g., ED + QED, as in Mandelbaum and Zeltyn [24]). But QED queues also arise beyond call centers. To wit, waiting time in QED call centers is naturally measured in seconds and service times in minutes. This one-order time reduction (from minutes to seconds in the case of call centers) is a QED characteristic; indeed, it also arises in transportation (searching for parking takes minutes whereas parking time takes hours) and in healthcare (sojourn times in emergency departments take hours whereas hospitalization is days). Significantly, the abandonment phenomenon is relevant in all these examples, which is perhaps surprising for the latter; yet, a nonnegligible fraction of patients leave emergency departments without being seen by a doctor (Green et al. [13]).

1.2. Related research. Although the QED regime (without abandonments) can be traced back to Erlang [9] and Jagerman [16], the regime was first formalized by Halfin and Whitt [14]; for recent results on the QED regime, see Mandelbaum and Momčilović [21], Reed [27], Kaspi and Ramanan [19], Puhalskii and Reed [26], and references therein, with Reed's framework (Reed [27]) for the $G/GI/N$ queue being especially relevant. The $M/M/N+M$ (Erlang-A) system in the QED regime (with abandonments) was considered in Garnett et al. [12]. Extensions to the model with generally distributed abandonments can be found in Zeltyn and Mandelbaum [34], Zeltyn [33], Mandelbaum and Zeltyn [24], and Reed and Tezcan [29]. The $M/M/N+G$ system in the efficiency-driven regime was analyzed by Whitt [32]; for a summary of performance measures of this system, see Mandelbaum and Zeltyn [23]. Recently, fluid limits of many-server queues with abandonment were considered in Kang and Ramanan [18]. Independently of our work, many-server queues with customer abandonment were investigated by Dai and He [7], where the focus parallels our Lemma 3.8 (the authors establish a relation between the abandonment-count and queue-length processes). The literature on queues with

abandonments is extensive and includes models with various features; we refer the reader to the discussions in Garnett et al. [12] and Zeltyn [33].

1.3. Contributions. We consider a $G/GI/N+GI$ system in the QED regime. The limiting scaled number-of-customers-in-system process is described in terms of a nonlinear operator (Corollary 4.1); a corresponding result for the limiting scaled waiting-time processes (virtual and offered) is obtained as well (Corollary 4.2). In the case when there is no abandonment, our operator coincides with the one earlier obtained in Reed [27].

The proofs of our main results are based on two relations: (i) between corresponding systems with and without abandonment (Proposition 3.1), and (ii) between the queue-length and offered-waiting-time processes (Lemma 3.8). The first relation, in conjunction with results from Reed [27], is used to obtain upper bounds for the queue-length and waiting-time processes; the bounds are tight enough to yield the exact orders of magnitude for QED convergence. This enables us to approximate the abandonment process, properly centered and scaled (see (14) and Lemma 3.4). The approximation in (14), in turn, reveals the QED limit of the abandonment process, which paves the way to other QED limits. Relation (ii) is central in our paper because it allows one to circumvent the complex relation between the queue-length and the abandonment processes, which is necessary for obtaining limits of queue lengths. Indeed, the relation is complex because abandonments are determined by (offered) waiting times which, in turn, depend on queue length via a "first-passage-time" operator, as in Puhalskii [25] and Talreja and Whitt [30]. Still, one cannot directly use Puhalskii [25] or Talreja and Whitt [30] to deduce limits of waiting times from those of queue lengths since we could not analyze the latter in isolation. This is in contrast to Reed [27], which first derives limits of queue length, then uses Puhalskii [25] to deduce directly limits of waiting times.

To summarize, Lemma 3.8 shows that queue length and waiting times are asymptotically "close." Therefore, via Lemma 3.4 and (14), we express the abandonment process in terms of queue length. This results in representing the queue length in terms of itself, which is resolved through a sample-path mapping, as introduced in (25).

The above provides a justification for the need to develop techniques and tools that are not required in Reed [27], who analyzed the QED $G/GI/N$ queue. More specifically, for our QED $G/GI/N+GI$ queue, queue length and waiting times must be analyzed jointly, as already discussed. Then, the mapping (25) that characterizes the limiting queue-length processes generalizes the one in Reed [27]. (A mapping corresponding to waiting times is introduced in (34).) Our general result can be made explicit (see §5) in two special cases that have not yet been analyzed: Example 5.2 provides functional limit theorems that correspond to and further generalize Zeltyn [33] and Zeltyn and Mandelbaum [34], which considered $M/M/N+GI$ in steady state; Example 5.3 adds abandonments to Mandelbaum and Momčilović [21], which considered process limits for $G/D/N$. Finally, our results demonstrate that the role of the patience distribution in the QED regime is captured merely by the value of its density at the origin. This is practically important since patience data is censored (only lower bounds for patience are available for served customers) and possibly highly censored (e.g., 3% abandoning). We thus suggest an estimator for the patience density at the origin, based on a transient analogue of the steady-state relation between the probability of abandonment and the expected waiting time (Mandelbaum and Zeltyn [22]).

1.4. Contents. The paper is organized as follows. In the next section we describe the model and introduce the QED regime. Section 3 contains preliminaries; in particular, we discuss a relationship between the systems with and without abandonment, the infinite-service process associated with the arrival and service processes, processes associated with initial conditions and abandonment, the queue-lenth process, and a relationship between the queue length and offered waiting time. The main results of the paper are presented in §4. Examples are discussed in §5, future research is outlined in §6, and some proofs are provided in §7.

1.5. Notation. Denote by $D[0,\infty)$ the space of all real-valued functions on $[0,\infty)$ that are right continuous with left limits (r.c.l.l.), endowed with the standard Skorohod J_1 topology. The J_1 metric is denoted by $d_{J_1}(\cdot,\cdot)$ and the uniform metric u is defined by the uniform norm

$$\|x\|_T = \sup_{0\le t\le T} |x(t)|,$$

for $x\in D[0,\infty)$ and $T\ge 0$; similarly, the L^1 metric is defined by the L^1 norm

$$d^T_{L^1}(x,y) = \int_0^T |x(t)-y(t)|\,dt, \tag{1}$$

for $x, y \in D[0,\infty)$ and $T \geq 0$. Product metric spaces $(D^k[0,\infty), d_{J_1}^k)$ and $(D^k[0,\infty), u^k)$ are defined by $(D[0,\infty) \times \cdots \times D[0,\infty), d_{J_1} \times \cdots \times d_{J_1})$ and $(D[0,\infty) \times \cdots \times D[0,\infty), u \times \cdots \times u)$, respectively; $d_{J_1} \times \cdots \times d_{J_1}$ and $u \times \cdots \times u$ refer to the corresponding maximum metrics. Let $\Rightarrow$ denote convergence in distribution—for stochastic processes in $D[0,\infty)$, as well as for random variables in $\mathbb{R}$. Let $1_{\{\cdot\}}$ be the usual indicator function and $e = \{e(t) = t, t \geq 0\}$ be the identity map. The composition map is denoted by $\circ$; i.e., for $(x, y) \in D[0,\infty) \times D[0,\infty)$, $x \circ y$ is defined by $(x \circ y)(t) = x(y(t))$, $t \geq 0$. For $x, y \in \mathbb{R}$, x^+ denotes the positive part of x, and $x \wedge y = \min\{x, y\}$.

2. Assumptions

2.1. The model. We consider a sequence of first-come, first-served (FCFS) $G/GI/N+GI$ queues indexed by the number of servers N. Customers arriving after $t = 0$ are indexed by natural numbers in an increasing order of their arrival times. Customer i arrives to the system at time $t_i > 0$ and two quantities are associated with it: the service requirement s_i and patience p_i. The service requirements of customers, $\{s_i, i \geq 1\}$, are independent and identically distributed (i.i.d.), characterized by a distribution function F, which does not vary with N (set $\bar{F} = 1 - F$). The sequence $\{p_i, i \geq 1\}$ is i.i.d. with a distribution G^N for the Nth system. For simplicity of notation, we shall not index arrival times, service requirements, and customers' patience by N—this dependency will be implicit.

Define $A^N(t)$, $t \geq 0$, to be the number of arrivals in the Nth system over the time interval $[0, t]$. The process $A^N = \{A^N(t), t \geq 0\}$ is r.c.l.l., nondecreasing, nonnegative, integer valued, with jumps of size 1 such that $A^N(0) = 0$ and $A^N(t) < \infty$ for all $t \geq 0$, almost surely (a.s.). The arrival process is related to the customer arrival times $\{t_i, i \geq 1\}$ by $t_i = \inf\{t \geq 0\colon A^N(t) \geq i\}$, $i \geq 1$. Define $\tau^N = \{\tau^N(t), t \geq 0\}$ by $\tau^N(t) = t_{A^N(t)}$ for $t \geq t_1$ and $\tau^N(t) = 0$ for $t \leq t_1$; $\tau^N(t)$ is the time of the last arrival prior to t.

At time $t = 0$, there are q_0^N initial customers in the system, labeled by $-q_0^N, -q_0^N + 1, \ldots, -1$. Those with indices $-q_0^N, -q_0^N + 1, \ldots, -(q_0^N - N)^+ - 1$ are in service with i.i.d. service requirements drawn from the distribution F_*, the residual distribution associated with F:

$$F_*(x) = \mu \int_0^x \bar{F}(u)\, du, \tag{2}$$

where $\mu^{-1} = \mathbb{E}\, s_1$ is the mean service, which we assume exists (also set $\bar{F}_* = 1 - F_*$). The remaining $(q_0^N - N)^+$ initial customers (indexed by $-(q_0^N - N)^+, -(q_0^N - N)^+ + 1, \ldots, -1$, if exist) have independent service requirements distributed according to F. However, their patience is infinite, i.e., $p_{-i} \equiv \infty$ for $i = 1, 2, \ldots, (q_0^N - N)^+$. This assumption is convenient for the analysis while being nonrestrictive, as argued at the end of this section.

Let v_i denote the *offered* waiting time of the ith customer—the amount of time the customer awaits service if the customer would have been infinitely patient ($p_i = \infty$). The *virtual* waiting time $V^N(t)$ at time $t \geq 0$ is the amount of time (measured beyond t) until one of the servers becomes idle, provided no new arrivals would have occurred after time t; by definition, $V^N(t) = 0$ if there exists an idle server at time t. The random variable $V^N(t)$ captures the amount of work in the queue at time t. (Note that a service completion that is immediately followed by a new service initiation does not render a server idle.) We set $V^N = \{V^N(t), t \geq 0\}$. The *actual* waiting time of the ith customer is then given by $v_i \wedge p_i$. That is, if customer i eventually enters service then v_i is equal to its actual waiting time and $p_i > v_i$; on the other hand, if customer i abandons the system then $v_i = V^N(t_i-)$ (note that only customers with positive indices can abandon) and $v_i \geq p_i$. We use $V_{\leftarrow}^N = \{V_{\leftarrow}^N(t), t \geq 0\}$ to denote the offered-waiting-time process, with $V_{\leftarrow}^N(t) = v_{A^N(t)}$, for $t \geq t_1$, and $V_{\leftarrow}^N(t) = v_{-q_0^N}$ for $0 \leq t < t_1$. The offered-waiting-time process is defined in such a way that if customer i arrives at time t_i then $v_i = V_{\leftarrow}^N(t_i)$ rather than $v_i = V^N(t_i-)$. Both V^N and $V_{\leftarrow}^N$ are r.c.l.l. processes.

Define $Q^N = \{Q^N(t), t \geq 0\}$, where $Q^N(t)$ is the total number of customers in the system at time $t \geq 0$; this number includes customers receiving service, customers awaiting service that eventually receive service, and customers awaiting service that eventually abandon. For the purpose of analysis, it is convenient to consider an alternative model in which customers who abandon the system, do so upon arrival (based on p_i's). Namely, customers, upon arrival, "compare" their p_i with v_i and immediately abandon the system if $p_i \leq v_i$; in this model, all customers awaiting service receive service eventually. Such dynamics are easier to analyze, and it turns out asymptotically equivalent to the original system. To distinguish between the two models, we introduce $H^N = \{H^N, t \geq 0\}$, where $H^N(t)$ is the number of customers at time $t \geq 0$ in the system with abandonment upon arrival.

2.2. The QED regime. We assume that the sequence of processes $\{A^N\}$ satisfies (i) a functional strong law of large numbers (FSLLN):

$$A^N/\lambda^N \to e \tag{3}$$

u.o.c. a.s., as $N \to \infty$, where λ^N is the arrival rate in the Nth system, and (ii) a functional central limit theorem (FCLT):

$$\hat{A}^N := \frac{1}{\sqrt{N}}(A^N - \lambda^N e) \;\Rightarrow\; \hat{A}, \tag{4}$$

as $N \to \infty$, where $\hat{A}$ is a stochastic process with a.s. continuous sample paths.

The offered load to the Nth system is λ^N/μ and the traffic intensity is $\rho^N = \lambda^N/(\mu N)$. In the QED regime, the number of servers N and traffic intensity ρ^N are related, in the limit as $N \to \infty$, via

$$\sqrt{N}(1-\rho^N) \to \beta, \tag{5}$$

for some $-\infty < \beta < \infty$. In this regime, it is expected that the (virtual) waiting time vanishes as $N \to \infty$, hence only the behavior of G^N around the origin is relevant in the limit. To this end, we assume $G^N(0) = 0$ and

$$\hat{G}^N \to \theta e, \tag{6}$$

u.o.c., as $N \to \infty$, for some $0 \le \theta < \infty$, where $\hat{G}(t) := \sqrt{N}G^N(t/\sqrt{N})$. The condition (6) is satisfied, for example, when $G^N = G$ for all N, and $G(t)/t \to \theta$ as $t \downarrow 0$ (or, equivalently, θ is the right-hand derivative of G at the origin).

The scaled and centered versions of Q^N and H^N are defined by

$$\hat{Q}^N = \{\hat{Q}^N(t), t \ge 0\} = \frac{1}{\sqrt{N}}(Q^N - N)$$

and

$$\hat{H}^N = \{\hat{H}^N(t), t \ge 0\} = \frac{1}{\sqrt{N}}(H^N - N),$$

respectively. As will be shown (see Theorem 4.1 and Corollary 4.1 in §4), the difference between $\hat{Q}^N$ and $\hat{H}^N$ vanishes in the limit, as $N \to \infty$. The scaled versions of the waiting time processes are given by

$$\hat{V}^N = \{\hat{V}^N(t), t \ge 0\} = \mu\sqrt{N}V^N$$

and

$$\hat{V}^N_{\leftarrow} = \{\hat{V}^N_{\leftarrow}(t), t \ge 0\} = \mu\sqrt{N}V^N_{\leftarrow};$$

note that we use μ in the scaling for waiting time processes, which amounts to measuring wait in units of average service time.

2.3. Initial conditions. The number of customers in the system, at time $t = 0$, is given by $Q^N(0) = H^N(0) = q_0^N$. It is assumed that a scaled and centered version of q_0^N converges in distribution:

$$\hat{q}_0^N = \frac{1}{\sqrt{N}}(q_0^N - N) \;\Rightarrow\; \hat{q}_0, \tag{7}$$

as $N \to \infty$. This condition (together with the assumption that the residual service times of customers in service at $t = 0$ are i.i.d. with distribution F_*) is identical to the assumptions made in Reed [27]. Although our initial condition is appealing in its simplicity, it is not the unique initial condition that induces the QED regime; e.g., see Mandelbaum and Momčilović [21].

