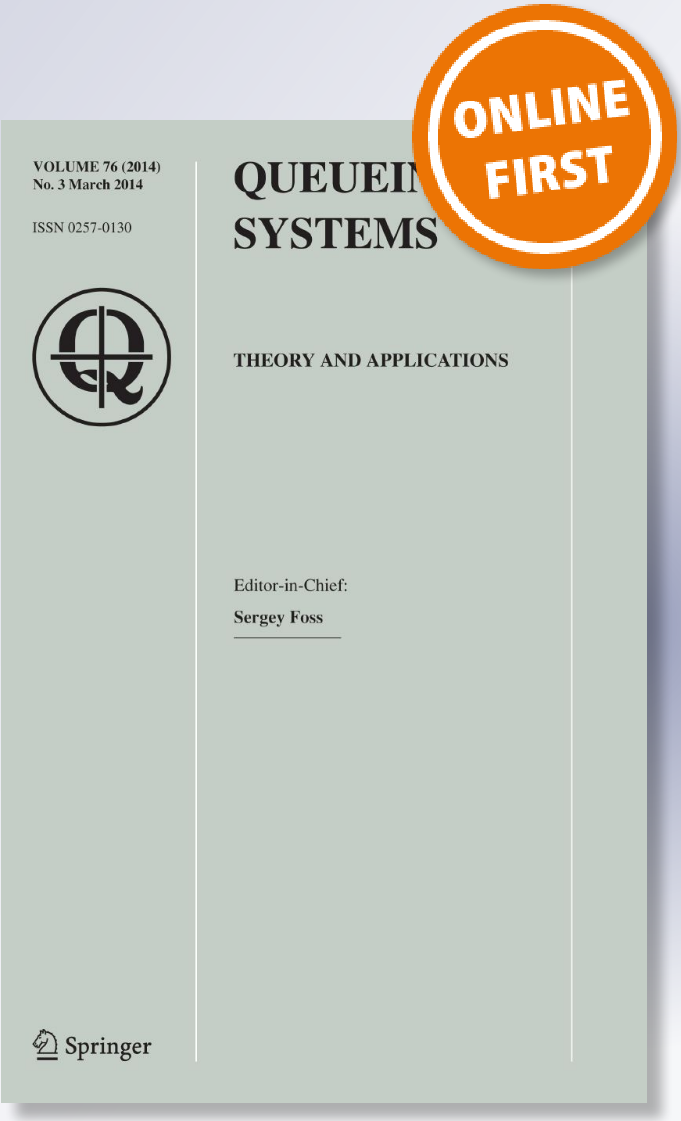
# *Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment*

**A. Korhan Aras, Xinyun Chen & Yunan Liu**

**Queueing Systems**
Theory and Applications

ISSN 0257-0130

Queueing Syst
DOI 10.1007/s11134-018-9575-0



Springer

Queueing Syst
https://doi.org/10.1007/s11134-018-9575-0

# Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment

**A. Korhan Aras**[1] · **Xinyun Chen**[2] · **Yunan Liu**[3]

Received: 2 August 2017 / Revised: 30 November 2017

**Abstract** Extending Ward Whitt's pioneering work "Fluid Models for Multiserver Queues with Abandonments, Operations Research, 54(1) 37–54, 2006," this paper establishes a many-server heavy-traffic functional central limit theorem for the overloaded $G/GI/n + GI$ queue with stationary arrivals, nonexponential service times, $n$ identical servers, and nonexponential patience times. Process-level convergence to non-Markovian Gaussian limits is established as the number of servers goes to infinity for key performance processes such as the waiting times, queue length, abandonment and departure processes. Analytic formulas are developed to characterize the distributions of these Gaussian limits.

**Keywords** Many-server queues · Many-server heavy-traffic limits · Nonexponential service times · Efficiency-driven regime · Customer abandonment · Gaussian approximation · Functional central limit theorem

**Mathematics Subject Classification** 60B12 · 60B05 · 60F17 · 60K25 · 60M20 · 90B22

✉ Yunan Liu
yliu48@ncsu.edu

A. Korhan Aras
akaras@ncsu.edu

Xinyun Chen
xinyun.chen@whu.edu.cn

[1] SAS Institute, Cary, NC, USA

[2] School of Economics and Management, Wuhan University, Wuhan, China

[3] Industrial and Systems Engineering Department, North Carolina State University, Raleigh, NC, USA

Published online: 28 March 2018

## 1 Introduction

Many empirical studies have revealed that service-time and abandonment-time distributions in service systems (for example, call centers and health care) are far from exponentially distributed, and yet researchers prior to 2006 had to assume Markovian structure (with exponential distributions) in order to gain mathematical tractability. In 2006, Ward Whitt [38] introduced a new framework to study many-server queues having nonexponential service times and abandonment times. Specifically, Whitt developed the fluid model, which is proven to be the *many-server heavy-traffic* (MSHT) *functional weak law of large numbers* (FWLLN) limit, for the $G/GI/n+GI$ queueing model having stationary arrivals (the $G$), *independent and identically distributed* (i.i.d.) nonexponential service times (the $GI$), $n$ servers, customer abandonment according to i.i.d. patience times following a nonexponential distribution (the $+GI$), and a *first-come first-served* (FCFS) service rule.

Since 2006, Whitt's pioneering work [38] opened a new line of research on non-Markovian queues which successfully brought more practical models within the reach of tractability. As his academic descendants, we are pleased to be able to contribute to this special issue. In this paper, we will extend results in [38] by developing an MSHT *functional central limit theorem* (FCLT) for the $G/GI/n+GI$ model operating in the *efficiency-driven* (ED) regime. We will prove that properly scaled performance functions, such as the waiting time, number in system, and queue length, converge in distribution to Gaussian processes as $n$ increases.

**MSHT literature on ED models** There is a large body of literature on MSHT limits for queueing models. We hereby only review the related work on the ED regime, or equivalently, the *overloaded* case. A fluid model for the $G/GI/n+GI$ queue is developed by Whitt [38] using two-parameter performance functions keeping track of elapsed service and waiting time of customers; in addition, an FWLLN is established in [38] in a discrete-time framework. This fluid model has subsequently been extended to incorporate time-varying arrivals and staffing levels in [23] and network fluid models [22,25]. FWLLN results for the $G/GI/n+GI$ queue have been obtained in [15,24,39].

We next review FCLT results for queueing systems in the ED regime. Whitt [37] showed that the queue-length process of the Markovian $M/M/s+M$ queue has an Ornstein–Uhlenbeck (OU) FCLT limit. Mandelbaum et al. [28] developed the FWLLN and FCLT limits for queueing networks having Markovian probability structure. Dai et al. [8] established a multidimensional diffusion FCLT limit for the $GI/Ph/n+M$ queue with exponential abandonment times and phase-type service times (the $Ph$). A stochastic partial differential equation (SPDE) limit was established in [17] for general many-server queueing models using measure-valued processes. Liu and Whitt [26] studied the time-varying $G_t/M/s_t+GI$ system alternating between *underloaded* and overloaded time intervals. They have obtained a stochastic differential equation driven by independent Brownian motions for the waiting time process; they also established limits for the number in system, the queue length, the virtual waiting time, and the number of abandonments. In a recent paper by Huang et al. [14], the authors developed an FCLT limit for the ED $G/M/n+GI$ model under the hazard rate scaling and applied their FCLT results in the context of delay announcements. It is evident that the FCLT limits for the performance processes are bound to become non-Markovian for fully non-Markovian queueing systems (having especially nonexponential service times), under the standard FCLT scaling. See [13,27] for performance approximation formulas for the overloaded $G/GI/n + GI$ queue. In particular, He [13] obtained an OU process FCLT limit for the queue-length process, with the mean patience time going to infinity. This mean-patience-time scaling has been proposed and studied by [37] for the Markovian $M/M/n + M$ model (see Sect. 4 therein). In [13], the customers' individual behavior disappears in the limit (for example, patience time approaches infinity, service time distribution function beyond its first two moments no longer plays a role). In addition, the approximation formulas based on the new patience-time-scaled FCLT limits in [13] may become ineffective for systems having small or medium mean patience times.

In this paper, we develop an FCLT for the $G/GI/n + GI$ queue under the conventional scaling. Specifically, we only apply spatial scaling (no temporal scaling); we scale the queue length, but we do not scale the waiting times (nor the distribution functions of service and patience times) so that customer behavior (characterized by their distribution functions) can be fully preserved in the limit; the full distribution of service (patience) time plays a role in the MSHT FCLT limit beyond its first and second moments. Compared to [13], our FCLT limits may provide performance formulas for models that are more customized to the customer behavior and those with a wider range of model inputs (especially when the mean abandonment time is not too large). However, the trade-off here is that our FCLT limits is more complex than those of [13]. (Our limits are not Markovian.)

**Main difficulty of $GI$ service** For models with exponential service times as in [14, 26], the service-completion process can be formulated as a nonhomogeneous Poisson process (NHPP) which nicely converges to a time-changed Brownian motion. This helps develop convenient FCLT limits for other performance functions. For instance, the FCLT limit for the waiting-time process solves a simple Brownian driven *stochastic differential equation* (SDE) with a linear drift:

$$d\widehat{W}(t) = h(t)\widehat{W}(t)\mathrm{d}t + I_\lambda(t)\mathrm{d}\mathcal{B}_\lambda(t) + I_a(t)\mathrm{d}\mathcal{B}_a(t) + I_s(t)\mathrm{d}\mathcal{B}_s(t), \tag{1.1}$$

where $\mathcal{B}_\lambda$, $\mathcal{B}_a$ and $\mathcal{B}_s$ are three independent Brownian motions corresponding to the FCLT limits of the arrivals, abandonments and service completions, and $h$, $I_\lambda$, $I_a$ and $I_s$ are determinisitic functions of the model inputs. See (4.9) and (6.64) in [26] for details.

For $GI$ service, the main difficulty is that the service-completion process is no longer an NHPP so it does not converge to a convenient Brownian limit. In this paper, we show that the service-completion process converges to a non-Brownian zero-mean Gaussian process with a known covariance function. Unlike (1.1), we will obtain an SDE for the waiting-time process driven by both Brownian motions and a Gaussian process.

We prove an FCLT for key performance functions of the $G/GI/n + GI$ queueing model; we identify the FCLT limits and fully describe their distributions. To characterize the limiting processes we construct a stochastic integral with respect to centered

Gaussian processes with almost surely Hölder continuous sample paths where the integrand is a two-parameter *deterministic* function. We show that such stochastic integrals can be defined pathwise, and they satisfy an integration-by-parts formula. Integrals with respect to non-Brownian Gaussian processes have been studied by [5] (fractional Brownian motions) and [1,19] (Volterra processes).

Our proofs are based on the careful analysis of the number of customers entering service from the queue. Unlike [14,26,31], we introduce a new representation for the enter-service process, based on which we derive an SDE, indexed by $n$, for the prelimit waiting-time process, and we prove its convergence to a limiting SDE. The main steps of our proof involves a martingale FCLT, Gronwall's inequality, and the continuous mapping theorem. Unlike [14,26], we do not take the commonly used compactness approach (see, for example, [36]) to prove weak convergence. An advantage of our new proof is that we can avoid having to prove tightness, which is often quite complicated (for example, see [14,26] for the complex treatment of tightness, even for $M$ service). In particular, using the $n$-indexed SDE, we prove stochastic boundedness of the waiting time and then prove the weak convergence by applying Gronwall's inequality. We further characterize the FCLT limits by computing the covariance function of the Gaussian solution to the limiting SDE. Convergence of other processes is established by applying the continuous mapping theorem. In addition, the martingale FCLT in this paper is different from those in [9,14].

**Organization of the paper** In Sect. 2 we describe a sequence of the $G/GI/n + GI$ queueing systems and specify all model assumptions. In Sect. 3 we give some preliminary results that are building blocks of the main results. In Sect. 4, we present our main results: We first give our FCLT limits in Sect. 4.1; we next characterize the distributions of the FCLT limits in Sects. 4.2 and 4.3. Proofs of the main results are given in Sect. 5. Additional proofs appear in Appendix A. Generalizing the main results in Sect. 4, we consider a more general staffing function in Appendix B. Additional results dealing with positive initial queue content appear in a longer online appendix [2].

## 2 A sequence of overloaded $G/GI/n + GI$ queues

We consider a sequence of $G/GI/n + GI$ queueing systems, which is indexed by the number of servers $n$, having i.i.d. nonexponential service times with *cumulative distribution function* (cdf) $G$, *complementary cdf* (ccdf) $G^c = 1-G$, *probability density function* (pdf) $g$ and mean service time $1/\mu < \infty$, and non-exponential patience times (the $+GI$) with cdf $F$, ccdf $F^c = 1 - F$, pdf $f$, and hazard rate $h_F = f/F^c$. All random variables and processes are defined on a common probability space. We next define relevant system functions and give assumptions on our model primitives. These assumptions will be enforced throughout the paper.

*Service times and patience times* We impose the following regularity conditions:

(i) The patience-time pdf $f$ satisfies

$$0 < f(x) \le f^{\uparrow} \equiv \sup_{x \ge 0} f(x) < \infty, \qquad x \ge 0.$$

(ii) The service-time cdf $G$ and pdf $g$ satisfy

$$\limsup_{t \downarrow 0} \frac{G(t) - G(0)}{t} < \infty, \qquad \text{and} \qquad g^{\uparrow} \equiv \sup_{x \ge 0} g(x) < \infty. \tag{2.1}$$

*Remark 2.1* (Necessity of assumptions on service and patience distributions) The first condition in (2.1) is necessary to obtain Gaussian limits for service-completion processes; see Theorem 5 in [35], see also [11,36]. Existence of densities $g$ and $f$ are required to obtain appropriate fluid limits (see [38]); and their finiteness is required to facilitate our proofs, specifically in Proposition 4.2 and (5.14).

*Arrival process* Let $N_n(t)$ be the number of customer arrivals in the interval $[0, t]$. We assume $N_n$ satisfies an FCLT

$$\widehat{N}_n(t) \equiv n^{-1/2}(N_n(t) - n\Lambda(t)) \Rightarrow \widehat{N}(t) = c_\lambda \mathcal{B}_\lambda(\Lambda(t)) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{2.2}$$

where $\mathcal{B}_\lambda$ is a standard Brownian motion (BM), $\Lambda(t) = \lambda t$, $\lambda$ and $c_\lambda > 0$ measure the (average) arrival rate and stochastic variability of the arrival process $N_n$ asymptotically. One way to construct an $N_n$ satisfying (2.2) is to simply apply a time change with function $n\Lambda(t)$ to a rate-1 renewal process with interrenewal times having variance $c_\lambda^2$ (see [12,20] for examples). A benchmark case is the Poisson arrival with $c_\lambda = 1$. Here the notation "⇒" denotes *weak convergence* (i.e., convergence in distribution). We denote by $\mathcal{D} \equiv \mathcal{D}([0, T]; \mathbb{R})$ the space of real-valued right-continuous functions with left limits on the interval $[0, T]$, and by $\mathcal{C} \equiv \mathcal{C}([0, T]; \mathbb{R})$ the subset of $\mathcal{D}$ consisting of continuous functions. Convergence in $\mathcal{D}$ is characterized through the Skorokhod $J_1$-topology; $J_1$-convergence to a continuous limit implies uniform convergence over all compact sets. See [6,36] for details of weak convergence in $\mathcal{D}$ and $\mathcal{C}$.

We remark that our analysis allows the FCLT limit $\widehat{N}$ to be a more general process having a continuous sample-path and independent increments (so $\widehat{N}$ is not restricted to a Brownian motion). An immediate corollary of the FCLT (2.2) is an FWLLN. In particular,

$$\bar{N}_n(t) \equiv n^{-1} N_n(t) \Rightarrow \Lambda(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.$$

Since our model operates in the ED regime, we assume the traffic intensity $\rho \equiv \lambda/\mu > 1$.

*System functions* Let $E_n(t)$, $D_n(t)$ and $A_n(t)$ be the total number of customers who have entered service, completed service, and abandoned from the queue in $[0, t]$, respectively. Let the two-parameter process $B_n(t, y)$ ($Q_n(t, y)$) denote the number of customers in service (in queue) at time $t$ for at most $y$ units of time in the $n$th system. In addition, let $B_n(t) \equiv B_n(t, \infty)$, $Q_n(t) \equiv Q_n(t, \infty)$, and $X_n(t) = B_n(t) + Q_n(t)$ be the number of customers in service, number waiting in the queue, and total number in the system at time $t$. Let $W_n(t)$ denote the *head-of-line waiting time* (HWT), i.e., the elapsed waiting time of the customer at the head of line at time $t$, i.e., the waiting time of the customer who has been waiting the longest (if there is any); $W_n(t) = 0$ if there is no customer waiting in the queue. Finally, we let $V_n(t)$ be the *potential waiting time* (PWT) at time $t$, i.e., the waiting time of an arriving customer at $t$ assuming the customer has infinite patience. When the system is overloaded, $W_n(t)$ and $V_n(t)$ informally satisfy the relations

$$V_n(t - W_n(t)) = W_n(t) + O(1/n) \tag{2.3}$$

$$V_n(t) = W_n(t + V_n(t) + O(1/n)) + O(1/n), \tag{2.4}$$

where (2.3) suggests that the virtual waiting time at the time of arrival of the head-of-line customer at time $t$ is the head-of-line customer's elapsed waiting time in line at time $t$ plus the additional time until one of the $n$ busy servers becomes idle. (The additional time is $O(1/n)$ if there are $n$ busy servers.) The equality in (2.4) is obtained by a change of variable.

*Initial content* At time 0, we assume the system is initially critically loaded, that is, $Q_n(0) = W_n(0) = 0$ and $X_n(0) = B_n(0) = n$ for all $n \geq 1$. Letting $\nu$ be a generic service time, we assume that customers initially in service at time 0 have i.i.d. remaining service times $\nu_1^{(0)}, \ldots, \nu_n^{(0)}$ following cdf $G_e$, the equilibrium version of $G$, given by

$$G_e(x) = \frac{\int_0^x \bar{G}(s)\,\mathrm{d}s}{\mathbb{E}[\nu]}, \qquad x \geq 0. \tag{2.5}$$

The above assumption has been commonly used in the literature [11,13,17,29,33]. Because the system is asymptotically overloaded for all $t \geq 0$, the service-completion process associated with each server forms an independent equilibrium renewal process. The assumption is not too restrictive because we plan to later focus on characterizing the long-run behavior on which initial conditions have little impact.

*MSHT scalings* For $E_n$, $D_n$, $A_n$, $B_n$, $Q_n$ and $X_n$, we define their corresponding LLN-scaled versions

$$\bar{E}_n \equiv \frac{E_n}{n}, \quad \bar{D}_n \equiv \frac{D_n}{n}, \quad \bar{A}_n \equiv \frac{A_n}{n}, \quad \bar{B}_n \equiv \frac{B_n}{n}, \quad \bar{Q}_n \equiv \frac{Q_n}{n}, \quad \bar{X}_n \equiv \frac{X_n}{n}, \tag{2.6}$$

and their CLT-scaled versions

$$\widehat{E}_n \equiv \frac{E_n - nE}{\sqrt{n}}, \qquad \widehat{D}_n \equiv \frac{D_n - nD}{\sqrt{n}}, \qquad \widehat{A}_n \equiv \frac{A_n - nA}{\sqrt{n}},$$

$$\widehat{B}_n \equiv \frac{B_n - nB}{\sqrt{n}}, \qquad \widehat{Q}_n \equiv \frac{Q_n - nQ}{\sqrt{n}}, \qquad \widehat{X}_n \equiv \frac{X_n - nX}{\sqrt{n}}. \tag{2.7}$$

For PWT $V_n$ and HWT $W_n$, we define their CLT-scaled version as

$$\widehat{W}_n \equiv \sqrt{n}\,(W_n - w) \qquad \text{and} \qquad \widehat{V}_n \equiv \sqrt{n}\,(V_n - v). \tag{2.8}$$

The centering terms $E$, $D$, $A$, $B$, $Q$, $X$, $w$ and $v$ are the fluid limits, given in Sect. 4.

