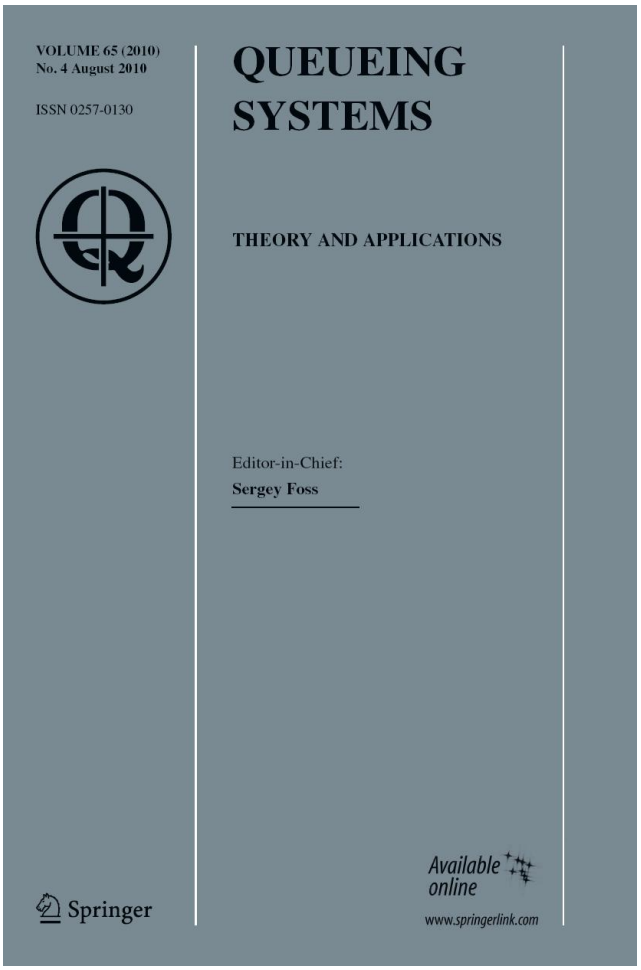
*ISSN 0257-0130, Volume 65, Number 4*




**This article was published in the above mentioned Springer issue.
The material, including all portions thereof, is protected by copyright;
all rights are held exclusively by Springer Science + Business Media.
The material is for personal use only;
commercial use is not permitted.
Unauthorized reproduction, transfer and/or use
may be a violation of criminal as well as civil law.**

Queueing Syst (2010) 65: 325–364
DOI 10.1007/s11134-010-9184-z
Author's personal copy

# Two-parameter heavy-traffic limits for infinite-server queues

**Guodong Pang · Ward Whitt**

Received: 27 December 2008 / Revised: 4 October 2009 / Published online: 27 July 2010
© Springer Science+Business Media, LLC 2010

**Abstract** In order to obtain Markov heavy-traffic approximations for infinite-server queues with general non-exponential service-time distributions and general arrival processes, possibly with time-varying arrival rates, we establish heavy-traffic limits for two-parameter stochastic processes. We consider the random variables $Q^e(t, y)$ and $Q^r(t, y)$ representing the number of customers in the system at time $t$ that have elapsed service times less than or equal to time $y$, or residual service times strictly greater than $y$. We also consider $W^r(t, y)$ representing the total amount of work in service time remaining to be done at time $t + y$ for customers in the system at time $t$. The two-parameter stochastic-process limits in the space $D([0, \infty), D)$ of $D$-valued functions in $D$ draw on, and extend, previous heavy-traffic limits by Glynn and Whitt (Adv. Appl. Probab. 23, 188–209, 1991), where the case of discrete service-time distributions was treated, and Krichagina and Puhalskii (Queueing Syst. 25, 235–280, 1997), where it was shown that the variability of service times is captured by the Kiefer process with second argument set equal to the service-time c.d.f.

**Keywords** Infinite-server queues · Heavy-traffic limits for queues · Markov approximations · Two-parameter processes · Measure-valued processes · Time-varying arrivals · Martingales · Functional central limit theorems · Invariance principles · Kiefer process

**Mathematics Subject Classification (2000)** Primary 60K25 · 60F17 · 60J25 · 60G60 · 60G44 · 60H05 · 60G57 · Secondary 90B22 · 60H99

G. Pang · W. Whitt (✉)
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA
e-mail: ww2040@columbia.edu

G. Pang
e-mail: gp2224@columbia.edu

## 1 Introduction

One reason heavy-traffic limits for queueing systems are useful is that they show that non-Markov stochastic processes describing system performance can be approximated by Markov stochastic processes under heavy loads. For a Markov process, it suffices to know the present state of that stochastic process in order to determine the distribution of the stochastic process at future times; we need no additional information from the past. With Markov approximations, that remains true approximately. In applications, the Markov property shows that the proper state has been identified and shows what needs to be measured in order to understand system performance.

The classic example is the conventional heavy-traffic limit for the $GI/GI/s$ queue, having $s$ servers, unlimited waiting room, and independent and identically distributed (i.i.d.) service times independent of a renewal arrival process. The standard description of system state is the number of customers in the system at time $t$, which we will call the queue length and denote by $Q(t)$. With non-exponential interarrival and service times, the stochastic process $\{Q(t)\colon t \geq 0\}$ is not Markov. Then the future evolution at any time depends on the elapsed interarrival time and the elapsed service times of all customers being served. However, the conventional heavy-traffic limit, in which the traffic intensity approaches the critical value 1 from below while the number of servers remains fixed, shows that the queue-length process $\{Q(t)\colon t \geq 0\}$ is approximately equal to a Markov process, in particular, reflected Brownian motion, under heavy loads [20, 21, 48]. In fact, the interarrival times and service times need not come from independent sequences of i.i.d. random variables. Instead, it suffices to have the associated partial sums, or equivalently, the associated counting processes satisfy a FCLT. Moreover, the Markov property of the limit extends to conventional heavy-traffic limits for networks of queues [17].

The situation is very different for many-server heavy-traffic limits when the service-time distribution is non-exponential, either with $s = \infty$ or $s \to \infty$. In this paper, we will consider the case in which $s = \infty$, i.e., the $G/GI/\infty$ model with i.i.d. service times independent of a general arrival process, where heavy traffic is achieved by letting $\lambda \to \infty$, while the service-time distribution is held fixed. However, the problem is relevant more generally with many servers, where $s \to \infty$ as $\lambda \to \infty$ with $s - \lambda = O(\sqrt{\lambda})$, as in [16]. For infinite-server models, we index the stochastic processes by the arrival rate $\lambda$. We are interested in the infinite-server model both for its own sake and as an approximation for many-server queues. In fact, heavy-traffic limits for infinite-server models can play a role in characterizing the heavy-traffic limits for corresponding many-server models, as shown by Reed [39, 40] and [32, 38].

With infinitely many exponential servers, we again obtain Markov diffusion limits, as first shown by Iglehart [18] for the $M/M/\infty$ model; see [37] for a review. A systematic way to extend the limit to general arrival processes is given in Sect. 7.3 of [37]. However, with non-exponential service times, the established heavy-traffic limit for $Q(t)$ is *not* Markov. As first shown by Borovkov [3], and further discussed in [14, 19, 27, 30, 47], the limit process is Gaussian, which implies that the distribution of $Q(t)$ itself is approximately normal, but the limiting Gaussian stochastic process is non-Markov, unless the service times are exponential (plus a minor additional case, [13, 27]).

We consider a stochastic process characterizing the system state for which the associated heavy-traffic limit process is Markov. We do so in two ways: First, we consider the two-parameter stochastic process $\{Q^e(t, y)\colon\ t \geq 0,\ \ y \geq 0\}$, where $Q^e(t, y)$ represents the number of customers in the system at time $t$ with elapsed service times less than or equal to time $y$. We do not pay attention to specific customers or servers but only count the total numbers. The random quantity $Q^e(t, y)$ is an *observable* quantity given the system history up to time $t$. We recommend that the stochastic process $\{Q^e(t, y)\colon\ t \geq 0,\ \ y \geq 0\}$ be used in models and measured in practice. Ways to exploit such ages for control are discussed in [9].

So far, we have used elapsed service times, because they are directly observable. We can equally well work with residual service times, and consider the process $Q^r(t, y)$ counting the number of customers in the system at time $t$ with residual service times strictly greater than $y$. With i.i.d. service times having c.d.f. $F$, we can go from one formulation to the other. If the elapsed service time is $y$, then the residual service time has distribution $F_y(x) \equiv F(x + y)/F^c(y)$ for $x \geq 0$, where $F^c(y) \equiv 1 - F(y)$. If the service times are learned when service begins, then both $Q^r(t, y)$ and $Q^e(t, y)$ are directly observable. Otherwise, elapsed service times correspond to what we observe, while residual service times represent the future load, whose distribution we may want to describe.

We regard $\{Q^e(t, \cdot)\colon\ t \geq 0\}$ and $\{Q^r(t, \cdot)\colon\ t \geq 0\}$ as function-valued stochastic processes, in particular, random elements of the function space $D_D$; see Sect. 2.3. Since the functions $Q^e(t, y)$ ($Q^r(t, y)$) are nondecreasing (nonincreasing) in $y$, we can also regard $Q^e(t, \cdot)$ and $Q^r(t, \cdot)$ as measure-valued processes, but we will work in the framework $D_D$.

For the $M/GI/\infty$ model, it is easy to see that the stochastic process $\{Q^e(t, \cdot) : t \geq 0\}$ is a Markov process; [10, 34] (where references to earlier work are given). The key idea, expressed in the proof of Theorem 1 of [10], is a Poisson-random-measure representation. For the more general $GI/GI/\infty$ model, having a non-Poisson renewal arrival process, the stochastic process $\{Q^e(t, \cdot)\colon\ t \geq 0\}$ is in general not Markov from that perspective, because the future evolution also depends on the elapsed interarrival time. The Markov property is violated more severely when the arrival process is not renewal. However, just as for the $G/GI/s$ model discussed above, the heavy-traffic limit for the arrival process typically does have independent increments, so this non-Markovian aspect disappears in the heavy-traffic limit. In the limit, $Q^e(t, y)$ for the $G/GI/\infty$ model is asymptotically equivalent to what it would be in the corresponding $M/GI/\infty$ model, except for a constant factor $c_a^2$ to account for the different variance; see Theorem 4.2 and Corollaries 4.1 and 4.2.

*Proof strategy* Our proof builds on previous work by Glynn and Whitt [14] and Krichagina and Puhalskii [27]. First, a restricted form of the desired two-parameter stochastic-process limit was already established in Theorem 3 of [14] for the case of service-time distributions with finite support. That result is only stated in $D$ for arbitrary fixed second parameter $y$, but it can be extended quite easily to the function space $D_D$. Since distributions with finite support are dense in the space of all probability distributions, one might consider the matter settled. However, much depends on the precise assumptions made about the service-time distribution. The goal

should be to treat general service-time distributions without any extra conditions. We should not need to assume that any moments are finite or that the c.d.f. is continuous or absolutely continuous.

One important feature of [27] is that they treat completely general service-time distributions. However, they do not state limits for two-parameter queueing processes. It might seem that it should be a routine extension to do so, but we show that is *not* so, because a candidate limit process is not a random element of the space $D_D$ for discontinuous service-time c.d.f.'s, as we explain in Remark 3.3. Fortunately, however, the argument in [27] can be extended to the two-parameter case if we restrict attention to continuous service-time c.d.f.'s, which we do.

A key idea in [14] is to treat service-time distributions with finite support by representing them as finite mixtures of deterministic service-time c.d.f.'s, and then split the arrival process into corresponding arrival processes associated with each deterministic service time; see Sect. 3 of [14], especially, Proposition 3.1. That step relies on the FCLT for split counting processes, as in Sect. 9.5 of [48]. The mixture argument extends quite directly to treat arbitrary discrete distributions. It also extends to arbitrary distributions if we can treat continuous service-time c.d.f.'s, but the proof in [14] does not seem to extend naturally to continuous service-time c.d.f.'s.

Hence, for the final case of a continuous service-time c.d.f., we draw heavily on [27]. Our limits for continuous service-time c.d.f.'s are extensions of theirs, obtained using the same function space and the same martingale arguments. The proof in [27] already took a two-parameter approach and, following Louchard [30], showed that it is fruitful to view the service times through the associated sequential empirical process (in (2.3) below). They showed that a scaled version of the sequential empirical process converges to the two-parameter standard Kiefer process, with the service time c.d.f. in the second argument (see (2.6) below). This convergence was established in the space $D_D$; see Sect. 2.3.

*Other related literature* As noted in [27], the relevance of the two-parameter Kiefer process for the infinite-server queue was first observed by Louchard [30]. The results here were briefly outlined in Sect. 6.4 in our survey [37]. (The first drafts of this paper were written at that time.) Related fluid limits for measure-valued processes have since been obtained in [23, 25, 53]. However, the first fluid limit for two-parameter processes for this model evidently was the fluid limit in Sect. 6 of [49] for the discrete-time version of that more general $G_t(n)/GI/s + GI$ model, having both time-dependent and state-dependent arrivals. Decreusefond and Moyal [8] established a FCLT for the $M/GI/\infty$ model. In contemporaneous work, Reed and Talreja [41] extend the result in [8] to the $G/GI/\infty$ model and show that the limit process $\hat{Q}^e$ can be regarded as an infinite-dimensional (distribution-valued) OU process, thus proving that the limit process $\{\hat{Q}^e(t,\cdot)\colon\ t \geq 0\}$ is a Markov process. In these other papers, like [14], there are extra regularity conditions on the service-time c.d.f. Moreover, the alternative spaces admit fewer continuous functions.

*Organization of this paper* We start with preliminaries in Sect. 2. In Sect. 3 we state our main results, focusing only on new arrivals (ignoring any customers initially in the system). In Sect. 4 we characterize the limit processes. In Sect. 5 we treat the

initial conditions, and treat all customers in the system. In Sect. 6 we prove the main theorem: Theorem 3.2, focusing on the case of continuous service-time distribution. In Sect. 6.1 we prove the continuity of the representation of some key processes in the space $D_D$. In Sect. 6.2 we continue the proof by establishing tightness of the key processes. In Sect. 6.3 we complete the proof by establishing convergence of the finite-dimensional distributions. There is also a longer version of the present paper [36] available from the authors' web sites. It has a longer introduction; it shows how known results for the special case of exponential service times can be derived from our formulas; it presents supporting technical details, including basic facts about the Brownian sheet, the Kiefer process, two-parameter stochastic integrals, tightness criteria in the space $D_D$ and some detailed calculations.