Next, we discuss an alternative model for patience of the initial customers. Namely, suppose that initial customers (at $t = 0$) do not have infinite patience but rather the sequence $\{p_{-i}, 1 \le i \le (q_0^N - N)^+\}$ is i.i.d., drawn from G^N. We argue that this variation does not impact our asymptotic results. To this end, let r_0^N be the number of initial customers that abandon the system:

$$r_0^N = \sum_{i=1}^{(q_0^N - N)^+} 1_{\{p_{-i} \le v_{-i}\}};$$

i.e., $(q_0^N - r_0^N)$ initial customers awaiting service end up receiving service. Then, the following lemma holds:

LEMMA 2.1. *$r_0^N/\sqrt{N} \Rightarrow 0$, as $N \to \infty$.*

PROOF. See §7.1. □

As a consequence, we have

$$(q_0^N - N - r_0^N)/\sqrt{N} \;\Rightarrow\; \hat{q}_0,$$

as $N \to \infty$. Thus, the two models are asymptotically equivalent since (7) is the only assumption on the initial number of customers in the system and our results depend on the limit $\hat{q}_0$ only (see Theorem 4.1 and Corollaries 4.1 and 4.2 in §4).

3. Preliminaries

3.1. No abandonment. Consider the sequence of queues indexed by N, as introduced in the previous section. We next describe a corresponding sequence of systems without customer abandonment; entities associated with the systems without abandonment are appended by the "dot" symbol. Namely, for the Nth system without abandonment, we set the initial and input parameters equal to those of the Nth system with abandonment, except that all customers have infinite patience in the new system: $\dot{A}^N = A^N$; $\dot{q}_0^N = q_0^N$; $\dot{s}_i = s_i$, $i \geq -\dot{q}_0^N$; and $\dot{p}_i = \infty$, $i \geq -\dot{q}_0^N$. To obtain upper bounds on the offered waiting times $\{v_i, i \geq 1\}$, the following proposition (Bhattacharya and Ephremides [2]) is used in conjunction with the results for the system without abandonment (Reed [27]) (see Proposition 3.2 below). The process $\dot{V}_{\leftarrow}^N = \{\dot{V}_{\leftarrow}^N(t), t \geq 0\}$ is now a waiting-time process (as opposed to $V_{\leftarrow}^N$, which is an offered wait): if $\dot{v}_i$ is the waiting time of customer i, then $\dot{V}_{\leftarrow}^N(t) = \dot{v}_{\dot{A}^N(t)}$, for $t \geq t_1$, and $\dot{V}_{\leftarrow}^N(t) = \dot{v}_{-1}$, for $0 \leq t < t_1$.

PROPOSITION 3.1 (BHATTACHARYA AND EPHREMIDES [2]). *$V_{\leftarrow}^N(t) \leq \dot{V}_{\leftarrow}^N(t)$ and $H^N(t) \leq \dot{H}^N(t) = \dot{Q}^N(t)$, for $t \geq 0$.*

PROOF. For completeness, we provide a proof in §7.2, which is verified within the setup of the present paper. □

The following result is a consequence of the preceding proposition and Proposition 5.3 in Reed [27].

PROPOSITION 3.2. *$V_{\leftarrow}^N \Rightarrow 0$, as $N \to \infty$.*

3.2. Infinite-server processes. For each N, consider a corresponding infinite-server process $X^N = \{X^N(t), t \geq 0\}$, defined by the original arrival process A^N and the sequence of service times $\{s_i, i \geq 1\}$, as follows:

$$\begin{aligned} X^N(t) &= \sum_{i=1}^{A^N(t)} 1_{\{t_i + s_i > t\}} \\ &= \sum_{i=1}^{A^N(t)} (1_{\{s_i > t - t_i\}} - \bar{F}(t - t_i)) + \int_0^t \bar{F}(t-s)\, dA^N(s). \end{aligned}$$

In addition, introduce $\hat{X}^N = \{\hat{X}^N(t), t \geq 0\}$ to be a scaled and centered version of X^N:

$$\hat{X}^N = \frac{1}{\sqrt{N}}(X^N - N\rho^N F_*), \tag{8}$$

namely, for $t \geq 0$,

$$\hat{X}^N(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N(t)} (1_{\{s_i > t - t_i\}} - \bar{F}(t - t_i)) + \int_0^t \bar{F}(t-s)\, d\hat{A}^N(s). \tag{9}$$

The following lemma, due to Krichagina and Puhalskii (see Theorem 3 in Krichagina and Puhalskii [20]), characterizes the limiting infinite-server process. Earlier results on the infinite-server process were obtained by Borovkov [4] and Iglehart [15]; for a recent measure-valued approach, see Decreusefond and Moyal [8] and Reed and Talreja [28]. Define $U = \{U(t,x), t \geq 0, x \in [0,1]\}$ to be a Kiefer process, that is a two-parameter continuous centered Gaussian process on $\mathbb{R}_+ \times [0,1]$, with covariance function $\mathbb{E}[U(s,x)U(t,y)] = (s \wedge t)(x \wedge y - xy)$.

LEMMA 3.1 (KRICHAGINA AND PUHALSKII [20]). *The sequence of infinite-server processes $\{\hat{X}^N\}$ converges in distribution in $D[0,\infty)$, as $N \to \infty$, to the process $\hat{X} = \{\hat{X}(t), t \geq 0\}$ defined by*

$$\hat{X}(t) = \int_0^t \bar{F}(t-s)\, d\hat{A}(s) + \int_0^t \int_0^t 1_{\{s+x \leq t\}}\, dU(\mu s, F(x)), \quad t \geq 0;$$

here U is a Kiefer process, $\hat{A}$ and U are independent, and the first integral is to be interpreted as the result of integration by parts.

Recall the definition of offered waiting times $\{v_i, i \geq 1\}$ from §2. It will turn out convenient to define a (scaled and centered) process $\hat{X}^N_\Delta = \{\hat{X}^N_\Delta(t), t \geq 0\}$ by

$$\hat{X}^N_\Delta(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N(t)} (1_{\{t-t_i-v_i<s_i\leq t-t_i\}} - \bar{F}(t-t_i-v_i) + \bar{F}(t-t_i)), \quad t \geq 0, \tag{10}$$

because this process relates to the (scaled and centered) number of customers with positive indices (those with arrival times $t_i > 0$) in the system at time $t \geq 0$, via the following equality:

$$\hat{X}^N(t) + \hat{X}^N_\Delta(t) - \int_0^t \bar{F}(t-s)\, d\hat{A}^N(s) = \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N(t)} (1_{\{s_i>t-t_i-v_i\}} - \bar{F}(t-t_i-v_i)).$$

3.3. Initial-customers processes. In this subsection, we consider the infinite-server processes associated with the customers initially in the system (at time $t = 0$). The process $I^N = \{I^N(t), t \geq 0\}$ is defined by

$$I^N(t) = \sum_{i=(q_0^N-N)^+ +1}^{q_0^N} 1_{\{s_{-i}>t\}} + \sum_{i=1}^{(q_0^N-N)^+} 1_{\{s_{-i}>t\}},$$

for $t \geq 0$; recall that the random variables s_{-i} in the two sums are distributed according to F_* and F, respectively. Hence, the scaled and centered version $\hat{I}^N = \{\hat{I}(t), t \geq 0\}$ is defined by

$$\hat{I}^N = \frac{1}{\sqrt{N}}(I^N - (q_0^N \wedge N)\bar{F}_* - (q_0^N - N)^+\bar{F}),$$

namely, for $t \geq 0$,

$$\hat{I}^N(t) = \frac{1}{\sqrt{N}} \sum_{i=(q_0^N-N)^+ +1}^{q_0^N} (1_{\{s_{-i}>t\}} - \bar{F}_*(t)) + \frac{1}{\sqrt{N}} \sum_{i=1}^{(q_0^N-N)^+} (1_{\{s_{-i}>t\}} - \bar{F}(t)); \tag{11}$$

recall that $\mathbb{E}\, 1_{\{s_{-i}>t\}} = \bar{F}_*(t)$, $(q_0^N - N)^+ < i \leq q_0^N$, and $\mathbb{E}\, 1_{\{s_{-i}>t\}} = \bar{F}(t)$, $1 \leq i \leq (q_0^N - N)^+$. The following lemma characterizes the limiting behavior (as $N \to \infty$) of $\hat{I}^N$.

Lemma 3.2. *$\hat{I}^N \Rightarrow \hat{I} = W \circ F_*$, as $N \to \infty$, where $W = \{W(t), t \in [0,1]\}$ is a (standard) Brownian bridge, that is, a centered Gaussian process with covariance function $\mathbb{E}[W(t)W(s)] = t \wedge s - ts$.*

Proof. Define $\hat{I}^N_1 = \{\hat{I}^N_1(t), t \geq 0\}$ and $\hat{I}^N_2 = \{\hat{I}^N_2(t), t \geq 0\}$ such that $\hat{I}^N = \hat{I}^N_1 + \hat{I}^N_2$; i.e., $\hat{I}^N_1(t)$ and $\hat{I}^N_2(t)$ correspond to the first and second summand in (11), respectively. From Lemma 3.1 in Krichagina and Puhalskii [20], the random time change theorem and (7), it follows that $\hat{I}^N_1 \Rightarrow \hat{I}$, as $N \to \infty$. By the same argument $\hat{I}^N_2 \Rightarrow 0$, as $N \to \infty$, since (7) implies $(q_0^N - N)^+/N \Rightarrow 0$, as $N \to \infty$. □

Next we introduce $\hat{I}^N_\Delta = \{\hat{I}^N_\Delta(t), t \geq 0\}$, where

$$\hat{I}^N_\Delta(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{(q_0^N-N)^+} (1_{\{t-v_{-i}<s_{-i}\leq t\}} - \bar{F}(t-v_{-i}) + \bar{F}(t)), \quad t \geq 0. \tag{12}$$

The relationship between $\hat{I}^N_\Delta$ and $\hat{I}^N$ is similar to the relationship between $\hat{X}^N_\Delta$ and $\hat{X}^N$:

$$\hat{I}^N(t) + \hat{I}^N_\Delta(t) = \frac{1}{\sqrt{N}} \sum_{i=(q_0^N-N)^+ +1}^{q_0^N} (1_{\{s_{-i}>t\}} - \bar{F}_*(t)) + \frac{1}{\sqrt{N}} \sum_{i=1}^{(q_0^N-N)^+} (1_{\{s_{-i}>t-v_i\}} - \bar{F}(t-v_i)), \tag{13}$$

for $t \geq 0$, which is the (scaled and centered) number of customers with negative indices that are in the system at time $t \geq 0$. Note that the sum in (12) consists of elements corresponding to customers awaiting service at time $t = 0$ only; this is due to the fact that $v_{-i} = 0$ for $(q_0^N - N)^+ < i \leq q_0^N$ by definition. The following lemma states that the process $\hat{X}^N_\Delta + \hat{I}^N_\Delta$ vanishes as $N \to \infty$.

Lemma 3.3. *$\hat{X}^N_\Delta + \hat{I}^N_\Delta \Rightarrow 0$, as $N \to \infty$.*

Proof. See §7.3. □

3.4. Abandonment. Throughout the present section, we consider processes that correspond to the system with *abandonment upon arrival* (see the discussion in §2). This system is easier to analyze than the one where customers abandon after waiting. However, as already noted, the two systems are equivalent in the QED regime.

The infinite-server process X^N was constructed from all arriving customers. Now, let $Z^N = \{Z^N(t), t \geq 0\}$ be the infinite-server process induced only by arrivals that do abandon, namely,

$$Z^N(t) = \sum_{i=1}^{A^N(t)} 1_{\{t_i+s_i>t\}} 1_{\{p_i \leq v_i\}}, \quad t \geq 0.$$

Consequently, the scaled and centered version $\hat{Z}^N = \{\hat{Z}^N(t), t \geq 0\}$ is defined by

$$\hat{Z}^N(t) = \frac{1}{\sqrt{N}} \left(Z^N(t) - \int_0^t \bar{F}(t-s) G^N(V^N_{\leftarrow}(s))\, dA^N(s) \right), \quad t \geq 0; \tag{14}$$

the independence of service requirements, customer patience, and the arrival process, together with the independence of (s_i, p_i) and v_i, yield

$$\begin{aligned} \mathbb{E}\,\hat{Z}^N(t) &= \frac{1}{\sqrt{N}} \mathbb{E} \sum_{i=1}^{A^N(t)} (1_{\{t_i+s_i>t\}} 1_{\{p_i \leq v_i\}} - \bar{F}(t-t_i) G^N(v_i)) \\ &= \frac{1}{\sqrt{N}} \mathbb{E} \sum_{i=1}^{A^N(t)} (\mathbb{E}[1_{\{s_i>t-t_i\}} \mid t_i] \mathbb{E}[1_{\{p_i \leq v_i\}} \mid v_i] - \bar{F}(t-t_i) G^N(v_i)) = 0, \end{aligned} \tag{15}$$

where the second equality is due to $\mathbb{E}[1_{\{s_i>t-t_i\}} \mid t_i] = \bar{F}(t-t_i)$ and $\mathbb{E}[1_{\{p_i \leq v_i\}} \mid v_i] = G^N(v_i)$. The next lemma states that the process $\hat{Z}^N$ is negligible in the limit, as $N \to \infty$. The lemma is based on the assumptions $G^N(0) = 0$ and $\theta < \infty$. During time intervals when the offered waiting time is positive, the rate at which customers abandon is proportional to $\sqrt{N}$ (for large N), which is negligible relative to the total arrival rate λ^N, the latter being linear in N.

Lemma 3.4. *$\hat{Z}^N \Rightarrow 0$, as $N \to \infty$.*

Proof. See §7.4. □

Similarly, the infinite-server process due to customers who do not abandon will be denoted by $Y^N = \{Y^N(t), t \geq 0\} = X^N - Z^N$, with

$$Y^N(t) = \sum_{i=1}^{A^N(t)} 1_{\{t_i+s_i>t\}} 1_{\{p_i>v_i\}}.$$

Because customers abandon the system at a rate proportional to $\sqrt{N}$, the scaling and centering for Y^N is the same as for the process X^N in (8). Thus, $\hat{Y}^N = \{\hat{Y}^N(t), t \geq 0\}$ with

$$\hat{Y}^N = \frac{1}{\sqrt{N}} (Y^N - N\rho^N F_*), \tag{16}$$

which yields

$$\hat{Y}^N(t) = \hat{X}^N(t) - \hat{Z}^N(t) - \int_0^t \bar{F}(t-s) \sqrt{N} G^N(V^N_{\leftarrow}(s))\, d\check{A}^N(s), \quad t \geq 0, \tag{17}$$

where $\check{A}^N = \{\check{A}^N(t), t \geq 0\}$ is a linearly-scaled arrival process:

$$\check{A}^N = A^N/N.$$

In parallel with $\hat{X}^N_\Delta$ and $\hat{I}^N_\Delta$, define $\hat{Y}^N_\Delta = \{\hat{Y}^N_\Delta(t), t \geq 0\}$ by $\hat{Y}^N_\Delta = \hat{X}^N_\Delta - \hat{Z}^N_\Delta$, where $\hat{Z}^N_\Delta = \{\hat{Z}^N_\Delta(t), t \geq 0\}$ is given by

$$\hat{Z}^N_\Delta(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N(t)} (1_{\{t-t_i-v_i<s_i \leq t-t_i\}} - \bar{F}(t-t_i-v_i) + \bar{F}(t-t_i)) 1_{\{p_i \leq v_i\}}, \quad t \geq 0. \tag{18}$$

Lemma 3.5. *$\hat{Y}^N_\Delta + \hat{I}^N_\Delta \Rightarrow 0$, as $N \to \infty$.*

Proof. See §7.5. □

Finally, we introduce $A^N_{\triangleright} = \{A^N_{\triangle}(t), t \geq 0\}$—the arrival process of customers that do not abandon the system, i.e., the customers that are eventually served; this process, at time $t \geq 0$, is given by

$$A^N_{\triangleright}(t) = \sum_{i=1}^{A^N(t)} 1_{\{p_i > v_i\}}.$$

A corresponding scaled and centered version $\hat{A}^N_{\triangleright} = \{\hat{A}^N_{\triangleright}(t), t \geq 0\}$ is defined by

$$\hat{A}^N_{\triangleright} = \frac{1}{\sqrt{N}}(A^N_{\triangleright} - \lambda^N e);$$

the latter process is also used in the proof of Lemma 3.3 (see §7.3). The last lemma in this subsection stems from the fact that $\hat{A}$ has a.s. continuous sample paths and V^N vanishes in the limit, as $N \to \infty$. The process $(\tau^N + V^N_{\leftarrow})$ arises when the relation between H^N and $V^N_{\leftarrow}$ is considered. In particular, $V^N_{\leftarrow}(t) = V^N_{\leftarrow}(\tau^N(t)) = V^N_{\leftarrow} \circ \tau^N(t)$, for $t \geq 0$, is defined not by $H^N \circ \tau^N(t)$ only but rather by $H^N \circ \tau^N(t)$ and $H^N \circ (\tau^N + V^N_{\leftarrow})(t)$ jointly (see the proof of Lemma 3.8).