## 3 Preliminaries

We now present some preliminary results which are the building blocks of our analysis. In Sect. 3.1, we first provide convenient representations for prelimit processes. Next, in Sect. 3.2 we define a class of stochastic integrals with respect to Gaussian processes which will be used to analyze our FCLT limits.

### 3.1 Prelimit processes

First define the LLN- and CLT-scaled sequential empirical processes

$$\begin{aligned}
\bar{U}_n(t,x) &\equiv \frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor} \mathbf{1}(\xi_i \le x), \qquad t \ge 0,\ 0 \le x \le 1,\\
\widehat{U}_n(t,x) &\equiv \sqrt{n}\left(\bar{U}_n(t,x) - \mathbb{E}\left[\bar{U}_n(t,x)\right]\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor} \left(\mathbf{1}(\xi_i \le x) - x\right),
\end{aligned} \tag{3.1}$$

where $\xi_1, \xi_2, \ldots$ are i.i.d. random variables uniformly distributed on [0, 1]. It has been shown in [18] that $\widehat{U}_n \Rightarrow \widehat{U}$ in $\mathcal{D}_{\mathcal{D}} \equiv \mathcal{D}([0,\infty); \mathcal{D}([0,1]; \mathbb{R}))$, as $n \to \infty$, where the two-parameter process $\widehat{U}$ is the standard Kiefer process. See [18,31] and references therein for more details.

***Enter-service process*** Based on the sequential empirical process in (3.1), we give a stochastic integral representation for $E_n$, the number of customers entering service in the interval $[0, t]$. Let random variables $0 \le \tau_1^n \le \tau_2^n \le \cdots$ denote the customers' arrival times, and $\gamma_1, \gamma_2, \ldots$ denote the i.i.d. patience times with cdf $F$. Then, the enter-service process is given by

$$\begin{aligned}
E_n(t) &\equiv \sum_{i=1}^{N_n(t-W_n(t))} \mathbf{1}(\gamma_i > V_n(\tau_i^n -))\\
&= n\int_0^{t-W_n(t)}\int_0^1 \mathbf{1}(y > F(V_n(s-)))\, d\bar{U}_n(\bar{N}_n(s), y).
\end{aligned} \tag{3.2}$$

Our new representation in (3.2) is more convenient than those in [26]; it helps simplify the proofs (see Sect. 5 for details). To see why Equation (3.2) holds: First, by the definition of $W_n(t)$, all arrivals before time $t - W_n(t)$ have already entered service provided that they do not abandon; Next, the condition $\gamma_i > V_n(\tau_i^n -)$ guarantees that

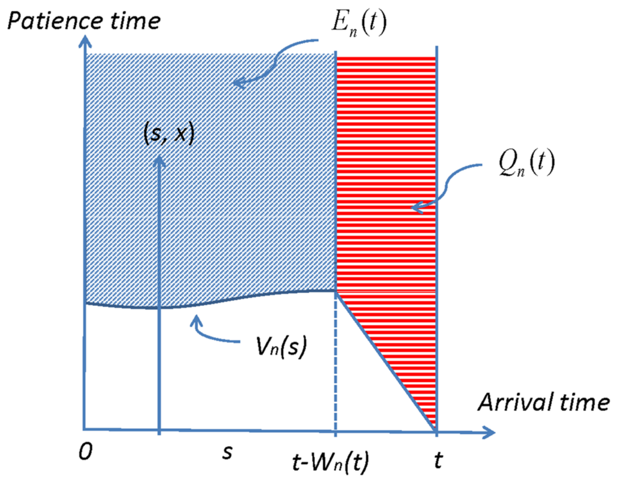


**Fig. 1** Graphic demonstration of $E_n(t)$

customer $i$ arriving time $\tau_i^n$ will not abandon (because its patience time $\gamma_i$ is bigger than its offered waiting time $w_i^n = V_n(\tau_i^n-)$). Here $E_n(t)$ is the random measure counting the number of points in the top left shaded area in Fig. 1. Following [18,31], we obtain the equivalent stochastic integral representation in (3.2).

The process $E_n(t)$ in (3.2) can be decomposed into the sum of three terms, which can be quickly verified by the definitions of $\bar{U}_n$ and $\widehat{U}_n$ (also see [31]). Specifically,

$$E_n(t) \equiv E_{n,1}(t) + E_{n,2}(t) + E_{n,3}(t), \tag{3.3}$$

where

$$E_{n,1}(t) \equiv \sqrt{n} \int_0^{t-W_n(t)} F^c(V_n(s-))\, \mathrm{d}\widehat{N}_n(s), \tag{3.4}$$

$$E_{n,2}(t) \equiv \sqrt{n} \int_0^{t-W_n(t)} \int_0^1 \mathbf{1}(y > F(V_n(s-)))\, \mathrm{d}\widehat{U}_n(\bar{N}_n(s), y), \tag{3.5}$$

$$E_{n,3}(t) \equiv n\lambda \int_0^{t-W_n(t)} F^c(V_n(s-))\, \mathrm{d}s. \qquad t \geq 0. \tag{3.6}$$

We remark that the decomposition (3.3) nicely separates the variability of the $n$th system: Given the waiting times $V_n$ and $W_n$, (3.4) captures the variability in the arrival process through $\widehat{N}_n$; (3.5) includes the variability in the abandonment times through $\widehat{U}_n$; and (3.6) represent the average value of $E_n$. In addition, the "$-$" in $V_n(\cdot)$ will disappear as $n \to \infty$, because both the FWLLN limit $v(s)$ and FCLT limit $\widehat{V}(s)$ are continuous in time $s$.

***Queue-length process*** Similar to $E_n$, the number of customer waiting in line at time $t$ can be represented as

$$\begin{aligned} Q_n(t) &= \sum_{i=N_n(t-W_n(t))+1}^{N_n(t)} \mathbf{1}(\gamma_i + \tau_i^n > t) \\ &= n \int_{t-W_n(t)}^{t} \int_0^1 \mathbf{1}(y > F(t-s))\, d\bar{U}_n(\bar{N}_n(s), y). \end{aligned} \tag{3.7}$$

To wit, a customer $i$ is waiting in queue at $t$ if it arrives after time $t - W_n(t)$ and its patience time $\gamma_i > t - \tau_i^n$. See the shaded area on the right in Fig. 1. Similar to (3.2), (3.7) can be represented as the sum of three terms, i.e,

$$Q_n(t) \equiv \; Q_{n,1}(t) + Q_{n,2}(t) + Q_{n,3}(t), \qquad t \geq 0,$$

$$\text{where} \quad Q_{n,1}(t) \equiv \sqrt{n} \int_{t-W_n(t)}^{t} F^c(t-s)\, \mathrm{d}\widehat{N}_n(s), \tag{3.8}$$

$$Q_{n,2}(t) \equiv \sqrt{n} \int_{t-W_n(t)}^{t} \int_0^1 \mathbf{1}(x > F(t-s))\, \mathrm{d}\widehat{U}_n(\bar{N}_n(s), x), \tag{3.9}$$

$$Q_{n,3}(t) \equiv n\lambda \int_{t-W_n(t)}^{t} F^c(t-s)\, \mathrm{d}s, \qquad t \geq 0. \tag{3.10}$$

### 3.2 Gaussian integrals

**Definition 3.1** (*Hölder continuity*) A real-valued function $\phi$ defined on $[a, b]$ is Hölder continuous of order $0 < \alpha < 1$ if there is a constant $c$ such that

$$|\phi(s) - \phi(t)| \leq c|s-t|^{\alpha} \quad \text{for all} \quad a < s < t < b.$$

Let $Z(\omega, t)$ be a Gaussian process with Hölder continuous sample paths for almost all $\omega \in \Omega$, zero mean, and covariance function $C_Z(s, t) \equiv \mathrm{Cov}(Z(s), Z(t))$. We next consider the stochastic integral

$$L(\omega, t) \equiv \int_0^t J(t, u)\, \mathrm{d}Z(\omega, u), \quad t \geq 0, \tag{3.11}$$

where $J(t, u)$ is a two-parameter function which is differentiable with respect to $u$. We next illustrate that (3.11) can be defined via an integration-by-parts formula for a Riemann–Stieltjes integral, following (2.3) of Chapter IV in [34].

**Proposition 3.1** (A Gaussian integral) *Suppose* $(Z(\omega, t) : t \geq 0)$ *is a centered Gaussian process on the interval* $[0, T]$ *with almost surely Hölder continuous sample paths, and, for each fixed* $t$*, the deterministic integrand* $J(t, u)$ *is continuously differentiable with respect to* $u$*. Then the stochastic integral* $L(\omega, t)$ *in* (3.11) *is well defined and*

$$L(\omega, t) = \int_0^t J(t, u)\, \mathrm{d}Z(\omega, u) = J(t, t)Z(\omega, t) - J(t, 0)Z(\omega, 0)$$

$$- \int_0^t Z(\omega, u) J_u(t, u)\, \mathrm{d}u, \tag{3.12}$$

*where $J_u(t, u) \equiv \partial J(t, u)/\partial u$, the equality holds almost surely, and the integral on the right-hand side is understood as the Riemann–Stieltjes integral.*

*In addition, for $0 \le t_1 < t_2$, $L(\omega, t)$ in (3.11) is a centered Gaussian process with the covariance function*

$$\begin{aligned} C_L(t_1, t_2) &\\ = J(t_1, t_1)J(t_2, t_2)C_Z(t_1, t_2) &+ J(t_1, 0)J(t_2, 0)C_Z(0, 0) \\ &- J(t_2, t_2)J(t_1, 0)C_Z(0, t_2) - J(t_1, t_1)J(t_2, 0)C_Z(0, t_1) \\ &- \int_0^{t_1} J(t_2, t_2)J_s(t_1, s)C_Z(s, t_2)\,\mathrm{d}s + \int_0^{t_1} J(t_2, 0)J_s(t_1, s)C_Z(0, s)\,\mathrm{d}s \\ &- \int_0^{t_2} J(t_1, t_1)J_s(t_2, s)C_Z(t_1, s)\,\mathrm{d}s + \int_0^{t_2} J(t_1, 0)J_s(t_2, s)C_Z(0, s)\,\mathrm{d}s \\ &+ \int_0^{t_1}\int_0^{t_2} J_s(t_1, s)J_r(t_2, r)C_Z(s, r)\,\mathrm{d}s\mathrm{d}r. \end{aligned} \tag{3.13}$$

The proof of Proposition 3.1 is given in Appendix A.

## 4 Main results

We present our FCLT results for the overloaded $G/GI/n + GI$ model in Sect. 4.1; we establish the process-level convergence of the CLT-scaled system functions. In Sect. 4.2, we further characterize the distributions of the Gaussian FCLT limits and give steady-state performance approximation formulas.

### 4.1 An FCLT for the $G/GI/n + GI$ queue and Gaussian limits

We first give an FWLLN for the overloaded $G/GI/n + GI$ model; we show that the LLN-scaled processes in (2.6) converge to their fluid limits.

**Theorem 4.1** (FWLLN for the overloaded $G/GI/n + GI$ model) *If all assumptions in Sect. 2 hold, then, as $n \to \infty$,*

$$(\bar{W}_n, \bar{V}_n, \bar{D}_n, \bar{E}_n, \bar{B}_n, \bar{Q}_n, \bar{X}_n, \bar{N}_n, \bar{A}_n) \Rightarrow (w, v, D, E, B, Q, X, \Lambda, A) \quad \text{in} \quad \mathcal{D}^9, \tag{4.1}$$

*where $B(t) = 1$, $\Lambda(t) = \lambda t$, $w$ and $v$ satisfy*

$$w(t) = \int_0^t \left(1 - \frac{\mu}{\lambda F^c(w(u))}\right) \mathrm{d}u, \tag{4.2}$$

$$v(t) = w(t + v(t)) \quad \text{and} \quad w(t) = v(t - w(t)), \tag{4.3}$$

*and $D$, $E$, $Q$, $X$ and $A$ are given by*

$$D(t) = E(t) = \mu t, \quad Q(t) = \lambda \int_{t-w(t)}^{t} F^c(t-s)\,\mathrm{d}s, \quad X(t) = Q(t) + 1, \tag{4.4}$$

*and $A(t) = \Lambda(t) - E(t) - Q(t)$.*

The limiting fluid functions here are special cases of those of the more general $G_t/GI/s_t + GI$ model in [23]. We give the proof of Theorem 4.1 in Appendix A; the proof follows from the proofs in [24,26].

Next we establish an FCLT result showing that the CLT-scaled system functions in (2.7) and (2.8) converge to their corresponding Gaussian FCLT limits.

**Theorem 4.2** (FCLT for the overloaded $G/GI/n + GI$ model) *If all assumptions in Sect. 2 hold, as $n \to \infty$,*

$$(\widehat{W}_n, \widehat{V}_n, \widehat{D}_n, \widehat{E}_n, \widehat{B}_n, \widehat{Q}_n, \widehat{X}_n, \widehat{N}_n, \widehat{A}_n) \Rightarrow (\widehat{W}, \widehat{V}, \widehat{E}, \widehat{E}, \widehat{B}, \widehat{Q}, \widehat{Q}, \widehat{N}, \widehat{A}) \text{ in } \mathcal{D}^9, \tag{4.5}$$

*where $\widehat{B}(t) = 0$, and $\widehat{A}(t) = \widehat{N}(t) - \widehat{Q}(t) - \widehat{E}(t)$.*

*The FCLT limit for the enter-service process $\widehat{E}(t)$ is a centered Gaussian process with covariance*

$$C_E(s,t) \equiv Cov(\widehat{E}(s), \widehat{E}(t)) = \mathbb{E}[S_0(s)S_0(t)] - \mu^2 s\, t, \quad s,t \geq 0, \tag{4.6}$$

*where $S_0$ is an equilibrium renewal process (ERP) with interrenewal cdf $G$ and the first renewal cdf $G_e$ in (2.5).*

*The FCLT limit for the HWT $\widehat{W}(t)$ uniquely solves the SDE*

$$\widehat{W}(t) = -\frac{1}{F^c(w(t))} \int_0^t f(w(s))\,\widehat{W}(s)\,\mathrm{d}s + \frac{1}{\lambda F^c(w(t))}\widehat{G}(t), \tag{4.7}$$

*where $w$ is given in (4.2), $f$ is the pdf of $F$,*

$$\begin{aligned} \widehat{G}(t) &\equiv \int_0^t F^c(w(s))\,\mathrm{d}\widehat{N}(s - w(s)) + \mathcal{B}_a\left(\lambda \int_0^t F^c(v(u))F(v(u))\,\mathrm{d}u\right) - \widehat{E}(t) \\ &= \int_0^t c_\lambda F^c(w(s))\mathrm{d}\mathcal{B}_\lambda(\Lambda(s - w(s))) \\ &\quad + \mathcal{B}_a\left(\lambda \int_0^t F^c(v(u))F(v(u))\,\mathrm{d}u\right) - \widehat{E}(t), \end{aligned} \tag{4.8}$$

*with $\mathcal{B}_a$ being a standard Brownian motion independent of the processes $\widehat{N}$ and $\widehat{E}$.*

*The FCLT limit for the PWT satisfies*

$$\widehat{V}(t) = \frac{\widehat{W}(t + v(t))}{1 - \dot{w}(t + v(t))}, \quad t \geq 0, \tag{4.9}$$

*where $\dot{w}$ is the derivative of $w$.*

*The FCLT limit for the queue-length process is the sum of three terms, i.e.,*

$$
\begin{aligned}
\widehat{Q}(t) &\equiv \widehat{Q}_1(t) + \widehat{Q}_2(t) + \widehat{Q}_3(t),\\
\widehat{Q}_1(t) &\equiv \int_{t-w(t)}^{t} F^c(t-s)\,\mathrm{d}\widehat{N}(s) = \int_{t-w(t)}^{t} c_\lambda F^c(t-s)\,\mathrm{d}\mathcal{B}_\lambda(\Lambda(s)),\\
\widehat{Q}_2(t) &\equiv \int_{t-w(t)}^{t}\int_0^1 \mathbf{1}(x > F(t-s))\,\mathrm{d}\widehat{U}(\lambda s, x)\\
&\overset{\mathrm{d}}{=} \int_{t-w(t)}^{t} \sqrt{F^c(t-s))F(t-s)}\,\mathrm{d}\mathcal{B}_a\,(\Lambda(s))\\
&\overset{\mathrm{d}}{=} \mathcal{B}_a\left(\int_{t-w(t)}^{t} F^c(t-s)F(t-s)\lambda\,\mathrm{d}s\right),\\
\widehat{Q}_3(t) &\equiv \lambda F^c(w(t))\widehat{W}(t),
\end{aligned}
\tag{4.10}
$$

*where $\widehat{U}$ is a standard Kiefer process.*

*Remark 4.1* (Separation of variability and special case of $M$ service) The process $\widehat{G}$ in (4.8) is characterized by three independent terms. The first term captures the variability of the arrival process (as a function of $\widehat{N}$); the second term accounts for the randomness of the patience times of customers waiting in queue; and the third term stems from the variability of the service process (through $\widehat{E}$). Independence of the three processes follows from mutual independence of arrivals, service times, and patience times.

For the $G/M/n+GI$ model with exponential service, we remark that the FCLT limit of the enter-server process is a Brownian motion, i.e., $\widehat{E}(t) = \mathcal{B}_s(\mu t)$, where $\mathcal{B}_s$ is an independent standard BM, because $S_0$ becomes a Poisson process with rate $\mu$. This is consistent with the SDE (1.1) and results in [26].

### 4.2 Characterizing the distributions of the FCLT limits

We now further characterize the distributions of the FCLT limits given in Theorem 4.2. We first give a Gaussian integral representation for the FCLT limit for HWT $\widehat{W}$ which is an integral of $\widehat{E}$. Because the covariance of $\widehat{E}$ is related to the covariance of an ERP (4.6), we first discuss how to compute the variance and covariance for ERPs.