## 2 Preliminaries

### 2.1 Initial conditions and assumptions

It is convenient to treat the congestion experienced by customers initially in the system separately from the congestion experienced by new arrivals, because they usually can be regarded as being asymptotically independent. Thus we first focus only on new arrivals and then later treat the initial conditions in Sect. 5.

*Assumptions for the arrival processes* We consider a sequence of $G/GI/\infty$ queues indexed by $n$, where the arrival rate is increasing in $n$. For the $n$th system, let $A_n(t)$ be the number of arrivals by time $t$ and $\tau_i^n$ the time of the $i$th arrival.

We assume that the sequence of arrival processes satisfy a FCLT, specified below. All single-parameter continuous-time stochastic processes are assumed to be random elements of the function space $D \equiv D([0,\infty),\mathbb{R})$ with the usual Skorohod $J_1$ topology [2, 48]. Convergence $x_n \to x$ as $n \to \infty$ in the $J_1$ topology is equivalent to uniform convergence on compact subsets (u.o.c.) when the limit function $x$ is continuous. Throughout, we will have a bar, as in $\bar{A}_n(t)$, to denote the law of large number (LLN) scaling (as in (2.2) below) and a hat, as in $\hat{A}_n(t)$, to denote the central limit theorem (CLT) scaling (as in (2.1) below).

**Assumption 1** (FCLT) There exist: (i) a *continuous* nondecreasing deterministic real-valued function $\bar{a}$ on $[0,\infty)$ with $\bar{a}(0) = 0$ and (ii) a stochastic process $\hat{A}$ in $D$ with continuous sample paths, such that

$$\hat{A}_n(t) \equiv n^{-1/2}\big(A_n(t) - n\bar{a}(t)\big) \Rightarrow \hat{A}(t) \quad \text{in } D \text{ as } n \to \infty. \tag{2.1}$$

As an immediate consequence of Assumption 1, we have an associated functional weak law of large numbers (FWLLN)

$$\bar{A}_n(t) \equiv \frac{A_n(t)}{n} \Rightarrow \bar{a}(t) \quad \text{in } D \text{ as } n \to \infty. \tag{2.2}$$

In order to obtain a limiting Markov process we will also assume that the limiting stochastic process $\hat{A}$ has independent increments, but we will obtain limits more generally.

*The standard case* The standard case in Assumption 1 has special $\bar{a}$ and $\hat{A}$. For the FWLLN limit, the standard case is $\bar{a}(t) = \lambda t, t \geq 0$ for some positive constant $\lambda$, which corresponds to an arrival rate of $\lambda_n \equiv \lambda n$ in the $n$th system, but our more general form allows for time-varying arrival rates as in [10, 33, 34].

For the FCLT limit $\hat{A}$, the standard case is BM. That occurs when the arrival processes are scaled versions of a common renewal process with interarrival times having mean $\lambda^{-1}$ and SCV $c_a^2$. Then $\hat{A}(t) = \sqrt{\lambda c_a^2} B_a(t)$, where $B_a$ is a standard BM. Of course, the convergence to BM in (2.1) holds much more generally, e.g., see Chap. 4 of [48]. Except for the SCV $c_a^2$, in the standard case Assumption 1 makes the arrival process asymptotically equivalent to a Poisson process. Thus, in the standard case, the limiting results will be identical to the limit for the $M/GI/\infty$ model when $c_a^2 = 1$, and very similar for $c_a^2 \neq 1$. Actually, there is an important structural difference when $c_a^2 \neq 1$, which we discuss in Sect. 4.

*Assumptions for the service times and the empirical process*

**Assumption 2** (A sequence of i.i.d. random variables) We assume that the service times of new arrivals come from a sequence of i.i.d. nonnegative random variables $\{\eta_i\colon\ i \geq 1\}$ with a *general* c.d.f. $F$, independent of $n$ and the arrival processes.

As in [27], it is significant that our queue-length heavy-traffic limits over finite time intervals do not require more assumptions about the service-time c.d.f. $F$. It need not have a finite mean. However, for subsequent results we will need to assume in addition that $F$ has a finite mean $\mu^{-1}$ and even a finite second moment with SCV $c_s^2$.

Krichagina and Puhalskii [27] observed that it is fruitful to view the service times through the two-parameter *sequential empirical process*

$$\bar{K}_n(t,x) \equiv \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}(\eta_i \leq x), \quad t \geq 0,\ x \geq 0, \tag{2.3}$$

which is directly expressed in the LLN scaling. Here $\mathbf{1}(A)$ is the indicator function. Since the service times are i.i.d. (without any imposed moment conditions), we have a FWLLN for $\bar{K}_n$ itself and a FCLT for the scaled process

$$\hat{K}_n(t,x) \equiv \sqrt{n}\big(\bar{K}_n(t,x) - E\big[\bar{K}_n(t,x)\big]\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor}\big(\mathbf{1}(\eta_i \leq x) - F(x)\big), \tag{2.4}$$

for $t \geq 0$ and $x \geq 0$.

These stochastic-process limits are based on corresponding limits in the case of random variables uniformly distributed on $[0,1]$. Let $\hat{U}_n(t,x)$ denote the stochastic process $\hat{K}_n(t,x)$ when $\eta_i$ is uniformly distributed on $[0,1]$, so that $F(x) = x$, $0 \leq x \leq 1$. Extending previous results by Bickel and Wichura [1], Krichagina and Puhalskii [27] showed that

$$\hat{U}_n(t,x) \Rightarrow U(t,x) \quad \text{in } D\big([0,\infty), D\big([0,1],\mathbb{R}\big)\big) \text{ as } n \to \infty, \tag{2.5}$$

where $U(t,x)$ is the *standard Kiefer process*; see Csörgö and Révész [7], Gaenssler and Stute [12], and van der Vaart and Wellner [45]. In particular, $U(t,x) = W(t,x) - xW(t,1)$, where $W(t,x)$ is a two-parameter BM (Brownian sheet), so that $U(\cdot,x)$ is a BM for each fixed $x$, while $U(t,\cdot)$ is a Brownian bridge for each fixed $t$. The Brownian bridge $B^0$ can be defined in terms of a standard BM $B$ by $B^0(t) \equiv B(t) - tB(1)$, $0 \le t \le 1$; it corresponds to BM conditional on having $B(1) = 0$.

It is significant that $\hat{K}_n$ can be expressed as a simple composition of $\hat{U}_n$ with the c.d.f. $F$ in the second component. We thus have

$$\hat{K}_n(t,x) = \hat{U}_n\big(t, F(x)\big) \Rightarrow \hat{K}(t,x) \equiv U\big(t, F(x)\big) \quad \text{in } D\big([0,\infty), D\big([0,\infty), \mathbb{R}\big)\big), \tag{2.6}$$

as $n \to \infty$ without imposing any conditions upon $F$, because $F$ is not dependent on $n$. Moreover, the convergence is with respect to a stronger topology on $D_D \equiv D([0,\infty), D([0,\infty), \mathbb{R}))$; convergence is uniform over sets of the form $[0,T] \times [0,\infty)$; we have uniformity over $[0,\infty)$ in the second argument. That will turn out to be important when we treat the remaining-workload process. As a consequence of the FCLT in (2.6), we immediately obtain the associated FWLLN

$$\bar{K}_n(t,x) \Rightarrow \bar{k}(t,x) \equiv tF(x) \quad \text{in } D_D \text{ as } n \to \infty, \tag{2.7}$$

where again there is uniformity in $x$ over $[0,\infty)$.

## 2.2 Prelimit processes

Let $Q_n^e(t,y)$ represent the number of customers in the $n$th queueing system at time $t$ that have *elapsed* service times less than or equal to $y$; let $Q_n^r(t,y)$ represent the corresponding number that have *residual* service times strictly greater than $y$. Let $Q_n^t(t)$ represent the *total number* (the superscript $t$) of customers in the $n$th queueing system at time $t$. Clearly, $Q_n^t(t) = Q_n^e(t,t) = Q_n^r(t,0)$, and

$$\begin{aligned} Q_n^r(t,y) &= Q_n^e(t+y,t+y) - Q_n^e(t+y,y) = Q_n^t(t+y) - Q_n^e(t+y,y), \\ Q_n^e(t,y) &= Q_n^r(t,0) - Q_n^r(t-y,y) = Q_n^t(t) - Q_n^r(t-y,y). \end{aligned} \tag{2.8}$$

From (2.8), it is evident that we can construct all three processes $Q_n^e$, $Q_n^r$ and $Q_n^t$ from either $Q_n^e$ or $Q_n^r$. Observe that $Q_n^r$ and $Q_n^e$ can be expressed as

$$\begin{aligned} Q_n^r(t,y) &= \sum_{i=1}^{A_n(t)} \mathbf{1}\big(\tau_i^n + \eta_i > t + y\big), \quad t \ge 0,\ y \ge 0, \\ Q_n^e(t,y) &= \sum_{i=A_n(t-y)}^{A_n(t)} \mathbf{1}\big(\tau_i^n + \eta_i > t\big), \quad t \ge 0,\ 0 \le y \le t. \end{aligned} \tag{2.9}$$

From (2.9), we see the connection to the sequential empirical process $\bar{K}_n$ in (2.3). Indeed, the key observation (following [27]) is that we can rewrite the random sums

in (2.9) as integrals with respect to the random field $\bar{K}_n$ by

$$Q_n^r(t,y) = n\int_0^t \int_0^\infty \mathbf{1}(s+x>t+y)\,d\bar{K}_n\big(\bar{A}_n(s),x\big), \quad t,y\ge 0,$$
$$Q_n^e(t,y) = n\int_{t-y}^t \int_0^t \mathbf{1}(s+x>t)\,d\bar{K}_n\big(\bar{A}_n(s),x\big), \quad t\ge 0,\ 0\le y\le t, \tag{2.10}$$

for $\bar{K}_n$ in (2.3). These two-dimensional integrals in (2.10) are two-dimensional Stieltjes integrals. In the present context, the integrals in (2.10) are understood to be (defined as) the random sums in (2.9).

**Lemma 2.1** (Representation of $Q_n^r$ and $Q_n^e$) *The processes $Q_n^r$ and $Q_n^e$ defined in* (2.9) *and* (2.10) *can be represented as*

$$Q_n^r(t,y) = n\int_0^t F^c(t+y-s)\,d\bar{a}(s) + \sqrt{n}\big(\hat{X}_{n,1}^r(t,y)+\hat{X}_{n,2}^r(t,y)\big), \quad t,y\ge 0, \tag{2.11}$$

$$Q_n^e(t,y) = n\int_{t-y}^t F^c(t-s)\,d\bar{a}(s) + \sqrt{n}\big(\hat{X}_{n,1}^e(t,y)+\hat{X}_{n,2}^e(t,y)\big),$$
$$t\ge 0,\ 0\le y\le t, \tag{2.12}$$

*where*

$$\hat{X}_{n,1}^r(t,y) \equiv \int_0^t F^c(t+y-s)\,d\hat{A}_n(s), \qquad \hat{X}_{n,1}^e(t,y) \equiv \int_{t-y}^t F^c(t-s)\,d\hat{A}_n(s), \tag{2.13}$$

$$\begin{aligned}\hat{X}_{n,2}^r(t,y) &\equiv \int_0^t \int_0^\infty \mathbf{1}(s+x>t+y)\,d\hat{R}_n(s,x)\\ &= -\int_0^t \int_0^\infty \mathbf{1}(s+x\le t+y)\,d\hat{R}_n(s,x),\end{aligned} \tag{2.14}$$

$$\begin{aligned}\hat{X}_{n,2}^e(t,y) &\equiv \int_{t-y}^t \int_0^t \mathbf{1}(s+x>t)\,d\hat{R}_n(s,x)\\ &= -\int_{t-y}^t \int_0^t \mathbf{1}(s+x\le t)\,d\hat{R}_n(s,x),\end{aligned} \tag{2.15}$$

$$\begin{aligned}\hat{R}_n(t,y) &\equiv \hat{K}_n\big(\bar{A}_n(t),y\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{A_n(t)}\big(\mathbf{1}(\eta_i\le y)-F(y)\big)\\ &= \sqrt{n}\bar{K}_n\big(\bar{A}_n(t),y\big) - \hat{A}_n(t)F(y) - \sqrt{n}\bar{a}(t)F(y),\end{aligned} \tag{2.16}$$

*with the integrals in* (2.13), (2.14) *and* (2.15) *all defined as Stieltjes integrals for functions of bounded variation as integrators.*

*Proof* Apply (2.4) to get the first relation in (2.16). (Right away, from (2.6), we see that $\hat{R}_n(t,x) \Rightarrow \hat{K}(\bar{a}(t),x)$.) Use (2.4) and (2.3) to get the rest of (2.16) and

$$\begin{aligned}\bar{K}_n\big(\bar{A}_n(t),x\big) &= \frac{1}{n}\sum_{i=1}^{A_n(t)} \mathbf{1}(\eta_i \le x) \\ &= \frac{1}{\sqrt{n}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{A_n(t)}\big(\mathbf{1}(\eta_i \le x) - F(x)\big)\right] \\ &\quad + \frac{1}{\sqrt{n}}\sqrt{n}\big(\bar{A}_n(t) - \bar{a}(t)\big)F(x) + \bar{a}(t)F(x) \\ &= \frac{1}{\sqrt{n}}\hat{R}_n(t,x) + \frac{1}{\sqrt{n}}\hat{A}_n(t)F(x) + \bar{a}(t)F(x). \end{aligned} \tag{2.17}$$

Combine (2.10) and (2.17) to get (2.11). The alternative representation for $\hat{X}_{n,2}(t,y)$ holds because $\hat{K}_n(t,\infty) = 0$ and thus $\hat{R}_n(t,\infty) = 0$ for all $t$. □

We will also consider several related processes. Let $F_n^e(t,\cdot)$ and $F_n^r(t,\cdot)$ represent the *empirical age distribution* and the *empirical residual distribution* at time $t$ in the $n$th system, respectively, i.e.,

$$F_n^e(t,y) \equiv Q_n^e(t,y)/Q_n^t(t), \quad t \ge 0,\ 0 \le y \le t, \tag{2.18}$$

and

$$F_n^{r,c}(t,y) \equiv 1 - F_n^r(t,y) \equiv Q_n^r(t,y)/Q_n^t(t), \quad t \ge 0,\ y \ge 0. \tag{2.19}$$

For each $n$ and $t$, $F_n^e(t,\cdot)$ and $F_n^r(t,\cdot)$ are proper c.d.f.'s. Let $D_n(t)$ count the number of departures in the interval $[0,t]$; clearly, $D_n(t) \equiv A_n(t) - Q_n^t(t)$ for $t \ge 0$.