Lemma 3.6. *$\hat{A}^N_{\triangleright} \circ (\tau^N + V^N_{\leftarrow}) - \hat{A}^N_{\triangleright} \circ \tau^N \Rightarrow 0$, as $N \to \infty$.*

Proof. The value of the process $\hat{A}^N_{\triangleright}$, at time $t \geq 0$, is given by

$$\hat{A}^N_{\triangleright}(t) = \hat{A}^N(t) - \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N(t)} 1_{\{p_i \leq v_i\}},$$

and, thus, $\hat{A}^N_{\triangleright} \circ (\tau^N + V^N_{\leftarrow}) - \hat{A}^N_{\triangleright} \circ \tau^N = \hat{A}^N \circ (\tau^N + V^N_{\leftarrow}) - \hat{A}^N \circ \tau^N - \hat{A}^N_{\Delta}$, where $\hat{A}^N_{\Delta} = \{\hat{A}^N_{\Delta}(t), t \geq 0\}$ which, for $t \geq 0$, satisfies

$$\begin{aligned}\hat{A}^N_{\Delta}(t) &= \frac{1}{\sqrt{N}} \sum_{i=A^N(\tau^N(t))+1}^{A^N(\tau^N(t)+V^N_{\leftarrow}(t))} 1_{\{p_i \leq v_i\}} \\ &\leq \frac{1}{\sqrt{N}} \sum_{i=A^N(t)+1}^{A^N(t+V^N_{\leftarrow}(t))} 1_{\{p_i \leq v_i\}};\end{aligned}$$

the inequality follows from the monotonicity of $A^N(\cdot)$, $\tau^N(t) \leq t$, and $A^N(\tau^N(t)) = A^N(t)$. Assumption (4) and Proposition 3.2 imply $\hat{A}^N \circ (e + V^N_{\leftarrow}) - \hat{A}^N \Rightarrow 0$ and $\hat{A}^N_{\Delta} \Rightarrow 0$, as $N \to \infty$. The statement of the lemma follows. □

3.5. Queue length. The number of customers in the system at time $t \geq 0$ can be expressed as the sum of indicator functions (Borovkov [4], Krichagina and Puhalskii [20], Reed [27]):

$$H^N(t) = \sum_{i=1}^{A^N(t)} 1_{\{t_i + s_i + v_i > t\}} 1_{\{p_i > v_i\}} + \sum_{i=(q_0^N - N)^+ + 1}^{q_0^N} 1_{\{s_{-i} > t\}} + \sum_{i=1}^{(q_0^N - N)^+} 1_{\{s_{-i} + v_{-i} > t\}}. \tag{19}$$

On the other hand, Proposition 2.1 in Reed [27] renders

$$\int_0^t (H^N(t-s) - N)^+ \, dF(s) = \sum_{i=1}^{A^N(t)} (\bar{F}(t - t_i - v_i) - \bar{F}(t - t_i)) 1_{\{p_i > v_i\}} + \sum_{i=1}^{(q_0^N - N)^+} (\bar{F}(t - v_{-i}) - \bar{F}(t)).$$

Then, combining the preceding equality and (19) yields, for $t \geq 0$,

$$\begin{aligned}H^N(t) = &\sum_{i=1}^{A^N(t)} (1_{\{t_i + s_i + v_i > t\}} - \bar{F}(t - t_i - v_i) + \bar{F}(t - t_i)) 1_{\{p_i > v_i\}} + \sum_{i=(q_0^N - N)^+ + 1}^{q_0^N} (1_{\{s_{-i} > t\}} - \bar{F}_*(t)) \\ &+ \sum_{i=1}^{(q_0^N - N)^+} (1_{\{s_{-i} + v_{-i} > t\}} - \bar{F}(t - v_{-i})) + (q_0^N - N)^+ \bar{F}(t) + (q_0^N \wedge N) \bar{F}_*(t) + \int_0^t (H^N(t-s) - N)^+ \, dF(s),\end{aligned}$$

or, equivalently, in terms of scaled processes (see (7), (10), (13), (16), and (18)), for $t \geq 0$:

$$\hat{H}^N(t) = (\hat{q}_0^N)^+(\bar{F}(t) - \bar{F}_*(t)) + \hat{q}_0^N \bar{F}_*(t) + \hat{I}^N(t) + \hat{I}_\Delta^N(t) + \hat{Y}^N(t) + \hat{Y}_\Delta^N(t) + \int_0^t (\hat{H}^N(t-s))^+ \, dF(s) - \sqrt{N}(1-\rho^N)F_*(t). \tag{20}$$

We now recall an operator, φ: $D[0,\infty) \to D[0,\infty)$, which was introduced in Reed [27]; it plays a fundamental role in the analysis of QED queues without abandonment.

Definition 3.1 (Reed [27]). For each $x \in D[0,\infty)$, let $\varphi(x)$ be the unique solution y to

$$y(t) = x(t) + \int_0^t y^+(t-s)\, dF(s), \quad t \geq 0.$$

Then, (20) can be rewritten in terms of the operator φ:

$$\hat{H}^N = \varphi\big((\hat{q}_0^N)^+(\bar{F} - \bar{F}_*) + \hat{q}_0^N \bar{F}_* + \hat{I}^N + \hat{I}_\Delta^N + \hat{Y}^N + \hat{Y}_\Delta^N - \sqrt{N}(1-\rho^N)F_*\big). \tag{21}$$

The next proposition establishes L^1-continuity of φ. In Reed [27], only continuity of φ in the topology of uniform convergence was considered. The additional mode of L^1-continuity is needed in order to relate $\hat{H}^N$ and $\hat{V}_\leftarrow^N$ in Lemma 3.8 (via Lemma 3.7). In particular, due to (14) (see also (17)), rather than approximating $\hat{V}_\leftarrow^N$ by $\hat{H}^N$ directly, it suffices to only relate integrals of these processes over finite time intervals.

Proposition 3.3. *The function* φ: $D[0,\infty) \to D[0,\infty)$ *is Lipschitz continuous in the* L^1 *topology over bounded intervals.*

Proof. See §7.6. □

We now proceed to show that the scaled number-in-system process $\hat{H}^N$ does not change significantly (in the L^1 sense, as $N \to \infty$) over time intervals during which individual customers await service. Note that $t = \tau^N(s) + V_\leftarrow^N(s)$ is the time when the last arriving customer before $t = s$ were to enter service if it had infinite patience (recall that $V_\leftarrow^N$ is the offered-waiting-time process).

Lemma 3.7. *We have, as* $N \to \infty$,

$$\left\{ \int_0^t \left| \hat{H}^N \circ (\tau^N + V_\leftarrow^N)(s) - \hat{H}^N(s) \right| ds,\ t \geq 0 \right\} \ \Rightarrow\ 0.$$

Proof. See §7.7. □

3.6. Offered waiting time. The following lemma relates the (limiting and scaled) queue-length and offered-waiting-time processes in the QED regime. Recall that waiting is measured in units of average service time.

Lemma 3.8. *We have, as* $N \to \infty$,

$$\left\{ \int_0^t \left| (\hat{H}^N(s))^+ - \hat{V}_\leftarrow^N(s) \right| ds,\ t \geq 0 \right\} \ \Rightarrow\ 0.$$

Remark 3.1. The lemma relates the queue-length and offered-waiting-time processes without a priori requiring that either of the processes converges weakly.

Proof. For $t \geq 0$, let $D^N(t)$ be the number of service completions during the time interval $[0,t]$. First, by definition, $V_\leftarrow^N(t)$ satisfies, for $t \geq 0$,

$$(H^N(\tau) + 1_{\{p \leq V_\leftarrow^N(t)\}} - N)^+ = D^N(\tau + V_\leftarrow^N(t)) - D^N(\tau), \tag{22}$$

where $\tau \equiv \tau^N(t)$ is the time of the last arrival prior to time t and $p \equiv p^N(t) = p_{A^N(t)}$ is the patience of the corresponding customer (set $p_0 = \infty$). The presence of the indicator function in (22) is due to the fact that the customer arriving at time τ might abandon the system on arrival (if $p \leq V_\leftarrow^N(\tau)$). Recall that, by definition, $V_\leftarrow^N(t) = V_\leftarrow^N(\tau)$ is the offered waiting time of the customer with index $A^N(t)$, i.e., the waiting time this customer would experience if it were not to abandon. The sum $H^N(\tau) + 1_{\{p \leq V_\leftarrow^N(t)\}}$ represents the number of customers in the system at time τ if the patience of the arriving customer is infinite. Second, the number of the customers in the system at time $\tau + V_\leftarrow^N(t) = \tau + V_\leftarrow^N(\tau)$ can be expressed as a linear combination of arrivals and departures:

$$\begin{aligned} H^N(\tau + V_\leftarrow^N(t)) &= H^N(\tau) + A_\triangleright^N(\tau + V_\leftarrow^N(t)) - A_\triangleright^N(\tau) - D^N(\tau + V_\leftarrow^N(t)) + D^N(\tau) \\ &= H^N(\tau) + A_\triangleright^N(\tau + V_\leftarrow^N(t)) - A_\triangleright^N(\tau) - (H^N(\tau) - N + 1_{\{p \leq V_\leftarrow^N(t)\}})^+, \end{aligned}$$

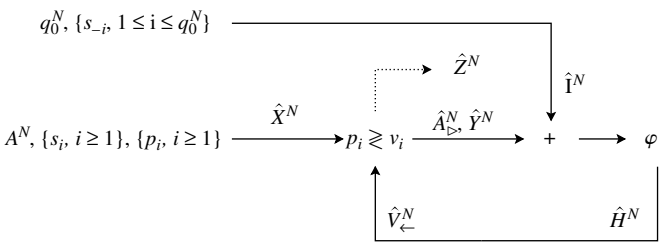

FIGURE 1. Relations between various processes/variables.

where the second equality is due to (22). Considering whether $V^N(t) > 0$ or $V^N(t) = 0$ in the preceding equation results in

$$(H^N(\tau + V^N_{\leftarrow}(t)) - N)^+ = A^N_{\triangleright}(\tau + V^N_{\leftarrow}(t)) - A^N_{\triangleright}(\tau) - 1_{\{p \le V^N_{\leftarrow}(t)\}} 1_{\{H^N(\tau + V^N_{\leftarrow}(t)) = N\}}. \tag{23}$$

Third, centering and rescaling the quantities in (23) gives rise to

$$\begin{aligned}(\hat{H}^N(t))^+ - \hat{V}^N_{\leftarrow}(t) = (\hat{H}^N(t))^+ - (\hat{H}^N(\tau + V^N_{\leftarrow}(t)))^+ + \hat{A}^N_{\triangleright}(\tau + V^N_{\leftarrow}(t)) - \hat{A}^N_{\triangleright}(\tau)\\ - (1 - \rho^N)\hat{V}^N_{\leftarrow}(t) - 1_{\{p \le V^N_{\leftarrow}(t)\}} 1_{\{H^N(\tau + V^N_{\leftarrow}(t)) = N\}} / \sqrt{N}.\end{aligned} \tag{24}$$

Next, note that (5) and Proposition 3.2 yield, as $N \to \infty$,

$$(1 - \rho^N)\hat{V}^N_{\leftarrow} \;\Rightarrow\; 0.$$

Finally, the statement follows from (24), the preceding limit and Lemmas 3.6 and 3.7. □

3.7. Summary of notation. We find it helpful to summarize, in Figure 1, various relations among the processes that have been introduced in this section. Process $\hat{I}^N$ corresponds to customers that are initially in the system at time $t = 0$, and $\hat{X}^N$ is the infinite-server process that corresponds to the customers that arrive after $t = 0$. Based on a comparison of customer patience and offered waiting times, $\hat{X}^N$ splits into the abandonment process $\hat{I}^N$ and the infinite-server process $\hat{Y}^N$ due to customers that receive service (do not abandon). Reed's operator φ provides a description of the queue-length process $\hat{H}^N$ in terms of $\hat{Y}^N$ and $\hat{I}^N$. Finally, the queue-length process is closely related to the (offered) waiting-time process $\hat{V}^N_{\leftarrow}$.

4. Results. This section contains the main results of the paper. A central role is played by a mapping ϕ, applicable to the model with abandonment, which is a generalized version of the mapping φ in Reed [27]. The two mappings coincide for $\theta = 0$ (no abandonment in the limit). Recall that the waiting time vanishes in the limit (Proposition 3.2), and hence, the sequence of patience distributions $\{G^N\}$ manifests itself only through the parameter θ (cf. (6)).

DEFINITION 4.1. The mapping ϕ: $D[0,\infty) \to D[0,\infty)$ is such that $\phi(x)$, for each $x \in D[0,\infty)$, is the unique solution y to

$$y(t) = x(t) + \int_0^t y^+(t-s)\,dF(s) - \frac{\theta}{\mu}\int_0^t y^+(t-s)\,dF_*(s), \quad t \ge 0. \tag{25}$$

The next proposition guarantees that ϕ is well defined and summarizes some of its properties.

PROPOSITION 4.1. *For each $x \in D[0,\infty)$ there exists a unique solution $\phi(x)$ to (25). The function ϕ: $D[0,\infty) \to D[0,\infty)$ is Lipschitz continuous in the topology of uniform convergence over bounded intervals and it is measurable with respect to the Borel σ-field generated by the Skorohod J_1 topology.*

PROOF. See §7.8. □

4.1. Queue length. The following is the main result of our paper.

THEOREM 4.1. *For the QED $G/GI/N + GI$ queue, with abandonments upon arrivals, as $N \to \infty$ we have*

$$\hat{H}^N \;\Rightarrow\; \phi(\hat{q}_0^+(\bar{F} - \bar{F}_*) + \hat{q}_0 \bar{F}_* + \hat{I} + \hat{X} - \beta F_*).$$

Remark 4.1. In the context of Theorem 4.1, the last term in (25) captures the effect of customers abandonment in the QED regime; note that the integration is with respect to the residual distribution F_* rather than the service distribution F. Namely, ϕ quantifies the negative feedback due to abandonment ($\theta/\mu > 0$): the higher the number in the system, the higher the offered waiting time, the higher the abandonment rate, the lower the effective arrival rate of customers that eventually receive service, and the lower the number in the system; on the other hand, the lower the number in the system, the lower the offered waiting time, the lower the abandonment rate, the higher the arrival rate of customers that eventually receive service, and the higher the number in the system.