**Proposition 4.1** (Covariance of an equilibrium renewal process) *Suppose $N^0$ is a stationary renewal counting process (having stationary increments with $N^0(0) = 0$) with interrenewal-time cdf $G$ having pdf $g$ and mean $\mu^{-1}$. Then, for $t < u$,*

$$
\mathrm{Cov}(N^0(t), N^0(u)) = \mathrm{Var}(N^0(t)) + \mathrm{Cov}(N^0(t), N^0(u) - N^0(t)), \tag{4.11}
$$

*where*

$$\mathrm{Var}(N^0(t)) = 2\mu \int_0^t (M(s) - \mu s + 0.5)\,\mathrm{d}s = 2\mu \int_0^t M(s)\,\mathrm{d}s - \mu^2 t^2 + \mu t, \tag{4.12}$$

$$\begin{aligned}&\mathrm{Cov}(N^0(t), N^0(u) - N^0(t))\\ &\quad = \mu \int_0^t da \int_0^{u-t} db\, g(a+b)[1 + M(t-a)][1 + M(u-t-b)] - \mu^2 t(u-t),\end{aligned} \tag{4.13}$$

*where $M(t)$ is the renewal function of the associated ordinary renewal process, satisfying the renewal equation*

$$M(t) = G(t) + \int_0^t M(t-x)g(x)\,\mathrm{d}x. \tag{4.14}$$

*Proof* The proof of (4.12) is given on p. 57 of [7]. See also Theorem 7.2.4 of [36]. (We point out that there is a mistake in the proof of Theorem 7.2.4 of [36] so the covariance formulas there are incorrect. We give the correct versions here.) For (4.13), consider the first renewal occurring after $t$; it falls at $t + b$ with the stationary-excess pdf $g_e(b) \equiv \mu G^c(b)$. Conditional on that renewal being at $t + b$, the last renewal by $t$ occurs at $t - a$ with pdf $g(a+b)/G^c(b)$. Conditional on the time of these two renewals at $t - a$ and $t + b$, we have

$$\begin{aligned}&\mathbb{E}[N^0(t)(N^0(u) - N^0(t))]\\ &\quad = \int_0^t da \int_0^{u-t} db\, \mathbb{E}\Big[N^0(t)(N^0(u) - N^0(t))|S_{N(t)}\\ &\qquad = t - a, S_{N(t)+1} = t + b\Big]\mu f(a+b)\\ &\quad = \mu \int_0^t da \int_0^{u-t} db\, [M(t-a) + 1][M(u-t-b) + 1]g(a+b).\end{aligned}$$

Because $N^0$ is an ERP, $\mathbb{E}[N^0(t)]\mathbb{E}[(N^0(u) - N^0(t))] = (\mu t) \cdot (\mu(u-t))$, which yields (4.13). □

To prove that the Gaussian integral is a well-defined stochastic integral, we justify that $\widehat{E}$ in (4.6) has almost surely Hölder continuous sample paths.

**Proposition 4.2** *The FCLT limit for the enter-service process $\{\widehat{E}(t) : t \geq 0\}$ is Hölder continuous for almost all $\omega \in \Omega$ for some $\theta > 0$.*

*Proof* Let $Y(t)$ be a centered Gaussian process with covariance function $\phi(s,t) \equiv \mathbb{E}[Y(s)Y(t)]$. According to Corollary 25.6 of [34], we know that a sufficient condition for $Y(t)$ to be continuous is that $\phi$ should be locally Hölder continuous, that is, for each $N \in \mathbb{N}$ there exists $\theta = \theta(N) > 0$ and $C = C(N)$ such that, for $|s|, |t| \leq N$,

$$|\phi(s,t)-\phi(t,t)| \le C|s-t|^{\theta}. \tag{4.15}$$

We next validate the condition (4.15) for $\widehat{E}(t)$. Combining (4.11) and (4.13), we obtain, for $u > t$,

$$\begin{aligned}
&|\phi(u,t)-\phi(t,t)| = |\mathrm{Cov}(\widehat{E}(t), \widehat{E}(u)-\widehat{E}(t))| \\
&\quad = \left|\mu \int_0^t da \int_0^{u-t} db\, g(a+b)[1+M(t-a)][1+M(u-t-b)] - \mu^2 t(u-t)\right|,
\end{aligned}$$

where $\mu$ is the service rate and $g$ is service-time pdf. Then, for any $0 \le t < u \le N$ and $a \in [0,t]$, $b \in [0, u-t]$,

$$\begin{aligned}
\phi(u,t)-\phi(t,t) &\le \mu \int_0^t da \int_0^{u-t} db g^{\uparrow}[1+M(N)]^2 + \mu^2 t(u-t) \\
&= \mu t(u-t) g^{\uparrow}[1+M(N)]^2 + \mu^2 t(u-t) \\
&\le \left(\mu N g^{\uparrow}[1+M(N)]^2 + \mu^2 N\right)(u-t),
\end{aligned}$$

which satisfies the sufficient condition (4.15) after taking the absolute value of both sides. Hence, $\widehat{E}$ has a version with continuous sample paths. Then, by Kolmogorov's continuity theorem, we deduce that the version is Hölder continuous of some order $\theta > 0$. □

Having proved that the Gaussian process $\widehat{E}$ has the desired sample-path properties, we next provide a closed-form solution $\widehat{W}$ to the SDE (4.7).

**Corollary 4.1** (Gaussian integrals for $\widehat{W}$) *The FCLT limit for the HWT is given by*

$$\begin{aligned}
\widehat{W}(t) \equiv& \widehat{W}_1(t) + \widehat{W}_2(t) + \widehat{W}_3(t) \\
=& \int_0^t c_{\lambda} \frac{F^c(w(u))H(t,u)}{q(t,w(t))}\, \mathrm{d}\mathcal{B}_{\lambda}(\Lambda(u-w(u))) \\
&+ \int_0^t \frac{\sqrt{\lambda F^c(v(u))F(v(u))}H(t,u)}{q(t,w(t))}\, \mathrm{d}\mathcal{B}_a(u) - \int_0^t \frac{H(t,u)}{q(t,w(t))}\, \mathrm{d}\widehat{E}(u),
\end{aligned} \tag{4.16}$$

*where the first term on the right-hand side is defined in Lemma 5.1, the third term is defined in Sect. 3.2; $q(t,w(t)) = \lambda F^c(w(t))$, and*

$$H(t,u) \equiv e^{\int_u^t h(r)\,\mathrm{d}r} \quad \textit{with} \quad h(r) \equiv -\frac{\lambda f(w(r))}{q(t,w(t))} = -\frac{f(w(r))}{F^c(w(t))}, \quad 0 \le r \le t. \tag{4.17}$$

*Proof* To verify (4.16) is indeed the unique strong solution to (4.7), we apply Itô's rule. Note that the non-Brownian integrals are Riemann–Stieltjes integrals with deterministic integrands for almost all $\omega \in \Omega$. Moreover, the integrators $\mathcal{B}_{\lambda}$, $\mathcal{B}_a$ and $\widehat{E}$ are all independent. Letting $q(t,w(t)) = \lambda F^c(w(t))$, we rewrite (4.7) as

$$\begin{aligned}\widehat{W}(t) = &-\int_0^t \frac{\lambda f(w(u))}{q(u, w(u))}\,\widehat{W}(u)\,\mathrm{d}u + \int_0^t \frac{\sqrt{\lambda F^c(v(u))F(v(u))}}{q(u, w(u))}\,\mathrm{d}\mathcal{B}(u)\\ &-\int_0^t \frac{1}{q(u, w(u))}\,\mathrm{d}\widehat{E}(u) + \int_0^t c_\lambda \frac{F^c(w(u))}{q(u, w(u))}\,\mathrm{d}\mathcal{B}_\lambda(\Lambda(u - w(u))), \quad (4.18)\end{aligned}$$

where the time-changed Brownian term in (4.7) is replaced with an equivalent Itô integral. The uniqueness and existence of a solution to (4.18) immediately follows from Itô theory because the last two terms of (4.18) do not involve $\widehat{W}(t)$ and are independent of the Brownian motion in the second term. Furthermore, the last two integrals are Riemann–Stieltjes integrals for almost all $\omega \in \Omega$. Consequently, we can use Itô's formula to solve (4.18). In particular, using the differential form, we have

$$\mathrm{d}\widehat{W}(t) = h(t)\widehat{W}(t)\,\mathrm{d}t + K_1(t)\mathrm{d}\mathcal{B}_a(t) + K_2(t)\mathrm{d}\widehat{E}(t) + K_3(t)\,\mathrm{d}\mathcal{B}_\lambda(t - w(t)),$$

which implies

$$d\left(e^{-\int_0^t h(r)\,\mathrm{d}r}\,\widehat{W}(t)\right) = \widetilde{K}_1(t)\mathrm{d}\mathcal{B}_a(t) + \widetilde{K}_2(t)\mathrm{d}\widehat{E}(t) + \widetilde{K}_3(t)\,\mathrm{d}\mathcal{B}_\lambda(t - w(t)),$$

where $\widetilde{K}_i(t) \equiv H(t, 0)K_i(t)$, $i = 1, 2, 3$,

$$K_1(t) \equiv \frac{\sqrt{\lambda F^c(v(u))F(v(u))}}{q(u, w(u))}, \quad K_2(t) \equiv -\frac{1}{q(u, w(u))}, \quad K_3(t) \equiv c_\lambda \frac{F^c(w(u))}{q(u, w(u))}.$$

Integrating both sides and multiplying through by $H(t, 0)$ yields (4.16). □

***Covariance formulas for the Gaussian limits*** Since all the FCLT limits are Gaussian processes, it suffices to compute their means and covariances. We next give closed-form covariance formulas for the FCLT limits $\widehat{Q}$, $\widehat{W}$ and $\widehat{X}$. Our covariance formulas are explicit functions of the covariance of $\widehat{E}$. The proof of the next theorem is given in Appendix A.

**Theorem 4.3** (Further characterization of the FCLT limits) *The FCLT limits $\widehat{W}$, $\widehat{V}$ and $\widehat{Q}$ are all centered Gaussian processes with the covariance functions*

$$\begin{aligned}\mathrm{Cov}(\widehat{W}(t), \widehat{W}(t')) &\equiv C_{\widehat{W}}(t, t') = C_{\widehat{W}_1}(t, t') + C_{\widehat{W}_2}(t, t') + C_{\widehat{W}_3}(t, t')\\ \mathrm{Cov}(\widehat{V}(t), \widehat{V}(t')) &\equiv C_{\widehat{V}}(t, t') = \frac{C_{\widehat{W}}(t, t')}{(1 - \dot{w}(t))(1 - \dot{w}(t'))},\\ \mathrm{Cov}(\widehat{Q}(t), \widehat{Q}(t')) &\equiv C_{\widehat{Q}}(t, t') = C_{\widehat{Q}_1}(t, t') + C_{\widehat{Q}_2}(t, t') + C_{\widehat{Q}_3}(t, t'),\end{aligned}$$

*for* $t,\ t' \geq 0$*, where*

$$\begin{aligned}
C_{\widehat{W}_1}(t,t') &= \lambda c_\lambda^2 \int_0^{t\wedge t'} \frac{F^c(w(u))^2 H(t,u)^2}{q(u,w(u))^2}\,(1-\dot{w}(u))\,\mathrm{d}u,\\
C_{\widehat{W}_2}(t,t') &= \lambda \int_0^{t\wedge t'} \frac{F^c(v(u))F(v(u))H(t,u)^2}{q(u,w(u))^2}\,\mathrm{d}u,\\
C_{\widehat{W}_3}(t,t') &= C_{\widehat{E}}(t,t') - \int_0^{t'} J_u(t',u)C_{\widehat{E}}(u,t)\mathrm{d}u - \int_0^{t} J_u(t,u)C_{\widehat{E}}(u,t')\mathrm{d}u\\
&\qquad + \int_0^{t}\int_0^{t'} J_u(t,u)J_u(t',v)C_{\widehat{E}}(u,v)\mathrm{d}v\mathrm{d}u,\\
C_{\widehat{Q}_1}(t,t') &= \lambda c_\lambda^2 \int_{(t-w(t))\vee(t'-w(t'))}^{t\wedge t'} F^c(t-s)F^c(t'-s)\mathrm{d}s,\\
C_{\widehat{Q}_2}(t,t') &= \lambda \int_{(t-w(t))\vee(t'-w(t'))}^{t\wedge t'} F(t\wedge t'-s)F^c(t\vee t'-s)\mathrm{d}s,\\
C_{\widehat{Q}_3}(t,t') &= \lambda^2\, F^c(w(t))\, F^c(w(t'))\, C_{\widehat{W}}(t,t'), \qquad (4.19)
\end{aligned}$$

*where* $C_{\widehat{E}}(u,v) \equiv \mathrm{Cov}(\widehat{E}(u),\widehat{E}(v))$ *is the covariance function for* $\widehat{E}$ *in* (4.6)*,* $J(t,u) \equiv H(t,u)/q(u,w(u))$*,* $H(t,u)$ *and* $q(u,w(u))$ *are as in* (4.17)*, and*

$$\begin{aligned}
J_u(t,u) \equiv & \frac{\partial J(t,u)}{\partial u} = -\frac{h(u)H(t,u)}{q(u,w(u))} - \frac{q(u,w(u))H(t,u)}{q(u,w(u))^2}\\
= & \frac{(1-\dot{w}(u))h_F(w(u))e^{-\int_u^t (1-\dot{w}(s))h_F(w(s))\,\mathrm{d}s}}{\lambda F^c(w(u))}\\
& + \frac{f(w(u))e^{-\int_u^t (1-\dot{w}(s))h_F(w(s))\,\mathrm{d}s}}{\lambda F^c(w(u))^2}. \qquad (4.20)
\end{aligned}$$

### 4.3 Steady-state distributions of the FCLT limits

We now characterize the steady-state distributions of the FCLT limits as $t \to \infty$. In particular, we show that the steady state of the FCLT limits are centered Gaussian random variables, and we compute their variances to fully characterize their distributions. Consequently, we obtain approximations for the long-run performance of the $n$th queueing system, i.e.,

$$Q_n(\infty) \approx nQ(\infty) + \sqrt{n}\widehat{Q}(\infty), \quad W_n(\infty) \approx w(\infty) + \frac{1}{\sqrt{n}}\widehat{W}(\infty), \qquad (4.21)$$

where $Q(\infty)$ and $w(\infty)$ are the steady states of the fluid functions in Theorem 4.1, $\widehat{Q}(\infty)$ and $\widehat{W}(\infty)$ are the steady states of $\widehat{Q}(t)$ and $\widehat{W}(t)$. A more rigorous argument to support (4.21) involves the validation of the interchange of the two limits "$n \to \infty$" and "$t \to \infty$". However, this is beyond the scope of this paper.

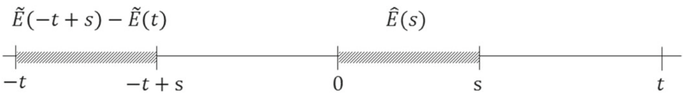


**Fig. 2** Extension of $\widehat{E}$ on $(-\infty, 0]$

First, Theorem 3.1 of [38] and Theorem 4.1 of [23] show that the overloaded fluid queueing system has the following steady state functions:

$$w \equiv w(\infty) = F^{-1}\left(1 - \frac{1}{\rho}\right), \quad v \equiv v(\infty) = w, \quad q(t, w(t)) = \lambda F^c(w) \quad \text{and}$$
$$Q \equiv Q(\infty) = \lambda \int_0^w F^c(x)\mathrm{d}x. \tag{4.22}$$

Next, we characterize the steady state of the FCLT limits in Theorem 4.2 by letting $t \to \infty$. We first obtain the variance for $\widehat{W}(\infty)$, which is necessary for $\widehat{V}(\infty)$ and $\widehat{Q}(\infty)$.

The convergence, as $t \to \infty$, of the variance function of $\widehat{W}_i(t)$ in Theorem 4.3 is straightforward for $i = 1, 2$. However, the treatment of $\widehat{W}_3(t)$ is not straightforward because the covariance formula for $\widehat{W}_3(t)$ in (4.19) involves several integrals; the negative term in (4.19) goes to $-\infty$ and the positive term goes to $\infty$ as $t \to \infty$ (because $\mathrm{Var}(\widehat{E}(t)) = C_{\widehat{E}}(t,t) \to \infty$ as $t \to \infty$).

To derive convenient steady-state formulas, we propose a technique which *extends* the Gaussian limit $\widehat{E}$ given in Theorem 4.2 to the interval $(-\infty, 0]$. Specifically, we define another Gaussian process $\widetilde{E}$ in Lemma 4.1 that can be understood as a two-sided extension of $\widehat{E}$ to the negative half line to resolve this issue (see Fig. 2). We imagine the stationary FCLT limit is associated with a queueing system starting infinitely far in the past (which is in steady state at time 0).

The proof of Lemma 4.1 is given in Appendix A.

**Lemma 4.1** (Extending $\widehat{E}$ to the negative half line) *There exists a Gaussian process* $\{\widetilde{E}(t) : -\infty < t \le 0\}$ *such that* $(i)$ $\widetilde{E}(0) = 0$*;* $(ii)$ $\mathbb{E}[\widetilde{E}(t)] = 0$ *for all* $-\infty < t \le 0$*; and* $(iii)$ *the covariance function*

$$\mathrm{Cov}\left(\widetilde{E}(-x), \widetilde{E}(-y)\right) = \widetilde{C}(-x,-y) \equiv C_{\widehat{E}}(x \vee y, x \vee y) - C_{\widehat{E}}(x \vee y, |x-y|) \tag{4.23}$$

*for* $x \ge 0$, $y \ge 0$, *with* $C_{\widehat{E}}$ *being the covariance function in* (4.6)*. In addition,* $\widetilde{E}$ *has the same stationary increment distribution as* $\widehat{E}$*. In particular, for any* $t > 0$,

$$\{\widetilde{E}(s-t) - \widetilde{E}(-t) : 0 \le s \le t\} \overset{d}{=} \{\widehat{E}(s) : 0 \le s \le t\}. \tag{4.24}$$

Using the extended version $\widetilde{E}$, we next obtain a more convenient expression for $\mathrm{Var}(\widehat{W}_3(t))$ in Theorem 4.4. The proof of Theorem 4.4 is in Appendix A.

**Theorem 4.4** (Steady state of the FCLT limits) *The steady-state versions $\widehat{W}(\infty)$, $\widehat{V}(\infty)$ and $\widehat{Q}(\infty)$ for the FCLT limits of $\widehat{W}(t)$, $\widehat{V}(t)$ and $\widehat{Q}(t)$ are Gaussian random variables with means 0 and variances*

$$\begin{aligned}\sigma_W^2 &\equiv \mathrm{Var}(\widehat{W}(\infty)) = \mathrm{Var}(\widehat{V}(\infty))\\ &\equiv \frac{c_\lambda^2}{2h_F(w)\lambda} + \frac{F(w)}{2\lambda f(w)} + 2\frac{h_F(w)^2}{\lambda^2 F^c(w)^2}\int_0^\infty \int_0^x e^{-h_F(w)(x+y)}\widetilde{C}(-x,-y)\,\mathrm{d}y\mathrm{d}x\end{aligned} \tag{4.25}$$

$$\sigma_Q^2 \equiv \mathrm{Var}(\widehat{Q}(\infty)) \equiv \lambda c_\lambda^2 \int_0^w F^c(u)^2\mathrm{d}u + \lambda\int_0^w F(u)F^c(u)\,\mathrm{d}u + \lambda^2 F^c(w)^2\sigma_W^2, \tag{4.26}$$

*where $\widetilde{C}(-x,-y)$ is as in* (4.23)*, and $w$ is given in* (4.22).

We next show that our formulas degenerate to special cases of $M$ service. The proof of Corollary 4.2 is in Appendix A.

**Corollary 4.2** (The $M$ service special cases)

(i) *For the $G/M/n+GI$ queue having exponential service times, steady-state variances of $\widehat{W}$ and $\widehat{Q}$ reduce to*

$$\begin{aligned}\sigma_W^2 &= \frac{(c_\lambda^2-1)+2\rho}{2\lambda h_F(w)} \quad \textit{and}\\ \sigma_Q^2 &= \lambda\int_0^w \left((c_\lambda^2-1)F^c(u)+1\right)F^c(u)\mathrm{d}u + \lambda^2 F^c(w)^2\frac{(c_\lambda^2-1)+2\rho}{2\lambda h_F(w)}.\end{aligned} \tag{4.27}$$

(ii) *For the fully Markovian $M/M/n+M$ queue, the steady-state variance of $\widehat{W}$ and $\widehat{Q}$ reduce to*

$$\sigma_W^2 = 1/\mu\theta \quad \textit{and} \quad \sigma_Q^2 = \lambda/\theta, \tag{4.28}$$

*where $\theta > 0$ is the abandonment rate ($1/\theta$ is the mean abandonment time).*

## 5 Proofs

The proof of Theorem 4.1 is given in Sect. A.2. To prove Theorem 4.2, we first prove the FCLT for $\widehat{W}_n$ (Sect. 5.1). Given the FCLT for $\widehat{W}_n$, we establish FCLTs for the other processes in Sect. 5.2. Proofs of all other results are given in Appendix A.