We will also consider several processes characterizing the workload in total service time. For these limits, we will assume that we are in the standard case for the arrival process and impose extra moment conditions on the service-time c.d.f. $F$. The total input of work over $[0,t]$ is

$$I_n(t) \equiv \sum_{i=1}^{A_n(t)} \eta_i, \quad t \ge 0. \tag{2.20}$$

The amount of the workload to have arrived by time $t$ that will be remaining after time $t+y$ is

$$W_n^r(t,y) \equiv \int_y^\infty Q_n^r(t,x)\,dx, \quad t \ge 0,\ y \ge 0. \tag{2.21}$$

Then the total (remaining) workload at time $t$ is $W_n^t(t) \equiv W_n^r(t,0)$. Finally, the total amount of completed service work by time $t$ is $C_n(t) \equiv I_n(t) - W_n^t(t)$.

### 2.3 The space $D_D$

Our limits for two-parameter processes will be in the space $D_D$, which we regard as a subset of $D([0,\infty), D([0,\infty),\mathbb{R}))$, where $D \equiv D([0,\infty), S)$, for a separable metric space $S$, is the space of all right-continuous $S$-valued functions with left limits in $(0,\infty)$; see [2, 48] for background. We will be considering the subset of functions $x(t,y)$ which have finite limits as the second argument $y \to \infty$. For example, we have $Q_n^e(t,y) = Q_n^e(t,t)$ for all $y > t$ and $Q_n^r(t,y) \to 0$ as $y \to \infty$. We will be using the standard Skorohod [42] $J_1$ topology on all $D$ spaces, but since all limit processes will have continuous sample paths, convergence in our space $D_D$ is equivalent to uniform convergence over subsets of the form $[0,T] \times [0,\infty)$. (We already observed that we have such stronger uniform convergence over that non-compact set for $\hat{K}_n$ to the Kiefer process in (2.5).) We refer to [44] for the convergence preservation of various functions in $D_D$.

For two-parameter processes, one might consider using generalizations of the spaces of two-parameter real-valued functions considered by Straf [43] and Neuhaus [35], but those spaces require limits to exist at each point in the domain (subset of $\mathbb{R}^2$) through all paths lying in each of the four quadrants centered at that point. That works fine for the sequential empirical process $K_n$, but *not* for $Q_n^r(t,y)$. For example, suppose that the first two arrivals occur at times 1 and 3, and that the arrival at time 1 has a service time of 2. Then limits do not exist along all paths in the southeast and northwest quadrants at the point $(t,y) = (2,1)$, because there are discontinuities along a negative 45° line running through that point. The value shifts from 0 to 1 at that line. However, there is no difficulty in the larger space $D_D$.

### 2.4 The service-time distribution as a mixture

The general service-time c.d.f. $F$ has at most countably many discontinuity points. Let $p_d$ ($p_c$) be the total probability mass at the discontinuity (continuity) points, i.e., $p_d \equiv \sum_{x\ge 0} \Delta F(x) \le 1$ and $p_c = 1 - p_d \le 1$, where $\Delta F(x) \equiv F(x) - F(x-)$. To focus on the interesting case, suppose that $0 < p_d < 1$. We order the discontinuity points by the size of their probability mass in decreasing order (using the natural order in case of ties); i.e., let $\{\bar{x}_1, \bar{x}_2, \ldots\}$ be such that $\Delta F(\bar{x}_i) \ge \Delta F(\bar{x}_{i+1})$. Define two proper c.d.f.'s $F_c$ and $F_d$ for a continuous random variable $\eta^c$ and a discrete random variable $\eta^d$, respectively, by

$$F_c(x) \equiv P\big(\eta^c \le x\big) \equiv \frac{1}{p_c}\left(F(x) - \sum_{y\le x} \Delta F(y)\right), \quad x \ge 0,$$

and

$$F_d(x) \equiv \sum_{j:\, \bar{x}_j \le x} P\big(\eta^d = \bar{x}_j\big), \quad \text{and} \quad p_{d,i} \equiv P\big(\eta^d = \bar{x}_i\big) \equiv \frac{\Delta F(\bar{x}_i)}{p_d}, \quad x \ge 0.$$

Note that $F$ can be represented as the mixture $F = p_c F_c + p_d F_d$.

Let $A_n^c(t)$, $A_n^d(t)$ and $A_{n,i}^d(t)$ count the number of arrivals by time $t$ with continuous service time, with a discrete service time, and with a deterministic service time $\bar{x}_i$,

$i = 1, 2, \ldots$, respectively. Clearly, $A_n^d(t) = \sum_{i=1}^{\infty} A_{n,i}^d(t)$ and $A_n(t) = A_n^d(t) + A_n^c(t)$ for $t \geq 0$. Define the LLN-scaled processes $\bar{A}_n^c \equiv n^{-1} A_n^c$, $\bar{A}_n^d \equiv n^{-1} A_n^d$, and $\bar{A}_{n,i}^d \equiv n^{-1} A_{n,i}^d$.

Under Assumptions 1 and 2, for a general service-time c.d.f., we can decompose the system into two subsystems, one with arrival processes $A_n^c$ and service-time distribution $F_c$ and the other with arrival processes $A_n^d$ and discrete service times $\{\bar{x}_i\colon i \geq 1\}$ with distribution $F_d$. We will adopt the method in [27] to analyze the first subsystem in the space $D_D$, then the method in [14] to analyze the second subsystem, and then we put them together to obtain the limits for the whole system.

## 3 Main results

In this section, we state the main results of this paper: the FWLLN and FCLT for the scaled processes associated with $Q_n^r$ and $W_n^r$, along with the closely related processes. We give the proofs in Sect. 6. Define the LLN-scaled processes $\bar{Q}_n^r \equiv \{\bar{Q}_n^r(t, y), t \geq 0, y \geq 0\}$ by

$$\bar{Q}_n^r(t, y) \equiv \frac{Q_n^r(t, y)}{n}, \tag{3.1}$$

and similarly for the processes $\bar{Q}_n^e$, $\bar{Q}_n^t$, $\bar{D}_n$, $\bar{W}_n^r$, $\bar{W}_n^t$, $\bar{I}_n$ and $\bar{C}_n$. Define the LLN-scaled processes $\bar{F}_n^e \equiv \{\bar{F}_n^e(t, y), t \geq 0, 0 \leq y \leq t\}$ and $\bar{F}_n^{r,c} \equiv \{\bar{F}_n^{r,c}(t, y), t \geq 0, y \geq 0\}$ by

$$\bar{F}_n^e(t, y) \equiv \bar{Q}_n^e(t, y)/\bar{Q}_n^t(t) \quad \text{and} \quad \bar{F}_n^{r,c}(t, y) \equiv \bar{Q}_n^r(t, y)/\bar{Q}_n^t(t), \tag{3.2}$$

where $\bar{F}_n^e(t, y)$ and $\bar{F}_n^{r,c}(t, y)$ are defined to be 0 if $\bar{Q}_n^t(t) = 0$ for some $t$. By Lemma 2.1,

$$\bar{Q}_n^r(t, y) = \int_0^t F^c(t + y - s)\, d\bar{a}(s) + \frac{1}{\sqrt{n}}\big(\hat{X}_{n,1}(t, y) + \hat{X}_{n,2}(t, y)\big), \quad t, y \geq 0. \tag{3.3}$$

When we focus on the amount of work, as in the workload processes, we use the *stationary-excess* (or residual-lifetime) c.d.f. associated with the service-time c.d.f. $F$ (assumed to have finite mean $\mu^{-1}$), defined by

$$F_e(x) \equiv \mu \int_0^x F^c(s)\, ds, \quad x \geq 0. \tag{3.4}$$

The mean of $F_e$ is $E[\eta^2]/2E[\eta] = (c_s^2 + 1)/2\mu$; that will be used in part (c) of Theorem 3.1 below.

**Theorem 3.1** (FWLLN)

(a) *Under Assumptions 1 and 2,*

$$\begin{aligned}&\big(\bar{A}_n, \bar{A}_n^c, \bar{A}_n^d, \big\{\bar{A}_{n,i}^d\colon i \ge 1\big\}, \bar{K}_n, \bar{Q}_n^r, \bar{Q}_n^t, \bar{Q}_n^e, \bar{F}_n^e, \bar{F}_n^{r,c}, \bar{D}_n\big)\\ &\quad\Rightarrow \big(\bar{a}, \bar{a}^c, \bar{a}^d, \big\{\bar{a}_i^d\colon i \ge 1\big\}, \bar{k}, \bar{q}^r, \bar{q}^t, \bar{q}^e, \bar{f}^e, \bar{f}^{r,c}, \bar{d}\big)\end{aligned} \tag{3.5}$$

*in* $D^3 \times D^\infty \times D_D^2 \times D \times D_D^3 \times D$ *as* $n \to \infty$ *w.p.*1, *where the limits are deterministic functions*: $\bar{a}$ *is the limit in* (2.2), $\bar{a}^c \equiv p_c \bar{a}$, $\bar{a}^d \equiv p_d \bar{a}$, $\bar{a}_i^d \equiv p_{d,i} \bar{a}^d$, *for* $i \ge 1$, $\bar{k}(t,x) \equiv tF(x)$ *in* (2.7),

$$\bar{q}^r(t,y) \equiv \int_0^t F^c(t+y-s)\, d\bar{a}(s), \quad t \ge 0,\ y \ge 0, \tag{3.6}$$

$$\bar{q}^e(t,y) \equiv \int_{t-y}^t F^c(t-s)\, d\bar{a}(s), \quad t \ge 0,\ 0 \le y \le t, \tag{3.7}$$

$\bar{q}^t(t) \equiv \bar{q}^r(t,0) = \bar{q}^e(t,t)$, $\bar{f}^e(t,y) \equiv \bar{q}^e(t,y)/\bar{q}^t(t)$, $\bar{f}^{r,c}(t,y) \equiv \bar{q}^r(t,y)/\bar{q}^t(t)$ *and* $\bar{d} = \bar{a} - \bar{q}^t$.

(b) *If, in addition to the assumptions in part* (a), $\bar{a}(t) = \lambda t$, $t \ge 0$, *and the service-time c.d.f.* $F$ *has finite mean* $\mu^{-1}$, *then*

$$\big(\bar{W}_n^r, \bar{W}_n^t, \bar{I}_n, \bar{C}_n\big) \Rightarrow \big(\bar{w}^r, \bar{w}^t, \bar{i}, \bar{c}\big) \quad \text{in } D_D \times D^3 \text{ as } n \to \infty \text{ w.p.1}, \tag{3.8}$$

*jointly with the limits in* (3.5), *where*

$$\begin{aligned}\bar{w}^r(t,y) &\equiv \lambda \int_y^\infty \bar{q}^r(t,x)\, dx, \quad t \ge 0,\ y \ge 0,\\ &= \lambda \int_y^\infty \left( \int_0^t F^c(t+x-s)\, ds \right) dx = \frac{\lambda}{\mu} \int_0^t F_e^c(y+s)\, ds,\\ \bar{w}^t(t) &\equiv \bar{w}^r(t,0) = \frac{\lambda}{\mu} \int_0^t F_e^c(s)\, ds,\\ \bar{i}(t) &\equiv \frac{\lambda t}{\mu} \quad \text{and} \quad \bar{c}(t) \equiv \bar{i}(t) - \bar{w}^t(t) = \frac{\lambda}{\mu} \int_0^t F_e(s)\, ds,\end{aligned} \tag{3.9}$$

*for* $F_e$ *in* (3.4).

(c) *If, in addition to the assumptions of parts* (a) *and* (b), $E[\eta^2] < \infty$, *then*

$$\bar{w}^r(t,y) \to \frac{\lambda}{\mu} \int_0^\infty F_e^c(y+s)\, ds < \infty \quad \text{and} \quad \bar{w}^t(t) \to \frac{\lambda(c_s^2+1)}{2\mu^2} \quad \text{as } t \to \infty. \tag{3.10}$$

We obtain Theorem 3.1 as an immediate corollary to the following FCLT, which exploits centering by the deterministic limits above. For the FCLT, define the normalized processes

$$\hat{Q}_n^r(t,y) \equiv \sqrt{n}\big(\bar{Q}_n^r(t,y) - \bar{q}^r(t,y)\big), \tag{3.11}$$

for $t \geq 0$ and $y \geq 0$, and similarly for the other processes, using the centering terms above. By (3.3) and (3.6),

$$\hat{Q}_n^r(t,y) = \hat{X}_{n,1}^r(t,y) + \hat{X}_{n,2}^r(t,y), \quad t \geq 0,\ y \geq 0. \tag{3.12}$$

Moreover,

$$\begin{aligned}\hat{F}_n^{r,c}(t,y) &\equiv \sqrt{n}\big(\bar{F}_n^{r,c}(t,y) - \bar{f}^{r,c}(t,y)\big) \\ &= \bar{Q}_n^t(t)^{-1}\big(\hat{Q}_n^r(t,y) - \hat{Q}_n^t(t)\bar{f}^{r,c}(t,y)\big), \quad t \geq 0,\ y \geq 0,\end{aligned}$$

and

$$\begin{aligned}\hat{F}_n^e(t,y) &\equiv \sqrt{n}\big(\bar{F}_n^e(t,y) - \bar{f}^e(t,y)\big) \\ &= \bar{Q}_n^t(t)^{-1}\big(\hat{Q}_n^e(t,y) - \hat{Q}_n^t(t)\bar{f}^e(t,y)\big), \quad t \geq 0,\ 0 \leq y \leq t.\end{aligned}$$

Define the CLT-scaled processes $\hat{A}_n^c \equiv \{\hat{A}_n^c(t):\ t \geq 0\}$, $\hat{A}_n^d \equiv \{\hat{A}_n^d(t):\ t \geq 0\}$ and $\hat{A}_{n,i}^d \equiv \{\hat{A}_{n,i}^d(t):\ t \geq 0\}$ by

$$\begin{aligned}\hat{A}_n^c(t) &\equiv n^{1/2}\big(\bar{A}_n^c(t) - \bar{a}^c(t)\big), \qquad \hat{A}_n^d(t) \equiv n^{1/2}\big(\bar{A}_n^d(t) - \bar{a}^d(t)\big), \\ \hat{A}_{n,i}^d(t) &\equiv n^{1/2}\big(\bar{A}_{n,i}^d(t) - \bar{a}_i^d(t)\big),\end{aligned}$$

for $t \geq 0$ and $i \geq 1$.