Proof. Using (17) and Definition 4.1, equality (21) can be rewritten as

$$\hat{H}^N = \phi(\hat{M}^N + \hat{I}^N + \hat{X}^N + \hat{\Delta}^N),$$

where

$$\hat{M}^N = (\hat{q}_0^N)^+(\bar{F} - \bar{F}_*) + \hat{q}_0^N \bar{F}_* - \sqrt{N}(1-\rho^N)F_*, \tag{26}$$

and $\hat{\Delta}^N = \{\hat{\Delta}^N(t), t \geq 0\}$ is given by

$$\hat{\Delta}^N(t) = \hat{I}_\Delta^N(t) + \hat{Y}_\Delta^N(t) - \hat{Z}^N(t) - \int_0^t \bar{F}(t-s)\sqrt{N}G^N(V_{\leftarrow}^N(s))\, d\check{A}^N(s) + \frac{\theta}{\mu}\int_0^t (\hat{H}^N(t-s))^+\, dF_*(s).$$

Combining Lemmas 3.4 and 3.5 together with (2) and Lemma 3.8 yields, as $N \to \infty$,

$$\hat{\Delta}^N \;\Rightarrow\; 0. \tag{27}$$

From (5) and (7) it follows that, as $N \to \infty$,

$$\hat{M}^N \;\Rightarrow\; \hat{M} = \hat{q}_0^+(\bar{F} - \bar{F}_*) + \hat{q}_0\bar{F}_* - \beta F_*. \tag{28}$$

Now, we argue that, as $N \to \infty$, jointly

$$(\hat{M}^N, \hat{I}^N, \hat{X}^N, \hat{\Delta}^N) \;\Rightarrow\; (\hat{M}, \hat{I}, \hat{X}, 0); \tag{29}$$

note that the convergence of marginals is due to (28), Lemmas 3.1 and 3.2, and (27). To this end, introduce $\check{I}^N = \{\check{I}^N(t), t \geq 0\}$ with

$$\check{I}^N(t) = \frac{1}{\sqrt{N}} \sum_{i=q_0^N-N+1}^{q_0^N} (1_{\{\check{s}_{-i} > t\}} - \bar{F}_*(t)),$$

where $\check{s}_{-i} = s_{-i}$ for $(q_0^N - N)^+ < i \leq q_0^N$, and $\{\check{s}_{-i}, q_0^N - N < i \leq (q_0^N - N)^+\}$ is an i.i.d. sequence drawn from F_* and independent of all service requirements, arrival processes, and q_0^N. Observe that the preceding sum contains exactly N elements (rather than a random number that depends on q_0^N), and the N-element sequence $\{\check{s}_{-i}, q_0^N - N < i \leq q_0^N\}$ is independent of q_0^N by construction (q_0^N is just an index in this case, and the elements of the sequence are independent of q_0^N); as a consequence, $\check{I}^N$ and q_0^N are independent. Then the definitions of $\check{I}^N$ and $\hat{I}^N$ imply, for $t \geq 0$,

$$\check{I}^N(t) - \hat{I}^N(t) = \frac{1}{\sqrt{N}} \sum_{i=q_0^N-N+1}^{(q_0^N-N)^+} (1_{\{\check{s}_{-i} > t\}} - \bar{F}_*(t)) - \frac{1}{\sqrt{N}} \sum_{i=1}^{(q_0^N-N)^+} (1_{\{s_{-i} > t\}} - \bar{F}(t)),$$

that, in turn, leads to (see the proof of Lemma 3.2)

$$(\check{I}^N, \hat{I}^N) \;\Rightarrow\; (\hat{I}, \hat{I}), \tag{30}$$

as $N \to \infty$. The limit $(\hat{M}^N, \check{I}^N, \hat{X}^N, 0) \Rightarrow (\hat{M}, \hat{I}, \hat{X}, 0)$, as $N \to \infty$, is due to the convergence of marginals and the independence of the prelimit processes $\hat{M}^N$, $\check{I}^N$, and $\hat{X}^N$ (Whitt [31, Theorem 11.4.4]); the independence is due to the fact that $\hat{M}^N$ depends on q_0^N only (see (26)), $\hat{X}^N$ depends only on the quantities associated with customers that are not initially in the system (see (9)), and $\check{I}$ is independent of both q_0^N and A^N, $\{s_i, i \geq 1\}$. Furthermore, the following holds:

$$d_{J_1}((\hat{M}^N, \check{I}^N, \hat{X}^N, 0), (\hat{M}^N, \hat{I}^N, \hat{X}^N, \hat{\Delta}^N)) \;\leq\; d_{J_1}(\check{I}^N, \hat{I}^N) + d_{J_1}(0, \hat{\Delta}^N)$$
$$\Rightarrow 0,$$

as $N \to \infty$, where the limit is due to (30), (27), and Theorem 11.4.8 in Whitt [31]. Finally, (29) follows from the preceding limit and Theorem 11.4.7 in Whitt [31].

The rest of the proof is almost identical to the corresponding part of the proof of Theorem 5.1 in Reed [27]. Specifically, the space $D^4[0,\infty)$ is separable under the product topology (e.g., see Theorem 11.4.1 in Whitt [31]); therefore, due to (29) and the Skorohod representation theorem (e.g., see Teorem 3.2.2 in Whitt [31]), there exists an alternative probability space with $\{(\tilde{M}^N, \tilde{I}^N, \tilde{X}^N, \tilde{\Delta}^N)\}_N$ and $(\tilde{M}, \tilde{I}, \tilde{X}, 0)$ defined on it with the following properties:

$$\begin{aligned}(\tilde{M}^N, \tilde{I}^N, \tilde{X}^N, \tilde{\Delta}^N) &\stackrel{d}{=} (\hat{M}^N, \hat{I}^N, \hat{X}^N, \hat{\Delta}^N),\\ (\tilde{M}, \tilde{I}, \tilde{X}, 0) &\stackrel{d}{=} (\hat{M}, \hat{I}, \hat{X}, 0),\\ \tilde{M}^N, \tilde{I}^N, \tilde{X}^N, \tilde{\Delta}^N) &\to (\tilde{M}, \tilde{I}, \tilde{X}, 0) \quad \text{a.s.},\end{aligned} \tag{31}$$

as $N \to \infty$. It should be noted that the last limit also holds under the uniform metric (not just J_1 metric) since both $\hat{I}$ and $\hat{X}$ have continuous sample paths and the set of discontinuity points of M^N is a subset of discontinuity points of F, for all N. Hence, we have, as $N \to \infty$,

$$\tilde{M}^N + \tilde{I}^N + \tilde{X}^N + \tilde{\Delta}^N \to \tilde{M} + \tilde{I} + \tilde{X} \quad \text{a.s.} \tag{32}$$

under the uniform metric.

Define $\tilde{H}^N = \phi(\tilde{M}^N + \tilde{I}^N + \tilde{X}^N + \tilde{\Delta}^N)$ and note that, because of the measurability property of ϕ (Proposition 4.1) and (31), we have

$$\tilde{H}^N \stackrel{d}{=} \hat{H}^N. \tag{33}$$

Moreover, (32) and Proposition 4.1 (continuity part) yield, as $N \to \infty$,

$$\tilde{H}^N = \phi(\tilde{M}^N + \tilde{I}^N + \tilde{X}^N + \tilde{\Delta}^N) \to \phi(\tilde{M} + \tilde{I} + \tilde{X}) \quad \text{a.s.}$$

The fact that almost sure convergence implies convergence in distribution and convergence in the uniform metric implies convergence in the J_1 metric, together with (33), Proposition 4.1 (the measurability part) and the preceding limit yield

$$\hat{H}^N \;\Rightarrow\; \phi(\hat{M} + \hat{I} + \hat{X}),$$

as $N \to \infty$. The statement of the theorem now follows. □

Recall that Q^N is the process of the total number of customers in the system when abandonments occur after waiting (as opposed to upon arrival). In view of Theorem 4.1, the following result indicates that, in the QED regime, the scaled number of customers awaiting service that eventually abandon becomes negligible (relative to the scaled total number of customers awaiting service) as the number of servers increases.

COROLLARY 4.1. *For the QED $G/GI/N+GI$ queue, with abandonments after waiting, we have, as $N \to \infty$,*

$$\hat{Q}^N \;\Rightarrow\; \phi(\hat{q}_0^+(\bar{F} - \bar{F}_*) + \hat{q}_0\bar{F}_* + \hat{I} + \hat{X} - \beta F_*),$$

where the limit coincides with that in Theorem 4.1.

PROOF. The processes $\hat{Q}^N$ and $\hat{H}^N$ are related via $\hat{Q}^N = \hat{H}^N + \hat{R}^N$, where $\hat{R}^N = \{\hat{R}^N(t), t \ge 0\}$ is given by

$$\hat{R}^N(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N(t)} 1_{\{t_i + p_i > t\}} 1_{\{p_i \le v_i\}}.$$

Thus, in view of Theorem 4.1, it is sufficient to prove $\hat{R}^N \Rightarrow 0$, as $N \to \infty$. To this end, for any positive c and δ, the following inequality holds for all sufficiently large N:

$$\mathbb{P}[\|\hat{R}^N\|_T > \epsilon] \le \mathbb{P}[\|\hat{R}^N_{(c,\delta)}\|_T > \epsilon] + \mathbb{P}[\|\hat{V}^N_{\leftarrow}\|_T > c],$$

where $\hat{R}^N_{(c,\delta)} = \{\hat{R}^N_{(c,\delta)}(t), t \ge 0\}$ is an infinite-server process with deterministic service times and

$$\hat{R}^N_{(c,\delta)}(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N(t)} 1_{\{t_i + \delta > t\}} 1_{\{p_i \le c/(\mu\sqrt{N})\}}.$$

Now, Theorem 3 in Krichagina and Puhalskii [20] implies $\hat{R}^N_{(c,\delta)} \Rightarrow \{c\theta(t \wedge \delta), t \ge 0\}$, as $N \to \infty$. On the other hand, Proposition 3.2 yields

$$\lim_{c\to\infty} \limsup_{N\to\infty} \mathbb{P}[\|\hat{V}^N_{\leftarrow}\|_T > c] = 0.$$

Therefore, given $\epsilon > 0$, for any $\xi > 0$ it is possible to select c and δ such that $\mathbb{P}[\|\hat{R}^N\|_T > \epsilon] < \xi$, for all N large enough. Consequently, $\hat{R}^N \Rightarrow 0$, as $N \to \infty$, and the corollary follows. □

4.2. Waiting time. Now we introduce a mapping ψ that is the analogue of ϕ for the virtual-waiting-time process.

DEFINITION 4.2. The mapping ψ: $D[0,\infty) \to D[0,\infty)$ is such that $\psi(x)$, for each $x \in D[0,\infty)$, is the unique solution y to

$$y(t) = \left(x(t) + \int_0^t y(t-s)\, dF(s) - \frac{\theta}{\mu} \int_0^t y(t-s)\, dF_*(s) \right)^+, \quad t \geq 0. \tag{34}$$

REMARK 4.2. Note that, for $x \in D[0,\infty)$, if $y = \phi(x)$ then $y^+ = \psi(x)$, i.e., $\psi(x) = (\phi(x))^+$.

The next corollary characterizes the limiting waiting-time processes. Let $L^N = \{L^N(t), t \geq 0\}$ be the abandonment process in the Nth system, that is $L^N(t)$ is the number of customers that abandon by time t.

COROLLARY 4.2. *For the QED $G/GI/N+GI$ queue we have, as $N \to \infty$,*

$$\hat{V}^N \;\Rightarrow\; \hat{V} = \psi(\hat{q}_0^+(\bar{F} - \bar{F}_*) + \hat{q}_0 \bar{F}_* + \hat{I} + \hat{X} - \beta F_*)$$

and

$$\hat{V}^N_{\leftarrow} \;\Rightarrow\; \hat{V}.$$

REMARK 4.3. Note that, in view of Remark 4.2, the virtual waiting time $\hat{V}$ and the queue length $\hat{Q}$ are related via

$$\hat{V} = \hat{Q}^+.$$

(Recall that $\hat{V}^N = \mu\sqrt{N} V^N$ and $\hat{V}^N_{\leftarrow} = \mu\sqrt{N} V^N_{\leftarrow}$. In some of the literature, scaling that does not include the prefactor μ is used, resulting in $\hat{V} = \hat{Q}^+/\mu$ rather than $\hat{V} = \hat{Q}^+$ as in the present paper.)

PROOF. Recall that, from the definition of the process $A_{\triangleright}$, it follows that, for $t \geq 0$,

$$\hat{A}^N_{\triangleright}(t) = \hat{A}^N(t) - \frac{1}{\sqrt{N}} \int_0^t G^N(V^N_{\leftarrow}(s))\, dA^N(t).$$

The preceding limits, (3), (4), (6), Lemma 3.8, and Theorem 4.1 yield, as $N \to \infty$,

$$\hat{A}^N_{\triangleright} \;\Rightarrow\; \left\{ \hat{A}(t) - \theta \int_0^t \hat{H}^+(s)\, ds,\ t \geq 0 \right\}.$$

Now, let $D^N = \{D^N(t), t \geq 0\}$ be the departure process in the Nth system, i.e., $D^N(t)$ is the number of customers that receive service by time t. Then D^N and L^N can be expressed in terms of the arrival and queue-length processes:

$$D^N(t) = A^N_{\triangleright}(t) - H^N(t) + q_0^N$$

and

$$L^N(t) = A^N(t) - A^N_{\triangleright}(t) + H^N(t) - Q^N(t),$$

where $t \geq 0$. These representations, (4.2), (4), Theorem 4.1, and Corollary 4.1, imply, as $N \to \infty$, that $(D^N - \lambda^N e)/\sqrt{N} \Rightarrow \{\hat{D}(t), t \geq 0\}$, $L^N/\sqrt{N} \Rightarrow \{\hat{L}(t), t \geq 0\}$ and $L^N/N \to 0$ u.o.c. a.s., where

$$\hat{D}(t) = \hat{A}(t) - \theta \int_0^t \hat{H}^+(s)\, ds - \hat{H}(t) + \hat{q}_0, \quad t \geq 0,$$

and

$$\hat{L}(t) = \theta \int_0^t \hat{H}^+(s)\, ds, \quad t \geq 0. \tag{35}$$

Given the preceding limits, (3), (4), Corollary 4.1, the continuity of sample paths of $\hat{A}$, and the Lipschitz continuity of ϕ (Proposition 4.1), the virtual-waiting-time process $\hat{V}^N$ converges due to Talreja and Whitt [30]: $\hat{V}^N \Rightarrow \hat{Q}^+ = \{Q^+(t), t \geq 0\}$, as $N \to \infty$, where $\hat{Q}$ is such that $\hat{Q}^N \Rightarrow \hat{Q}$, as $N \to \infty$. However, from Corollary 4.1 it follows that $\hat{Q} = \phi(\hat{q}_0^+(\bar{F} - \bar{F}_*) + \hat{q}_0 \bar{F}_* + \hat{I} + \hat{X} - \beta F_*)$. The convergence of the offered-waiting-time processes $\hat{V}^N_{\leftarrow}$ can be deduced from Puhalskii [25] because, in addition to convergence of the queue-legth process, we have convergence of the arrival processes of customers that eventually receive service. □

5. Examples

Example 5.1 (Estimating Patience). In any application of models with abandonment, there is the need to estimate the patience distribution (Gans et al. [11]). Our results indicate that, in the QED regime, it suffices to merely estimate θ, the density of patience at the origin. The following corollary provides a theoretical justification for our proposed estimator.

Corollary 5.1. *For the $G/GI/N+GI$ queue we have, as $N\to\infty$,*

$$\left\{\sqrt{N}\frac{L^N(t)}{A^N(t)},\ t\ge 0\right\}\ \Rightarrow\ \left\{\frac{\theta}{\mu t}\int_0^t \hat{Q}^+(s)\,ds,\ t\ge 0\right\}.$$

Proof. The statement follows from (4), Corollary 4.1, and (35). □

The corollary suggests that an estimator for θ, $\hat{\theta}$, can be obtained in the following manner:

$$\hat{\theta}=\frac{L^N(t)/A^N(t)}{(1/(\mu Nt))\int_0^t (Q^N(s)-N)^+\,ds}.$$

The numerator is simply the fraction of customers abandoning up to time t; a practical approximation for the denominator can be the average waiting time up to time t. The accuracy of such estimators remains an interesting open problem.

We next consider two specific examples; both correspond to systems that have not yet been analyzed. The first example generalizes Zeltyn [33] and Zeltyn and Mandelbaum [34], and the second example expands on Mandelbaum and Momčilović [21].