*Remark 5.1* (Extending to nonstationary arrivals) Although both Theorems 4.1 and 4.2 are stated under the assumption of stationary arrivals with $\Lambda(t) = \lambda t$, our proof can be easily generalized to the case of nonstationary arrivals with a time-varying arrival rate $\lambda(t)$, as long as the system is asymptotically overloaded. Because real service systems (such as health care) are often overloaded with time-varying arrivals, the more general FCLT results with time-varying arrivals may stimulate future research (for example, conducting transient analysis and controls).

### 5.1 FCLT for HWT

It might be possible to prove the FCLT for $\widehat{W}_n$ using the standard *compactness approach*: (i) tightness (which implies that every subsequence has a further convergent subsubsequence); and (ii) uniqueness of the limit of every convergent subsequence [13,14,26,31]. But that approach would involve a complicated treatment of tightness. For example, see [26,31] for details (the tightness proofs for the CLT-scaled processes there are somewhat tricky and lengthy). We hereby adopt a new approach: (i) we show that the prelimit $\widehat{W}_n$ satisfies an SDE indexed by $n$; (ii) using the prelimit SDE, we establish the full convergence $\widehat{W}_n \Rightarrow \widehat{W}$ using the continuous mapping theorem, martingale convergence theorem and Gronwall's inequality. We show that the limit process $\widehat{W}$ uniquely solves the SDE in (4.7), which generalizes the SDE given in (6.64) of [26]. The extension from $M$ service to $GI$ service replaces the Brownian motion $\mathcal{B}_s$ therein by a centered Gaussian process. Our new approach has two advantages: First, it is simpler (because it nicely avoids having to prove the tightness in space $\mathcal{D}$); Second, this method may be used to treat other processes and models in future research.

#### 5.1.1 Overview of the proof

The FCLT of $\widehat{W}_n$ draws heavily on the careful analysis of $\widehat{E}_n$ and its convergence as $n \to \infty$. To wit, the HWT $W_n$ ought to increase (decrease) if the flow-into-service $E_n$ is big (small); and the variability of HWT (represented by $\widehat{W}_n$) largely depends on the variability of $E_n$ (i.e., $\widehat{E}_n$). On the one hand, we will prove that $\widehat{E}_n$ converges to a Gaussian process $\widehat{E}$ by taking advantage of the structure of the superposition of many ERPs.

On the other hand, following the decomposition given in (3.3)–(3.6), we write

$$\widehat{E}_n(t) = \frac{E_{n,1}(t)}{\sqrt{n}} + \frac{E_{n,2}(t)}{\sqrt{n}} + \frac{E_{n,3}(t) - nE_3(t)}{\sqrt{n}} \equiv \widehat{E}_{n,1}(t) + \widehat{E}_{n,2}(t) + \widehat{E}_{n,3}(t). \tag{5.1}$$

We establish the convergence of $\widehat{E}_{n,1}$ and $\widehat{E}_{n,2}$ separately, and we express the third term $\widehat{E}_{n,3}$ as a function of the desired $\widehat{W}_n$. This will result in an SDE for $\widehat{W}_n$ (this is our key step). To establish the convergence of (5.1), we will show their joint convergence and apply the continuous mapping theorem with addition. We know that joint convergence of two random elements is equivalent to the individual convergence of both terms if they are independent. Although $\widehat{E}_{n,1}$, $\widehat{E}_{n,2}$ and $\widehat{E}_{n,3}$ are not independent, because they all involve the arrival-time sequence $\tau_i^n$ (or equivalently $N_n$), HWT $W_n$ and PWT $V_n$, they are conditionally independent given $\big(\bar{N}_n, W_n, V_n\big)$. Hence, in order to treat the three terms separately, we can condition upon $(\bar{N}_n, W_n, V_n)$ (which converges according to the FWLLN result) and then uncondition. See Lemma 4.1 of [3] for a reference, which is a variant of Theorem 7.6 of [32]. The proof of the next lemma is given in Sect. A.7 of Appendix A.

**Lemma 5.1** (Convergence of the first two terms in (5.1)) *As $n \to \infty$,*

$$\begin{aligned} \widehat{E}_{n,1}(t) \Rightarrow \widehat{E}_1(t) &\equiv \int_0^{t-w(t)} c_\lambda F^c(v(u))\, \mathrm{d}\mathcal{B}_\lambda(\Lambda(u)) \\ &= \int_0^t c_\lambda F^c(w(s))\, \mathrm{d}\mathcal{B}_\lambda(\Lambda(s-w(s))), \end{aligned} \tag{5.2}$$

$$\begin{aligned} \widehat{E}_{n,2}(t) \Rightarrow \widehat{E}_2(t) &\equiv \int_0^{t-w(t)} \int_0^1 \mathbf{1}(y > F(v(s)))\, \mathrm{d}\widehat{U}(\Lambda(s), y) \\ &\overset{\mathrm{d}}{=} \int_0^{t-w(t)} \sqrt{F^c(v(u))F(v(u))}\, \mathrm{d}\mathcal{B}_a\,(\Lambda(u)) \\ &\overset{\mathrm{d}}{=} \mathcal{B}_a\left(\int_0^{t-w(t)} F^c(v(u))F(v(u))\lambda\, \mathrm{d}u\right) \\ &= \mathcal{B}_a\left(\int_0^t F^c(w(s))F(w(s))\,(1-\dot{w}(s))\,\lambda\, \mathrm{d}s\right). \end{aligned} \tag{5.3}$$

### *5.1.2 Treating the third term in (5.1)*

According to the FWLLN, we have $\bar{E}_{n,i}(t) \equiv (1/n)E_{n,i}(t) \Rightarrow 0$ for $i = 1, 2$, and

$$\bar{E}_{n,3}(t) \equiv \frac{1}{n}E_{n,3}(t) \Rightarrow E(t) = E_3(t) \equiv \int_0^{t-w(t)} F^c(v(s))\, \mathrm{d}\Lambda(s) \quad \text{as} \quad n \to \infty. \tag{5.4}$$

Following (3.6) and (5.4), we have

$$\begin{aligned} &E_{n,3}(t) - nE_3(t) \\ &\quad= n\int_0^{t-W_n(t)} F^c(V_n(s-))\, \mathrm{d}\Lambda(s) - n\int_0^{t-w(t)} F^c(V_n(s-))\, \mathrm{d}\Lambda(s) \\ &\qquad+ n\int_0^{t-w(t)} [F^c(V_n(s-)) - F^c(v(s-))]\, \mathrm{d}\Lambda(s) \\ &\quad= n\int_{t-w(t)}^{t-W_n(t)} F^c(V_n(s-))\, \mathrm{d}\Lambda(s) + n\int_0^{t-w(t)} [F^c(V_n(s-)) - F^c(v(s-))]\, \mathrm{d}\Lambda(s) \\ &\quad= -\sqrt{n}F^c(\theta_{1,n}(t))\lambda\widehat{W}_n(t) - \sqrt{n}\int_0^{t-w(t)} f(\theta_{2,n}(s))\widehat{V}_n(s-)\, \mathrm{d}\Lambda(s) + o(\sqrt{n}) \\ &\quad= -\sqrt{n}F^c(\theta_{1,n}(t))\lambda\widehat{W}_n(t) - \sqrt{n}\int_0^{t-w(t)} f(\theta_{2,n}(s))\widehat{V}_n(s)\, \mathrm{d}\Lambda(s) + o(\sqrt{n}), \end{aligned} \tag{5.5}$$

where $f$ is the pdf of $F$, the last equality holds because $v(s-) = v(s)$ and $|V_n(s) - V_n(s-)| = O(1/n)$ (note there are $n$ busy servers), and $\theta_{1,n}(t)$ and $\theta_{2,n}(t)$ satisfy

$$V_n(t - W_n(t)) \wedge V_n(t - w(t)) \le \theta_{1,n}(t) \le V_n(t - W_n(t)) \vee V_n(t - w(t)), \quad (5.6)$$

$$V_n(t) \wedge v(t) \le \theta_{2,n}(t) \le V_n(t) \vee v(t). \quad (5.7)$$

From Lemma 5.1 and (5.5), we have

$$\begin{aligned} E_n(t) &= \sqrt{n}\widehat{E}_{n,1}(t) + \sqrt{n}\widehat{E}_{n,2}(t) + \big(E_{n,3}(t) - nE_3(t)\big) + nE_3(t) \\ &= \sqrt{n} \int_0^t c_\lambda F^c(w(s))\, \mathrm{d}\mathcal{B}_\lambda(\Lambda(s - w(s))) \\ &\quad + \sqrt{n}\, \mathcal{B}_a \left( \int_0^t F^c(v(u)) F(v(u))\, \mathrm{d}\Lambda(u) \right) - \sqrt{n} F^c(\theta_{1,n}(t)) \lambda \widehat{W}_n(t) \\ &\quad - \sqrt{n} \int_0^{t-w(t)} f(\theta_{2,n}(s)) \widehat{V}_n(s)\, \mathrm{d}\Lambda(s) + nE_3(t) + o(\sqrt{n}). \end{aligned} \quad (5.8)$$

***Deriving an SDE for*** $\widehat{W}_n$ We observe that the desired $\widehat{W}_n$ now appears in (5.8). To derive an SDE for $\widehat{W}_n$, it remains to relate the PWT $\widehat{V}_n$ in (5.7) to $\widehat{W}_n$. Let $\Delta V_n(t) \equiv V_n(t) - v(t)$ and $\Delta W_n(t) \equiv W_n(t) - w(t)$, where $w(t)$ and $v(t)$ are as the fluid limits given in Theorem 4.1. Using (2.4) we write

$$\begin{aligned} \Delta V_n(t) &= \Delta W_n(t + V_n(t) + O(1/n)) + w(t + V_n(t)) - w(t + v(t)) + O(1/n) \\ &= \Delta W_n(t + V_n(t) + O(1/n)) + \dot{w}(t + v(t)) \Delta V_n(t) + O(1/n), \end{aligned}$$

where the last equality holds because

$$\begin{aligned} w(t + V_n(t)) = w(t + v(t)) &+ \dot{w}(t + v(t)) \Delta V_n(t) \\ &+ \frac{1}{2} \ddot{w}(t + v(t)) \Delta V_n^2(t) + o(\Delta V_n^2(t)), \end{aligned}$$

and $\ddot{w}(t) \equiv d^2 w(t)/\mathrm{d}t^2$, which exists by (4.2) and the smoothness of $F$. As $\Delta V_n(t) = O(1/\sqrt{n})$, we have $w(t + V_n(t)) - w(t + v(t)) = \dot{w}(t + v(t)) \Delta V_n(t) + O(1/n)$. Hence, we have

$$\Delta V_n(t) = \frac{\Delta W_n(t + V_n(t) + O(1/n))}{1 - \dot{w}(t + v(t))} + o(\Delta V_n(t)) + O(1/n),$$

which implies that

$$\sup_{0 \le t \le T} \left| \widehat{V}_n(t) - \frac{\widehat{W}_n(t + v(t))}{1 - \dot{w}(t + v(t))} \right| = \sqrt{n}\, o(1/n) = o(1/\sqrt{n}). \quad (5.9)$$

Note that $o(\Delta V_n(t)) = o(1/n)$ since $\Delta V_n(t)$ is of $O(1/n)$. This provides a formula to switch between the two waiting times $\widehat{V}_n(t)$ and $\widehat{W}_n(t)$.

Applying the change-of-variable formula (5.9), the last integral in (5.8) becomes

$$\begin{aligned}
&\sqrt{n}\int_0^{t-w(t)} f(\theta_{2,n}(s))\widehat{V}_n(s)\,\mathrm{d}\Lambda(s)\\
&\quad= \sqrt{n}\int_0^{t-w(t)} f(\theta_{2,n}(s))\left(\frac{\widehat{W}_n(s+v(s))}{1-\dot{w}(s+v(s))}\right)\mathrm{d}\Lambda(s)+o(1)\\
&\quad= \sqrt{n}\int_0^{t} f(\theta_{3,n}(u))\left(\frac{\widehat{W}_n(u)}{1-\dot{w}(u)}\right)(1-\dot{w}(u))\lambda\,\mathrm{d}u+o(1)\\
&\quad= \sqrt{n}\int_0^{t} f(\theta_{3,n}(u))\,\widehat{W}_n(u)\,\lambda\,\mathrm{d}u+o(1),
\end{aligned} \tag{5.10}$$

where the second equality holds by applying the second formula in (4.3) and a change of variable $u \equiv s + v(s)$. To wit, first, the second equality in (4.3) implies that $t - w(t) + v(t - w(t)) = t - w(t) + w(t) = t$; second, the first equality in (4.3) implies that $w(u) = w(s + v(s)) = v(s)$, so that $s = u - v(s) = u - w(u)$ and $\mathrm{d}s = (1 - \dot{w}(u))\mathrm{d}u$. Here $\theta_{3,n}(t)$ satisfies

$$V_n(t-w(t)) \wedge v(t-w(t)) \le \theta_{3,n}(t) \le V_n(t-w(t)) \vee v(t-w(t)). \tag{5.11}$$

***FCLT limits for*** $\widehat{E}_n$ Label all servers from 1 to $n$. Let $D_j(t)$ count the number of service completions at server $j$ by time $t$, $1 \le j \le n$. Because the system operates in the ED regime with $\rho > 1$, all servers will be busy at all times with probability 1 as $n \to \infty$. Hence, the total number of service completions in $[0, t]$ is given by $D_n(t) = \sum_{j=1}^{n} D_j(t)$ for $t \ge 0$, where $D_1(t), D_2(t), \ldots$ are I.I.D. ERPs. (That is, $D_n$ is asymptotically equivalent to the superposition of $n$ ERPs.) By Theorem 2 of [35], we have $(\widehat{E}_n, \widehat{D}_n) \Rightarrow (\widehat{E}, \widehat{E})$ as $n \to \infty$, where the limiting Gaussian process $\widehat{E}$ is given by (4.6). Hence, we can write

$$E_n(t) = nE(t) + \sqrt{n}\widehat{E}(t) + o(\sqrt{n}). \tag{5.12}$$

Combining (5.10), (5.8), and (5.12) yields an SDE

$$\widehat{W}_n(t) = -\frac{1}{F^c(\theta_{1,n}(t))}\int_0^t f(\theta_{3,n}(s))\,\widehat{W}_n(s)\;\mathrm{d}s + \frac{1}{F^c(\theta_{1,n}(t))\lambda}\widehat{G}(t) + o(1), \tag{5.13}$$

where

$$\begin{aligned}
\widehat{G}(t) \equiv &\int_0^t c_\lambda F^c(w(s))\,\mathrm{d}\mathcal{B}_\lambda(\Lambda(s-w(s)))\\
&+ \mathcal{B}_a\left(\int_0^t F^c(v(u))F(v(u))\,\mathrm{d}\Lambda(u)\right) - \widehat{E}(t).
\end{aligned}$$

To complete the proof of the FCLT for $\widehat{W}_n$, we first revise (5.13) to obtain a much neater SDE for $\widehat{W}_n$. To do so, we apply Gronwall's inequality. See Sect. 1 in the online appendix and [28] for a reference. We first apply Gronwall's inequality to show that $\widehat{W}_n$ is *stochastically bounded*. The SDE (5.13) implies that

$$\left|\widehat{W}_n(t)\right| \leq \frac{1}{F^c(\theta_{1,n}(t))} \int_0^t f(\theta_{3,n}(s))\,|\widehat{W}_n(s)|\,\mathrm{d}s + \frac{1}{F^c(\theta_{1,n}(t))\lambda}\,|\widehat{G}(t)| + o(1). \tag{5.14}$$

Applying Gronwall's inequality leads to

$$\begin{aligned}\left|\widehat{W}_n(t)\right| &\leq \frac{|\widehat{G}(t)|}{\lambda F^c(\theta_{1,n}(t))} + \int_0^t \frac{|\widehat{G}(u)|}{\lambda F^c(\theta_{1,n}(u))} e^{\int_u^t \frac{f(\theta_{3,n}(r))}{F^c(\theta_{1,n}(u))}\mathrm{d}r}\, \frac{f(\theta_{3,n}(u))}{F^c(\theta_{1,n}(t))}\,\mathrm{d}u + o(1)\\ &\leq \frac{|\widehat{G}(t)|}{\lambda F^c(\theta_{1,n}(t))} + \frac{e^{\frac{t}{F^c(\theta_{1,n}(t))}}}{F^c(\theta_{1,n}(t))} \int_0^t \frac{|\widehat{G}(u)|f(\theta_{3,n}(u))}{\lambda F^c(\theta_{1,n}(u))}\,\mathrm{d}u + o(1).\end{aligned}$$

Hence, the stochastic boundedness of $\widehat{W}_n$ follows from the stochastic boundedness of $\widehat{G}$.

Inequalities (5.7) and (5.11) imply that

$$\begin{aligned}&\theta_{1,n}(t) = v(t - w(t)) + o(1) = w(t) + o(1) \quad \text{and}\\ &\theta_{3,n}(t) = v(t - w(t)) + o(1) = w(t) + o(1),\end{aligned} \tag{5.15}$$

where the second and last equality follows from (4.3). Replacing $\theta_{1,n}(t)$ and $\theta_{3,n}(t)$ by $w(t)$ in the SDE (5.13) yields the much cleaner SDE

$$\widehat{W}_n(t) = -\frac{1}{F^c(w(t))} \int_0^t f(w(s))\,\widehat{W}_n(s)\,\mathrm{d}s + \frac{1}{F^c(w(t))\lambda}\,\widehat{G}(t) + o(1). \tag{5.16}$$

Note that the *stochastic boundedness* of $\widehat{W}_n$ plays a key role here because it keeps the errors caused by the approximations in (5.15) under control.

*Remark 5.2* (Formulas for time-varying arrival rate $\lambda(t)$) If the arrival rate $\lambda(t)$ is a time-varying function, then the SDEs (5.13) and (5.16) generalize to

$$\begin{aligned}\widehat{W}_n(t) &= -\frac{1}{\widetilde{g}_n(t)} \int_0^t f(\theta_{3,n}(s))\,\widehat{W}_n(s)\,\lambda(s - w(s))\,\mathrm{d}s + \frac{1}{\widetilde{g}_n(t)}\,\widehat{G}(t) + o(1)\\ &= -\frac{1}{F^c(w(t))\lambda(t - w(t))} \int_0^t f(w(s))\,\widehat{W}_n(s)\,\lambda(s - w(s))\,\mathrm{d}s\\ &\quad + \frac{1}{F^c(w(t))\lambda(t - w(t))}\,\widehat{G}(t) + o(1),\end{aligned}$$

where $\widetilde{g}_n(t) \equiv F^c(\theta_{1,n}(t))\lambda(t - w(t))$.

***Finishing the proof of the FCLT of $\widehat{W}_n$*** In order to show that $\widehat{W}_n \Rightarrow \widehat{W}$, where $\widehat{W}$ satisfies the SDE (4.7), we match both sides of the two SDEs (5.16) and (4.7):

$$\begin{aligned}\left|\widehat{W}_n(t) - \widehat{W}(t)\right| &\le \frac{1}{F^c(w(t))}\int_0^t f(w(s))\,|\widehat{W}_n(s) - \widehat{W}(s)|\,\mathrm{d}s + o(1)\\ &\equiv \int_0^t |\widehat{W}_n(s) - \widehat{W}(s)|\,\widetilde{\mu}(s)\,\mathrm{d}s + o(1),\end{aligned}$$

with $\widetilde{\mu}(s) = f(w(s))/F^c(w(t))$. Applying Gronwall's inequality once again yields that

$$\left|\widehat{W}_n(t) - \widehat{W}(t)\right| \le e^{\frac{\int_0^t f(w(s))\mathrm{d}s}{F^c(w(t))}}\int_0^t o(1)\frac{f(w(u))}{F^c(w(t))}\,\mathrm{d}u + o(1),$$

which implies that $\left\|\widehat{W}_n - \widehat{W}\right\|_T \Rightarrow 0$ as $n \to \infty$.