The joint deterministic limits in Theorem 3.1 are equivalent to the separate one-dimensional limits, but that is not true for the FCLT generalization below. Let $\circ$ be the composition function, i.e., $(x \circ y)(t) \equiv x(y(t))$, $t \geq 0$. Let $\stackrel{\mathrm{d}}{=}$ mean equality in distribution.

**Theorem 3.2** (FCLT)

(a) *Under Assumptions 1 and 2,*

$$\begin{aligned}&\big(\hat{A}_n, \hat{A}_n^c, \hat{A}_n^d, \{\hat{A}_{n,i}^d:\ i \geq 1\}, \hat{K}_n, \hat{Q}_n^r, \hat{Q}_n^t, \hat{Q}_n^e, \hat{F}_n^{r,c}, \hat{F}_n^e, \hat{D}_n\big) \\ &\quad \Rightarrow \big(\hat{A}, \hat{A}^c, \hat{A}^d, \{\hat{A}_i^d:\ i \geq 1\}, \hat{K}, \hat{Q}^r, \hat{Q}^t, \hat{Q}^e, \hat{F}^{r,c}, \hat{F}^e, \hat{D}\big)\end{aligned} \tag{3.13}$$

*in* $D^3 \times D^\infty \times D_D^2 \times D \times D_D^3 \times D$ *as* $n \to \infty$, *where* $\hat{A}$ *is the limit in* (2.1),

$$\begin{aligned}&\hat{A}^c = p_c\hat{A} + S^c \circ \bar{a}, \qquad \hat{A}^d = p_d\hat{A} + S^d \circ \bar{a}, \qquad \hat{A}_i^d = p_d p_{d,i}\hat{A} + S_i^d \circ \bar{a}, \\ &S^c = -S^d, \qquad S^c \stackrel{\mathrm{d}}{=} \sqrt{p_c(1-p_c)}B, \qquad S^d \stackrel{\mathrm{d}}{=} \sqrt{p_d(1-p_d)}B, \\ &S_i^d \stackrel{\mathrm{d}}{=} \sqrt{p_d p_{d,i}(1-p_d p_{d,i})}B, \quad i \geq 1,\end{aligned} \tag{3.14}$$

*where* $B$ *is a standard BM, independent of* $\hat{A}$, *and the process* $(S^c, S^d, \{S_i^d:\ i \geq 1\})$ *is an infinite-dimensional BM with mean* 0 *and covariance matrix* $\mathbf{C}$ *where* $\mathbf{C}_{c,c} = p_c(1-p_c)$, $\mathbf{C}_{d,d} = p_d(1-p_d)$, $\mathbf{C}_{c,d} = \mathbf{C}_{d,c} = -p_c p_d$, $\mathbf{C}_{i,i} =$

$p_d p_{d,i}(1-p_d p_{d,i})$ *for* $i \geq 1$, $\mathbf{C}_{i,c} = \mathbf{C}_{c,i} = -p_c p_d p_{d,i}$, $\mathbf{C}_{i,d} = \mathbf{C}_{d,i} = -p_d^2 p_{d,i}$ *and* $\mathbf{C}_{i,j} = -p_d^2 p_{d,i} p_{d,j}$ *for* $i \neq j$, *and the representations for* $\hat{Q}^r$ *and* $\hat{Q}^e$ *are*

$$\begin{aligned} \hat{Q}^r(t,y) &= \hat{X}_1^{c,r}(t,y) + \hat{X}_2^{c,r}(t,y) + \hat{X}^{d,r}(t,y), \quad t \geq 0,\ y \geq 0, \\ \hat{Q}^e(t,y) &= \hat{X}_1^{c,e}(t,y) + \hat{X}_2^{c,e}(t,y) + \hat{X}^{d,e}(t,y), \quad t \geq 0,\ 0 \leq y \leq t, \end{aligned} \tag{3.15}$$

*where*

$$\begin{aligned} \hat{X}_1^{c,r}(t,y) &\equiv \int_0^t F_c^c(t+y-s)\, d\hat{A}^c(s), \qquad \hat{X}_1^{c,e}(t,y) \equiv \int_{t-y}^t F_c^c(t-s)\, d\hat{A}^c(s), \\ \hat{X}_2^{c,r}(t,y) &\equiv \int_0^t \int_0^\infty \mathbf{1}(s+x > t+y)\, d\hat{K}^c\big(\bar{a}^c(s), x\big), \\ \hat{X}_2^{c,e}(t,y) &\equiv \int_{t-y}^t \int_0^t \mathbf{1}(s+x > t)\, d\hat{K}^c\big(\bar{a}^c(s), x\big), \\ \hat{X}^{d,r}(t,y) &\equiv \sum_{i=1}^\infty \big(\hat{A}_i^d(t) - \hat{A}_i^d\big(t - (\bar{x}_i - y)^+\big)\big), \\ \hat{X}^{d,e}(t,y) &\equiv \sum_{i=1}^\infty \big(\hat{A}_i^d(t) - \hat{A}_i^d\big(t - (\bar{x}_i \wedge y)\big)\big), \end{aligned} \tag{3.16}$$

*with* $\hat{K}^c(\bar{a}^c(s), x) = U(\bar{a}^c(s), F_c(x))$, *which is independent of* $\hat{A}$. $\hat{Q}^t(t) \equiv \hat{Q}^r(t,0)$, $\hat{Q}^e(t,y) \equiv \hat{Q}^t(t) - \hat{Q}^r(t-y,y)$, $\hat{F}^{r,c}(t,y) \equiv \bar{q}^t(t)^{-1}(\hat{Q}^r(t,y) - \hat{Q}^t(t) f^{r,c}(t,y))$, $\hat{F}^e(t,y) \equiv \bar{q}^t(t)^{-1}(\hat{Q}^e(t,y) - \hat{Q}^t(t) f^e(t,y))$, *and* $\hat{D} = \hat{A} - \hat{Q}^t$. *All these limit processes are continuous. If, in addition,* $\hat{A} = B_a \circ \bar{a}$, *as when* $A_n$ *is nonhomogeneous Poisson, then* $\hat{A}^d$ *and* $\hat{A}^c$ *are independent, and thus* $\hat{X}_1^{c,r}$, $\hat{X}_2^{c,r}$ *and* $\hat{X}^{d,r}$ *are mutually independent.*

(b) *If, in addition to the assumptions in part* (a), $\bar{a}(t) = \lambda t$, $t \geq 0$, *and the service-time c.d.f.* $F$ *has finite mean* $\mu^{-1}$, *then* $(\hat{W}_n^r, \hat{W}_n^t) \Rightarrow (\hat{W}^r, \hat{W}^t)$ *in* $D_D \times D$ *as* $n \to \infty$ *jointly with the limits in* (3.13), *where*

$$\hat{W}^r(t,y) \equiv \int_y^\infty \hat{Q}^r(t,x)\, dx, \quad \textit{and} \quad \hat{W}^t(t) \equiv \hat{W}^r(t,0) = \int_0^\infty \hat{Q}^r(t,x)\, dx. \tag{3.17}$$

(c) *If, in addition to the assumptions in parts* (a) *and* (b), $E[\eta^2] < \infty$, *then* $(\hat{I}_n, \hat{C}_n) \Rightarrow (\hat{I}, \hat{C})$ *in* $D^2$ *as* $n \to \infty$ *jointly with the limits above, where*

$$\hat{I}(t) \equiv \sqrt{\lambda c_s^2} B_s(t) + \mu^{-1}\hat{A} \quad \textit{and} \quad \hat{C}(t) \equiv \hat{I}(t) - \hat{W}^t(t), \quad t \geq 0, \tag{3.18}$$

*with* $B_s$ *being a standard BM independent of* $\hat{A}$.

*Remark 3.1* The limit processes $\hat{Q}^r$ and $\hat{Q}^e$ can also be expressed as the sum of the following three mutually independent processes

$$\begin{aligned}\hat{Q}^r(t,y) &= \hat{X}_1^r(t,y) + \hat{X}_2^{c,r}(t,y) + \hat{X}_3^r(t,y), \quad t \geq 0,\ y \geq 0,\\ \hat{Q}^e(t,y) &= \hat{X}_1^e(t,y) + \hat{X}_2^{c,e}(t,y) + \hat{X}_3^e(t,y), \quad t \geq 0,\ 0 \leq y \leq t,\end{aligned} \tag{3.19}$$

where

$$\begin{aligned}\hat{X}_1^r(t,y) &\equiv \int_0^t F^c(t+y-s)\,d\hat{A}(s), \qquad \hat{X}_1^e(t,y) \equiv \int_{t-y}^t F^c(t-s)\,d\hat{A}(s),\\ \hat{X}_3^r(t,y) &\equiv \int_0^t F_c^c(t+y-s)\,dS^c\big(\bar{a}(s)\big)\\ &\quad + \sum_{i=1}^{\infty}\big(S_i^d\big(\bar{a}(t)\big) - S_i^d\big(\bar{a}\big(t-(\bar{x}_i-y)^+\big)\big)\big),\\ \hat{X}_3^e(t,y) &\equiv \int_{t-y}^t F_c^c(t-s)\,dS^c\big(\bar{a}(s)\big) + \sum_{i=1}^{\infty}\big(S_i^d\big(\bar{a}(t)\big) - S_i^d\big(\bar{a}\big(t-(\bar{x}_i\wedge y)\big)\big)\big).\end{aligned} \tag{3.20}$$

The asymptotic variability of the arrival process is captured by $\hat{A}$, which appears only in $\hat{X}_1^r$ and $\hat{X}_1^e$; the asymptotic variability of the service process is captured by $\hat{K}^c$, which appears only in $\hat{X}_2^{c,r}$ and $\hat{X}_2^{c,e}$; while the asymptotic variability of service-time splitting is captured by $S^c$ and $S_i^d$, which appears only in $\hat{X}_3^r$ and $\hat{X}_3^e$. Thus, in some sense, there is additivity of stochastic effects, as pointed out in [27, 30], but this might be misinterpreted. Notice that *both* $\hat{X}_1^r$ and $\hat{X}_2^r$ depend on the full service-time c.d.f. $F$, not just its mean. On the other hand, the arrival process beyond its deterministic rate only appears in $\hat{X}_1^r$ and $\hat{X}_1^e$, so that there is a genuine asymptotic insensitivity to the arrival process beyond its rate in $\hat{X}_2^{c,r}$ and $\hat{X}_2^{c,e}$.

This claim holds because, by (3.14), we can write $\hat{X}_1^c(t,y)$ and $\hat{X}^d(t,y)$ in (3.15) as

$$\begin{aligned}\hat{X}_1^{c,r}(t,y) &= \int_0^t F_c^c(t+y-s)\,d\big(p_c\hat{A}(t) + S^c\big(\bar{a}(s)\big)\big)\\ &= \int_0^t \Big(F^c(t+y-s) - \sum_{u>t+y-s}\Delta F(u)\Big)\,d\big(\hat{A}(t) + p_c^{-1}S^c\big(\bar{a}(s)\big)\big),\end{aligned}$$

and

$$\begin{aligned}\hat{X}^{d,r}(t,y) &= \sum_{i=1}^{\infty}\big[p_d p_{d,i}\big(\hat{A}(t) - \hat{A}\big(t-(\bar{x}_i-y)^+\big)\big)\\ &\quad + \big(S_i^d\big(\bar{a}(t)\big) - S_i^d\big(\bar{a}\big(t-(\bar{x}_i-y)^+\big)\big)\big)\big]\\ &= \int_0^t\Big(\sum_{u>t+y-s}\Delta F(u)\Big)\,d\hat{A}(t) + \sum_{i=1}^{\infty}\big(S_i^d\big(\bar{a}(t)\big) - S_i^d\big(\bar{a}\big(t-(\bar{x}_i-y)^+\big)\big)\big),\end{aligned}$$

which implies that $\hat{X}_1^{c,r}(t,y) + \hat{X}^{d,r}(t,y) = \hat{X}_1^r(t,y) + \hat{X}_3^r(t,y)$ for each $t \geq 0$ and $y \geq 0$. Similarly, $\hat{X}_1^{c,e}(t,y) + \hat{X}^{d,e}(t,y) = \hat{X}_1^e(t,y) + \hat{X}_3^e(t,y)$ holds.

*Remark 3.2* The two integrals in the expression for $\hat{Q}^r$ are stochastic integrals. The first integral for $\hat{X}_1^{c,r}$ (or $\hat{X}_1^r$) is a standard Ito integral if $\hat{A}$ is a (time-changed) Brownian motion; otherwise, the expression for $\hat{X}_1^{c,r}$ (or $\hat{X}_1^r$) is interpreted as the form after integration by parts. The relevant version of integration by parts for $\hat{X}_{n,1}^r$ and $\hat{X}_1^r$ is given in Bremaud [4], p. 336. For $\hat{X}_{n,1}^r$, it yields

$$\hat{X}_{n,1}^r(t,y) = F^c(y)\hat{A}_n(t) - \int_0^t \hat{A}_n(s-)\,dF^c(t+y-s), \tag{3.21}$$

and similarly for $\hat{X}_1$. The left limit $\hat{A}_n(s-)$ in (3.21) is only needed if the functions $F$ and $\hat{A}_n$ have common discontinuities with positive probability. The second integral for $\hat{X}_2$ is either understood as the stochastic integrals with respect to two-parameter processes of the first type, or in the mean-square sense, as in [27]; see Sect. 6.3.

In the literature, several types of stochastic integrals with respect to two-parameter processes have been defined. The first type of integral was first defined for two-parameter Brownian sheets by Cairoli [5] (see also [46]), generalizing the definition of Ito's integral directly. It was generalized to $n$-parameter Brownian sheets by Wong and Zakai [51] and to general martingales by Cairoli and Walsh [6]. Even more generalization appears in Wong and Zakai [52]. We refer to Koshnevisan [26] for a relatively complete review. The important property we apply here is the isometry property, analogous to the Ito isometry property.