Example 5.2 ($G/M/N+GI$). Consider a system with exponential service times, noting that $F_*=F$. In addition, suppose that the sequences of random arrival times $\{t_i^N\}$ satisfy, for some $c>0$,

$$\left\{\frac{t^N_{\lfloor Nt\rfloor}-t/\mu}{\sqrt{N}c/\mu},\ t\ge 0\right\}\ \Rightarrow\ \hat{B}, \tag{36}$$

as $N\to\infty$, where $\hat{B}$ is a standard Brownian motion. (Note that there exists a sequence of arrival times for each N, namely, the jump times of A^N.) Then, $\hat{H}^N\Rightarrow\hat{H}$ and $\hat{Q}^N\Rightarrow\hat{Q}$, as $N\to\infty$, due to Theorem 4.1 and Corollary 4.1, respectively. Here, $\hat{Q}=\hat{H}$ is the unique solution to

$$\hat{Q}(t)=\hat{I}(t)+\hat{X}(t)+(\hat{q}_0+\beta)\exp\{-\mu t\}-\beta+(\mu-\theta)\int_0^t \hat{Q}^+(t-s)\exp\{-\mu s\}\,ds,\quad t\ge 0, \tag{37}$$

in which $\hat{q}_0$ is given in (7), $\hat{I}$ in Lemma 3.2, and $\hat{X}$ in Lemma 3.1. Similarly, due to Corollary 4.2, $\hat{V}^N\Rightarrow\hat{V}$, as $N\to\infty$, where $\hat{V}$ is the unique solution to

$$\hat{V}(t)=(\hat{I}(t)+\hat{X}(t)+(\hat{q}_0+\beta)\exp\{-\mu t\}-\beta+(\mu-\theta)\int_0^t \hat{V}(t-s)\exp\{-\mu s\}\,ds)^+,\quad t\ge 0.$$

The definitions of I^N and X^N give rise to

$$\frac{1}{\sqrt{N}}(I^N+X^N-\rho^N N)=\hat{I}^N+\hat{X}^N+(\hat{q}_0^N+\sqrt{N}(1-\rho^N))\bar{F}, \tag{38}$$

where $\bar{F}(t)=\exp\{-\mu t\}$, $t\ge 0$. Because the service times are exponential, the process on the left-hand side of the preceding equality weakly converges to $\hat{S}=\{\hat{S}(t),t\ge 0\}$, which satisfies $\hat{S}(0)=\hat{q}_0+\beta$ and

$$d\hat{S}(t)=-\mu\hat{S}(t)dt+\sqrt{\mu(1+c^2)}\,dB(t),\quad t\ge 0; \tag{39}$$

here $\{B(t),t\ge 0\}$ is a standard Brownian motion (Krichagina and Puhalskii [20]). Now, (37) and (38) result in

$$d\hat{Q}(t)=d\hat{S}(t)-(\mu(\hat{Q}(t)-\hat{S}(t)+\beta)-(\mu-\theta)\hat{Q}^+(t))\,dt,$$

which, combined with (39), yields

$$\begin{aligned} d\hat{Q}(t)&=\begin{cases}(\mu(\hat{S}(t)-\beta)-\theta\hat{Q}(t))dt+d\hat{X}(t) & \hat{Q}(t)>0,\\ (\mu(\hat{S}(t)-\beta)-\mu\hat{Q}(t))dt+d\hat{X}(t) & \hat{Q}(t)\le 0,\end{cases}\\ &=\begin{cases}-(\mu\beta+\theta\hat{Q}(t))dt+\sqrt{\mu(1+c^2)}\,dB(t) & \hat{Q}(t)>0,\\ -(\mu\beta+\mu\hat{Q}(t))dt+\sqrt{\mu(1+c^2)}\,dB(t) & \hat{Q}(t)\le 0;\end{cases}\end{aligned} \tag{40}$$

the initial condition for $\hat{Q}$ is $\hat{Q}(0)=\hat{q}_0$.

Finally, in the special case $\mu=\theta$, the operator ϕ simplifies to $\phi(x)=x$ (see (25)) and, therefore, $\hat{Q}=\hat{H}=\hat{q}_0+\hat{I}+\hat{X}-(\hat{q}_0+\beta)F$, with $F(t)=1-\exp\{-\mu t\}$, $t\geq 0$. Note that $\hat{q}_0+\hat{I}+\hat{X}-(\hat{q}_0+\beta)F=\hat{S}-\beta$ is the limiting scaled and centered infinite-server process with the initial condition taken to be $\hat{q}_0$; for the QED $M/M/N+M$ system, this relation holds even in the prelimit.

EXAMPLE 5.3 ($G/D/N+GI$). The deterministic service distribution $F(s)=1_{\{s\geq 1/\mu\}}$ implies a uniform residual distribution $F_*(s)=(\mu s)^+\wedge 1$. Theorem 4.1 and Corollary 4.1 guarantee $\hat{H}^N\Rightarrow\hat{H}$ and $\hat{Q}^N\Rightarrow\hat{Q}$, as $N\to\infty$, where $\hat{Q}=\hat{H}$ satisfies, for $0\leq t<1/\mu$,

$$\hat{Q}(t)=\hat{q}_0^+\mu t+\hat{q}_0(1-\mu t)+\hat{I}(t)+\hat{X}(t)-\beta\mu t-\theta\int_0^t\hat{Q}^+(t-s)\,ds,$$

while, for $t\geq 1/\mu$,

$$\hat{Q}(t)=\hat{X}(t)-\beta+\hat{Q}^+(t-1/\mu)-\theta\int_0^{1/\mu}\hat{Q}^+(t-s)\,ds;$$

as in the previous example, $\hat{q}_0$ is given in (7), $\hat{I}$ in Lemma 3.2, and $\hat{X}$ in Lemma 3.1. On the other hand, Corollary 4.2 implies $\hat{V}^N\Rightarrow\hat{V}$, as $N\to\infty$, where $\hat{V}$ is the unique solution to

$$\hat{V}(t)=(\hat{q}_0^+\mu t+\hat{q}_0(1-\mu t)+\hat{I}(t)+\hat{X}(t)-\beta\mu t-\theta\int_0^t\hat{V}(t-s)\,ds)^+,\quad t\in[0,1/\mu),$$

and

$$\hat{V}(t)=(\hat{X}(t)-\beta+\hat{V}(t-1/\mu)-\theta\int_0^{1/\mu}\hat{V}(t-s)\,ds)^+,\quad t\geq 1/\mu.$$

When comparing the present example with the QED $G/D/N$ queue (no abandonment) (Jelenković et al. [17]), one observes that having abandonments results in more complex dynamics. Specifically, whereas in Jelenković et al. [17] the distribution of $\hat{Q}(t)$ depends only on $\hat{Q}(t-1/\mu)$ (as far as $\hat{Q}$ is concerned), here $\hat{Q}(t)$ depends on all values of $\hat{Q}$ during the time interval $[t-1/\mu,t)$. This is due to the presence of the residual service distribution in the operators ϕ and ψ.

6. Future research: Stationary distribution. Our analysis addresses the transient behavior of a QED system with impatient customers. The *stationary* distributions of the queue length and the waiting time remain unknown, as is the case for the corresponding system without abandonment; note that the system with impatient customers remains stable (as $t\to\infty$) for all finite values of the capacity parameter β. (A large-deviation characterization of the stationary distributions for a QED queue without abandonments can be found in Gamarnik and Momčilović [10].)

We observe that Example 5.2 is consistent with the results in Garnett et al. [12] on the *stationary* number-in-system process (for the $M/M/N+M$ system in the QED regime). Based on (37), it is thus tempting to conjecture that, for the $G/M/N+GI$ system, the stationary versions of number-in-system processes converge weakly, in the QED regime, as $N\to\infty$, to the process $\tilde{Q}=\{\tilde{Q}(t),t\in\mathbb{R}\}$, where $\tilde{Q}$ is the unique stationary process that solves

$$\tilde{Q}(t)=\tilde{X}(t)-\beta+(\mu-\theta)\int_{-\infty}^t\tilde{Q}^+(s)\exp\{-\mu(t-s)\}\,ds;$$

here $\tilde{X}=\{\tilde{X}(t),t\in\mathbb{R}\}$ is the stationary version of the infinite-server process $\hat{X}$ (see also Lemma 3.2). Under assumption (36), $\tilde{X}$ satisfies $d\tilde{X}(t)=-\mu\tilde{X}(t)\,dt+\sqrt{\mu(1+c^2)}\,dB(t)$, where $\{B(t),t\in\mathbb{R}\}$ is a standard Brownian motion (since $\hat{I}$ vanishes as $t\to\infty$; see Lemma 3.2, and Example 5.2 in §5). An example where these assumptions ($\tilde{X}$ stationary and (36)) prevail is when the arrival process is stationary renewal and $\hat{q}_0$ has the corresponding stationary distribution. A conjecture for the stationary distribution of $\hat{q}_0$ is provided in (41) below; in the case of Poisson arrivals the (diffusion) stationary distribution of $\hat{q}_0$ was calculated in Zeltyn [33]. Consequently, $\tilde{Q}$ is a (piecewise) Ornstein–Uhlenbeck process ($\tilde{Q}$ satisfies (40), where $\tilde{Q}$ substitutes for $\hat{Q}$), as derived earlier in Garnett et al. [12] for the case $c=1$ (Poisson arrivals). Based on the preceding and Browne and Whitt [5], one can calculate the probability density function of $\tilde{Q}(t)$ (see also Garnett et al. [12], Zeltyn [33]):

$$f_{\tilde{Q}(t)}(q)=\chi f_-(q)1_{\{q\leq 0\}}+(1-\chi)f_+(q)1_{\{q>0\}},\tag{41}$$

where $f_-(q)=\tilde{c}\Phi'(\tilde{c}(q+\beta))/\Phi(\tilde{c}\beta)$, $f_+(q)=\tilde{c}\sqrt{\theta/\mu}\Phi'(\tilde{c}(q\sqrt{\theta/\mu}+\beta\sqrt{\mu/\theta}))/\Phi(-\tilde{c}\beta\sqrt{\mu/\theta})$,

$$\tilde{c}=\sqrt{\frac{2}{1+c^2}},$$

Φ and Φ' are the distribution and density functions of the standard normal random variable, respectively, and $\chi = f_+(0)/(f_+(0)+f_-(0))$. Furthermore, from the stochastic differential equation for $\tilde{Q}$, one deduces directly that the stationary distribution of $\tilde{c}\tilde{Q}(t)$ is equal to the stationary distribution of the identically scaled limiting queue length $\tilde{Q}_A(t)$ in the Erlang-A model, but with the load parameter $\tilde{c}\beta$. This makes results on the Erlang-A model, documented for example in Mandelbaum and Zeltyn [23], directly applicable to the QED $G/M/N+GI$ queue. For example,

$$\mathbb{P}[\text{wait} > 0] = \mathbb{P}[\tilde{Q}(t) > 0] = \left(1 + \sqrt{\frac{\theta}{\mu}}\frac{h(\tilde{c}\beta\sqrt{\mu/\theta})}{h(-\tilde{c}\beta)}\right)^{-1}$$

and

$$\mathbb{E}\,\tilde{Q}^+(t) = \frac{\mu}{\tilde{c}\theta}\left(h\left(\tilde{c}\beta\sqrt{\mu/\theta}\right) - \tilde{c}\beta\sqrt{\mu/\theta}\right)\left(\sqrt{\frac{\mu}{\theta}} + \frac{h(\tilde{c}\beta\sqrt{\mu/\theta})}{h(-\tilde{c}\beta)}\right)^{-1},$$

where $h(q) = \Phi'(q)/(1-\Phi(q))$ is the hazard rate of the standard normal distribution. Corollary 4.2 and Remark 4.3 now provide a recipe for calculating also performance measures that involve waiting time. In particular, it is well known that $\mathbb{P}[\text{abandon}] = \theta\,\mathbb{E}[\text{wait}]$ when the patience distribution is exponential.

7. Proofs

7.1. Proof of Lemma 2.1. Let $\{\hat{s}_{-i}, i \geq 1\}$ and $\{\hat{p}_{-i}, i \geq 1\}$ be two i.i.d. sequences defined by distributions F_* and G^N, respectively. The FCFS policy implies $v_{-i-1} \leq v_{-i}$, for $1 \leq i < (q_0^N - N)^+$ and, hence, for $\epsilon > 0$, $v \geq 0$ and $c \geq 0$, we have

$$\begin{aligned}
\mathbb{P}[r_0^N/\sqrt{N} > \epsilon] &\leq \mathbb{P}\left[\sum_{i=1}^{(q_0^N-N)^+} 1_{\{p_{-i}\leq v\}} > \epsilon\sqrt{N}\right] + \mathbb{P}[v_{-1} > v] \\
&\leq \mathbb{P}\left[\sum_{i=1}^{\lceil c\sqrt{N}\rceil} 1_{\{\hat{p}_{-i}\leq v\}} > \epsilon\sqrt{N}\right] + \mathbb{P}[\hat{q}_0^N > c] + \mathbb{P}\left[\sum_{i=(q_0^N-N)^++1}^{q_0^N} 1_{\{s_{-i}\leq v\}} < q_0^N - N\right] \\
&\leq \frac{\lceil c\sqrt{N}\rceil}{\epsilon\sqrt{N}} G^N(v) + \mathbb{P}\left[\sum_{i=1}^{N}(F_*(v) - 1_{\{\hat{s}_{-i}\leq v\}}) > NF_*(v) - c\sqrt{N}\right] + 2\,\mathbb{P}[\hat{q}_0^N > c],
\end{aligned}$$

where the second inequality is due to the fact that the event $\{v_{-1} > v\}$ implies that the number of service completions in the time interval $[0, v]$ is less than $(q_0^N - N)$; in addition, the number of service completions in $[0, t]$ is lower bounded by the sum in the last term in the second inequality; Markov inequality is used to obtain the third inequality. Setting $v = d/\sqrt{N}$, with $d = d(c)$ large enough such that $\sqrt{N}F_*(d/\sqrt{N}) - c > \epsilon$ for all N large enough (which is feasible due to definition (2) of F_*) and applying Markov inequality result in

$$\mathbb{P}[r_0^N/\sqrt{N} > \epsilon] \leq \frac{\lceil c\sqrt{N}\rceil}{\epsilon\sqrt{N}} G^N(d/\sqrt{N}) + \frac{F_*(d/\sqrt{N})}{(\sqrt{N}F_*(d/\sqrt{N}) - c)^2} + 2\,\mathbb{P}[\hat{q}_0^N > c].$$

Finally, letting first $N \to \infty$, recalling (2), (6), and (7), and then letting $c \to \infty$ yields the statement of the lemma. □

7.2. Proof of Proposition 3.1. It is sufficient to prove the statement for offered waiting times because it implies the result for queue lengths:

$$\begin{aligned}
H^N(t) &= \sum_{i=1}^{A^N(t)} 1_{\{t_i+s_i+V_{\leftarrow}^N(t_i)>t\}} 1_{\{V_{\leftarrow}^N(t_i)\leq p_i\}} + \sum_{i=1}^{q_0^N\wedge N} 1_{\{s_{-i}>t\}} + \sum_{i=q_0^N\wedge N+1}^{q_0^N} 1_{\{s_{-i}+v_{-i}>t\}} \\
&\leq \sum_{i=1}^{A^N(t)} 1_{\{t_i+s_i+\dot{V}_{\leftarrow}^N(t_i)>t\}} + \sum_{i=1}^{q_0^N\wedge N} 1_{\{s_{-i}>t\}} + \sum_{i=q_0^N\wedge N+1}^{q_0^N} 1_{\{s_{-i}+\dot{v}_{-i}>t\}} = \dot{H}^N(t),
\end{aligned}$$

for $t \geq 0$; note that $\dot{v}_i = v_i$ and $\dot{p}_i = p_i = \infty$, for $-q_0^N \leq i < -q_0^N \wedge N$, by construction. Furthermore, one can consider $V_{\leftarrow}^N$ and $\dot{V}_{\leftarrow}^N$ only at the moments of arrivals ($t = t_i$ for some $i \geq 0$) and $t = 0$, because, between arrivals, both $V_{\leftarrow}^N$ and $\dot{V}_{\leftarrow}^N$ remain constant.