*Remark 5.3* (If we were to take the compactness approach) The key step in our new approach is the development of the convenient SDEs (5.13) and (5.16). We remark that our new SDE representation will provide a simple proof even if we were to take the conventional compactness approach. The first step of the compactness approach requires tightness of $\widehat{W}_n$, which can be shown by establishing *(i)* that $\widehat{W}_n$ is stochastically bounded (already shown here) and *(ii)* that $\widehat{W}_n$ has controlled modulus of continuity (see [36] for the necessary and sufficient condition for tightness in $\mathcal{D}$). However, our new SDE (5.16) provides a simple proof for step *(ii)*. Indeed, with the integral representation (5.16) for $\widehat{W}_n$, the stochastic boundedness can be used to control the modulus of continuity, that is, we can show that

$$\left|\widehat{W}_n(t+\delta) - \widehat{W}_n(t)\right| \le C(t)\int_t^{t+\delta} f(w(s))\left|\widehat{W}(s)\right|\mathrm{d}s + o(1),$$

for some finite $C(t)$. The stochastic boundedness of $\widehat{W}_n$ implies that $\left|\widehat{W}_n(t+\delta) - \widehat{W}_n(t)\right|$ is asymptotically bounded by $\widetilde{C}\delta$ for some $\widetilde{C} < \infty$ for all $0 \le t \le T$, which concludes the $\mathcal{C}$-tightness for $\widehat{W}$.

Next, given tightness for $\widehat{W}_n$, we assume that there exists a convergent subsequence $\widehat{W}_{n_k}$. We can easily use the SDE (5.16) to show that the subsequence $\widehat{W}_{n_k}$ converges to some $\widehat{W}^*$ which solves the SDE (4.7).

### 5.2 FCLT for other processes

That $(\widehat{N}_n, \widehat{D}_n, \widehat{E}_n, \widehat{W}_n) \Rightarrow (\widehat{N}, \widehat{E}, \widehat{E}, \widehat{W})$ follows from the convergence-together theorem (see Theorem 11.4.7. of [36]) and the continuous mapping theorem. The convergence $\widehat{W}_n \Rightarrow \widehat{W}$ and (5.9) implies that $\widehat{V}_n \Rightarrow \widehat{V}$ with

$$\widehat{V}(t) = \frac{\widehat{W}(t + v(t))}{1 - \dot{w}(t + v(t))}.$$

We next prove the FCLT for the queue-length process $\widehat{Q}_n$ based on the FCLT for $\widehat{W}_n$ and the continuous mapping theorem. First, the FWLLN implies that

$$Q_1(t) = Q_2(t) = 0, \quad Q_3(t) = \int_0^{t-w(t)} F^c(t-s)\lambda \, \mathrm{d}s.$$

Following (3.8)–(3.10), as $n \to \infty$,

$$\widehat{Q}_{n,1}(t) \equiv \frac{1}{\sqrt{n}} Q_{n,1}(t) \Rightarrow \widehat{Q}_1(t) \equiv \int_{t-w(t)}^{t} c_\lambda F^c(t-s) \, \mathrm{d}\mathcal{B}_\lambda(\Lambda(s)), \tag{5.17}$$

$$\begin{aligned} \widehat{Q}_{n,2}(t) &\equiv \frac{1}{\sqrt{n}} Q_{n,2}(t) \Rightarrow \widehat{Q}_2(t) \equiv \int_{t-w(t)}^{t} \int_0^1 \mathbf{1}(x > F(t-s)) \, \mathrm{d}\widehat{U}(\Lambda(s), x) \\ &\overset{\mathrm{d}}{=} \int_{t-w(t)}^{t} \sqrt{F^c(t-s)F(t-s)} \, \mathrm{d}\mathcal{B}_a\left(\Lambda(s)\right) \\ &\overset{\mathrm{d}}{=} \mathcal{B}_a\left(\int_{t-w(t)}^{t} F^c(t-s)F(t-s)\lambda \, \mathrm{d}s\right) = \mathcal{B}_a\left(\int_0^{w(t)} F^c(u)F(u)\lambda \, \mathrm{d}u\right), \end{aligned} \tag{5.18}$$

$$\begin{aligned} \widehat{Q}_{n,3}(t) &\equiv \frac{1}{\sqrt{n}} \left(Q_{n,3}(t) - n\, Q_3(t)\right) = \sqrt{n} \int_{t-W_n(t)}^{t-w(t)} F^c(t-s)\lambda \, \mathrm{d}s \\ &= \widehat{W}_n(t) F^c(w(t))\lambda + o(1) \Rightarrow \widehat{Q}_3(t) \equiv \widehat{W}(t) F^c(w(t))\lambda. \end{aligned} \tag{5.19}$$

Here the proofs for the convergence in (5.17) and (5.18) are similar to the proof of Lemma 5.1. Note that $\widehat{Q}_3$ in (5.19) involves $\widehat{W}$ given in (4.16), which involves stochastic integrals with respect to $\mathcal{B}_\lambda$ and $\widehat{E}$, and the Kiefer integral of $\widehat{U}$ (or Brownian motion $\mathcal{B}_a$). A careful analysis reveals that the Kiefer integral of $\widehat{E}_2$ in Lemma 5.1 involves $\widehat{U}$ in the time interval $[0, t - w(t)]$, while $\widehat{Q}_2$ in (5.18) involves $\widehat{U}$ in $[t - w(t), t]$. So $\widehat{Q}_2$ and $\widehat{Q}_3$ are independent because a Kiefer process has independent increments with respect to the first (time) component. Similarly, because $\widehat{E}_1$ in Lemma 5.1 involves $\mathcal{B}_\lambda$ in $[0, t - w(t)]$, while $\widehat{Q}_1$ in (5.17) involves $\mathcal{B}_\lambda$ in $[t - w(t), t]$, $\widehat{Q}_1$ and $\widehat{Q}_3$ are independent. In summary, all three terms $\widehat{Q}_1$, $\widehat{Q}_2$ and $\widehat{Q}_3$ are independent.

The above analysis enables us to obtain an alternative representation for $\widehat{Q}$ by regrouping the integrals, writing $\widehat{Q}$ as a sum of three new independent integrals:

$$\widehat{Q}(t) = \int_0^t K_\lambda(t,u) \, \mathrm{d}\mathcal{B}_\lambda(\Lambda(u)) + \int_0^t K_a(t,u) \, \mathrm{d}\mathcal{B}_a(u) + \int_0^t K_s(t,u) \, \mathrm{d}\widehat{E}(u),$$

where the integrands $K_\lambda(t,u)$, $K_a(t,u)$ and $K_s(t,u)$ are analytic functions, with $K_\lambda(t,u)$ and $K_a(t,u)$ being piecewise functions (having different forms for $0 \le u \le t - w(t)$ and $t - w(t) \le u \le t$). This alternative formula nicely separates the variabilities in the arrival process (through $\mathcal{B}_\lambda$), abandonment times (through $\mathcal{B}_a$ or $\widehat{U}$) and service times (through $\widehat{E}$).

**Acknowledgements** We thank editors Amy Ward and Guodong Pang for inviting us to contribute to this special issue and anonymous referees for providing constructive comments. The third author would like to thank Ward Whitt for his support and guidance through the years, and for being a tremendous source of inspiration. The first and third authors acknowledge supports from NSF Grant CMMI 1362310.

# Appendix

## A Additional Proofs

### A.1 Proof of Proposition 3.1

To make sure (3.11) is well defined and to be able to characterize its distribution, we define a sequence of discrete versions of (3.11), that is, $\{L^{(m)} : m \geq 1\}$, where

$$L^{(m)}(t) \equiv \sum_{i=0}^{m-1} J(t, u_i)\,(Z(\omega, u_{i+1}) - Z(\omega, u_i))\,, \quad t \geq 0, \tag{A.1}$$

for a given partition on the interval $[0, t]$, $0 = u_0 < u_1 < \cdots < u_m = t$. Suppose $\omega \in \Omega$ is such that $(Z(\omega, t) : t \geq 0)$ has Hölder continuous sample paths. For simplicity, we suppress $\omega$ hereafter. For $m > 0$, consider the partition $0 = u_0 < u_1 < \cdots < u_m = t$ and

$$\begin{aligned} L^{(m)}(t) &= \sum_{i=0}^{m-1} J(t, u_i)\,(Z(u_{i+1}) - Z(u_i)) \\ &= \sum_{i=1}^{m} J(t, u_{i-1}) Z(u_i) - \sum_{i=0}^{m-1} J(t, u_i) Z(u_i) \\ &= J(t, u_{m-1}) Z(u_m) - J(t, 0) Z(0) \\ &\quad - \sum_{i=1}^{m-1} \big[\, J(t, u_i) - J(t, u_{i-1}) \,\big]\, Z(u_i). \end{aligned} \tag{A.2}$$

Because $Z(t)$ is continuous, the summation converges to the Riemann–Stieltjes integral as the partition mesh goes to 0 if $J(t, u)$ is monotone in the second component for each $t$. Moreover, if $J(t, u)$ is differentiable for each $t$, we can replace the integrator $\mathrm{d}J(t, u)$ of the Riemann–Stieltjes integral with $J_u(t, u)\mathrm{d}u$, where the subscript denotes derivative with respect to the second component. The Riemann–Stieltjes integral is well defined if the derivative as a function of $u$ for fixed $t$ is continuous. (In general, finitely many jumps are allowed.) Therefore,

$$\begin{aligned} \int_0^t J(t, u)\,\mathrm{d}Z(u) \equiv \lim_{m \to \infty} L^{(m)}(t) &\overset{\text{a.s.}}{=} J(t, t) Z(\omega, t) - J(t, 0) Z(\omega, 0) \\ &\quad - \int_0^t Z(\omega, u)\,\mathrm{d}J(t, u). \end{aligned}$$

Moreover, with $\Delta \equiv \max\{u_i - u_{i-1} : 1 \le i \le m\}$, we have

$$\begin{aligned}
&\sum_{i=1}^{m-1} \left[ J(t, u_i) - J(t, u_{i-1}) \right] Z(u_i) - \int_0^t J_u(t, u) Z(u)\, \mathrm{d}u \\
&\quad = \sum_{i=1}^{m-1} \int_{u_{i-1}}^{u_i} \left[ \frac{J(t, u_i) - J(t, u_{i-1})}{u_i - u_{i-1}} - J_u(t, u) \right] (Z(u) - Z(u_i))\, \mathrm{d}u \\
&\quad \le \frac{1}{\Delta} \cdot c_1 \Delta \cdot c_2 \Delta^{\alpha} \to 0
\end{aligned}$$

as $\Delta \to 0$, where the inequality holds because $\hat{Z}$ has Hölder continuous sample paths and $J(t, u)$ is differentiable with respect to the second component.

We prove Proposition 3.1 in two steps. First, we show in Lemma A.1 that if the sequence of covariance functions associated with the processes $\{L^{(m)} : m \ge 1\}$ converges to some limit function, then the sequence $\{L^{(m)} : m \ge 1\}$ converges in distribution to a Gaussian process. Moreover, the covariance function of the limit Gaussian process coincides with the limit of the covariance function associated with $\{L^{(m)} : m \ge 1\}$. Then, in the second step, we show that the covariance functions associated with $\{L^{(m)} : m \ge 1\}$ indeed converge.

**Lemma A.1** *Let $X^{(m)} \equiv \left(X_1^{(m)}, \ldots, X_l^{(m)}\right)$ be a sequence of centered Gaussian random vector in $\mathbb{R}^l$ and let $\Sigma^{(m)}$ be the covariance matrix of $X^{(m)}$. If $\Sigma^{(m)} \to \Sigma$ as $m \to \infty$, then $X^{(m)} \Rightarrow X$, where the limit $X$ is Gaussian with mean zero and covariance $\Sigma$.*

*Proof* Consider the characteristic function $\phi_m(\theta) \equiv \mathbb{E}\left[e^{i\theta^T X^{(m)}}\right]$ of the vector $X^{(m)}$. The convergence $\Sigma^{(m)} \to \Sigma$ implies the convergence of characteristic functions

$$\phi_m(\theta) = e^{-\frac{1}{2}\theta^T \Sigma^{(m)} \theta} \to \phi(\theta) \equiv e^{-\frac{1}{2}\theta^T \Sigma \theta}$$

due to continuity of $\phi_m$. Then the result follows from Lévy's continuity theorem. □

We next show that the covariance functions associated with the sequence $\{L^{(m)} : m \ge 1\}$ in (A.1) converge. We consider a partition of the interval $[0, t_2]$ such that there are a total of $m_2$ intervals partitioning $[0, t_2]$ and $m_1$ intervals partitioning $[0, t_1]$. We use the form in (A.2) to compute the covariance of $L^{(m)}(t)$. Let $C_Z(\cdot, \cdot)$ be the covariance function associated with the process $Z$. Then, for $0 \le t_1 < t_2$ and the partition $0 = s_0 < s_1 < \cdots < s_{m_1 - 1} < s_{m_1} = t_1 < s_{m_1+1} < \cdots < s_{m_2-1} < s_{m_2} = t_2$,

$$\begin{aligned}
&\mathbb{E}[L^{(m)}(t_1) L^{(m)}(t_2)] \\
&\quad = \mathbb{E}\Bigg[\left(J(t_1, s_{m_1-1}) Z(t_1) - J(t_1, 0) Z(0) - \sum_{i=1}^{m_1-1} (J(t_1, s_i) - J(t_1, s_{i-1})) Z(s_i)\right) \\
&\qquad \times \left(J(t_2, s_{m_2-1}) Z(t_2) - J(t_2, 0) Z(0) - \sum_{i=1}^{m_2-1} (J(t_2, s_i) - J(t_2, s_{i-1})) Z(s_i)\right)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&= J(t_1, s_{m_1-1})J(t_2, s_{m_2-1})C_Z(t_1, t_2) + J(t_1, 0)J(t_2, 0)C_Z(0, 0)\\
&\quad - J(t_2, s_{m_2-1})J(t_1, 0)C_Z(0, t_2) - J(t_1, s_{m_1-1})J(t_2, 0)C_Z(0, t_1)\\
&\quad - \sum_{i=1}^{m_1-1} J(t_2, s_{m_2-1})(J(t_1, s_i) - J(t_1, s_{i-1}))C_Z(s_i, t_2)\\
&\quad + \sum_{i=1}^{m_1-1} J(t_2, 0)(J(t_1, s_i) - J(t_1, s_{i-1}))C_Z(0, s_i)\\
&\quad - \sum_{i=1}^{m_2-1} J(t_1, s_{m_1-1})(J(t_2, s_i) - J(t_2, s_{i-1}))C_Z(t_1, s_i)\\
&\quad + \sum_{i=1}^{m_2-1} J(t_1, 0)(J(t_2, s_i) - J(t_2, s_{i-1}))C_Z(0, s_i)\\
&\quad + \sum_{i=1}^{m_1-1}\sum_{j=1}^{m_2-1} (J(t_1, s_i) - J(t_1, s_{i-1}))(J(t_2, s_j) - J(t_2, s_{j-1}))C_Z(s_i, s_j).
\end{aligned}
$$

Convergence of the first four terms follows from continuity of $u \mapsto J(t, u)$ for each fixed $t$ as $s_{m_1-1} \to t_1$ and $s_{m_2-1} \to t_2$ as $m \to \infty$. Convergence of the last four summations follows from the fact that $C_Z$ is bounded over compact intervals and $J(t, u)$ is differentiable and, therefore, bounded for each $t$ over compact intervals. Hence the limits of these terms are the Riemann–Stieltjes integrals given in (3.13). Finally, the last summation term converges to the two-dimensional Riemann–Stieltjes integral in (3.13) due to similar reasoning. □

### A.2 Proof of Theorem 4.1

We first establish a FWLLN for $W_n$ following the compactness approach, i.e., $(i)$ the sequence $W_n$ is $\mathcal{C}$-tight, which implies that every subsequence has a convergent subsequence with a limit in $\mathcal{C}$; and $(ii)$ every convergent subsequence converges to the same limit, which in our case uniquely solves the ODE in (4.2). Finally, we establish convergence for the other processes and characterize their limits. We remark that the tightness for $W_n$ is quite straightforward, but the tightness for the CLT-scaled processes (for example, $\widehat{W}_n$) is complicated (which is why we adopt a new approach to prove the FCLT).

The proof closely follows the arguments in [24] and Sect. 6.6 of [26]. We, hereby, redo the steps therein for the new representation of the enter-service process $E_n$; we use the decomposition in (3.3)–(3.6) that is different than the expressions for $E_n$ in [26].

*Tightness of* $\{W_n\}$ To prove tightness, first we show that $W_n$ is stochastically bounded and then show that $W_n$ has controlled modulus of continuity, that is, for each $T > 0$ and $\epsilon > 0$,

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty} \mathbb{P}(w(W_n, \delta, T) > \epsilon) = 0, \tag{A.3}$$

where $w(W_n, \delta, T)$ is the modulus of continuity of $W_n$, i.e., $\sup\{w(W_n, [t_1, t_2]) : 0 \leq t_1 < t_2 \leq (t_1 + \delta) \wedge T\}$ with $w(W_n, A) \equiv \sup\{W_n(s_1) - W_n(s_2) : s_1, s_2 \in A\}$.

The *stochastic boundedness* is obvious, because, in any finite interval $[0, T]$, we immediately see that HOL satisfies $0 \leq W_n(t) \leq T$ for all $n \geq 1$, $t \in [0, T]$.

To treat the *modulus of continuity*, we first see that $W_n(t + \delta) - W_n(t) \leq \delta$ for $\delta > 0$ and $0 \leq t \leq T$, because the HWT can increase at most at rate 1. Therefore, it remains to find a bound on $W_n(t) - W_n(t + \delta)$. To this end, define

$$\bar{E}_{n,3}(t, \delta) \equiv \bar{E}_{n,3}(t + \delta) - \bar{E}_{n,3}(t) = \int_{t-W_n(t)}^{t+\delta-W_n(t+\delta)} F^c(V_n(s))\lambda \, ds. \qquad (A.4)$$

Because the ccdf $F^c(x) > 0$ for all $x \geq 0$, let $c \equiv \inf_{x\in[0,T]}\{F^c(x)\} > 0$. Hence, the integrand in (A.4) is bounded below by a constant $c\lambda > 0$, which yields a lower bound on $\bar{E}_{n,3}(t, \delta)$:

$$W_n(t) - W_n(t + \delta) + \delta \leq \frac{\bar{E}_{n,3}(t, \delta)}{c\lambda}, \quad t \geq 0.$$

From the FCLT in Theorem 2 of [35], we know that $\bar{D}_n(t) \Rightarrow D(t) = \mu t$ so that $\bar{E}_{n,3}(t) \to E_3(t) = D(t) = \mu t$. Therefore, we have $\limsup_{n\to\infty}\{W_n(t) - W_n(t + \delta)\} \leq \mu\delta/c\lambda$ so that

$$\limsup_{n\to\infty} |W_n(t + \delta) - W_n(t)| \leq c^*\delta, \qquad c^* \equiv \max\left(\mu/c\lambda, 1\right). \qquad (A.5)$$

Hence, $W_n$ is tight. In addition, (A.5) also implies that the limit of every convergent subsequence of $W_n$ is in $\mathcal{C}$ and is Lipschitz continuous.