*Remark 3.3* We remark that if the service-time c.d.f. $F$ is discontinuous, the process $\hat{X}_2$ defined by

$$\hat{X}_2(t,y) \equiv \int_0^t \int_0^\infty \mathbf{1}(s+x > t+y)\,d\hat{K}\big(\bar{a}(s),x\big)$$

is only continuous in $t$, but not in $y$, and in fact, it is not even in the space $D_D$. The continuity of $\hat{X}_2$ and $\hat{Q}^t$ in $t$ can be obtained as in Lemma 5.1 of [27]. To see that $\hat{X}_2$ need not be in $D_D$, suppose that $F$ is the mixture of two point masses $y_1 > 0$ and $y_2 > 0$. Then, applying (4.1) below, we see that, for each $t \geq 0$, $\hat{X}_2(t,y) = 0$ for all $y \geq 0$ except $y_1$ and $y_2$, so that $\hat{X}_2(t,\cdot) \notin D$. That property follows from (4.1) because $\Delta_{\hat{K}}(t_1,t_2,x_1,x_2) = 0$ for $0 < x_1 < x_2$ unless either $y_1 < x_1 < y_2$ or $x_1 < y_1 < x_2 < y_2$. That means that the random measure attaches all mass on the strips $x = y_1$ and $x = y_2$. Incidentally, in this example, $\hat{X}_2(t,\cdot)$ is an element of the space $E$ in Chap. 15 of [48]. That explains why we split the general distribution into a mixture of a discrete distribution and a continuous distribution.

We now establish additional results in the standard case for the fluid limits. In particular, we will obtain an analog of the classic result for the $M/GI/\infty$ model, stating that in steady state both the elapsed service times and the residual service times are distributed as mutually independent random variables, each with c.d.f. $F_e$ in (3.4). We will see that the limiting empirical age distribution is precisely $F_e$, just as is true for the prelimit processes with a Poisson arrival process.

**Corollary 3.1** (The standard case) *Consider the standard case in which* $\bar{a}(t) = \lambda t$, $t \geq 0$, *and* $\hat{A} = \sqrt{\lambda c_a^2} B_a$, *where* $B_a$ *is a standard BM. Assume that the service-time distribution* $F$ *has finite mean* $\mu^{-1}$. *Under Assumptions 1 and 2, the limits in (3.5) hold with*

$$\begin{aligned}
\bar{q}^r(t,y) &\equiv \lambda \int_0^t F^c(t+y-s)\,ds = \lambda \int_0^t F^c(y+s)\,ds \\
&\to (\lambda/\mu) F_e^c(y) \quad \text{as } t \to \infty, \\
\bar{q}^e(t,y) &\equiv \lambda \int_{t-y}^t F^c(t-s)\,ds \\
&= \lambda \int_0^y F^c(s)\,ds = (\lambda/\mu) F_e(y), \quad \text{for } t \geq 0, \\
\bar{f}^e(t,y) &\equiv \bar{q}^e(t,y)/\bar{q}^t(t) \to F_e(y) \quad \text{as } t \to \infty, \\
\bar{f}^{r,c}(t,y) &\equiv \bar{q}^r(t,y)/\bar{q}^t(t) \to F_e^c(y) \quad \text{as } t \to \infty.
\end{aligned} \tag{3.22}$$

## 4 Characterizing the limit processes

We now show that the two-parameter queue-length limit processes, $\hat{Q}^r(t,y)$ and $\hat{Q}^e(t,y)$, constitute continuous Brownian analogs of the Poisson-random-measure representation for the $M/GI/\infty$ model [10]. (But the limit is only identical to the limit for the $M/GI/\infty$ model when $c_a^2 = 1$.) A key role here is played by the *transformed Kiefer process* $\hat{K}(t,x) \equiv U(t,F(x)) = W(t,F(x)) - F(x)W(t,1)$. Any finite number of $\hat{K}$-increments,

$$\begin{aligned}
\Delta_{\hat{K}}(t_1,t_2,x_1,x_2) &\equiv \hat{K}(t_2,x_2) - \hat{K}(t_2,x_1) - \hat{K}(t_1,x_2) + \hat{K}(t_1,x_1) \\
&= \Delta_W\big(t_1,t_2,F(x_1),F(x_2)\big) \\
&\quad - \big(F(x_2) - F(x_1)\big)\big(W(t_2,1) - W(t_1,1)\big)
\end{aligned} \tag{4.1}$$

for $0 \leq t_1 < t_2$ and $0 \leq x_1 < x_2$, are independent random variables provided that the rectangles $(t_1,t_2] \times (x_1,x_2]$ have disjoint horizontal time intervals $(t_1,t_2]$.

We only treat $\hat{Q}^r$ here. If the limit process $\hat{A}$ has independent increments, then so does $\hat{Q}^r$, provided that it is viewed as a function-valued process with the argument $t$. The limit processes $\hat{Q}^r$ is then a Markov process in $D_D$ (only considering the argument $t$). This result can be based on a basic decomposition, depicted in Fig. 1.

**Theorem 4.1** (Decompositions, independent increments and the Markov property for $\hat{Q}^r$) *The limiting random variables* $\hat{X}_1^{c,r}(t,y)$, $\hat{X}_2^{c,r}(t,y)$, $\hat{X}^{d,r}(t,y)$ *and* $\hat{Q}^r(t,y)$ *in Theorem 3.2 admit the decompositions*

$$\begin{aligned}
\hat{X}_i^{c,r}(t_2,y) &= \hat{X}_i^{c,r}(t_1,y+t_2-t_1) + Z_i^{c,r}(t_1,t_2,y), \quad \text{for } i = 1,2, \text{ and } t_2 > t_1 \geq 0, \\
\hat{X}^{d,r}(t_2,y) &= \hat{X}^{d,r}(t_1,y+t_2-t_1) + Z^{d,r}(t_1,t_2,y), \quad t_2 > t_1 \geq 0, \\
\hat{Q}^r(t_2,y) &= \hat{Q}^r(t_1,y+t_2-t_1) + Z^r(t_1,t_2,y), \quad t_2 > t_1 \geq 0,
\end{aligned} \tag{4.2}$$

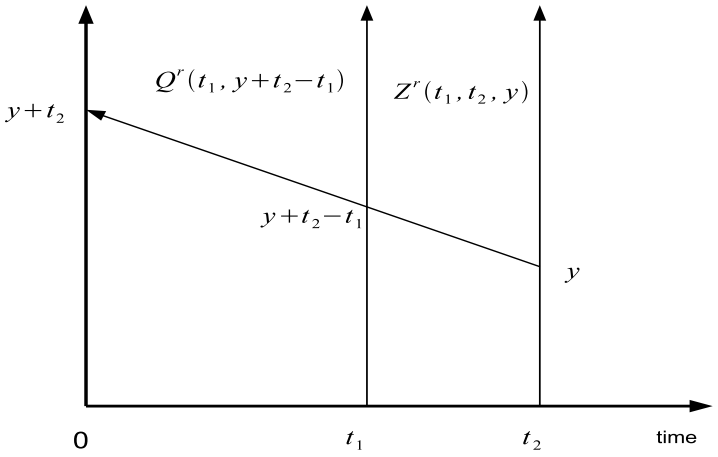



**Fig. 1** The basic decomposition for $Q^r(t, y)$

*where* $y \geq 0$, $Z^r \equiv Z_1^{c,r} + Z_2^{c,r} + Z^{d,r}$, *and*

$$Z_1^{c,r}(t_1, t_2, y) \equiv \int_{t_1}^{t_2} F_c^c(t + y - s)\, d\hat{A}^c(s),$$

$$Z_2^{c,r}(t_1, t_2, y) \equiv \int_{t_1}^{t_2} \int_0^{\infty} \mathbf{1}(s + x > t + y)\, d\hat{K}^c\big(\bar{a}^c(s), x\big),$$

$$Z^{d,r}(t_1, t_2, y) \equiv \sum_{i=1}^{\infty} \big[\big(\hat{A}_i^d(t_2) - \hat{A}_i^d(t_1)\big) - \big(\hat{A}_i^d\big(t_2 - (\bar{x}_i - y)^+\big) - \hat{A}_i^d\big(t_1 - (\bar{x}_i - y)^+\big)\big)\big].$$

*If, in addition to the assumptions of Theorem 3.2, the limit process* $\hat{A}$ *has independent increments, which occurs in the standard case of Corollary 3.1, where* $\hat{A}$ *is a BM, then the two random variables on the right in (4.2) are independent in each case. Moreover, the four processes* $\{\hat{X}_1^{c,r}(t, \cdot)\colon t \geq 0\}$, $\{\hat{X}_2^{c,r}(t, \cdot)\colon t \geq 0\}$, $\{\hat{X}^{d,r}(t, \cdot)\colon t \geq 0\}$ *and* $\{\hat{Q}^r(t, \cdot)\colon t \geq 0\}$ *all have independent increments, and are thus Markov processes (with respect to the argument* $t$*).*

*Proof* The decomposition for $\hat{X}_1^{c,r}(t, y)$, $\hat{X}_2^{c,r}(t, y)$, $\hat{X}^{d,r}(t, y)$ and $\hat{Q}^r(t, y)$ in (4.2) is by direct construction, as in Fig. 1. The independent-increments property is inherited from $\hat{K}^c$, $\hat{A}^c$ and $\hat{A}^d$. □

We now show that the limit processes are Gaussian if $\hat{A}$ is Gaussian, which again is the case if $\hat{A}$ is BM. For nonstationary non-Poisson arrival processes $(G_t)$, we can construct such $G_t$ processes (or just think of them) by letting the original arrival processes $\{A_n(t)\colon t \geq 0\}$ be defined by $A_n(t) \equiv \tilde{A}(n\bar{a}(t))$, $t \geq 0$, where $\tilde{A} \equiv \{\tilde{A}(t)\colon t \geq 0\}$ is a rate-1 stationary (or asymptotically stationary) stochastic point

process, such that $\tilde{A}$ satisfies a FCLT with limit $\sqrt{c_a^2}B_a$, where $B_a$ is a standard BM. As a consequence, a natural Gaussian limit process is $\hat{A}(t) \equiv \sqrt{c_a^2}B_a(\bar{a}(t))$, $t \geq 0$. Indeed, this occurs for the familiar $M_t$ case, for which $c_a^2 = 1$.

**Theorem 4.2** (Gaussian property) *If, in addition to the assumptions of Theorem 3.2, the limit process $\hat{A}$ is Gaussian, then the limit processes $\hat{Q}^t$, $\hat{Q}^e$, $\hat{Q}^r$, $\hat{D}$, $\hat{V}^r$, $\hat{V}^t$ in (3.13) are all continuous Gaussian processes. If $\hat{A}(t) = \sqrt{c_a^2}B_a(\bar{a}(t))$ for $t \geq 0$, where $B_a$ is a standard BM, then for each fixed $t \geq 0$ and $y \geq 0$,*

$$\begin{aligned} &\hat{Q}^r(t,y) \stackrel{\mathrm{d}}{=} N\big(0, \sigma_{q,r}^2(t,y)\big), \qquad \hat{Q}^e(t,y) \stackrel{\mathrm{d}}{=} N\big(0, \sigma_{q,e}^2(t,y)\big), \\ &\hat{W}^r(t,y) \stackrel{\mathrm{d}}{=} N\big(0, \sigma_w^2(t,y)\big), \end{aligned} \tag{4.3}$$

*where*

$$\begin{aligned} \sigma_{q,r}^2(t,y) &= \big(c_a^2 - 1\big)\int_0^t F^c(t+y-s)^2\,d\bar{a}(s) + \int_0^t F^c(t+y-s)\,d\bar{a}(s), \\ \sigma_{q,e}^2(t,y) &= \big(c_a^2 - 1\big)\int_{t-y}^t F^c(t-s)^2\,d\bar{a}(s) + \int_{t-y}^t F^c(t-s)\,d\bar{a}(s), \\ \sigma_w^2(t,y) &= c_a^2\int_y^\infty\int_y^\infty\int_0^t F^c(t+x-s)F^c(t+z-s)\,d\bar{a}(s)\,dx\,dz \\ &\quad + \int_y^\infty\int_y^\infty\int_0^t F(t+x\wedge z-s)F^c(t+x\vee z-s)\,d\bar{a}(s)\,dx\,dz. \end{aligned}$$

*Proof* It is obvious that the limit processes are Gaussian when the limit arrival process $\hat{A}$ is Gaussian. We only need to derive the variance formulas. We will use (3.19) to calculate them and the mutual independence between the three terms in the expression of $\hat{Q}^r$ gives $\sigma_{q,r}^2(t,y) = \sigma_{1,r}^2(t,y) + \sigma_{2,c,r}^2(t,y) + \sigma_{3,r}^2(t,y)$, where $\sigma_{1,r}^2(t,y) = E[(\hat{X}_1^r(t,y))^2]$, $\sigma_{2,c,r}^2(t,y) = E[(\hat{X}_2^{c,r}(t,y))^2]$ and $\sigma_{3,r}^2(t,y) = E[(\hat{X}_3^r(t,y))^2]$. By Ito's isometry, we have

$$\sigma_{1,r}^2(t,y) = c_a^2\int_0^t F^c(t+y-s)^2\,d\bar{a}(s),$$

and

$$\begin{aligned} \sigma_{3,r}^2(t,y) &= p_d p_c\int_0^t F_c^c(t+y-s)^2\,d\bar{a}(s) \\ &\quad + \sum_{i=1}^\infty\big(p_d p_{d,i}(1 - p_d p_{d,i})\big(\bar{a}(t) - \bar{a}\big(t - (\bar{x}_i - y)^+\big)\big)\big) \\ &\quad - 2p_d^2\sum_{i<j} p_{d,i}p_{d,j}\big(\bar{a}(t) - \bar{a}\big(t - \big((\bar{x}_i \wedge \bar{x}_j) - y\big)^+\big)\big) \end{aligned}$$

$$- 2\sum_{i=1}^{\infty} p_c p_d p_{d,i} \int_0^t F_c^c(t+y-s)\, d\big(\bar{a}(s) - \bar{a}\big(s - (\bar{x}_i - y)^+\big)\big).$$

Having $\hat{X}_2^{c,r}$ well-defined with continuous paths follows from the definition of stochastic integral with respect to the Brownian sheet of the first type. It clearly has mean 0. Its variance is given by

$$\begin{aligned}
\sigma_{2,c,r}^2(t,y) &= E\left[\left(\int_0^t \int_0^\infty \mathbf{1}(s+x > t+y)\, dU\big(\bar{a}^c(s), F_c(x)\big)\right)^2\right] \\
&= E\left[\left(\int_0^t \int_0^\infty \mathbf{1}(s+x > t+y)\, d\big(W\big(\bar{a}^c(s), F_c(x)\big)\right.\right. \\
&\quad \left.\left. - F_c(x) W\big(\bar{a}^c(s), 1\big)\big)\right)^2\right] \\
&= \int_0^t \int_0^\infty \mathbf{1}(s+x > t+y)\, dF_c(x)\, d\bar{a}^c(s) + \int_0^t F_c^c(t+y-s)^2\, d\bar{a}^c(s) \\
&\quad - 2\int_0^t \int_0^\infty \mathbf{1}(s+x > t+y) F_c(t+y-s)\, dF_c(x)\, d\bar{a}^c(s) \\
&= \int_0^t F_c(t+y-s) F_c^c(t+y-s)\, d\bar{a}^c(s),
\end{aligned}$$

where the second equality uses the identity $U(x,y) = W(x,y) - yW(x,1)$, and the third equality uses the isometry property of the stochastic integral of the first type with respect to two-parameter Brownian sheets and also the isometry property of the stochastic Ito's integral.