Now, consider the closely related shortest-workload-first routing policy (that can be conveniently described by the Kiefer–Wolfowitz recurrence; e.g., see Baccelli and Bremaud [1, p. 91]), and let $W_n^N(t)$ and $\dot{W}_n^N(t)$, $1 \le n \le N$, be the nth smallest server workload in the system with and without abandonment, respectively. Then, it is well known that $V_{\leftarrow}^N(t_i) = W_1^N(t_i-)$ and $\dot{V}_{\leftarrow}^N(t_i) = \dot{W}_1^N(t_i-)$. Starting an induction, assume

$$W_n^N(t_i) \le \dot{W}_n^N(t_i) \tag{42}$$

for some $i \ge 1$ and all $1 \le n \le N$; the base of the induction is due to the assumption on the initial states (at $t = 0$). Let $\mathcal{R}$ be the standard reorder operator. Then, because the vectors of W_n^N's and $\dot{W}_n^N$'s satisfy the Kiefer–Wolfowitz recurrence, it follows that

$$\begin{aligned}
&(W_1^N(t_{i+1}), W_2^N(t_{i+1}), \ldots, W_N^N(t_{i+1}))\\
&\quad = \mathcal{R}(W_1^N(t_i) + s_{i+1}1_{\{p_i \le W_1^N(t_{i+1}-)\}} - t_{i+1} + t_i, W_2^N(t_i) - t_{i+1} + t_i, \ldots, W_N^N(t_i) - t_{i+1} + t_i)^+\\
&\quad \le \mathcal{R}(\dot{W}_1^N(t_i) + s_{i+1} - t_{i+1} + t_i, \dot{W}_2^N(t_i) - t_{i+1} + t_i, \ldots, \dot{W}_N^N(t_i) - t_{i+1} + t_i)^+\\
&\quad = (\dot{W}_1^N(t_{i+1}), \dot{W}_2^N(t_{i+1}), \ldots, \dot{W}_N^N(t_{i+1})),
\end{aligned}$$

where the inequality is due to the inductive assumption (42); the operator $(\cdot)^+$ is applied element wise. Therefore, (42) holds for all $i \ge 1$ and the proposition prevails. □

7.3. Proof of Lemma 3.3. Let $A_{\triangleright}^N = \{A_{\triangleright}^N(t), t \ge 0\}$, where, for $t \ge 0$,

$$A_{\triangleright}^N(t) = \sum_{i=1}^{A^N(t)} 1_{\{p_i > v_i\}}$$

represents the number of customers with arrival times in $[0, t]$ that eventually receive service (do not abandon); the process $A_{\triangleright}^N$ was also considered in §3.4 (see Lemma 3.6). Define a two-dimensional process $\{E^N(t, s),\ t \ge 0,\ s \ge 0\}$ by

$$E^N(t, s) = \sum_{i=A_{\triangleright}^N(t)-(H^N(t)-N)^+ +1}^{A_{\triangleright}^N(t)} 1_{\{\tilde{s}_i \le s\}},$$

where $\tilde{s}_i = s_{i-1}$, $-(q_0^N - N)^+ < i \le 0$, and $\tilde{s}_i = s_{A^N(\tilde{t}_i)}$, $i \ge 1$ with $\tilde{t}_i = \inf\{t \ge 0\colon A_{\triangleright}^N(t) = i\}$. The value of $E^N(t, s)$ is equal to the number of customers awaiting service at time t with service requirement at most s (recall that customers abandon upon arrival, if at all). Let $w_i = v_i 1_{\{p_i > v_i\}}$ for $i \ge 0$, $w_{-i} = v_{-i}$ for $1 \le i \le (q_0^N - N)^+$, and $w_{-i} = 0$ for $(q_0^N - N)^+ < i \le q_0^N$; note that $w_i = 0$ for all customers that abandon the system. Alternatively, $E^N(t, s)$ can be expressed as a sum over all customer indices:

$$E^N(t, s) = \sum_{i=-q_0^N}^{A^N(t)} 1_{\{t_i \le t < t_i + w_i\}} 1_{\{s_i \le s\}}, \tag{43}$$

where $t_{-i} = 0$ for $i = 1, \ldots, q_0^N$, and the element of the sum corresponding to $i = 0$ does not exist. Furthermore, we define $\{F^N(t, s), t \ge 0, s \ge 0\}$ by

$$F^N(t, s) := 1_{\{H^N(t) > N\}} \frac{E^N(t, s)}{(H^N(t) - N)^+}, \tag{44}$$

and note that

$$E^N(t, s) = E^N(t, s)1_{\{H^N(t) > N\}} = (H^N(t) - N)^+ F^N(t, s); \tag{45}$$

on the event $\{H^N(t) > N\}$, $F^N(t, \cdot)$ can be interpreted as the (empirical) distribution function of service requirements for customers awaiting service at time t. Observe that, for $\delta > 0$, (43) renders

$$E^N(t - s, s + \delta) - E^N(t - s, s) = \sum_{i=-q_0^N}^{A^N(t)} 1_{\{t_i \le t - s < t_i + w_i\}} 1_{\{s_i - \delta \le s < s_i\}}.$$

In view of the preceding equality, the change in the order of summation results in

$$\int_0^t E^N(t-s,\,ds) = \sum_{i=-q_0^N}^{A^N(t)} \int_0^t 1_{\{t_i \le t-s < t_i+w_i\}}\, d1_{\{s_i \le s\}}$$
$$= \sum_{i=-q_0^N}^{A^N(t)} 1_{\{t_i \le t-s_i < t_i+w_i\}},$$

and, thus, due to (45), we have

$$\int_0^t (H^N(t-s)-N)^+ F^N(t-s,\,ds) = \sum_{i=-q_0^N}^{A^N(t)} 1_{\{t-t_i-w_i < s_i \le t-t_i\}}. \tag{46}$$

On the other hand, for any $t \ge 0$, Proposition 2.1 in Reed [27] yields

$$\int_0^t (H^N(t-s)-N)^+\, dF(s) = \sum_{i=-q_0^N}^{A^N(t)} (\bar{F}(t-t_i-w_i) - \bar{F}(t-t_i)), \tag{47}$$

because only customers that do not abandon potentially contribute to the sum on the right-hand side of (47). Therefore, (46) and (47) imply (see (10) and (12)), for $t \ge 0$,

$$(\hat{X}_\Delta^N + \hat{I}_\Delta^N)(t) = \int_0^t (\hat{H}^N(t-s))^+ (F^N(t-s,\,ds) - F(ds)). \tag{48}$$

Next, extend the i.i.d. sequence $\{\tilde{s}_i,\, i > -(q_0^N - N)^+\}$ to all integer indices (by letting $\{\tilde{s}_i,\, i \le -(q_0^N - N)^+\}$ be an i.i.d. sequence, independent of $\{\tilde{s}_i,\, i > -(q_0^N - N)^+\}$, with its elements distributed according to F); observe that $\{\tilde{s}_i,\, i \in \mathbb{Z}\}$ is an i.i.d. sequence because both subsequences are i.i.d. (defined by F) and independent of each other. Now, define a family of empirical distribution functions $F_{i,j} = \{F_{i,j}(s),\, s \ge 0\}$:

$$F_{i,j}(s) = \frac{1}{j} \sum_{k=i-j+1}^{i} 1_{\{\tilde{s}_k \le s\}}, \tag{49}$$

where $i \ge 0$ and $j \ge 1$. In what follows, we estimate $\|F_{i,j} - F\|_\infty$ for a range of indices i and j. To this end, for any $\epsilon > 0$ and $s \ge 0$, there exist constants $\theta(\epsilon, s) > 0$ and $\gamma(\epsilon, s) < \infty$ (e.g., see Billingsley [3, p. 151]) such that, for all $j \ge 1$ (and all i),

$$\mathbb{P}[|F_{i,j}(s) - F(s)| > \epsilon] \le \gamma(\epsilon, s) \exp\{-j\theta(\epsilon, s)\}. \tag{50}$$

Moreover, by the same argument, replacing $1_{\{\tilde{s}_k \le s\}}$ with $1_{\{\tilde{s}_k < s\}}$ in the definition of $F_{i,j}(s)$ yields

$$\mathbb{P}[|F_{i,j}(s-) - F(s-)| > \epsilon] \le \gamma(\epsilon, s) \exp\{-j\theta(\epsilon, s)\}, \tag{51}$$

where $F(s-) = \mathbb{E}\, 1_{\{s_i < s\}}$, $i \ge 1$; the constants in (50) and (51) may differ in general. Given the distribution function F, for any $\epsilon > 0$ there exists a finite sequence of nonnegative reals $\{a_l,\, 1 \le l \le L\}$ such that

$$\bigcap_{l=1}^{L} \{\{|F_{i,j}(a_l) - F(a_l)| \le \epsilon\} \bigcap \{|F_{i,j}(a_l-) - F(a_l-)| \le \epsilon\}\} \subseteq \{\|F_{i,j} - F\|_\infty \le 2\epsilon\}.$$

This relationship, (50), and (51) imply the existence of $\theta(\epsilon) > 0$ and $\gamma(\epsilon) < \infty$ such that

$$\mathbb{P}[\|F_{i,j} - F\|_\infty > \epsilon] \le \gamma(\epsilon) \exp\{-j\theta(\epsilon)\}, \tag{52}$$

for $i \ge 0$ and $j \ge 1$. Now, we introduce a nonnegative real that characterizes a distance between F and $F_{i,j}$ for multiple indices i and j:

$$f_{k,l,n} = \sup_{0 \le i \le k} \sup_{l \le j \le k+n} \|F_{i,j} - F\|_\infty, \tag{53}$$

where $l \geq 1$. Then, for $\epsilon > 0$, the union bound and (52) yield $\mathbb{P}[f_{k,l,n} > \epsilon] \leq (k+1)(k+n)\gamma(\epsilon)\exp\{-l\theta(\epsilon)\}$. Finally, for any $\epsilon > 0$, the last inequality, (3), (5), and (7) result in, as $N \to \infty$,

$$\mathbb{P}[f_{A^N(T),\epsilon\sqrt{N},q_0^N} > \epsilon] \to 0. \tag{54}$$

Next, considering whether $\{\hat{H}^N(t-s) \leq \epsilon\}$ or $\{\hat{H}^N(t-s) > \epsilon\}$ in (48) yields

$$\begin{aligned}
\|\hat{X}_\Delta^N + \hat{I}_\Delta^N\|_T &\leq \epsilon + \sup_{0\leq t\leq T}\left|\int_0^t 1_{\{\hat{H}^N(t-s)>\epsilon\}}(\hat{H}^N(t-s))^+(F^N(t-s,ds) - F(ds))\right| \\
&\leq \epsilon + \|\hat{H}^N\|_T \sup_{0\leq t\leq T}\sup_{0\leq t\leq s}|(F^N(t-s,s) - F(s))1_{\{\hat{H}(t-s)>\epsilon\}}| \\
&\leq \epsilon + \|\hat{H}^N\|_T \sup_{0\leq t\leq T}\sup_{0\leq t\leq s}|F_{A^N(t-s),H^N(t-s)-N}(s) - F(s)| \\
&\leq \epsilon + \|\hat{H}^N\|_T f_{A^N(T),\epsilon\sqrt{N},q_0^N},
\end{aligned}$$

where the third inequality is due to $F^N(t,s) = F_{A^N(t),H^N(t)-N}(s)$ on the event $\{H^N(t) > N\}$ (see (44) and (49)); the last inequality follows from (53). Now, for any $\delta > 0$ there exists $\epsilon > 0$, small enough, so that the preceding inequality results in

$$\begin{aligned}
\mathbb{P}[\|\hat{X}_\Delta^N + \hat{I}_\Delta^N\|_T > 2\delta] &\leq \mathbb{P}[\|\hat{H}^N\|_T f_{A^N(T),\epsilon\sqrt{N},q_0^N} > \delta] \\
&\leq \mathbb{P}[f_{A^N(T),\epsilon\sqrt{N},q_0^N} > \delta/c] + \mathbb{P}[\|\hat{H}^N\|_T > c],
\end{aligned} \tag{55}$$

where $c > 0$ is arbitrary. Finally, taking lim sup (as $N \to \infty$) on both sides of (55) yields, due to (54),

$$\limsup_{N\to\infty}\mathbb{P}[\|\hat{X}_\Delta^N + \hat{I}_\Delta^N\|_T > \delta] \leq \limsup_{N\to\infty}\mathbb{P}[\|\hat{H}^N\|_T > c].$$

The final statement follows from the preceding by letting $c \to \infty$, Proposition 3.1, and Theorem 5.1 in Reed [27]. □

7.4. Proof of Lemma 3.4. For fixed $T > 0$ and $\Delta > 0$, the following holds:

$$\|\hat{Z}^N\|_T \leq \max_{0\leq i\leq\lfloor T/\Delta\rfloor}|\hat{Z}^N(i\Delta)| + \max_{0\leq i\leq\lfloor T/\Delta\rfloor}\sup_{0\leq\delta\leq\Delta}|\hat{Z}^N(i\Delta+\delta) - \hat{Z}^N(i\Delta)|. \tag{56}$$

First, we argue that $\hat{Z}^N(t) \Rightarrow 0$, as $N \to \infty$, for any fixed $t \geq 0$. For notational purposes, it is convenient to define the random variables $z_i(t) = 1_{\{s_i > t-t_i\}}1_{\{p_i\leq v_i\}} - \bar{F}(t-t_i)G^N(v_i)$; observe that $\mathbb{E}\, z_i(t) = 0$ since s_i, p_i are independent of t_i, v_i, and, hence, $\mathbb{E}[z_i(t) \mid t_i, v_i] = 0$. From (15) we have that $\mathbb{E}\,\hat{Z}^N(t) = 0$ and the second moment is given by

$$\begin{aligned}
\mathbb{E}(\hat{Z}^N(t))^2 &= \frac{1}{N}\mathbb{E}\sum_{i=1}^{A^N(t)} z_i^2(t) + \frac{2}{N}\mathbb{E}\sum_{i=1}^{A^N(t)}\sum_{j=i+1}^{A^N(t)} z_i(t)z_j(t) \\
&= \frac{1}{N}\mathbb{E}\sum_{i=1}^{A^N(t)} \bar{F}(t-t_i)G^N(v_i)(1 - \bar{F}(t-t_i)G^N(v_i));
\end{aligned}$$

the expectation of the double sum equals 0 because the service requirement and patience of an arriving customer is independent of the state of the system. Then, given that F and G^N are distribution functions, it follows that, for $\epsilon > 0$,

$$\mathbb{P}[|\hat{Z}^N(t)| > \epsilon] \leq \frac{1}{\epsilon^2 N}\mathbb{E}\sum_{i=1}^{A^N(t)} G^N(v_i) \to 0,$$

as $N \to \infty$, due to (3), (5), (6), and Proposition 3.2; thus, for fixed t, as $N \to \infty$,

$$\hat{Z}^N(t) \;\Rightarrow\; 0. \tag{57}$$

Next, we consider the second term on the right-hand side of (56). To this end, for $t > 0$ and $\delta > 0$, we have (see (14))

$$\hat{Z}^N(t+\delta) - \hat{Z}^N(t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{A^N(t)}(z_i(t+\delta) - z_i(t)) + \frac{1}{\sqrt{N}}\sum_{i=A^N(t)+1}^{A^N(t+\delta)} z_i(t+\delta),$$

and upper and lower bounds follow:

$$\begin{aligned}\hat{Z}^N(t+\delta)-\hat{Z}^N(t) &\le \frac{1}{\sqrt{N}}\sum_{i=1}^{A^N(t)}(\bar{F}(t-t_i)-\bar{F}(t+\delta-t_i))G^N(v_i)+\frac{1}{\sqrt{N}}\sum_{i=A^N(t)+1}^{A^N(t+\delta)}1_{\{p_i\le v_i\}}\\ &:= \hat{Z}^N_{(\uparrow,\delta)}(t);\\ \hat{Z}^N(t+\delta)-\hat{Z}^N(t) &\ge -\frac{1}{\sqrt{N}}\sum_{i=1}^{A^N(t)}(1_{\{s_i>t-t_i\}}-1_{\{s_i>t+\delta-t_i\}})1_{\{p_i\le v_i\}}-\frac{1}{\sqrt{N}}\sum_{i=A^N(t)+1}^{A^N(t+\delta)}G^N(v_i)\\ &:= -\hat{Z}^N_{(\downarrow,\delta)}(t).\end{aligned}$$