*Limit of Convergent Subsequence of* $\{W_n\}$ The $C$-tightness implies that every subsequence of $W_n$ has a convergent subsequence. Let $W_{n_k}$ be a convergent subsequence with the limit $w^*$, i.e., $W_{n_k} \Rightarrow w^*$. From (2.3) and (2.4), we deduce that the PWT on the subsequence also converges, that is, $V_{n_k} \Rightarrow v^*$, with the limit $v^*$ satisfying

$$v^*(t) = w^*(t + v^*(t)) \quad \text{and} \quad v^*(t - w^*(t)) = w^*(t), \quad t \geq 0. \qquad (A.6)$$

We now show that $w^*$ solves the ODE (4.2). On the one hand, the FCLT in Theorem 2 of [35] implies that $(\bar{E}_n, \bar{D}_n) \Rightarrow (D, D)$ with $D(t) = \mu t$. On the other hand, (3.3) implies that $\bar{E}_n$ along the subsequence associated with $W_{n_k}$ and $V_{n_k}$ converges to a limit $E^*$. Specifically,

$$\bar{E}_{n_k}(t) \Rightarrow E^*(t) = E_3^*(t) \equiv \int_0^{t-w^*(t)} F^c(v^*(s))\lambda \, ds = D(t) = \mu t. \qquad (A.7)$$

Because the prelimit process is $\mathcal{C}$-tight, we know the derivative $\dot{w}^*(t)$ exists. Taking derivative in (A.7) yields

$$\mu = (1-\dot{w}^*(t))F^c(v^*(t-w^*(t)))\lambda = (1-\dot{w}^*(t))F^c(w^*(t))\lambda, \qquad \text{(A.8)}$$

which coincides with the ODE (4.2).

*FWLLN for the other processes* To prove full convergence of $V_n$, we write

$$\begin{aligned}
&|V_n(t-W_n(t)) - v(t-w(t))| \\
&\quad \le |V_n(t-W_n(t)) - V_n(t-w(t))| + |V_n(t-w(t)) - v(t-w(t))| \\
&\quad = |W_n(t) - w(t) + O(1/n)| + |w(t) + O(1/n) - w(t)| \\
&\quad \le |W_n(t) - w(t)| + O(1/n) \qquad \text{(A.9)}
\end{aligned}$$

Apply the change of variable to (A.9) with $u_n \equiv t - W_n(t)$ and $u \equiv t - w(t)$ to obtain

$$\|V_n - v\| \le \frac{\|W_n - w\|}{\gamma} + O(1/n) = O(1/n) \qquad \text{(A.10)}$$

for a constant $\gamma > 0$, where the equality holds because $u_n = u + o(1)$.

The limit of the sequences of processes (3.8)–(3.10) can be obtained the same way it is done in [26], which makes use of Theorem 3.1. of [31] and then applies the continuous mapping theorem given $W_n \Rightarrow w$. From (6.17) of [26], we immediately write

$$\bar{Q}_{n,i} \Rightarrow 0 \quad \text{for} \quad i=1,2; \quad \bar{Q}_{n,3} \Rightarrow Q_3(t) \equiv \int_{t-w(t)}^{t} F^c(t-s)\lambda\,\mathrm{d}s, \quad \text{as} \quad n \to \infty. \qquad \text{(A.11)}$$

### A.3 Proof of Theorem 4.3

The expressions for $C_{\widehat{W}_1}(t,t')$ and $C_{\widehat{W}_2}(t,t')$ are obtained by applying rules of the Itô integral. Derivation of these functions follows from standard arguments and therefore the details are omitted.

To compute $C_{\widehat{W}_3}(t,t')$ we make use of (3.13) with $J(t,u) \equiv H(t,u)/q(u,w(u))$, where $H(t,u)$ and $q(u,w(u))$ are as in (4.17). In particular, for $0 \le t < t'$,

$$\begin{aligned}
C_{\widehat{W}_3}(t,t') =& J(t,t)J(t',t')C_E(t,t') - \int_0^{t'} J(t,t)J_u(t',u)C_E(t,u)\mathrm{d}u \\
&- \int_0^{t} J(t',t')J_u(t,u)C_E(t',u)\mathrm{d}u \\
&+ \int_0^{t}\int_0^{t'} J_u(t,u)J_v(t',v)C_E(u,v)dv\mathrm{d}u \\
=& \frac{1}{\lambda^2 F^c(w(t))F^c(w(t'))} C_E(t,t') - \frac{1}{\lambda F^c(w(t))}\int_0^{t'} J_u(t',u)C_E(t,u)\mathrm{d}u
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{\lambda F^c(w(t))} \int_0^t J_u(t,u) C_E(t',u) \mathrm{d}u \\
&+ \int_0^t \int_0^{t'} J_u(t,u) J_v(t',v) C_E(u,v) \mathrm{d}v \mathrm{d}u,
\end{aligned}
$$

where $J_u(t,u)$ is as in (4.20).

We next derive the covariance function for the limit queue-length process. First, $C_{\widehat{Q}_1}(t,t')$ can be obtained from isometry property of the Itô integral. The function $C_{\widehat{Q}_2}(t,t')$ can be obtained by $\widehat{U}(\lambda s, y) = \mathcal{W}(\lambda s, y) - y\mathcal{W}(\lambda s, 1)$, where $\mathcal{W}(\cdot.\cdot)$ is a two-dimensional Brownian motion. We refer interested readers to the long version of [31] and the references therein for a definition of the Kiefer process and of stochastic integrals with respect to two-parameter martingales. The last term easily follows by definition. □

### A.4 Proof of Lemma 4.1

First, we prove the existence of $\widetilde{E}(t)$. It suffices to show that for any $n \geq 1$ and $-t_1 < -t_2 < \ldots < -t_n \leq 0$, the matrix $M = (\widetilde{C}(-t_i, -t_j))_{i,j=1}^n$ is nonnegative definite. Let $r_1 = t_1$ and $r_j = t_{j-1} - t_j$ for $j = 2, \ldots, n$, and define $N = (C_E(r_i, r_j))_{i,j=1}^n$. For any $z = (z_1, z_2, \ldots, z_n)^T \in \mathbb{R}^n$, define $y = (y_1, y_2, .., y_n)^T$ such that $y_1 = \sum_{i=1}^n z_i$ and $y_j = -\sum_{i=j}^n z_i$ for $j = 2, \ldots, n$. Given that $\widetilde{C}(t,s) = C_E(-t,-t) - C_E(-t, s-t)$, we can compute

$$z^T M z = \sum_{i=1}^n \widetilde{C}(-t_i, -t_i) z_i^2 + 2 \sum_{1 \leq i < j \leq n} \widetilde{C}(-t_i, -t_j) z_i z_j = y^T N y.$$

We shall explain how to derive the above equation for $n = 2$.

$$
\begin{aligned}
&z_1^2 \widetilde{C}(-t_1, -t_1) + 2 z_1 z_2 \widetilde{C}(-t_1, -t_2) + z_2^2 \widetilde{C}(-t_2, -t_2) \\
&\quad = z_1^2 C_E(r_1, r_1) + 2 z_1 z_2 (C_E(r_1, r_1) - C_E(r_1, r_2)) + z_2^2 C_E(r_1 - r_2, r_1 - r_2) \\
&\quad = z_1^2 C_E(r_1, r_1) + 2 z_1 z_2 (C_E(r_1, r_1) - C_E(r_1, r_2)) \\
&\quad\quad + z_2^2 (C_E(r_1, r_1) - 2 C_E(r_1, r_2) + C_E(r_2, r_2)) \\
&\quad = (z_1 + z_2)^2 C_E(r_1, r_1) - 2 z_2 (z_1 + z_2) C_E(r_1, r_2) + z_2^2 C_E(r_2, r_2) = y^T N y.
\end{aligned}
$$

Since $C_E$ is the covariance function of a Gaussian process, the matrix $N$ is nonnegative definite and hence $y^T N y \geq 0$. As the vector $z$ is any vector in $\mathbb{R}^n$, we can conclude that $M$ is also nonnegative definite and the existence of $\widetilde{E}$ follows. The argument is similar for $n \geq 3$, therefore, the details are omitted.

Next we show that (4.24) holds. Since a Gaussian process is fully characterized by its covariance function, it suffices to show that, for any fixed $t > 0$ and $0 \leq r < s \leq t$,

$$\mathrm{Cov}(\widetilde{E}(-t + r) - \widetilde{E}(-t), \widetilde{E}(-t + s) - \widetilde{E}(-t)) = C_E(r, s).$$

By our definition of $\widetilde{C}(t,s)$, we can compute

$$\begin{aligned}&\mathrm{Cov}(\widetilde{E}(-t+r)-\widetilde{E}(-t),\widetilde{E}(-t+s)-\widetilde{E}(-t))\\&\quad=C_E(t-r,t-r)-C_E(t-r,s-r)+C_E(t,s)+C_E(t,r)-C_E(t,t).\end{aligned}\tag{A.12}$$

By the stationary increments of $\widehat{E}$, we have

$$\begin{aligned}C_E(t-r,t-r)&=\mathrm{Var}(\widehat{E}(t-r))=\mathrm{Var}(\widehat{E}(t)-\widehat{E}(r))\\&=C_E(t,t)-2C_E(t,r)+C_E(r,r),\\C_E(t-r,s-r)&=\mathrm{Cov}(\widehat{E}(t-r),\widehat{E}(s-r))=\mathrm{Cov}(\widehat{E}(t)-\widehat{E}(r),\widehat{E}(s)-\widehat{E}(r))\\&=C_E(t,s)-C_E(t,r)-C_E(r,s)+C_E(r,r),\end{aligned}$$

which along with (A.12) implies that

$$\mathrm{Cov}(\widetilde{E}(-t+r)-\widetilde{E}(-t),\widetilde{E}(-t+s)-\widetilde{E}(-t))=C_E(r,s).$$

This completes the proof. □

### A.5 Proof of Theorem 4.4

**Steady state of $\widehat{W}$** Let $\mathcal{N}(0,\sigma^2)$ denote the normal distribution with mean 0 and variance $\sigma^2$. First, we treat $\widehat{W}_1(t)$ in (4.16) by applying a change of variable with $u=s+v(s)$. Let $\kappa(t)$ be the inverse of the function $\beta(t)=t+v(t)$. We write

$$\begin{aligned}\widehat{W}_1(t)&=\int_{\kappa(0)}^{\kappa(t)}\frac{F^c(w(s+v(s)))H(t,s+v(s))}{\lambda F^c(w(t))}c_\lambda\,\mathrm{d}\mathcal{B}_\lambda(\Lambda(s+v(s)-w(s+v(s))))\\&=\int_0^{\kappa(t)}\frac{F^c(v(s))H(t,s+v(s))}{\lambda F^c(w(t))}c_\lambda\,\mathrm{d}\mathcal{B}_\lambda(\Lambda(s))\\&\stackrel{\mathrm{d}}{=}\int_0^{\kappa(t)}\frac{c_\lambda}{\sqrt{\lambda}}e^{-h_F(w)(t-s-v(s))}\mathrm{d}\mathcal{B}_\lambda(s)\\&\stackrel{\mathrm{d}}{=}\widetilde{\mathcal{B}}_\lambda\left(\frac{c_\lambda^2}{\lambda}\int_0^{\kappa(t)}e^{-2h_F(w)(t-s-v(s))}\,\mathrm{d}s\right)\\&\stackrel{\mathrm{d}}{=}\widetilde{\mathcal{B}}_\lambda\left(\frac{c_\lambda^2}{\lambda}\int_0^{t}e^{-2h_F(w)(t-s)}\,\mathrm{d}\kappa(s)\right)\stackrel{\mathrm{d}}{=}\widetilde{\mathcal{B}}_\lambda\left(\frac{c_\lambda^2}{2h_F(w)\lambda}\left(1-e^{-2t\,h_F(w)}\right)\right)\\&\Rightarrow\widehat{W}_1(\infty)\stackrel{\mathrm{d}}{=}\mathcal{N}\left(0,\frac{c_\lambda^2}{2h_F(w)\lambda}\right),\qquad\text{as}\quad t\to\infty,\end{aligned}$$

where the second equality follows from (4.3). Similarly, an application of Theorem 3.4.6 of [16] yields

$$\begin{aligned}\widehat{W}_2(t) &= \int_0^t \frac{\sqrt{F(w)}}{\sqrt{\lambda F^c(w)}}\, e^{-h_F(w)(t-u)} \mathrm{d}\mathcal{B}_a(u) + o(1)\\ &\overset{\mathrm{d}}{=} \widetilde{\mathcal{B}}_a\left(\frac{F(w)}{2\lambda f(w)}\left(1 - e^{-2t\, h_F(w)}\right)\right)\\ &\Rightarrow \widehat{W}_2(\infty) \overset{\mathrm{d}}{=} \mathcal{N}\left(0, \frac{F(w)}{2\lambda f(w)}\right), \qquad \text{as} \quad t \to \infty.\end{aligned}$$

Next, (4.22) and (4.16) imply that

$$\begin{aligned}\mathrm{Var}(\widehat{W}_3(t)) &= \frac{1}{\lambda^2 F^c(w)^2}\mathrm{Var}\left(\int_0^t e^{-h_F(w)(t-u)}\, \mathrm{d}\widehat{E}(u)\right)\\ &= \frac{1}{\lambda^2 F^c(w)^2}\mathrm{Var}\left(-e^{-h_F(w)t}\widetilde{E}(-t) - \int_0^t h_F(w) e^{-h_F(w)s}\widetilde{E}(-s)\mathrm{d}s\right)\\ &= \frac{1}{\lambda^2 F^c(w)^2}\Bigg[e^{-2h_F(w)t}\widetilde{C}(-t,-t)\\ &\quad + 2h_F(w)e^{-h_F(w)t}\int_0^t e^{-h_F(w)s}\widetilde{C}(-s,-t)\mathrm{d}s\\ &\quad + 2h_F(w)^2\int_0^t\int_0^x e^{-h_F(w)(x+y)}\widetilde{C}(-x,-y)dy\mathrm{d}x\Bigg], \end{aligned} \tag{A.13}$$

where the second equality follows from (4.24). Note that $\widetilde{C}(-t,-s) = \mathrm{Cov}(\widetilde{E}(-s), \widetilde{E}(-t)) \leq \sqrt{\mathrm{Var}(\widetilde{E}(t))\mathrm{Var}(\widetilde{E}(s))} = \sqrt{\mathrm{Var}(\widehat{E}(t))\mathrm{Var}(\widehat{E}(s))}$. As $\mathrm{Var}(\widehat{E}(t)) = O(t^2)$ as $t \to \infty$, we can conclude that $\widetilde{C}(-t,-s) = O(st)$ as $s, t \to \infty$. As a result, the first term in (A.13) is $O(e^{-2h_F(w)t}t^2) \to 0$ and the second term is $O(e^{-h_F(w)t}t^2) \to 0$ as $t \to \infty$. Hence, we conclude that

$$\widehat{W}_3(t) \Rightarrow \widehat{W}_3(\infty) \overset{\mathrm{d}}{=} \mathcal{N}(0, \sigma^2_{W_3}), \quad \text{as} \quad t \to \infty,$$

where

$$\sigma^2_{W_3} \equiv 2\frac{h_F(w)^2}{\lambda^2 F^c(w)^2}\int_0^\infty\int_0^x e^{-h_F(w)(x+y)}\widetilde{C}(-x,-y)\mathrm{d}y\mathrm{d}x, \tag{A.14}$$

and $\widetilde{C}(\cdot,\cdot)$ is as defined in Lemma 4.1.

Finally, by independence, we conclude that

$$\widehat{W}(t) \Rightarrow \widehat{W}(\infty) \equiv \widehat{W}_1(\infty) + \widehat{W}_2(\infty) + \widehat{W}_3(\infty) \overset{\mathrm{d}}{=} \mathcal{N}\left(0, \sigma^2_W\right), \quad \text{as} \quad t \to \infty,$$
$$\text{where} \quad \sigma^2_W \equiv \frac{c^2_\lambda}{2h_F(w)\lambda} + \frac{F(w)}{2\lambda f(w)} + \sigma^2_{W_3}.$$

**Steady state of $\widehat{Q}$** We next characterize the steady state for the queue-length process.

$$\begin{aligned}\widehat{Q}_1(t) &= c_\lambda \int_{t-w}^{t} F^c(t-s)\mathrm{d}\mathcal{B}_\lambda(\Lambda(s)) \stackrel{\mathrm{d}}{=} c_\lambda\sqrt{\lambda}\int_0^w F^c(w-s)\mathrm{d}\mathcal{B}_\lambda(t-w+s)\\ &\stackrel{\mathrm{d}}{=} c_\lambda\sqrt{\lambda}\int_0^w F^c(w-s)\mathrm{d}\mathcal{B}_\lambda(s) \Rightarrow \widehat{Q}_1(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}(0,\sigma^2_{Q_1}), \quad \text{as} \quad t\to\infty,\\ &\text{where} \quad \sigma^2_{Q_1} \equiv \lambda\, c^2_\lambda \int_0^w F^c(u)^2\mathrm{d}u. \end{aligned} \tag{A.15}$$

Next, the expression in (5.18) implies that, as $t\to\infty$,

$$\widehat{Q}_2(t) \stackrel{\mathrm{d}}{=} \mathcal{B}_a\left(\int_0^{w(t)} F^c(u)F(u)\lambda\,\mathrm{d}u\right) \Rightarrow \widehat{Q}_2(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}(0,\sigma^2_{Q_2}),$$

where $\sigma^2_{Q_2} \equiv \lambda\int_0^w F(u)F^c(u)\,\mathrm{d}u$. Finally, (4.10) yields that

$$\begin{aligned}\widehat{Q}_3(t) = \lambda\, F^c(w)\,\widehat{W}(t) &\Rightarrow \widehat{Q}_3(\infty) \equiv \lambda\, F^c(w)\,\widehat{W}(\infty)\\ &\stackrel{\mathrm{d}}{=} \mathcal{N}\left(0,\sigma^2_{Q_3}\right), \quad \text{as} \quad t\to\infty,\end{aligned}$$

where $\sigma^2_{Q_3} \equiv \lambda^2 F^c(w)^2\sigma^2_W$.

The independence of $\widehat{Q}_1$,$\widehat{Q}_2$ and $\widehat{Q}_3$ yields

$$\widehat{Q}(t) \Rightarrow \widehat{Q}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}\left(0,\sigma^2_Q\right), \quad \text{as} \quad t\to\infty, \qquad \text{where} \quad \sigma^2_Q \equiv \sigma^2_{Q_1}+\sigma^2_{Q_2}+\sigma^2_{Q_3}.$$

### A.6 Proof of Corollary 4.2

Remaining service times are exponentially distributed due to lack of memory if the service-time distribution is exponential. Consequently, service completions at each server are a Poisson process with constant rate $\mu > 0$, which implies by [35] that the sequence $\widehat{E}_n$ converges to a centered Gaussian process with covariance function $C_E(s,t) = \mu(s\wedge t)$ for $s,t\geq 0$. Then (4.23) becomes $\widetilde{C}(-x,-y) = \mu(x\vee y) - \mu|x-y|$ for $x\geq 0$, $y\geq 0$. Consequently, (A.14) becomes

$$\begin{aligned}\sigma^2_{W_3} &= 2\frac{h_F(w)^2}{\lambda^2 F^c(w)^2}\int_0^\infty\int_0^x e^{-h_F(w)(x+y)}\mu y\,\mathrm{d}y\mathrm{d}x\\ &= 2\frac{h_F(w)^2}{\lambda^2 F^c(w)^2}\int_0^\infty \mu e^{-h_F(w)x}\int_0^x ye^{-h_F(w)y}\,\mathrm{d}y\mathrm{d}x\\ &= 2\frac{h_F(w)^2}{\lambda^2 F^c(w)^2}\int_0^\infty \mu e^{-h_F(w)x}\left(-\frac{x}{h}e^{-h_F(w)x}+\frac{1}{h^2}\left(1-e^{-h_F(w)x}\right)\right)\mathrm{d}x\\ &= 2\frac{h_F(w)^2}{\lambda^2 F^c(w)^2}\left(\frac{-\mu}{h_F(w)}\int_0^\infty xe^{-2h_F(w)x}\mathrm{d}x\right.\end{aligned}$$

$$
\begin{aligned}
&+\frac{\mu}{h_F(w)^2}\int_0^\infty e^{-h_F(w)x}\left(1-e^{-h_F(w)x}\right)\mathrm{d}x\Bigg)\\
&=2\frac{h_F(w)^2}{\lambda^2 F^c(w)^2}\left(\frac{-\mu}{2h_F(w)^2}\frac{1}{2h_F(w)}+\frac{\mu}{h_F(w)^3}-\frac{\mu}{2h_F(w)^3}\right)=\frac{1}{2\lambda f(w)}.
\end{aligned}
$$

Summing up $\sigma^2_{W_3}$ with $\sigma^2_{W_i}(\infty)$ for $i=1,2$ yields (4.27).