Notice that

$$\begin{aligned}
&p_d p_c \int_0^t F_c^c(t+y-s)^2\, d\bar{a}(s) + \int_0^t F_c(t+y-s) F_c^c(t+y-s)\, d\bar{a}^c(s) \\
&\quad = \int_0^t p_c F_c^c(t+y-s)\big(1 - p_c F_c^c(t+y-s)\big)\, d\bar{a}(s).
\end{aligned}$$

Moreover, $F^c = p_c F_c^c + p_d F_d^c$ and $FF^c = (1 - p_c F_c^c - p_d F_d^c)(p_c F_c^c + p_d F_d^c) = p_c F_c^c (1 - p_c F_c^c) + p_d F_d^c (1 - p_d F_d^c) - 2 p_c F_c^c p_d F_d^c$. Then, simple algebra calculation gives the final expression for $\sigma_{q,r}^2(t,y)$. Similar argument applies to the calculation of $\sigma_{q,e}^2(t,y)$.

For the variance of $\hat{W}^r(t,y)$, by the independence of $\hat{X}_1^r(t,y)$, $\hat{X}_2^{c,r}(t,y)$ and $\hat{X}_3^r(t,y)$, we have

$$\begin{aligned}
E\big[\hat{W}^r(t,y)^2\big] &= E\left[\left(\int_y^\infty \hat{X}_1^r(t,x)\, dx\right)^2\right] + E\left[\left(\int_y^\infty \hat{X}_2^{c,r}(t,x)\, dx\right)^2\right] \\
&\quad + E\left[\left(\int_y^\infty \hat{X}_3^r(t,x)\, dx\right)^2\right].
\end{aligned}$$

Then by an analogous argument, we obtain the variance of $\hat{W}^r(t,y)$. □

We remark that, for Theorem 4.2, we could also have used an argument analogous to Lemma 5.1 in [27] by understanding the integral in $\hat{X}_2^{c,r}$ as a mean-square limit (Sect. 6.3). However, our approach here by applying properties of stochastic integrals of the first type with respect to two-parameter Brownian sheets simplifies the proof. Paralleling the result in Lemma 5.1 [27], we can easily check that for $0 \le t \le t'$, $0 \le y \le y'$,

$$\begin{aligned} &E\big[\big(\hat{X}_2^{c,r}(t,y) - \hat{X}_2^{c,r}(t',y')\big)^2\big] \\ &\quad = \int_0^t \big(F_c(t'+y'-u) - F_c(t+y-u)\big) \\ &\qquad \times \big(1 + F_c(t+y-u) - F_c(t'+y'-u)\big)\, d\bar{a}^c(u). \end{aligned}$$

**Corollary 4.1** (The special case $c_a^2 = 1$) *If, in addition to the assumptions of Theorem 4.2, $\bar{a}(t) = \int_0^t \lambda(s)\,ds$ and $c_a^2 = 1$, then*

$$\mathrm{Var}\big(\hat{Q}^r(t,y)\big) = \int_0^t F^c(t+y-u)\lambda(s)\,ds, \tag{4.4}$$

*for $t \ge 0$ and $y \ge 0$. The limit $\hat{A}$ and all the other limits are the same as if the unscaled arrival processes $\{A_n(t)\colon t \ge 0\}$ are Poisson processes (possibly nonhomogeneous). (When $A_n$ is Poisson, the prelimit variables $Q_n^r(t,y)$ and $Q_n^e(t,y)$ are Poisson random variables for each $t$ and $y$.) Moreover, as in the Poisson arrival case, for each $t \ge 0$ and $y \ge 0$, $\hat{Q}^r(t,y)$ is distributed the same as the limit of*

$$\hat{\mathcal{Q}}_n^r(t,y) \equiv \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{Q_n^t(t)} \eta_i(t,y) - \bar{q}^r(t,y)\right), \tag{4.5}$$

*where $\{\eta_i(t,y)\colon i \ge 1\}$ is a sequence of i.i.d. Bernoulli random variables with*

$$P\big(\eta_i(t,y) = 1\big) = \bar{f}^{r,c}(t,y), \tag{4.6}$$

*which are independent of the total queue length $\hat{Q}_n^t(t)$.*

*Proof* We need to justify (4.5). First, we note that this is the asymptotic generalization of an exact relation for Poisson arrivals; e.g., see Theorem 2.1 of [15]. Here we start by defining

$$\mathcal{Q}_n^r(t,y) \equiv \sum_{i=1}^{Q_n^t(t)} \eta_i(t,y),$$

for each $t \ge 0$ and $y \ge 0$. (In passing, we remark that $\mathcal{Q}_n^r(t,y) \stackrel{\mathrm{d}}{=} Q_n^r(t,y)$ in the special case of a nonhomogeneous $(M_t)$ arrival process, but not more generally.) By

the FWLLN, the fluid scaled processes $\bar{\mathcal{Q}}_n^r(t,y)$ converge to the fluid limit $\bar{q}^r(t,y)$ as $n \to \infty$:

$$\bar{\mathcal{Q}}_n^r(t,y) \Rightarrow \bar{\mathcal{Q}}^r(t,y) \equiv E\big[\eta_i(t,y)\big]\bar{q}^t(t) = \bar{f}^{r,c}(t,y)\bar{q}^t(t) = \frac{\bar{q}^r(t,y)}{\bar{q}^t(t)}\bar{q}^t(t) = \bar{q}^r(t,y).$$

We can write $\hat{\mathcal{Q}}_n^r(t,y)$ in (4.5) as

$$\hat{\mathcal{Q}}_n^r(t,y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n\bar{Q}_n^t(t)} \big(\eta_i(t,y) - \bar{f}^{r,c}(t,y)\big) + \bar{f}^{r,c}(t,y)\hat{Q}_n^t(t).$$

By FCLT for random walks with i.i.d. increments of mean 0 and finite variance (Theorem 8.2, [2]), continuity of composition in $D$ (Theorem 13.2.2, [48]) and Theorems 3.1 and 3.2, we obtain the weak convergence of $\hat{\mathcal{Q}}_n^r(t,y)$:

$$\hat{\mathcal{Q}}_n^r(t,y) \Rightarrow \hat{\mathcal{Q}}^r(t,y) \quad \text{in } D_D \text{ as } n \to \infty,$$

where

$$\hat{\mathcal{Q}}^r(t,y) \equiv \sigma_3(t,y)B_3\big(\bar{q}^t(t)\big) + \bar{f}^{r,c}(t,y)\hat{Q}^t(t),$$

where $\sigma_3^2(t,y) \equiv \bar{f}^{r,c}(t,y)(1-\bar{f}^{r,c}(t,y))$ and $B_3$ is a standard Brownian motion, independent of $\hat{Q}^t(t)$. Thus, $\hat{\mathcal{Q}}^r(t,y)$ is Gaussian with mean 0 and variance

$$\begin{aligned}
\mathrm{Var}\big(\hat{\mathcal{Q}}^r(t,y)\big) &= \sigma_3^2(t,y)\bar{q}^t(t) + \bar{f}^{r,c}(t,y)^2\,\mathrm{Var}\big(\hat{Q}^t(t)\big) \\
&= \bar{f}^{r,c}(t,y)\big(1-\bar{f}^{r,c}(t,y)\big)\bar{q}^t(t) + \bar{f}^{r,c}(t,y)^2\int_0^t F^c(t-u)\lambda(s)\,ds \\
&= \frac{\bar{q}^r(t,y)}{\bar{q}^t(t)}\left(1-\frac{\bar{q}^r(t,y)}{\bar{q}^t(t)}\right)\bar{q}^t(t) + \frac{\bar{q}^r(t,y)^2}{\bar{q}^t(t)^2}\bar{q}^t(t) \\
&= \bar{q}^r(t,y) = \int_0^t F^c(t+y-u)\lambda(s)\,ds = \mathrm{Var}\big(\hat{\mathcal{Q}}^r(t,y)\big).
\end{aligned}$$

Since $\hat{\mathcal{Q}}^r(t,y)$ and $\hat{\mathcal{Q}}^r(t,y)$ are both Gaussian with the same mean and variance, $\hat{\mathcal{Q}}^r(t,y)$ and $\hat{\mathcal{Q}}^r(t,y)$ are equal in distribution. When the arrival process is $M_t$, $\mathcal{Q}_n^r(t,y)$ has a Poisson distribution for each $n$, $t$ and $y$, so that the variance equals the mean. Since $c_a^2 = 1$, the limit must be the same here as in the $M_t$ case. □

We emphasize that Corollary 4.1 is consistent with known results for the $M_t/GI/\infty$ model. The asymptotic equivalence to the random sum in (4.5) and (4.6) is the asymptotic analog of the property for the $M_t/GI/\infty$ model that, conditional on the number of customers in the system, the remaining service times are distributed as i.i.d. random variables with c.d.f. $\bar{f}^{r,c}(t,\cdot)$; e.g., see Theorem 2.1 of [15]. This property does not hold for $c_a^2 \neq 1$.

**Corollary 4.2** (The standard case) *If* $\bar{a}(t) = \lambda t$ *and* $\hat{A} = \sqrt{\lambda c_a^2} B_a$, *then the variances of* $\hat{Q}^r(t, y)$ *and* $\hat{Q}^e(t, y)$ *are*

$$\begin{aligned}\sigma_{q,r}^2(t, y) &= \lambda\big(c_a^2 - 1\big)\int_0^t F^c(y+s)^2\,ds + \lambda\int_0^t F^c(y+s)\,ds \\ &\to \lambda\big(c_a^2 - 1\big)\int_y^\infty F^c(s)^2\,ds \\ &\quad + \lambda\int_y^\infty F^c(s)\,ds \equiv \sigma_{q,r}^2(y) \quad \text{as } t \to \infty,\ y \ge 0,\end{aligned}$$

*and*

$$\sigma_{q,e}^2(t, y) = \lambda\big(c_a^2 - 1\big)\int_0^y F^c(s)^2\,ds + \lambda\int_0^y F^c(s)\,ds \equiv \sigma_{q,e}^2(y), \quad t, y \ge 0.$$

*Thus,* $\hat{Q}^r(t, y) \Rightarrow N(0, \sigma_{q,r}^2(y))$ *and* $\hat{Q}^e(t, y) \Rightarrow N(0, \sigma_{q,e}^2(y))$ *as* $t \to \infty$. *If, in addition,* $c_a^2 = 1$, *then*

$$\sigma_{q,r}^2(t, y) = \lambda\int_0^t F^c(y+s)\,ds \to \lambda\int_y^\infty F^c(s)\,ds = \frac{\lambda}{\mu}F_e^c(y), \quad \text{as } t \to \infty,\ y \ge 0,$$

$$\sigma_{q,e}^2(t, y) = \lambda\int_0^y F^c(s)\,ds = \frac{\lambda}{\mu}F_e(y), \quad t, y \ge 0,$$

*and* $\mathrm{Var}(\hat{Q}^t(t)) = \lambda\int_0^t F^c(s)\,ds \to \lambda/\mu$ *as* $t \to \infty$.

## 5 Initial conditions

So far, we considered only new arrivals. Now we consider customers in the system initially. Like the generality of the service-time c.d.f., the initial conditions present technical difficulties. Our assumptions will be similar to those made in [27] and to those for the new arrivals in Sect. 2. However, these assumptions are less realistic here. Thus, for applications, it is good that the relevance of the initial conditions decreases as time evolves, because we can think of the system starting in the distant past with just new arrivals, so that we will be able to approximate the two-parameter processes by the Markov limit processes.

We assume that the remaining service times of the customers initially in the system are i.i.d., distributed according to some new c.d.f., independent of the number of customers in the system and everything associated with new arrivals. That rather strong assumption will actually be justified if we assume that the initial state we see is the result of an $M_t/GI/\infty$ system, possible with different model parameters, that started empty at some previous time. As noted in Corollary 4.1 and the remark before Corollary 4.2, this strong independence property actually holds in an $M_t/GI/\infty$ model. Moreover, that representation is asymptotically correct more generally if $c_a^2 = 1$. Unfortunately, however, that representation is not asymptotically correct if $c_a^2 \neq 1$. Nevertheless, it is a natural candidate approximate initial condition.

Here is our specific framework: Let $Q_n^{i,r}(y)$ be the number of customers initially in the $n$th system at time 0, not counting new arrivals, who have residual service times strictly greater than $y$. Let $Q_n^{i,t} \equiv Q_n^{i,r}(0)$ be the total number of customers initially in the $n$th system and let $Q_n^{i,e}(y)$ be the number of customers initially in the $n$th system that have elapsed service times less than or equal to $y$. Let $W_n^{i,r}(y)$ and $W_n^{i,t}$ be the corresponding workload processes, defined as in (2.21).

Let $\bar{Q}_n^{i,r}(y)$ and $\hat{Q}_n^{i,r}(y)$ be the associated scaled processes, defined by

$$\bar{Q}_n^{i,r}(y) \equiv \frac{Q_n^{i,r}(y)}{n} \quad \text{and} \quad \hat{Q}_n^{i,r}(y) \equiv \sqrt{n}\big(\bar{Q}_n^{i,r}(y) - \bar{q}^{i,r}(y)\big), \quad y \geq 0, \tag{5.1}$$

where $\bar{q}^{i,r}$ is the FWLLN limit of $\bar{Q}_n^{i,r}$ to be established. Let other scaled processes be defined similarly. What we need are the FWLLN $\bar{Q}_n^{i,r} \Rightarrow \bar{q}^{i,r}$ and the associated FCLT $\hat{Q}_n^{i,r} \Rightarrow \hat{Q}^{i,r}$ in $D$ as $n \to \infty$, jointly with the limits in Theorem 3.2. The extension to joint convergence with the other processes will be immediate if the stochastic processes associated with new arrivals are independent of the initial conditions. Otherwise, we require that we have the joint convergence $(\hat{A}_n, \hat{Q}_n^{i,r}) \Rightarrow (\hat{A}, \hat{Q}^{i,r})$ in $D \times D$, with the service times of new arrivals coming from a sequence of i.i.d. random variables, which is independent of both the arrival processes and the initial conditions. We now give sufficient conditions to get these limits.