The nonnegativity of $\hat{Z}^N_{(\uparrow,\delta)}(t)$ and $\hat{Z}^N_{(\downarrow,\delta)}(t)$ and their monotonicity in δ imply

$$\sup_{0\le\delta\le\Delta}|\hat{Z}^N(t+\delta)-\hat{Z}^N(t)|\le\hat{Z}^N_{(\uparrow,\Delta)}(t)+\hat{Z}^N_{(\downarrow,\Delta)}(t). \tag{58}$$

For notational simplicity, introduce $A^N_{(c)}=\{A^N_{(c)}(t),t\ge0\}$ by

$$\begin{aligned}A^N_{(c)}(t) &:= \sum_{i=1}^{A^N(t)}1_{\{p_i\le c/(\mu\sqrt{N})\}}\\ &= \sum_{i=1}^{A^N(t)}(1_{\{p_i\le c/(\mu\sqrt{N})\}}-G^N(c/(\mu\sqrt{N})))-A^N(t)G^N(c/(\mu\sqrt{N}));\end{aligned} \tag{59}$$

also, set $\tilde{t}_i=\inf\{t\ge0\colon A^N_{(c)}(t)\ge i\}$ and $\tilde{s}_i=\{s_j\colon \tilde{t}_i=t_j\}$. The process $A^N_{(c)}$ is the arrival process of customers with patience at most $c/(\mu\sqrt{N})$, in the Nth system. Limits (3), (5), and (6) imply

$$\left\{\frac{A^N(t)}{N}\sqrt{N}G^N(c/(\mu\sqrt{N})),\ t\ge0\right\}\to c\theta e, \tag{60}$$

a.s. u.o.c., as $N\to\infty$, while the martingale inequality (Chung [6, Corollary 1, p. 331]) and (6) yield, for $T>0$,

$$\mathbb{P}\left[\sup_{1\le j\le2\mu TN}\left|\sum_{i=1}^{j}(1_{\{p_i\le c/(\mu\sqrt{N})\}}-G^N(c/(\mu\sqrt{N})))\right|>\epsilon\sqrt{N}\right]\le\frac{2\mu TG^N(c/(\mu\sqrt{N}))}{\epsilon^2}\to0, \tag{61}$$

as $N\to\infty$. Combining (59), (60), and (61) results in

$$A^N_{(c)}/\sqrt{N}\ \Rightarrow\ c\theta e, \tag{62}$$

as $N\to\infty$. Now, for $t\le T-\delta$, on the event $\{\|\hat{V}^N_{\leftarrow}\|_T\le c\}$ the first term on the right-hand side of (58) can be upper bounded by using monotonicity:

$$\begin{aligned}\hat{Z}^N_{(\uparrow,\Delta)}(t) &\le \frac{1}{\sqrt{N}}G^N(c/(\mu\sqrt{N}))\sum_{i=1}^{A^N(t)}(\bar{F}(t-t_i)-\bar{F}(t+\Delta-t_i))+\frac{1}{\sqrt{N}}(A^N_{(c)}(t+\Delta)-A^N_{(c)}(t))\\ &\Rightarrow c\theta\int_0^t(\bar{F}(t-s)-\bar{F}(t+\Delta-s))\,ds+c\theta\Delta,\end{aligned} \tag{63}$$

as $N\to\infty$, where the limit is due to (3), (5), and (6). Similarly, on the event $\{\|\hat{V}^N_{\leftarrow}\|_T\le c\}$, we have

$$\begin{aligned}\hat{Z}^N_{(\downarrow,\Delta)}(t) &\le \frac{1}{\sqrt{N}}\sum_{i=1}^{A^N_{(c)}(t)}(1_{\{\tilde{s}_i>t-\tilde{t}_i\}}-1_{\{\tilde{s}_i>t+\Delta-\tilde{t}_i\}})+\frac{1}{\sqrt{N}}(A^N(t+\Delta)-A^N(t))G^N(c/(\mu\sqrt{N}))\\ &\Rightarrow c\theta\int_0^t(\bar{F}(t-s)-\bar{F}(t+\Delta-s))\,ds+c\theta\Delta,\end{aligned} \tag{64}$$

as $N\to\infty$, where the limit is due to (3), (5), (6), and Theorem 3 in Krichagina and Puhalskii [20]. Now, for $c>0$, (58) implies

$$\mathbb{P}\left[\sup_{0\le\delta\le\Delta}|\hat{Z}^N(t+\delta)-\hat{Z}^N(t)|>\epsilon\right]\le\mathbb{P}[\hat{Z}^N_{(\uparrow,\Delta)}(t)+\hat{Z}^N_{(\downarrow,\Delta)}(t)>\epsilon,\|\hat{V}^N_{\leftarrow}\|_T\le c]+\mathbb{P}[\|\hat{V}^N_{\leftarrow}\|_T>c].$$

Selecting Δ small enough, letting $N \to \infty$ on both sides in the preceding inequality, using (63) and (64), and then increasing $c \to \infty$ yields (for fixed t)

$$\sup_{0 \le \delta \le \Delta} |\hat{Z}^N(t+\delta) - \hat{Z}^N(t)| \;\Rightarrow\; 0; \tag{65}$$

the limit is also due to Proposition 5.3 in Reed [27] and Proposition 3.1.

Finally, the lemma follows from (56), (57), and (65). □

7.5. Proof of Lemma 3.5. In view of Lemma 3.3, it is sufficient to prove $\hat{Z}^N_\Delta \Rightarrow 0$, as $N \to \infty$. Recall the definitions of $A^N_{(c)}$, $\{\tilde{t}_i, i \ge 1\}$ and $\{\tilde{s}_i, i \ge 1\}$ from the proof of Lemma 3.4. Now, for arbitrary $T > 0$ and $\epsilon > 0$, we have

$$\mathbb{P}[\|\hat{Z}^N_\Delta\|_T > \epsilon] \le \mathbb{P}[\|\hat{Z}^N_\Delta\|_T > \epsilon, \|\hat{V}^N_{\leftarrow}\|_T \le c] + \mathbb{P}[\|\hat{V}^N_{\leftarrow}\|_T > c].$$

On the event $\{\|\hat{V}^N_{\leftarrow}\|_T \le c\}$, the process $\hat{Z}^N_\Delta$, based on its definition, can be upper bounded as follows, for all $t \in [0, T]$ and all sufficiently large N:

$$\begin{aligned} |\hat{Z}^N_\Delta(t)| \;&\le\; \frac{1}{\sqrt{N}} \sum_{i=1}^{A^N_{(c)}(t)} (1_{\{\tilde{s}_i > t - \tilde{t}_i - \delta\}} - 1_{\{\tilde{s}_i > t - \tilde{t}_i\}}) + \frac{1}{\sqrt{N}} \int_0^t (\bar{F}(t-s-\delta) - \bar{F}(t-s))\, dA^N_{(c)}(s) \\ &=: \hat{Z}^N_{(c,\delta)}(t), \end{aligned}$$

where $\delta > 0$. The preceding two inequalities render

$$\mathbb{P}[\|\hat{Z}^N_\Delta\|_T > \epsilon] \le \mathbb{P}[\|\hat{Z}^N_{(c,\delta)}\|_T > \epsilon] + \mathbb{P}[\|\hat{V}^N_{\leftarrow}\|_T > c], \tag{66}$$

where $\hat{Z}^N_{(c,\delta)} = \{\hat{Z}^N_{(c,\delta)}(t), t \ge 0\}$.

Next, Theorem 3 in Krichagina and Puhalskii [20] and (62) yield, as $N \to \infty$,

$$\{\hat{Z}^N_{(c,\delta)}(t), t \ge 0\} \;\Rightarrow\; \left\{2c\theta \int_0^t (\bar{F}(t-s-\delta) - \bar{F}(t-s))ds,\ t \ge 0\right\}. \tag{67}$$

On the other hand, Proposition 3.2 implies

$$\lim_{c \to \infty} \limsup_{N \to \infty} \mathbb{P}[\|\hat{V}^N_{\leftarrow}\|_T > c] = 0. \tag{68}$$

Therefore, in view of (66), (67), and (68), given T and ϵ, for any $\xi > 0$, it is possible to select c and δ such that $\mathbb{P}[\|\hat{Z}^N_\Delta\|_T > \epsilon] < \xi$ for all N large enough. □

7.6. Proof of Proposition 3.3. Two cases are considered separately: (i) nondeterministic service times and (ii) deterministic service times. Let $y_1 = \varphi(x_1)$ and $y_2 = \varphi(x_2)$ for $x_1, x_2 \in D[0, \infty)$.

(i) Because service times are not single valued, there exist $\delta > 0$ and $0 < \epsilon < 1$ such that $F(x+\delta) - F(x) < \epsilon$, for all $x \ge 0$. Then it follows that

$$\begin{aligned} d^\delta_{L^1}(y_1, y_2) &\le d^\delta_{L^1}(x_1, x_2) + \int_0^\delta \int_0^t |y_1(t-s) - y_2(t-s)|\, dF(s)\, dt \\ &\le d^\delta_{L^1}(x_1, x_2) + \int_0^\delta d^\delta_{L^1}(y_1, y_2)\, dF(s) \\ &\le d^\delta_{L^1}(x_1, x_2) + \epsilon d^\delta_{L^1}(y_1, y_2), \end{aligned}$$

and, thus,

$$d^\delta_{L^1}(y_1, y_2) \le d^\delta_{L^1}(x_1, x_2)/(1-\epsilon). \tag{69}$$

Similarly, considering the time interval $[0, 2\delta]$ yields

$$\begin{aligned} d^{2\delta}_{L^1}(y_1, y_2) &\le d^{2\delta}_{L^1}(x_1, x_2) + \epsilon d^\delta_{L^1}(y_1, y_2) + \epsilon d^{2\delta}_{L^1}(y_1, y_2) \\ &\le d^{2\delta}_{L^1}(x_1, x_2)/(1-\epsilon) + \epsilon d^{2\delta}_{L^1}(y_1, y_2), \end{aligned}$$

where the second inequality is due to (69). From the preceding inequality one derives $d^{2\delta}_{L^1}(y_1, y_2) \le d^{2\delta}_{L^1}(x_1, x_2)/(1-\epsilon)^2$. The above argument can be applied l times iteratively to obtain $d^{l\delta}_{L^1}(y_1, y_2) \le d^{l\delta}_{L^1}(x_1, x_2)/(1-\epsilon)^l$. Therefore, for any T, there exists $c_T < \infty$ such that $d^T_{L^1}(y_1, y_2) \le c_T d^T_{L^1}(x_1, x_2)$.

(ii) Let a be such that $F(a-) = 0$ and $F(a) = 1$. Then $y_i(t) = x_i(t)$, $i = 1, 2$, for $t < a$ and $d^T_{L^1}(y_1, y_2) = d^T_{L^1}(x_1, x_2)$ for $T < a$. Next, assume that $d^T_{L^1}(y_1, y_2) \le c_T d^T_{L^1}(x_1, x_2)$ for some $T \ge a$ and $c_T < \infty$. Because of this assumption, since $y_i(t) = x_i(t) + y_i^+(t-a)$, $i = 1, 2$, for $t \ge a$, one has, for $0 < d \le a$,

$$\begin{aligned} d^{T+d}_{L^1}(y_1, y_2) &\le d^{T+d}_{L^1}(x_1, x_2) + d^T_{L^1}(y_1, y_2) \\ &\le d^{T+d}_{L^1}(x_1, x_2) + c_T d^T_{L^1}(x_1, x_2) \\ &\le (1 + c_T) d^{T+d}_{L^1}(x_1, x_2). \end{aligned}$$

The conclusion follows. □

7.7. Proof of Lemma 3.7. In view of Proposition 3.3, it is sufficient to consider the argument of the φ operator in (21). Recall the definition of $d_{J_1}(\cdot, \cdot)$ from the proof of Proposition 3.3.

The nondecreasing nature of distribution functions yields

$$\begin{aligned} d^T_{L^1}(F \circ (\tau^N + V^N_{\leftarrow}), F) &\le \int_0^T (F(t + \|V^N_{\leftarrow}\|_T) - F(t - \|e - \tau^N\|_T))dt \\ &\le \|V^N_{\leftarrow}\|_T + \|e - \tau^N\|_T, \end{aligned}$$

where the second inequality follows from $F(t) \le 1$ for all t; similarly,

$$d^T_{L^1}(F_* \circ (\tau^N + V^N_{\leftarrow}), F_*) \le \|V^N_{\leftarrow}\|_T + \|e - \tau^N\|_T.$$

For notational simplicity, let $\hat{J}^N := (\hat{q}^N_0)^+(\bar{F} - \bar{F}_*) + \hat{q}^N_0 \bar{F}_* - \sqrt{N}(1 - \rho^N)F_*$. The preceding two inequalities, jointly with (5) and (7), yield, as $N \to \infty$,

$$d^T_{L^1}(\hat{J}^N \circ (\tau^N + V^N_{\leftarrow}), \hat{J}^N) \;\Rightarrow\; 0. \tag{70}$$

The triangle inequality and the definition of $d^T_{L^1}$ (see (1)) result in

$$\begin{aligned} &d^T_{L^1}((\hat{I}^N_\Delta + \hat{Y}^N_\Delta - \hat{Z}^N) \circ (\tau^N + V^N_{\leftarrow}), \hat{I}^N_\Delta + \hat{Y}^N_\Delta - \hat{Z}^N) \\ &\quad\le d^T_{L^1}((\hat{I}^N_\Delta + \hat{Y}^N_\Delta - \hat{Z}^N) \circ (\tau^N + V^N_{\leftarrow}), 0) + d^T_{L^1}(\hat{I}^N_\Delta + \hat{Y}^N_\Delta - \hat{Z}^N, 0) \\ &\quad\le 2T \|\hat{I}^N_\Delta + \hat{Y}^N_\Delta - \hat{Z}^N\|_{T + V^N_{\leftarrow}(T)}, \end{aligned}$$

and, thus, invoking Lemmas 3.4 and 3.5, as well as Proposition 3.2, yields, as $N \to \infty$,

$$d^T_{L^1}((\hat{I}^N_\Delta + \hat{Y}^N_\Delta - \hat{Z}^N) \circ (\tau^N + V^N_{\leftarrow}), \hat{I}^N_\Delta + \hat{Y}^N_\Delta - \hat{Z}^N) \;\Rightarrow\; 0. \tag{71}$$

Next, for any $\epsilon > 0$ and $\delta > 0$, conditioning on the value of $\|\tau^N + V^N_{\leftarrow} - e\|_T$ results in

$$\begin{aligned} &\mathbb{P}[d^T_{L^1}((\hat{X}^N + \hat{I}^N) \circ (\tau^N + V^N_{\leftarrow}), \hat{X}^N + \hat{I}^N) > \epsilon] \\ &\quad\le \mathbb{P}[d^T_{L^1}((\hat{X}^N + \hat{I}^N) \circ (\tau^N + V^N_{\leftarrow}), \hat{X}^N + \hat{I}^N) > \epsilon, \|\tau^N + V^N_{\leftarrow} - e\|_T \le \delta] + \mathbb{P}[\|\tau^N + V^N_{\leftarrow} - e\|_T > \delta] \\ &\quad\le \mathbb{P}\left[\|\sup_{|s| \le \delta} |\hat{X}^N(t+s) - \hat{X}^N(t) + \hat{I}^N(t+s) - \hat{I}^N(t)|\|_T > \epsilon/T\right] + \mathbb{P}[\|\tau^N + V^N_{\leftarrow} - e\|_T > \delta]. \end{aligned} \tag{72}$$