The variance $\sigma^2_{W_3}$ for the $M/M/n+M$ queue can be immediately obtained by letting $c^2_\lambda=1$ and $h_F(w)=\theta$ in (4.27). Finally, we obtain $\sigma^2_Q$ in (4.28) as follows:

$$
\begin{aligned}
\sigma^2_Q(\infty)&=\lambda\int_0^w F^c(u)^2\,\mathrm{d}u+\lambda\int_0^w F(u)F^c(u)^2\,\mathrm{d}u+\lambda^2F^c(w)^2\sigma^2_W\\
&=\lambda\int_0^w F^c(u)\,\mathrm{d}u+\lambda^2F^c(w)^2\sigma^2_W=\frac{\lambda}{\theta}(1-e^{-\theta w})+\frac{\lambda}{\theta}\cdot\frac{1}{\rho}=\frac{\lambda}{\theta},
\end{aligned}
$$

where the last equality holds because $w=F^{-1}(1-1/\rho)$, so that $1-1/\rho=F(w)=1-e^{-\theta w}$. □

## A.7 Proof of Lemma 5.1

### *A.7.1 Proof of the convergence in* (5.2)

We consider the modified processes $\widehat{E}'_{n,1}(t)$ given below. We first prove convergence for the sequence $\widehat{E}'_{n,1}$ and then show that the difference between the modified sequence $\widehat{E}'_{n,1}$ and the desired sequence $\widehat{E}_{n,1}$ is asymptotically negligible (see (A.18)), which proves the desired convergence in (5.2).

Now define, for $t\ge 0$,

$$
\begin{aligned}
\widehat{E}'_{n,1}(t)&\equiv\frac{1}{\sqrt{n}}E'_{n,1}(t)=\int_0^{t-w(t)}F^c(v(s))\,\mathrm{d}\widehat{N}_n(s)\\
&=F^c(v(t-w(t)))\widehat{N}_n(t-w(t))-F^c(v(0))\widehat{N}_n(0)\\
&\quad-\int_0^{t-w(t)}\widehat{N}_n(s-)\,\mathrm{d}F^c(v(s))\\
&=F^c(w(t))\widehat{N}_n(t-w(t))-\widehat{N}_n(0)-\int_0^t\widehat{N}_n(s-w(s))\,\mathrm{d}F^c(w(s)).
\end{aligned}
\tag{A.16}
$$

The second equality holds by integration by parts. The last equality follows from (4.3). Next we define a mapping $\psi:\mathcal{D}\to\mathcal{D}$ such that, for $z\in\mathcal{D}$,

$$
\psi(z)(t)\equiv F^c(w(t))z(t)-z(0)-\int_0^t z(s)\,\mathrm{d}F^c(w(s)),\quad 0\le t\le T.
$$

We now prove that the mapping $\psi$ is continuous in $\mathcal{D}$. Let $\{x_n\}$ be a sequence in $\mathcal{D}$ such that $\|x_n - x\|_T \to 0$. Then

$$\begin{aligned}
&|\psi(x_n)(t) - \psi(x)(t)| \\
&= \left| F^c(w(t))x_n(t) - x_n(0) - \int_0^t x_n(s)\,\mathrm{d}F^c(w(s)) \right. \\
&\quad \left. - F^c(w(t))x(t) + x(0) + \int_0^t x(s)\,\mathrm{d}F^c(w(s)) \right| \\
&\le F^c(w(t))|x_n(t) - x(t)| + |x_n(0) - x(0)| \\
&\quad + \|x_n - x\|_T \left| \int_0^t \mathrm{d}F^c(w(s)) \right| \le 4\,\|x_n - x\|.
\end{aligned}$$

Hence the mapping $\psi$ is continuous. In general, proving convergence with respect to the uniform topology does not necessarily imply $J_1$ convergence because there may be measurability issues (see, for example, [6,36]). However, we will be interested in the case where the limit $x$ is continuous, i.e., $x \in \mathcal{C}$. Therefore, we will not have any measurability issues and obtain the desired convergence in $\mathcal{D}$ with respect to Skorokhod's $J_1$ metric.

Convergence of the modified process in (A.16) follows by the continuous mapping theorem with composition. In particular, let $Z_n(t) \equiv \widehat{N}_n(t - W_n(t))$. Then $Z_n : [0, T] \to \mathbb{R}$ and $Z_n \Rightarrow Z$, where $Z(t) \equiv \widehat{N}(t - w(t))$. Convergence of $\{Z_n\}$ follows from the continuous mapping theorem with composition. Then we have $n^{-1/2}\widetilde{E}_{n,1}(t) = \psi(Z_n)(t) \Rightarrow \psi(Z)(t)$ in $\mathcal{D}$ with

$$\begin{aligned}
\psi(Z) \equiv \int_0^t F^c(w(s))\,\mathrm{d}\widehat{N}(s - w(s)) &\equiv F^c(w(t))\widehat{N}(t - w(t)) - \widehat{N}(0) \\
&- \int_0^t \widehat{N}(s - w(s))\,\mathrm{d}F^c(w(s)) \\
&= F^c(w(t))c_\lambda \mathcal{B}_\lambda(\Lambda(t - w(t))) - c_\lambda \mathcal{B}_\lambda(0) \\
&- \int_0^t c_\lambda \mathcal{B}_\lambda(\Lambda(s - w(s)))\,\mathrm{d}F^c(w(s)).
\end{aligned}$$

Finally, to establish (5.2), we show that the difference between the processes $n^{-1/2}E_{n,1}(t)$ and $n^{-1/2}E^{'}_{n,1}(t)$ is asymptotically negligible. In particular,

$$\begin{aligned}
&\left|\widehat{E}_{n,1}(t) - \widehat{E}^{'}_{n,1}(t)\right| \qquad\qquad (A.17)\\
&= \frac{1}{\sqrt{n}} \left| \int_0^{t - W_n(t)} F^c(V_n(s-))\,\mathrm{d}\widehat{N}_n(s) - \int_0^{t - w(t)} F^c(v(s))\,\mathrm{d}\widehat{N}_n(s) \right| \\
&\le \frac{1}{\sqrt{n}} \left| \int_{t - w(t)}^{t - W_n(t)} F^c(V_n(s-))\,\mathrm{d}\widehat{N}_n(s) \right| \\
&\quad + \frac{1}{\sqrt{n}} \int_0^{t - w(t)} \left| F^c(V_n(s-)) - F^c(v(s)) \right| \,\mathrm{d}\widehat{N}_n(s)
\end{aligned}$$

$$\leq \frac{1}{\sqrt{n}} \left| \widehat{N}_n(t - W_n(t)) - \widehat{N}_n(t - w(t)) \right| + \frac{1}{\sqrt{n}} \left| \widehat{N}_n(t - w(t)) - \widehat{N}_n(0) \right| \Rightarrow 0. \tag{A.18}$$

### *A.7.2 Proof of the convergence in* (5.3)

To prove convergence in (5.3), we apply the martingale FCLT in [32] (see also [9,14] for applications of the martingale FCLT). First we define a sequence of discrete-time processes (see (A.19)) and argue that it is a sequence of martingales adapted to a specific filtration $\mathscr{H}_k^n$ as defined below. Next, we define continuous-time martingales using the discrete-time martingales in (A.19). Then we invoke Theorem 7.1.4. on p.339 in [10] to establish convergence and characterize the limit.

Consider the discrete-time processes

$$\widehat{H}_k^n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \left( \mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n) \right) \quad \text{for} \quad k = 1, 2, \ldots. \tag{A.19}$$

Also, consider the filtration $\mathscr{H}_k^n \equiv \sigma\{\tau_{i+1}^n, \nu_i^n, \gamma_i^n : 1 \leq i \leq k\}$. Then $\mathbb{E}[|\widehat{H}_k^n|] \leq k/\sqrt{n}$ and

$$\mathbb{E}[H_k^n - H_{k-1}^n | \mathscr{H}_{k-1}^n] = \frac{1}{\sqrt{n}} \left( \mathbb{E}[\mathbf{1}(\gamma_k^n > w_k^n) | \mathscr{H}_{k-1}^n] - F^c(w_k^n) \right) = 0,$$

which implies that the process $\{(\widehat{H}_k^n, \mathscr{H}_k^n) : k \geq 1\}$ is a discrete-time martingale for each $n \geq 1$.

Our next step is to replace $k$ with $\lfloor nt \rfloor$ for $t \geq 0$ to obtain a continuous-time martingale. By a direct application of Lemma 4.2 of [9], we deduce that the continuous-time process $(\widehat{H}^n(t), \mathscr{H}^n(t) : t \geq 0) \equiv (\widehat{H}_{\lfloor nt \rfloor}^n, \mathscr{H}_{\lfloor nt \rfloor}^n : t \geq 0)$ is a martingale with quadratic variation

$$\langle \widehat{H}^n \rangle(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \left( \mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n) \right)^2. \tag{A.20}$$

We next show that the sequence of martingales $(\widehat{H}^n(t), \mathscr{H}^n(t) : t \geq 0)$ satisfies the conditions of Theorem 7.1.4. of [10]. In particular, it is required that $(i)$ jumps of the processes $\widehat{H}^n(y)$ are asymptotically negligible and $(ii)$ the quadratic quadratic variation of the processes converges in probability to a limit characterized in Theorem 7.1.1. of [10].

$(i)$ *Negligibility of jumps.* We now show that condition $(a)$ of Theorem 7.1.4. holds. Let $\widehat{H}^n(t-) \equiv \lim_{s \uparrow t} \widehat{H}^n(s)$. Then, for each $T > 0$, we have $\sup_{0 \leq t \leq T} |\widehat{H}^n(t) - \widehat{H}^n(t-)| \leq 1/\sqrt{n}$ and hence

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{0\le t\le T} |\widehat{H}^n(t) - \widehat{H}^n(t-)|\right] = 0,$$

which is the desired condition.

(*ii*) *Convergence of quadratic variations.* We now prove that the quadratic variation processes given in (A.20) converges in the $L^2$ sense as $n \to \infty$. In particular,

$$\begin{align*}
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}\left(\mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n)\right)^2 - \int_0^{\Lambda^{-1}(t)} F^c(v(u))F^c(v(u))\,\mathrm{d}\Lambda(u)\right)^2\right]\\
&\le 2\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}\left[\left(\mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n)\right)^2 - F^c(w_i^n)F(w_i^n)\right]\right)^2\right]\\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}\left[F^c(w_i^n)F(w_i^n) - F^c(v(\tau_i^n-))F(v(\tau_i^n-))\right]\right)^2\right]\\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}F^c(v(\tau_i^n-))F(v(\tau_i^n-))\right.\right.\\
&\quad \left.\left. - \int_0^{\Lambda^{-1}(t)} F^c(v(u-))F(v(u-))\,\mathrm{d}\Lambda(u)\right)^2\right]\\
&\le \frac{2}{n^2}\sum_{i=1}^{\lfloor nt\rfloor}\mathbb{E}\left[\left(\mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n)\right)^2\left(F(w_i^n) - F^c(w_i^n)\right)^2\right]\\
&\quad + \frac{2}{n^2}\mathbb{E}\sum_{i\ne j}\left[\left(\mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n)\right)\left(\mathbf{1}(\gamma_j^n > w_j^n) - F^c(w_j^n)\right)\right.\\
&\quad \left.\times\left(F(w_i^n) - F^c(w_i^n)\right)\left(F(w_j^n) - F^c(w_j^n)\right)\right]\\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}\left[F^c(w_i^n)F(w_i^n) - F^c(v(\tau_i^n-))F(v(\tau_i^n-))\right]\right)^2\right] \tag{A.21}\\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}F^c(v(\tau_i^n-))F(v(\tau_i^n-))\right.\right.\\
&\quad \left.\left. - \int_0^{\Lambda^{-1}(t)} F^c(v(u)-)F(v(u)-)\,\mathrm{d}\Lambda(u)\right)^2\right]. \tag{A.22}
\end{align*}$$

The first sum vanishes as $n \to \infty$ because the summands are bounded by 1 and, therefore, the first term is bounded by $2\lfloor nt \rfloor / n^2 \to 0$ as $n \to \infty$. The summands of the second term are independent. Therefore, the second term is equal to 0.

To prove convergence of (A.21), we first rewrite the summands of (A.21) as

$$\begin{aligned} F^c(w_i^n)F(w_i^n) - F^c(v(\tau_i^n-))F(v(\tau_i^n-)) &= F^c(w_i^n) - F^c(v(\tau_i^n-)) \\ &- \left(F^c(w_i^n)^2 - F^c(v(\tau_i^n-))^2\right), \end{aligned} \tag{A.23}$$

Next we make use of the FWLLN for PWT $V_n(t)$, i.e., $V_n \Rightarrow v$ in $\mathcal{D}$, and continuity of the function $F$ to show that (A.21) converges to 0. In particular, for all $i \geq 1$,

$$F^c(w_i^n) = F^c(V_n(\tau_i^n-)) = F^c(v(\tau_i^n-) + o(1)).$$

Combining with (A.23), this implies that the summands in (A.21) can be bounded above by

$$\begin{aligned} &|F^c(v(\tau_i^n-) + o(1)) - F^c(v(\tau_i^n-))| + |F^c(v(\tau_i^n-) + o(1))^2 \\ &\quad - F^c(v(\tau_i^n-))^2| \leq |o(1)|, \end{aligned}$$

where the inequality holds by continuity of cdf $F$. This implies that the squared sum inside the expectation in (A.21) is bounded above by $(|o(1)|\lfloor nt \rfloor/n)^2 \leq t^2|o(1)| = o(1)$ for all $t \geq 0$. Convergence of (A.21) to 0 then follows from the dominated convergence theorem.

The summation in (A.22) can be alternatively represented as

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} F^c(v(\tau_i^n-))F(v(\tau_i^n-)) &= \int_0^{\Lambda_n^{-1}(t)} F^c(v(u-))F(v(u-))\, d\bar{N}_n(u) \\ &\Rightarrow \int_0^{\Lambda^{-1}(t)} F^c(v(u-))F(v(u-))\, \mathrm{d}\Lambda(u), \end{aligned} \tag{A.24}$$

where the convergence (A.24) follows from the continuous mapping theorem. Having established the convergence in (A.24), convergence in mean square is obtained by first applying the continuous mapping theorem with the function $f(x) = x^2$ and then applying the dominated convergence theorem by using the fact that both the summation and the limit integral in (A.24) are bounded by $t$. Hence (A.22) converges to 0. That completes the proof of convergence of the quadratic variation (A.20).

Having proved conditions $(i)$ and $(ii)$ are indeed satisfied, by Theorem 7.1.4 of [10], we deduce that $\widehat{H}^n \Rightarrow \widehat{H}$ in $\mathcal{D}$, where $\widehat{H}$ is a Gaussian process with independent increments and continuous sample paths. Moreover, as implied by the proof of Theorem 7.1.1. of [10], the limit $\widehat{H}$ is indeed a time-changed Brownian motion, where the time change is the limit of the quadratic variation, i.e.,

$$\widehat{H}(t) = \mathcal{B}_a(\langle \widehat{H} \rangle (t)) = \mathcal{B}_a\left( \int_0^{\Lambda^{-1}(t)} F^c(v(u-))F(v(u-))\,\mathrm{d}\Lambda(u) \right),$$

where $\mathcal{B}_a$ is the standard Brownian motion.

Finally, to complete the proof, we note that $\widehat{E}_{n,2}(t) = \widehat{H}^n(\bar{N}_n(t - W_n(t)))$. Then, by the convergence-together theorem, we have $\widehat{H}^n(\bar{N}_n) \Rightarrow \widehat{H}(\Lambda)$ in $\mathcal{D}$. Consequently, as $n \to \infty$,

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{N_n(t-W_n(t))} \left( \mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n) \right) \\ &\quad \Rightarrow \mathcal{B}_a\left( \int_0^{t-w(t)} F^c(v(u-))F(v(u-))\,\mathrm{d}\Lambda(u) \right). \end{aligned} \tag{A.25}$$

We next verify the other two expressions in (5.3). The last expression is obtained by a change of variable with $u = s - w(s)$. (Note that, according to (4.3), we have $v(s - w(s)) = w(s)$.) The Kiefer integral expression holds because it is a Gaussian process with zero mean and the same covariance function as the Brownian expression. Specifically, for $t, t' > 0$, the first Brownian expression has the covariance

$$\int_0^{(t-w(t))\wedge(t'-w(t'))} F^c(v(u))F(v(u))\,\mathrm{d}\Lambda(u). \tag{A.26}$$

On the other hand, the Kiefer integral in (5.3) has the covariance

$$\begin{aligned} &\mathbb{E}\left[ \int_0^{t-w(t)} \int_0^1 \mathbf{1}(y > F(v(s)))\mathrm{d}\widehat{U}(s, y) \right. \\ &\quad \left. \times \int_0^{t'-w(t')} \int_0^1 \mathbf{1}(y > F(v(s)))\mathrm{d}\widehat{U}(s, y) \right] \\ &= \mathbb{E}\left[ \int_0^{t-w(t)} \int_0^\infty \mathbf{1}(x > v(s))\mathrm{d}\widehat{U}(s, F(x)) \right. \\ &\quad \left. \times \int_0^{t'-w(t')} \int_0^\infty \mathbf{1}(x > v(s))\mathrm{d}\widehat{U}(s, F(x)) \right] \\ &= \int_0^{(t-w(t))\wedge(t'-w(t'))} \int_0^\infty \mathbf{1}(x > v(s))\mathrm{d}F(x)\mathrm{d}\Lambda(s) \\ &\quad + \int_0^{(t-w(t))\wedge(t'-w(t'))} F^c(v(s))F^c(v(s))\mathrm{d}\Lambda(s) \\ &\quad - 2\int_0^{(t-w(t))\wedge(t'-w(t'))} \int_0^\infty \mathbf{1}(x > v(s))F^c(v(s))\mathrm{d}F(x)\mathrm{d}\Lambda(s) \end{aligned}$$

$$
\begin{aligned}
&= \int_0^{(t-w(t))\wedge(t'-w(t'))} F^c(v(s))\mathrm{d}\Lambda(s) + \int_0^{(t-w(t))\wedge(t'-w(t'))} \left(F^c(v(s))\right)^2 \mathrm{d}\Lambda(s) \\
&\quad - 2\int_0^{(t-w(t))\wedge(t'-w(t'))} F^c(v(s))F^c(v(s))\mathrm{d}\Lambda(s) \\
&= \int_0^{(t-w(t))\wedge(t'-w(t'))} F^c(v(s))\mathrm{d}\Lambda(s) - \int_0^{(t-w(t))\wedge(t'-w(t'))} \left(F^c(v(s))\right)^2 \mathrm{d}\Lambda(s) \\
&= \int_0^{(t-w(t))\wedge(t'-w(t'))} F^c(v(s))\left(1 - F^c(v(s))\right)\mathrm{d}\Lambda(s),
\end{aligned}
$$

which coincides with (A.26).