*Assumptions for the initial conditions*

**Assumption 3** (i.i.d. service times) The service times of customers initially in the system come from a sequence $\{\eta_j^i\colon\ j \geq 1\}$ of i.i.d. nonnegative random variables with a *general* c.d.f. $F_i$ and $F_i(0) = 0$, independent of $n$ and independent of the total number of customers initially present and all random quantities associated with new arrivals.

**Assumption 4** (Independence and CLT for the initial number) The initial total number of customers in the system, $Q_n^{i,t}$, is independent of the service times of the initial customers and all random quantities associated with new arrivals. There exist (i) a nonnegative constant $\bar{q}^{i,t}$ and (ii) a random variable $\hat{Q}^{i,t}$ such that

$$\hat{Q}_n^{i,t} \equiv \frac{1}{\sqrt{n}}\big(Q_n^{i,t} - n\bar{q}^{i,t}\big) \Rightarrow \hat{Q}^{i,t} \quad \text{in } \mathbb{R} \text{ as } n \to \infty. \tag{5.2}$$

Paralleling Lemma 2.1, we have the representation result.

**Lemma 5.1** (Representation of $Q_n^{i,r}$) *The process $Q_n^{i,r}$ can be represented as*

$$Q_n^{i,r}(y) = \sum_{j=1}^{Q_n^{i,t}} \big(\mathbf{1}\big(\eta_j^i > y\big) - F_i^c(y)\big) + Q_n^{i,t} F_i^c(y), \quad y \geq 0. \tag{5.3}$$

**Theorem 5.1** (FWLLN and FCLT for the initial conditions) *Under Assumptions 3 and 4,*

$$\begin{aligned}
&\bar{Q}_n^{i,r}(y) \Rightarrow \bar{q}^{i,r}(y) \equiv F_i^c(y)\bar{q}^{i,t} \quad \text{in } D \text{ as } n \to \infty, \\
&\hat{Q}_n^{i,r}(y) \Rightarrow \hat{Q}^{i,r}(y) \equiv F_i^c(y)\hat{Q}^{i,t} + \sqrt{\bar{q}^{i,t}}B^0\big(F_i(y)\big) \quad \text{in } D \text{ as } n \to \infty,
\end{aligned} \tag{5.4}$$

*where $B^0$ is a Brownian bridge, independent of $\hat{Q}^{i,t}$.*

We can combine Theorems 3.1, 3.2 and 5.1 to treat the total number of customers in the system at time $t$ with residual service times strictly greater than $y$, which we denote by $Q_n^{T,r}(t,y)$. The key representation is

$$Q_n^{T,r}(t,y) = Q_n^r(t,y) + Q^{i,r}(t+y), \quad t \ge 0,\ y \ge 0. \tag{5.5}$$

**Corollary 5.1** (FWLLN and FCLT for all customers) *Under Assumptions 1–4,*

$$\begin{aligned}
\bar{Q}_n^{T,r}(t,y) &\equiv \bar{Q}_n^{i,r}(t+y) + \bar{Q}_n^r(t,y) \Rightarrow \bar{q}^{T,r}(t,y) \equiv \bar{q}^{i,r}(t+y) + \bar{q}^r(t,y) \\
&= F_i^c(t+y)\bar{q}^{i,t} + \int_0^t F^c(t+y-s)\,d\bar{a}(s), \\
\hat{Q}_n^{T,r}(t,y) &\equiv \hat{Q}_n^{i,r}(t+y) + \hat{Q}_n^r(t,y) \Rightarrow \hat{Q}^{T,r}(t,y) \equiv \hat{Q}^{i,r}(t+y) + \hat{Q}^r(t,y) \\
&= F_i^c(t+y)\hat{Q}^{i,t} + \sqrt{\bar{q}^{i,t}}B^0\big(F_i(t+y)\big) + \hat{X}_1^{c,r}(t,y) \\
&\quad + \hat{X}_2^{c,r}(t,y) + \hat{X}^{d,r}(t,y),
\end{aligned} \tag{5.6}$$

*in $D_D$ as $n \to \infty$, where $\hat{X}_1^{c,r}$, $\hat{X}_2^{c,r}$ and $\hat{X}^{d,r}$ are given in (3.16).*

## 6 Proof of the FCLT

We now prove the FCLT in Theorem 3.2. First, the joint convergence of the processes

$$\big(\hat{A}_n, \hat{A}_n^c, \hat{A}_n^d, \{\hat{A}_{n,i}^d\colon i \ge 1\}\big) \Rightarrow \big(\hat{A}, \hat{A}^c, \hat{A}^d, \{\hat{A}_i^d\colon i \ge 1\}\big)$$

follows from Theorem 9.5.1 in [48]. For the subsystem with discrete service-time distribution, the limits follow from an easy extension of [14]. In [14], the convergence to the limit $\hat{X}^{d,r}(t,y)$ is proved in the space $D$ for each fixed $y \ge 0$, however, the convergence can be easily generalized to be in the space $D_D$ since the limit process $\hat{A}$ is assumed to be continuous here (Assumption 1). Since the prelimit process of $\hat{X}^{d,r}$ is

$$\hat{X}_n^{d,r}(t,y) = \sum_{i=1}^{\infty}\big(\hat{A}_{n,i}^d(t) - \hat{A}_{n,i}^d\big(t - (\bar{x}_i - y)^+\big)\big), \quad t, y \ge 0,$$

it suffices to show that the mapping $\phi : D \to D_D$ defined by

$$\phi(z)(t, y) \equiv \sum_{i=1}^{\infty} \big(z(t) - z\big(t - (\bar{x}_i - y)^+\big)\big)$$

is continuous in the Skorohod $J_1$ topology and then apply the continuous mapping theorem. Moreover, in order to prove $\hat{W}_n^{r,d}(t, y) \Rightarrow \hat{W}^{r,d}(t, y)$ in $D_D$, where $\hat{W}_n^{r,d}(t, y)$ can be written as

$$\hat{W}_n^{r,d}(t, y) = \sum_{i=1}^{\infty} \int_y^{\bar{x}_i} \big(\hat{A}_{n,i}^d(t) - \hat{A}_{n,i}^d\big(t - (\bar{x}_i - x)^+\big)\big)\, dx, \quad t, y \geq 0,$$

we need to prove the continuity of the mapping $\psi : D \to D_D$ defined by

$$\psi(z)(t, y) = \int_y^{\bar{x}_i} \big(z(t) - z\big(t - (\bar{x}_i - x)^+\big)\big)\, dx, \quad z \in D,\ t, y \geq 0.$$

Since the limit $\hat{A}$ is continuous, it suffices to show the uniform continuity of the mapping $\psi$ on compact intervals, which follows from a direct argument. Thus, we will only focus on the subsystem with continuous service-time distributions. For notational convenience, we will simply suppose that $F$ in Assumption 2 is continuous such that $F_c = F$, $\hat{A}_n^c = \hat{A}_n$ and similarly for other processes. In particular, we write $\hat{X}_1^{c,r}$ and $\hat{X}_2^{c,r}$ simply as $\hat{X}_1$ and $\hat{X}_2$, respectively.

One might hope to obtain a very fast proof by applying the continuous mapping theorem with an appropriate continuous mapping. That would seem to be possible, because both the initial stochastic integral in (2.10) and the representation in Lemma 2.1 show that the scaled residual service queue-length process $\hat{Q}_n^r$ can be regarded as the image of a deterministic function $h : D \times D_D \to D_D$ mapping $(\hat{A}_n, \hat{K}_n)$ into $\hat{Q}_n^r$. Given that $(\hat{A}_n, \hat{K}_n) \Rightarrow (\hat{A}, \hat{K})$ under Assumptions 1 and 2, we would expect that corresponding limits for $\hat{Q}_n^r$ and the other processes would follow directly from an appropriate continuous mapping theorem. Unfortunately, the connecting map is complicated, being in the form of a stochastic integral, with the limit of the component $\hat{X}_{n,2}$ involving a two-dimensional stochastic integral. In fact, we will show below that we can easily treat the component $\hat{X}_{n,1}$ via the representation (3.21). However, $\hat{X}_{n,2}$ presents a problem. Unfortunately, the general results of weak convergence of stochastic integrals and differential equations in [28, 29, 31] do not seem to apply. Thus, instead, we will follow [27] and prove the convergence in the classical way, by proving tightness and convergence of the finite-dimensional distributions. (See [41] for a different way.)

For us, the first step is to get convergence for the process $\hat{R}_n$ jointly with $(\hat{A}_n, \hat{K}_n)$ by exploiting the composition map for a random time change, paralleling Sect. 13.2 of [48]; see [44] for extensions to $D_D$. Starting from $(\hat{A}_n, \hat{K}_n) \Rightarrow (\hat{A}, \hat{K})$, we first obtain $(\hat{A}_n, \bar{A}_n, \hat{K}_n) \Rightarrow (\hat{A}, \bar{a}, \hat{K})$ by applying (2.1) and Theorem 11.4.5 of [48]. We then apply the continuous mapping theorem for composition applied in the space $D_D$, where the composition is with respect to the first component of $\hat{K}_n$, and the limit $\bar{a}$

and $\hat{K}$ are both continuous (in the first component for $\hat{K}$). That yields

$$(\hat{A}_n, \bar{A}_n, \hat{K}_n, \hat{R}_n) \Rightarrow (\hat{A}, \bar{a}, \hat{K}, \hat{R}) \quad \text{in } D^2 \times D_D^2, \tag{6.1}$$

where $\hat{R}(t,x) = \hat{K}(\bar{a}(t), x) = U(\bar{a}(t), F(x))$ for $t \geq 0$ and $x \geq 0$. Since $\hat{R}$ does not involve $\hat{A}$, we see that $\hat{A}_n$ and $\hat{R}_n$ are asymptotically independent. Necessarily, then the processes $\hat{X}_{n,1}$ and $\hat{X}_{n,2}$ are asymptotically independent as well.

We use the classical method for establishing the limit

$$(\hat{A}_n, \bar{A}_n, \hat{K}_n, \hat{R}_n, \hat{X}_{n,1}, \hat{X}_{n,2}) \Rightarrow (\hat{A}, \bar{a}, \hat{K}, \hat{R}, \hat{X}_1, \hat{X}_2) \tag{6.2}$$

in $D^2 \times D_D^4$: We show convergence of the finite-dimensional distributions and tightness. We get tightness for $\{(\hat{A}_n, \bar{A}_n, \hat{K}_n, \hat{R}_n)\colon n \geq 1\}$ from the convergence in (6.1). We use the fact that tightness on product spaces is equivalent to tightness on each of the component spaces; see Theorem 11.6.7 of [48]. Since we can write $\hat{X}_{n,1}$ as (3.21), the tightness and convergence of $\hat{X}_{n,1} \Rightarrow \hat{X}_1$ in $D_D$ can be obtained directly by applying continuous mapping theorem if we can prove the mapping defined in (3.21) from $\hat{A}_n$ to $\hat{X}_{n,1}$ is continuous in $D_D$. We will prove the continuity of this mapping in $D_D$ in Sect. 6.1. We then establish tightness for $\{(\hat{X}_{n,1}, \hat{X}_{n,2})\colon n \geq 1\}$ in Sect. 6.2 and the required convergence of the finite-dimensional distributions associated with $\{(\hat{X}_{n,1}, \hat{X}_{n,2})\colon n \geq 1\}$ in Sect. 6.3. Given the limit in (6.2), the rest of the limits in parts (a) and (b) follows from the continuous mapping theorem. The limit in part (c) is an application of convergence preservation for composition with linear centering as in Corollary 13.3.2 of [48]. The component limits require finite second moments.

### 6.1 Continuity of the representation for $\hat{X}_{n,1}$ in $D_D$

In this section, we prove the continuity of the mapping $\phi\colon D \to D_D$ defined by

$$\phi(x)(t,y) \equiv F^c(y)x(t) - \int_0^t x(s-)\,dF(t+y-s), \tag{6.3}$$

for $x \in D$ and $t, y \geq 0$. By (3.21) and (3.16), we have $\hat{X}_{n,1}(t,y) = \phi(\hat{A}_n)(t,y)$ and $\hat{X}_1(t,y) = \phi(\hat{A})(t,y)$.