Lemmas 3.1 and 3.2, the continuous mapping theorem, the continuity of the sup operator, and the continuity of sample paths of $\hat{X}$ and $\hat{I}$ yield

$$\lim_{\delta \downarrow 0} \lim_{N \to \infty} \mathbb{P}\left[\|\sup_{|s| \le \delta} |\hat{X}^N(t+s) - \hat{X}^N(t) + \hat{I}^N(t+s) - \hat{I}^N(t)|\|_T > \epsilon/T\right] = 0.$$

The preceding limit, Proposition 3.2 and (72) imply, as $N \to \infty$,

$$d^T_{L^1}((\hat{X}^N + \hat{I}^N) \circ (\tau^N + V^N_{\leftarrow}), \hat{X}^N + \hat{I}^N) \;\Rightarrow\; 0. \tag{73}$$

Now, considering separately $s \in [0, t \wedge (\tau^N(t) + V^N_{\leftarrow}(t))]$ and $s \in (t \wedge (\tau^N(t) + V^N_{\leftarrow}(t)), t \vee (\tau^N(t) + V^N_{\leftarrow}(t))]$ we have

$$\begin{aligned}
&\int_0^T \left| \int_0^{\tau^N(t)+V^N_{\leftarrow}(t)} \bar{F}(\tau^N(t) + V^N_{\leftarrow}(t) - s)\sqrt{N}G^N(V^N_{\leftarrow}(s))\, d\check{A}^N(s) - \int_0^t \bar{F}(t-s)\sqrt{N}G^N(V^N_{\leftarrow}(s))\, d\check{A}^N(s) \right| dt \\
&\quad \le \sqrt{N}G^N(\|V^N_{\leftarrow}\|_T) \int_0^T \int_0^{t\wedge(\tau^N(t)+V^N_{\leftarrow}(t))} (F(t + \|V^N_{\leftarrow}\|_T - s) - F(t - \|\tau^N - e\|_T - s))\, d\check{A}^N(s)\, dt \\
&\qquad + \sqrt{N}G^N(\|V^N_{\leftarrow}\|_T) \int_0^T (\check{A}^N(t + \|V^N_{\leftarrow}\|_T) - \check{A}^N(t - \|\tau^N - e\|_T))\, dt \\
&\quad \le \sqrt{N}G^N(\|V^N_{\leftarrow}\|_T)(\|V^N_{\leftarrow}\|_T + \|\tau^N - e\|_T)\check{A}^N(T) + \sqrt{N}G^N(\|V^N_{\leftarrow}\|_T)(\check{A}^N(T + V^N_{\leftarrow}(T)) - \check{A}^N(T)), \qquad (74)
\end{aligned}$$

where the last inequality is due to a change in the order of integration. The preceding inequality, (6), (3), and Proposition 3.2 result in, as $N \to \infty$,

$$\sqrt{N}G^N(\|V^N_{\leftarrow}\|_T)((\|V^N_{\leftarrow}\|_T + \|\tau^N - e\|_T)\check{A}^N(T) + (\check{A}^N(T + V^N_{\leftarrow}(T)) - \check{A}^N(T))) \;\Rightarrow\; 0. \qquad (75)$$

Finally, the statement of the lemma follows from (17), (70), (71), (73), (74), (75), and Proposition 3.3. □

7.8. Proof of Proposition 4.1. The proof closely parallels the proof of Proposition 3.1 in Reed [27]. Two cases are considered separately: (i) deterministic and (ii) nondeterministic service times. The proof of measurability is the same for the two cases and is identical to the corresponding proof in Proposition 3.1 of Reed [27].

(i) Deterministic F. In this case, $F(t) = 1_{\{t \ge a\}}$, $F_*(t) = (t/a) \cdot 1_{\{0 \le t \le a\}}$, and $\mu = 1/a$.

Existence. First consider the interval $[0, a)$ only. Let $y_0 = 0$ and

$$y_{n+1}(t) = x(t) - \theta \int_0^t y_n^+(t-s)\, ds \qquad (76)$$

for $0 \le t < a$ and $n \ge 1$. Then for $\delta < a$ we have

$$\begin{aligned}
\|y_{n+1} - y_n\|_\delta &\le \delta\theta \|y_n - y_{n-1}\|_\delta \\
&\le (\delta\theta)^n \|x\|_\delta.
\end{aligned}$$

The preceding will serve as a base for an induction. Assume that

$$\|y_{n+1} - y_n\|_{k\delta} \le n^{k-1}(\delta\theta)^n \|x\|_{k\delta} \qquad (77)$$

for some k ($k\delta < a$). Then, for $(k+1)\delta < a$, the inductive assumption and (76) yield

$$\begin{aligned}
\|y_{n+1} - y_n\|_{(k+1)\delta} &\le \delta\theta \sum_{i=1}^k \|y_n - y_{n-1}\|_{i\delta} + \delta\theta \|y_n - y_{n-1}\|_{(k+1)\delta} \\
&\le (\delta\theta)^n \|x\|_{(k+1)\delta} \sum_{i=1}^k (n-1)^{i-1} + \delta\theta \|y_n - y_{n-1}\|_{(k+1)\delta} \\
&\le n^{k-1}(\delta\theta)^n \|x\|_{(k+1)\delta} + \delta\theta \|y_n - y_{n-1}\|_{(k+1)\delta}.
\end{aligned}$$

Iterating the argument from the preceding inequality results in

$$\|y_{n+1} - y_n\|_{(k+1)\delta} \le n^k (\delta\theta)^n \|x\|_{(k+1)\delta},$$

and hence (77) holds. In view of (77), selecting $\delta < 1/\theta$ implies that $\{y_n, n \ge 0\}$ is a Cauchy sequence and there exists y such that $y_n \to y$, as $n \to \infty$. Therefore, there exists a solution on the interval $[0, a)$.

Now consider the interval $[0, 2a)$. Let $y_0 = \{y_0(t) = y(t) 1_{\{0 \le t < a\}}, 0 \le t < 2a\}$ and

$$y_{n+1}(t) = \begin{cases} y(t) & 0 \le t < a, \\ x(t) + y(t-a) - \theta \displaystyle\int_{t-a}^t y_n^+(t-s)\, ds & a \le t < 2a, \end{cases}$$

where y is the solution on the interval $[0, a)$. By repeating the argument from the previous case, it is straightforward to show that there exists a solution on the interval $[0, 2a)$. Furthermore, by iterating the argument, one establishes the existence of a solution on an arbitrary interval of finite length.

Uniqueness. Let $\delta < a \wedge 1/\theta$. Suppose u and v are two solutions and consider

$$u(t) - v(t) = 1_{\{t \geq a\}}(u^+(t) - v^+(t)) - \theta \int_0^{t \wedge a} (u^+(t-s) - v^+(t-s))\, ds,$$

$t \geq 0$. For $0 \leq t \leq \delta$ we have $|u(t) - v(t)| \leq \delta\theta \|u - v\|_\delta$, and, therefore, $u(t) = v(t)$ for $0 \leq t \leq \delta$. Next $|u(t) - v(t)| \leq \delta\theta \|u - v\|_\delta + \delta\theta \|u - v\|_{2\delta}$ for $\delta < t \leq 2\delta$, yielding $u(t) = v(t)$ for $0 \leq 0 \leq 2\delta$. Repeating this argument multiple times leads to $u(t) = v(t)$ for $0 \leq t \leq a$.

Now, assume that $u(t) = v(t)$ for $0 \leq t \leq T$, where $T \geq a$. Then, for $T < t \leq T + \delta$, we have $|u(t) - v(t)| \leq \delta\theta \|u - v\|_{T+\delta}$, resulting in $u(t) = v(t)$ for $0 \leq t \leq T + \delta$. The uniqueness follows.

Lipschitz continuity. The definition of ϕ renders, for $y = \phi(x)$ and $t < a$,

$$y(t) = x(t) - \theta \int_0^t y^+(t-s)\, ds$$

and, thus, $\|\phi(x_1) - \phi(x_2)\|_\delta \leq \|x_1 - x_2\|_\delta + \delta\theta \|\phi(x_1) - \phi(y_2)\|_\delta$ if $\delta < t$. By selecting $\delta > 0$ small enough such that $\delta\theta < 1$, we have

$$\|\phi(x_1) - \phi(x_2)\|_\delta \leq \|x_1 - x_2\|_\delta / (1 - \delta\theta). \tag{78}$$

Considering the interval $[0, 2\delta]$ yields

$$\|\phi(x_1) - \phi(x_2)\|_{2\delta} \leq \|x_1 - x_2\|_{2\delta} + \delta\theta \|\phi(x_1) - \phi(x_2)\|_\delta + \delta\theta \|\phi(x_1) - \phi(x_2)\|_{2\delta},$$

which, upon combining with (78), results in

$$\|\phi(x_1) - \phi(x_2)\|_{2\delta} \leq \|x_1 - x_2\|_{2\delta} / (1 - \delta\theta)^2.$$

The preceding argument can be applied repeatedly to show that ϕ is Lipschitz continuous when the interval $[0, a)$ is considered.

For $t \geq a$, $y = \phi(x)$ renders

$$y(t) = x(t) + y^+(t-a) - \theta \int_0^a y^+(t-s)\, ds. \tag{79}$$

When $t = a$, we obtain

$$y(a) = x(a) + x^+(0) - \theta \int_0^a y^+(s)\, ds,$$

and, due to the case $t < a$, it follows that there exists $c_a < \infty$ such that $\|\phi(x_1) - \phi(x_2)\|_a \leq c_a \|x_1 - x_2\|_a$. This serves as the base for the induction. Now, suppose that for some $T \geq a$ there exists $c_T < \infty$ such that $\|\phi(x_1) - \phi(x_2)\|_T \leq c_T \|x_1 - x_2\|_T$. Now, for any $\delta < \min\{a, 1/\theta\}$, from (79) we have

$$\begin{aligned} \|\phi(x_1) - \phi(x_2)\|_{T+\delta} &\leq \|x_1 - x_2\|_{T+\delta} + (1 + a\theta)\|\phi(x_1) - \phi(x_2)\|_T + \delta\theta \|\phi(x_1) - \phi(x_2)\|_{T+\delta} \\ &\leq (1 + (1 + a\theta)c_T)\|x_1 - x_2\|_{T+\delta} + \delta\theta \|\phi(x_1) - \phi(x_2)\|_{T+\delta}, \end{aligned}$$

where the second inequality is due to the inductive assumption. Hence, $\|\phi(x_1) - \phi(x_2)\|_{T+\delta} \leq c_{T+\delta} \|x_1 - x_2\|_{T+\delta}$ with $c_{T+\delta} = (1 + (1 + a\theta)c_T)/(1 - \delta\theta) < \infty$.

(ii) Nondeterministcis F.

There exist $\delta > 0$ and $0 < \epsilon < 1$ such that

$$F(t+\delta) - F(t) + \theta F_*(t+\delta)/\mu - \theta F_*(t)/\mu < \epsilon, \tag{80}$$

for all $t \geq 0$, since F_* is absolutely continuous by definition. In view of this fact, the proof of existence, uniqueness, and Lipschitz continuity is almost identical to the proof of corresponding parts in Proposition 3.1 of Reed [27]. In particular, if $\tilde{F} := F - \theta F_*/\mu$ then

$$y(t) = x(t) + \int_0^t y^+(t-s)\, d\tilde{F}.$$

Note that the preceding relation can be written in terms of φ with F replaced by $\tilde{F}$, and , in view of (80), there exist $\delta > 0$ and $0 < \epsilon < 1$ such that

$$\tilde{F}(t+\delta) - \tilde{F}(t) < \epsilon, \tag{81}$$

for all $t \geq 0$. We can now apply directly the results in Reed [27, Proposition 3.1] because the analysis of φ in Reed [27] is based on (81). □

Acknowledgments. The authors thank an anonymous referee for careful comments that improved the presentation. Avishai Mandelbaum's research was supported in part by the Binational Science Foundation [Grant 2005175/2008480]; the Israeli Science Foundation [Grant 1357/08]; and by the Technion funds for the promotion of research and sponsored research.

References

[1] Baccelli, F., P. Bremaud. 2003. *Elements of Queueing Theory*, 2nd ed. Springer-Verlag, Berlin.
[2] Bhattacharya, P., A. Ephremides. 1991. Stochastic monotonicity properties of multiserver queues with impatient customers. *J. Appl. Probab.* **28**(3) 673–682.
[3] Billingsley, P. 1995. *Probability and Measure*, 3rd ed. Wiley, New York.
[4] Borovkov, A. A. 1967. On limit laws for service processes in multichannel systems. *Siberian Math. J.* **8**(5) 746–763.
[5] Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J. H. Dshalalow, ed. *Probability and Stochastic Series: Advances in Queueing: Theory, Methods and Open Problems*. CRC Press, Boca Raton, FL, 463–480.
[6] Chung, K. L. 1974. *A Course in Probability Theory*, 2nd ed. Academic Press, San Diego.
[7] Dai, J., S. He. 2010. Customer abandonment in many-server queues. *Math. Oper. Res.* **35**(2) 347–362.
[8] Decreusefond, L., P. Moyal. 2008. A functional central limit theorem for the M/GI/∞ queue. *Ann. Appl. Probab.* **18**(6) 2156–2178.
[9] Erlang, A. K. 1948. On the rational determination of the number of circuits. E. Brockmeyer, H. L. Halstrom, A. Jensen, eds. *The Life and Works of A. K. Erlang*. Copenhagen Telephone Company, Copenhagen, 216–221.
[10] Gamarnik, D., P. Momčilović. 2008. Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **40**(2) 548–577.
[11] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.
[12] Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.
[13] Green, R. A., P. C. Wyer, J. Giglio. 2002. ED walkout rate correlated with ED length of stay but not with ED volume or hospital census (abstract). *Acad. Emergency Medicine* **9**(5) 514.
[14] Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.
[15] Iglehart, D. L. 1973. Weak convergence of compound stochastic processes. *Stochastic Process. Appl.* **1**(1) 11–31.
[16] Jagerman, D. 1974. Some properties of the Erlang loss function. *Bell System Techn. J.* **53**(3) 525–551.
[17] Jelenković, P., A. Mandelbaum, P. Momčilović. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Syst. Theory Appl.* **47**(1–2) 53–69.
[18] Kang, W., K. Ramanan. 2010. Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* **20**(6) 2204–2260.
[19] Kaspi, H., K. Ramanan. 2011. Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* **21**(1) 33–114.
[20] Krichagina, E., A. Puhalskii. 1997. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Syst. Theory Appl.* **25**(1–4) 235–280.
[21] Mandelbaum, A., P. Momčilović. 2008. Queues with many servers: The virtual waiting-time process in the QED regime. *Math. Oper. Res.* **33**(3) 561–586.
[22] Mandelbaum, A., S. Zeltyn. 2004. The impact of customers' patience on delay and abandonment: Some empirically-driven experiments with the $M/M/N+G$ queue. *OR Spectrum* **26**(3) 377–411.
[23] Mandelbaum, A., S. Zeltyn. 2007. The $M/M/n+G$ queue: Summary of performance measures. http://iew3.technion.ac.il/serveng/References/.
[24] Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* **57**(5) 1189–1205.
[25] Puhalskii, A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* **19**(4) 946–954.
[26] Puhalskii, A., J. Reed. 2010. On many-server queues in heavy traffic. *Ann. Appl. Probab.* **20**(1) 129–195.
[27] Reed, J. 2009. The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* **19**(6) 2211–2269.
[28] Reed, J., R. Talreja. 2009. Distribution-valued heavy-traffic limits for the *G/GI/*∞ queue. Preprint, New York University, New York.
[29] Reed, J., T. Tezcan. 2009. Hazard rate scaling for the *GI/M/n + GI* queue. Preprint, New York University, New York.
[30] Talreja, R., W. Whitt. 2009. Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Probab.* **19**(6) 2137–2175.
[31] Whitt, W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York.
[32] Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.
[33] Zeltyn, S. 2005. Call centers with impatient customers: Exact analysis and many-server asymptotics of the $M/M/n+G$ queue. Ph.D. thesis, Technion, Haifa, Israel.
[34] Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Syst. Theory Appl.* **51**(3–4) 361–402.