## B Refined staffing levels

In this section, we consider a refined staffing function given by

$$s_n \equiv \lceil ns_1 + \sqrt{n}s_2 \rceil, \quad \text{where} \quad s_1,\ s_2 > 0. \tag{B.1}$$

The general form of (B.1) enables us to recover two of the staffing functions considered in [30] that, respectively, lead to the ED and ED+QED operating regimes. More specifically, the two staffing functions in [30] are given by

$$n_{\text{ED}} = \lceil (1-\gamma)R_n \rceil, \tag{B.2}$$

$$n_{\text{ED+QED}} = \lceil (1-\gamma)R_n + \delta\sqrt{R_n} \rceil, \tag{B.3}$$

where $0 < \gamma < 1$ and $R_n$ is the offered load, defined as $R_n = n\lambda/\mu$. Letting $s_1 = (1-\gamma)\lambda/\mu$ and $s_2 = 0$ yields (B.2), whereas letting $s_1 = (1-\gamma)\lambda/\mu$ and $s_2 = \delta\sqrt{\lambda/\mu}$ yields (B.3). See also [21] and Sect. 10 in [26] for time-varying versions of the refined staffing (B.1).

We next briefly discuss the changes resulting from considering the staffing function $s_n$ instead of $n$. In the previous sections, the staffing function happens to coincide with our scaling factor $n$, i.e., $s_n = n$. In this section, we let $n$ and $\sqrt{n}$ be the scaling factors for FWLLN and FCLT, respectively, and let the staffing function have a more general form $s_n = \lceil ns_1 + \sqrt{n}s_2 \rceil$, where $s_1, s_2 > 0$. To indicate the processes associated with the new staffing function, we use a superscript $r$, whereas to indicate the processes associated the case where $s_n = n$, we use notation without a superscript. Because the arrival process is independent of the staffing level, it holds that $\widehat{N}_n^r(t) = \widehat{N}_n(t)$ for all $n \geq 1$ and $t \geq 0$, and hence, $\widehat{N}^r(t) = \widehat{N}(t)$. The FWLLN limit and the FCLT limit for the service-completion process, on the other hand, becomes $D^r(t) = s_1 D(t)$, and $\widehat{D}^r(t) = \sqrt{s_1}\widehat{D}(t) + s_2 D(t)$, respectively, where $\widehat{D}(t)$ is a centered Gaussian process with covariance function $C_E$ in Theorem 4.2, and $D(t) = \mu t$ as in (4.4). Hence, $\widehat{D}^r(t)$ is a Gaussian process with covariance function $C^r(\cdot,\cdot) = s_1 C_E(\cdot,\cdot)$ and mean $s_2\mu t$. Consequently, the enter-service process satisfies $\widehat{E}^r(t) = \sqrt{s_1}\widehat{D}(t) + s_2\mu t$.

The following theorem is an analog of Theorems 4.1 and 4.2 for the $G/GI/n+GI$ model having the refined staffing function $s_n$ in (B.1).

**Theorem B.1** (FWLLN and FCLT with refined staffing) *Consider the $G/GI/n+GI$ with staffing level $s_n$ given by* (B.1) *and $\rho^r = \lambda/\mu s_1 > 1$.*

(i) *Under the conditions of Theorem* 4.1*, an analog of joint convergence in* (4.1) *holds as $n \to \infty$, where $\Lambda^r(t) = \lambda t$,*

$$\begin{aligned} &D^r(t) = E^r(t) = s_1\mu t, \quad w^r(t) = \int_0^t \left(1 - \frac{s_1\mu}{\lambda F^c(w^r(u))}\right) \mathrm{d}u, \\ &v^r(t) = w^r(t + v^r(t)). \end{aligned} \tag{B.4}$$

*The limits $Q^r(t)$, $X^r(t)$ and $A^r(t)$ have the same mathematical form as their counterparts in Theorem* 4.1 *with modified components.*

(ii) *Under the conditions of Theorem* 4.1*, an analog of joint convergence in* (4.5) *holds as $n \to \infty$, where*

$$\begin{aligned} \widehat{W}^r(t) = &\int_0^t \frac{F^c(w^r(u))H^r(t,u)}{q(t, w^r(t))} c_\lambda \, \mathrm{d}\mathcal{B}_\lambda(\Lambda(u - w^r(u))) \\ &+ \int_0^t \frac{\sqrt{\lambda F^c(v^r(u))F(v^r(u))}H^r(t,u)}{q(t, w^r(t))} \, \mathrm{d}\mathcal{B}_a(u) \\ &- \sqrt{s_1}\int_0^t \frac{H^r(t,u)}{q(t, w^r(t))} \, d\hat{E}(u) - s_2\mu \int_0^t \frac{H^r(t,u)}{q(t, w^r(t))} \, \mathrm{d}u, \end{aligned} \tag{B.5}$$

*$w^r(t)$ and $v^r(t)$ are as in* (B.4)*, $H^r(\cdot,\cdot)$ and $q(\cdot, w^r(\cdot))$ are as in* (4.17) *with $w(t)$ replaced by $w^r(t)$. The virtual waiting time $\widehat{V}^r(t)$ and the queue-length process $\widehat{Q}^r(t)$ have the same mathematical forms as in* (4.9) *and* (4.10)*, with $w(t)$, $v(t)$ and $\widehat{W}(t)$ replaced by their counterparts $w^r(t)$, $v^r(t)$ and $\widehat{W}^r(t)$. The FCLT limit for the abandonment process is $\widehat{A}^r(t) = \widehat{N}(t) - \widehat{Q}^r(t) - \widehat{E}^r(t)$.*

*Proof of Theorem B.1* The proof closely follows the arguments in the proofs of Theorem 4.1, Theorem 4.2, Corollary 4.1 and Theorem 4.4. Therefore, we mostly refer to proofs of those results in the proofs below and argue in what way the new staffing function $s_n = \lceil ns_1 + \sqrt{n}s_2 \rceil$ changes the arguments. We skip lengthy details. Throughout this subsection, the processes with a superscript $r$ correspond to those associated with staffing level $s_n$, whereas the processes without a superscript $r$ correspond to those associated with staffing level $n$.

The LLN- and CLT-scaled departure process

$$\bar{D}_n^r(t) \equiv \frac{\sum_{j=1}^{s_n} D_j(t)}{n} = \frac{s_n}{n} \cdot \frac{\sum_{j=1}^{s_n} D_j(t)}{s_n} \Rightarrow D^r(t) \equiv s_1 D(t) = s_1\mu t, \tag{B.6}$$

$$\begin{aligned} \widehat{D}_n^r(t) &\equiv \frac{\sum_{j=1}^{s_n} D_j(t) - nD^r(t)}{\sqrt{n}} = \frac{\sqrt{s_n}}{\sqrt{n}} \cdot \frac{\sum_{j=1}^{s_n} D_j(t) - s_n\mu t}{\sqrt{s_n}} + \frac{s_n\mu t - nD^r(t)}{\sqrt{n}} \\ &= \sqrt{s_1 + \frac{s_2}{\sqrt{n}}} \cdot \frac{\sum_{j=1}^{s_n} D_j(t) - s_n\mu t}{\sqrt{s_n}} + s_2\mu t + O(1/\sqrt{n}) \Rightarrow \widehat{D}^r(t) \\ &\equiv \sqrt{s_1}\widehat{D}(t) - s_2\mu t, \end{aligned} \tag{B.7}$$

where $O(1/\sqrt{n})$ in the second equality accounts for the error caused by dropping $\lceil\cdot\rceil$ in $s_n$, and $\widehat{D}(t)$ is the Gaussian process in Theorem 4.2. Hence, we deduce from (B.7) that $\hat{D}^r(t)$ is a Gaussian process with negative drift $-s_2\mu t$ and covariance function $C^r(\cdot,\cdot) = s_1 C_E(\cdot,\cdot)$, with $C_E$ being the covariance function in Theorem 4.2.

Having obtained the modified fluid limits in (B.6) and established the joint convergence, we deduce that the proof in Sect. A.2 continues to hold with minor modifications. But the limit in (A.7) changes because the fluid limit of the departure process is now given by $D^r(t) = s_1\mu t$. Consequently, the ODE in Theorem 4.1 has $s_1\mu t$ in the numerator instead of $\mu t$.

Similarly, given the joint convergence $(\widehat{N}_n^r, \widehat{D}_n^r, \widehat{E}_n^r) \Rightarrow (\widehat{N}^r, \widehat{D}^r, \widehat{E}^r)$, we can prove the FCLT with a slightly modified proof. The arguments in Sect. 5 continue to hold for modified fluid limits and cause only minor changes in the final expressions. In particular, (5.4)–(5.7) have the same mathematical form with fluid limits and prelimit stochastic process replaced with their counterparts with a superscript $r$. Hence the steps of the proof in Sect. 5.1.2 can be replicated with counterpart processes. The only step that requires careful treatment is that the limit of the enter-service process is now $\widehat{E}^r(t) \equiv \sqrt{s_1}\widehat{E}(t) - s_2\mu t$. Since the additional term $-s_2\mu t$ is deterministic and $\sqrt{s_1}\widehat{E}(t)$ is a centered Gaussian process, we can use similar arguments to proof of Corollary 4.1 to deduce that (B.5) is indeed the desired solution. □

Note that (B.5) is different than (4.16) in that the third term on the right-hand side is scaled by $\sqrt{s_1}$ and that there is an additional deterministic term. This implies that both the variance and mean of HWT change and so do those of the PWT and queue length. The corresponding steady-state formulas are given in the following corollary.

**Corollary B.1** (Steady state of limits with refined staffing) *Under the assumptions of Theorem 4.4, the steady-state random variables $\hat{W}^r(\infty)$, $\hat{V}^r(\infty)$ and $\hat{Q}^r(\infty)$ have Gaussian distributions with means and variances given below:*

$$\mu_{W^r} \equiv \mathbb{E}\left[\widehat{W}^r(\infty)\right] = \mathbb{E}\left[\widehat{V}^r(\infty)\right] = -\frac{s_2\mu}{\lambda f(w^r)},$$

$$\mathbb{E}\left[\widehat{Q}^r(\infty)\right] = \lambda F^c(w^r)\mathbb{E}\left[\widehat{W}^r(\infty)\right] = -\frac{s_2\mu}{h_F(w^r)},$$

$$\mathrm{Var}(\widehat{W}^r(\infty)) = \mathrm{Var}(\widehat{V}^r(\infty)) \equiv \sigma^2_{W^r} \equiv \frac{c_\lambda^2}{2h_F(w^r)\lambda} + \frac{F(w^r)}{2\lambda f(w^r)} + s_1\sigma^2_{W_3^r},$$

$$\textit{where} \quad \sigma^2_{W_3^r} \equiv 2\frac{h_F(w^r)^2}{\lambda^2 F^c(w^r)^2} \times \int_0^\infty \int_0^x e^{-h_F(w^r)(x+y)}\widetilde{C}(-x,-y)\,\mathrm{d}y\mathrm{d}x,$$

*the covariance function $\widetilde{C}(-x,-y)$ is as in (4.23), and $w^r = F^{-1}(1-1/\rho^r)$. The variance of the steady-state queue length $\widehat{Q}(\infty)$ has the same mathematical form with $w$ and $\widehat{W}$ replaced by $w^r$ and $\widehat{W}^r$.*

*Remark B.1 (Optimal staffing problems)* Heavy-traffic FWLLN and FCLT limits have proven useful in solving optimal staffing problems with respect to service-level constraints in large scale service systems [4,30]. A general framework for this type of approach has two steps: First, a corresponding optimal staffing problem is formulated and solved using analytic FWLLN or FCLT limits (which are often more convenient than their corresponding stochastic versions); Next, an asymptotic optimality result is established by showing that the FWLLN- or FCLT-based optimal staffing problem is asymptotically equivalent to its desired stochastic version as the scale $n \to \infty$. We advocate that our new FCLT limit with refined staffing functions provides a basis for solving optimal staffing problems in $G/GI/n + GI$ queueing systems (note that two control factors $s_1$ and $s_2$ for the staffing function are preserved in the limit). For example, in the FCLT-based optimal staffing problem, we may choose the optimal $s_1^*$ and $s_2^*$ in order to minimize certain performance functions, for example, the mean waiting time, queue length, or abandonment probability; see the formulation in [4]. We leave this to future research.

*Proof of Corollary B.1* We first derive the mean of $\widehat{W}(\infty)$ from (B.5). Since the first three terms in (B.5) have zero means, $\mathbb{E}[\widehat{W}(\infty)]$ is the limit of the last term in (B.5) as $t \to \infty$:

$$\begin{aligned}\mathbb{E}[\widehat{W}^r(\infty)] &= \lim_{t\to\infty} -s_2\mu \int_0^t \frac{H^r(t,u)}{q(t,w^r(t))}\,\mathrm{d}u = \lim_{t\to\infty} -s_2\mu \int_0^t \frac{e^{-h_F(w^r)(t-u)}}{\lambda F^c(w^r)}\,\mathrm{d}u \\ &= \lim_{t\to\infty} \frac{-s_2\mu}{\lambda f(w^r)}\left(1 - e^{-h_F(w^r)t}\right).\end{aligned}$$

Having established $\mathbb{E}[\widehat{W}(\infty)]$, it is easy to establish $\mathbb{E}[\widehat{V}(\infty)]$ and $\mathbb{E}[\widehat{Q}(\infty)]$ by letting $t \to \infty$ in

$$\widehat{V}^r(t) = \frac{\widehat{W}^r(t)}{1 - \dot{w}^r(t + v^r(t))} \quad \text{and} \quad \widehat{Q}^r(t) = \lambda F^c(w^r(t))\widehat{W}^r(t).$$

Computation of variance is standard and as given in Sect. A.4. □

## References

1. Alòs, E., Mazet, O., Nualart, D.: Stochastic calculus with respect to Gaussian processes. Ann. Probab. **29**, 766–801 (2001)
2. Aras, A.K., Chen, X., Liu, Y.: Longer online appendix: many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment (2017). https://yunanliu.wordpress.ncsu.edu/files/2017/11/ArasLiuChenOLGGnGApp11282017.pdf
3. Aras, A.K., Liu, Y., Whitt, W.: Heavy-traffic limit for the initial content process. Stoch. Syst. **7**, 95–142 (2017)
4. Bassamboo, A., Randhawa, R.S.: On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. Oper. Res. **58**, 1398–1413 (2010)
5. Biagini, F., Hu, Y., Øksendal, B., Zhang, T.: Stochastic Calculus for Fractional Brownian Motion and Applications. Springer, Berlin (2008)
6. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley-Interscience, New York (1999)
7. Cox, D.R.: Renewal Theory. Methuen, London (1962)
8. Dai, J., He, S., Tezcan, T.: Many server diffusion limits for $G/Ph/n + GI$ queues. Ann. Appl. Probab. **20**, 1854–1890 (2010)
9. Dai, J.G., He, S.: Customer abandonment in many-server queues. Math. Oper. Res. **35**(2), 347–362 (2010)

10. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. Wiley, New York (1986)
11. Gamarnik, D., Goldberg, D.: Steady-state $GI/GI/n$ queue in the Halfin–Whitt regime. Ann. Appl. Probab. **23**, 2382–2419 (2012)
12. He, B., Liu, Y., Whitt, W.: Staffing a service system with non-Poisson nonstationary arrivals. Probab. Eng. Inf. Sci. **30**, 593–621 (2016)
13. He, S.: Diffusion approximation for efficiency-driven queues: a space–time scaling approach. Working paper, National University of Singapore
14. Huang, J., Mandelbaum, A., Zhang, H., Zhang, J.: Refined models for efficiency-driven queues with applications to delay announcements and staffing. Working paper (2017)
15. Kang, W., Ramanan, K.: Fluid limits of many-server queues with reneging. Ann. Appl. Probab. **20**, 2204–2260 (2010)
16. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus. Springer, Berlin (1988)
17. Kaspi, H., Ramanan, K.: SPDE limits of many-server queue. Ann. Appl. Probab. **23**, 145–229 (2013)
18. Krichagina, E.V., Puhalskii, A.A.: A heavy-traffic analysis of a closed queueing system with a $GI/\infty$ service center. Queueing Syst. **25**, 235–280 (1997)
19. Lebovits, J.: Stochastic calculus with respect to Gaussian processes. Working paper (2017)
20. Liu, R., Kuhl, M.E., Liu, Y., Wilson, J.R.: Modeling and simulation of nonstationary non-Poisson processes. INFORMS J. Comput. (forthcoming)
21. Liu, Y.: Staffing to stabilize the tail probability of delay in service systems with time-varying demand. Oper. Res. (forthcoming)
22. Liu, Y., Whitt, W.: A network of time-varying many-server fluid queues with customer abandonment. Oper. Res. **59**, 835–846 (2011)
23. Liu, Y., Whitt, W.: The $G_t/GI/s_t+GI$ many-server fluid queue. Queueing Syst. **71**, 405–444 (2012)
24. Liu, Y., Whitt, W.: A many-server fluid limit for the $G_t/GI/s_t+GI$ queueing model experiencing periods of overloading. Oper. Res. Lett. **40**, 307–312 (2012)
25. Liu, Y., Whitt, W.: Algorithms for time-varying networks of many-server fluid queues. INFORMS J. Comput. **26**, 59–73 (2013)
26. Liu, Y., Whitt, W.: Many-server heavy-traffic limits for queues with time-varying parameters. Ann. Appl. Probab. **24**, 378–421 (2014)
27. Liu, Y., Whitt, W., Yu, Y.: Approximations for heavily-loaded $G/GI/n+GI$ queues. Naval Res. Logist. **63**, 187–217 (2016)
28. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. Queueing Syst. **30**, 149–201 (1998)
29. Mandelbaum, A., Momcilovic, P.: Queues with many servers and impatient customers. Math. Oper. Res. **37**, 41–65 (2012)
30. Mandelbaum, A., Zeltyn, S.: Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Oper. Res. **57**, 1189–1205 (2009)
31. Pang, G., Whitt, W.: Two-parameter heavy-traffic limits for infinite-server queues. Queueing Syst. **65**, 325–364 (2010)
32. Pang, G., Whitt, W., Talreja, R.: Martingale proofs of many-server heavy-traffic limits for Markovian queues. Probab. Surv. **4**, 193–267 (2007)
33. Reed, J.: The $G/GI/N$ queue in the Halfin–Whitt regime. Ann. Appl. Probab. **19**, 2211–2269 (2009)
34. Rogers, L.C.G., Williams, D.: Diffusions, Markov Processes, and Martingales. Wiley, New York (1994)
35. Whitt, W.: Queues with superposition arrival process in heavy traffic. Stoch. Process. Appl. **21**, 81–91 (1985)
36. Whitt, W.: Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and there Application to Queues. Springer, Berlin (2002)
37. Whitt, W.: Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Manag. Sci. **50**, 1449–1461 (2004)
38. Whitt, W.: Fluid models for multiserver queues with abandonments. Oper. Res. **54**, 37–54 (2006)
39. Zhang, J.: Fluid models of many-server queues with abandonment. Queueing Syst. **73**, 147–193 (2013)