**Lemma 6.1** *The mapping $\phi$ defined in (6.3) is continuous in $D_D$.*

*Proof* Suppose that $x_n \to x$ in $D$. We need to show that $d_{D_D}(\phi(x_n), \phi(x)) \to 0$ as $n \to \infty$. Let $T > 0$ be a continuity point of $x$ and consider the time domain $[0,T] \times [0,\infty)$. By the convergence $x_n \to x$ in $(D, J_1)$ as $n \to \infty$, there exist increasing homeomorphisms $\lambda_n$ of the interval $[0,T]$ such that $\|x_n - x \circ \lambda_n\|_T \to 0$ and $\|\lambda_n - e\|_T \to 0$ as $n \to \infty$, where $e(t) = t$ for all $t \geq 0$ and $\|y\|_T = \sup_{t\in[0,T]} |y(t)|$ for any $y \in D$. Let $M = \sup_{0\leq t\leq T} |x(t)| < \infty$. Since $F$ is continuous, it suffices to show that

$$\begin{aligned} &\big\|\phi(x_n)(\cdot,\cdot) - \phi(x)\big(\lambda_n(\cdot),\cdot\big)\big\|_T \\ &\quad = \sup_{(t,y)\in[0,T]\times[0,\infty)} \big|\phi(x_n)(t,y) - \phi(x)\big(\lambda_n(t),y\big)\big| \to 0, \quad \text{as } n \to \infty. \end{aligned}$$

Now, we have

$$
\begin{aligned}
&\bigl|\phi(x_n)(t,y)-\phi(x)\bigl(\lambda_n(t),y\bigr)\bigr| \\
&\quad=\biggl|F^c(y)x_n(t)-\int_0^t x_n(s-)\,dF(t+y-s) \\
&\qquad-F^c(y)x\bigl(\lambda_n(t)\bigr)+\int_0^{\lambda_n(t)} x(s-)\,dF\bigl(\lambda_n(t)+y-s\bigr)\biggr| \\
&\quad\le F^c(y)\bigl|x_n(t)-x\bigl(\lambda_n(t)\bigr)\bigr| \\
&\qquad+\biggl|\int_0^t x_n(s-)\,dF(t+y-s)-\int_0^{\lambda_n(t)} x(s-)\,dF\bigl(\lambda_n(t)+y-s\bigr)\biggr| \\
&\quad= F^c(y)\bigl|x_n(t)-x\bigl(\lambda_n(t)\bigr)\bigr| \\
&\qquad+\biggl|\int_0^t x_n(s-)\,dF(t+y-s)-\int_0^t x\bigl(\lambda_n(s)-\bigr)\,dF\bigl(\lambda_n(t)+y-\lambda_n(s)\bigr)\biggr| \\
&\quad\le F^c(y)\bigl|x_n(t)-x\bigl(\lambda_n(t)\bigr)\bigr|+\biggl|\int_0^t \bigl(x_n(s-)-x\bigl(\lambda_n(s)-\bigr)\bigr)\,dF(t+y-s)\biggr| \\
&\qquad+\biggl|\int_0^t x\bigl(\lambda_n(s)-\bigr)\,d\bigl(F\bigl(\lambda_n(t)+y-\lambda_n(s)\bigr)-F(t+y-s)\bigr)\biggr| \\
&\quad\le F^c(y)\bigl|x_n(t)-x\bigl(\lambda_n(t)\bigr)\bigr|+\|x_n-x\circ\lambda_n\|_T\bigl|F(y)-F(t+y)\bigr| \\
&\qquad+M\bigl|F\bigl(\lambda_n(t)+y\bigr)-F(t+y)\bigr| \\
&\quad\le 3\|x_n-x\circ\lambda_n\|_T+M\bigl|F\bigl(\lambda_n(t)+y\bigr)-F(t+y)\bigr|.
\end{aligned}
$$

The third term in the third inequality follows from the uniform continuity of the integrator because $F$ is continuous, monotone and bounded. By taking the supremum over $(t,y)\in[0,T]\times[0,\infty)$, the first term converges to 0 by the convergence of $x_n\to x$ in $D$, and the second term converges to 0 by the uniform convergence of $\lambda_n\to e$ in $[0,T]$ and the continuity of $F$. This implies the initial convergence to be shown, so that the mapping $\phi: D\to D_D$ is indeed continuous. □

## 6.2 Tightness

In this section, we establish tightness for the sequence of scaled processes in (3.13). It suffices to prove tightness of the sequences of processes $\{\hat{X}_{n,1}\colon n\ge 1\}$ and $\{\hat{X}_{n,2}\colon n\ge 1\}$ in $D_D$. By Assumption 1, the sequence of processes $\{\hat{A}_n\colon n\ge 1\}$ is tight. The tightness of $\{\hat{X}_{n,1}\}$ follows from the continuity of the mapping $\phi$ in $D_D$. It remains to show the tightness of $\{\hat{X}_{n,2}\}$ and then we obtain tightness of the sequences of processes $\{\bar{Q}_n^r\colon n\ge 1\}$ and $\{\hat{Q}_n^r\colon n\ge 1\}$ using the fact that tightness of product spaces is equivalent to the tightness on each of the component spaces.

**Theorem 6.1** *Under Assumptions 1 and 2 (F is continuous), the sequence of processes* $\{\hat{X}_{n,1}\colon n \geq 1\}$, $\{\hat{X}_{n,2}\colon n \geq 1\}$, $\{\bar{Q}_n^r\colon n \geq 1\}$ *and* $\{\hat{Q}_n^r\colon n \geq 1\}$ *are individually and jointly tight.*

In order to prove the tightness of $\{\hat{X}_{n,2}\colon n \geq 1\}$ defined in (2.14), we will closely follow the approach in [27], but we must adjust to the tightness criteria in $D_D$. The following tightness criteria are adapted to $D_D$ from Theorem 3.8.6 of Ethier and Kurtz [11]. For a review of tightness criteria for processes in the space $D$, see [50].

**Theorem 6.2** *A sequence of stochastic processes* $\{X_n\colon n \geq 1\}$ *in* $D_D$ *is tight if and only if*

(i) *the sequence* $\{X_n\colon n \geq 1\}$ *is stochastically bounded in* $D_D$, *i.e., for all* $\epsilon > 0$, *there exists a compact subset* $K \subset \mathbb{R}$ *such that*

$$P\big(\|X_n\|_T \in K\big) > 1 - \epsilon, \quad \text{for all } n \geq 1,$$

*where* $\|X_n\|_T = \sup_{s\in[0,T]}\{\sup_{t\in[0,T]} |X_n(s,t)|\}$; *and any one of the following*

(ii) *For all* $\delta > 0$, *and all uniformly bounded sequences* $\{\tau_n\colon n \geq 1\}$ *where for each* $n$, $\tau_n$ *is a stopping time with respect to the natural filtration* $\mathbf{F}_n = \{\mathcal{F}_n(t), t \in [0,T]\}$ *where* $\mathcal{F}_n(t) = \sigma\{X_n(s,\cdot)\colon 0 \leq s \leq t\}$, *there exists a constant* $\beta > 0$ *such that*

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \sup_{\tau_n} E\big[\big(1 \wedge d_{J_1}\big(X_n(\tau_n+\delta,\cdot), X_n(\tau_n,\cdot)\big)\big)^\beta\big] = 0;$$

*or*

(ii′) *For all* $\delta > 0$, *there exist a constant* $\beta$ *and random variables* $\gamma_n(\delta) \geq 0$ *such that for each* $n$, *w.p.*1,

$$E\big[\big(1 \wedge d_{J_1}\big(X_n(s+u,\cdot), X_n(s,\cdot)\big)\big)^\beta | \mathcal{F}_n\big]\big(1 \wedge d_{J_1}\big(X_n(s-v,\cdot), X_n(s,\cdot)\big)\big)^\beta \leq E\big[\gamma_n(\delta)|\mathcal{F}_n\big],$$

*for all* $0 \leq s \leq T$, $0 \leq u \leq \delta$ *and* $0 \leq v \leq s \wedge \delta$, *where* $\mathbf{F}_n = \{\mathcal{F}_n(t)\colon t \in [0,T]\}$ *with* $\mathcal{F}_n(t) = \sigma\{X_n(s,\cdot)\colon 0 \leq s \leq t\}$ *and*

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} E\big[\gamma_n(\delta)\big] = 0.$$

*Remark 6.1* The following condition is sufficient, but not necessary, for condition (ii) in Theorem 6.2:

*For all* $\delta_n \downarrow 0$ *and for all uniformly bounded sequences* $\{\tau_n\colon n \geq 1\}$, *where for each* $n$, $\tau_n$ *is a stopping time with respect to the natural filtration* $\mathbf{F}_n = \{\mathcal{F}_n(t)\colon t \in [0,T]\}$ *with* $\mathcal{F}_n(t) = \sigma\{X_n(s,\cdot)\colon 0 \leq s \leq t\}$,

$$d_{J_1}\big(X_n(\tau_n+\delta_n,\cdot), X_n(\tau_n,\cdot)\big) \Rightarrow 0, \quad \text{as } n \to \infty.$$

We will also need to generalize the tightness criteria in Lemma VI.3.32 in [22] for processes in the space $D$ to those in the space $D_D$ as in the following lemma, and its

proof also follows from that in [22] with inequalities for the modulus of continuity for functions in the space $D_D$.

**Lemma 6.2** *Suppose that a sequence of processes* $\{X_n\colon\ n \geq 1\}$ *in the space* $D_D$ *can be decomposed into two sequences* $\{Y_n^q\colon\ n \geq 1\}$ *and* $\{Z_n^q\colon\ n \geq 1\}$ *for some parameter* $q \in \mathbb{N}$, *i.e.,* $X_n = Y_n^q + Z_n^q$ *for each* $n \geq 1$, *and that* (i) *the sequence* $\{Y_n^q\colon\ n \geq 1\}$ *is tight in the space* $D_D$ *and* (ii) *for all* $T > 0$ *and* $\delta > 0$, $\lim_{q\to\infty} \limsup_{n\to\infty} P(\sup_{t,y\leq T} |Z_n^q(t,y)| > \delta) = 0$. *Then, the sequence* $\{X_n : n \geq 1\}$ *is tight in the space* $D_D$.

We first give a decomposition of the process $\hat{X}_{n,2}$ for each $n$. Following [27], we can write $\hat{R}_n(t,y)$ in (2.16) as

$$\hat{R}_n(t,y) = -\int_0^y \frac{\hat{R}_n(t,x)}{1-F(x)}\,dF(x) + \hat{L}_n(t,y),$$

where

$$\hat{L}_n(t,y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{A_n(t)} \left( \mathbf{1}(\eta_i \leq y) - \int_0^{y\wedge\eta_i} \frac{1}{1-F(x)}\,dF(x) \right).$$

We remark that we need not consider the left-hand limit of $\hat{R}_n$ in the second argument, as was done in [27], since the service-time c.d.f $F$ is assumed to be continuous, while $F$ is allowed to be discontinuous in [27]. Hence, $\hat{X}_{n,2}$ can be written as

$$\hat{X}_{n,2}(t,y) = \hat{G}_n(t,y) + \hat{H}_n(t,y), \quad \text{for } t \geq 0 \text{ and } y \geq 0, \tag{6.4}$$

where

$$\begin{aligned} \hat{G}_n(t,y) &\equiv \int_0^t \int_0^\infty \mathbf{1}(s+x \leq t+y)\,d\left( -\int_0^x \frac{\hat{R}_n(s,v)}{1-F(v)}\,dF(v) \right) \\ &= -\int_0^{t+y} \frac{\hat{R}_n(t+y-x,x)}{1-F(x)}\,dF(x), \end{aligned} \tag{6.5}$$

and

$$\hat{H}_n(t,y) \equiv \int_0^t \int_0^\infty \mathbf{1}(s+x \leq t+y)\,d\hat{L}_n(s,x). \tag{6.6}$$

Thus, the tightness of $\{\hat{X}_{n,2}\}$ follows from the tightness of $\{\hat{G}_n\}$ and $\{\hat{H}_n\}$. We will establish their tightness in the following two lemmas.

**Lemma 6.3** *Under Assumptions 1 and 2 (F is continuous), the sequence of processes* $\{\hat{G}_n\colon\ n \geq 1\} \equiv \{\{\hat{G}_n(t,y)\colon\ t \geq 0, y \geq 0\}, n \geq 1\}$ *is tight in the space* $D_D$.

*Proof* We will apply Lemma 6.2. We define the sequence of processes $\{\hat{G}_n^\epsilon\colon n \geq 1\}$, for some $\epsilon \in (0,1)$, by

$$\hat{G}_n^\epsilon(t,y) \equiv -\int_0^{t+y} \frac{\hat{R}_n(t+y-x,x)}{1-F(x)} \mathbf{1}\big(F(x) \leq 1-\epsilon\big)\, dF(x), \quad t,y \geq 0. \tag{6.7}$$

We will prove that $\{\hat{G}_n^\epsilon\colon n \geq 1\}$ is tight in $D_D$ and

$$\lim_{\epsilon \downarrow 0} \limsup_n P\left( \sup_{t,y \leq T} \left| \int_0^{t+y} \frac{\hat{R}_n(t+y-x,x)}{1-F(x)} \mathbf{1}\big(F(x) > 1-\epsilon\big)\, dF(x) \right| > \delta \right) = 0, \tag{6.8}$$

for each $\delta > 0$ and $T > 0$, and thus will conclude that the sequence $\{\hat{G}_n\}$ is tight in $D_D$ by Lemma 6.2. It is easy to see that (6.8) follows easily from (3.23) in [27]. So we only need to prove the tightness of the sequence of processes $\{\hat{G}_n^\epsilon\colon n \geq 1\}$.

Recall that $\hat{R}_n(t+y-x,x) = \hat{U}_n(\bar{A}_n(t+y-x), F(x))$. By (2.1) and $\hat{U}_n \Rightarrow U$ in (2.5) as $n \to \infty$, and by applying the continuous mapping theorem to the composition map of $\hat{U}_n$ with respect to the first argument (Theorem 13.2.2, [48]), we obtain

$$\hat{R}_n(t+y-x,x) = \hat{U}_n\big(\bar{A}_n(t+y-x), F(x)\big) \Rightarrow U\big(\bar{a}(t+y-x), F(x)\big) \quad \text{in } D_D,$$

as $n \to \infty$. The weak convergence of $\{\hat{R}_n : n \geq 1\}$ in $D_D$ implies that $\{\hat{R}_n : n \geq 1\}$ is stochastically bounded, so the integral representation of $\hat{G}_n^\epsilon$ in terms of $\hat{R}_n$ in (6.7) implies that $\{\hat{G}_n^\epsilon\colon n \geq 1\}$ is also stochastically bounded in $D_D$. We apply Theorem 6.2 to prove the tightness of $\{\hat{G}_n^\epsilon\colon n \geq 1\}$ in $D_D$. In this case, it is convenient to use the sufficient criterion in the remark right after Theorem 6.2.

Let $\mathbf{G}_n = \{\mathcal{G}_n(t)\colon t \in [0,T]\}$ be a filtration defined by

$$\begin{aligned} \mathcal{G}_n(t) &= \sigma\big\{\hat{R}_n(s,\cdot)\colon 0 \leq s \leq t\big\} \vee \mathcal{N} \\ &= \sigma\big\{\eta_i \leq x\colon 1 \leq i \leq A_n(t), x \geq 0\big\} \vee \sigma\big\{A_n(s)\colon 0 \leq s \leq t\big\} \vee \mathcal{N}, \end{aligned}$$

where $\mathcal{N}$ includes all the null sets. Note that the filtration $\mathbf{G}_n$ satisfies the usual conditions (Chap. 1, [24] and proof of Lemma 3.1 in [27]). Let $\delta_n \downarrow 0$ and $\{\tau_n\colon n \geq 1\}$ be a uniformly bounded sequence, where for each $n$, $\tau_n$ is a stopping times with respect to the filtration $\mathbf{G}_n$. Then, it suffices to show that

$$d_{J_1}\big(\hat{G}_n^\epsilon(\tau_n+\delta_n,\cdot), \hat{G}_n^\epsilon(\tau_n,\cdot)\big) \Rightarrow 0, \quad \text{as } n \to \infty.$$

Consider any sequence of nondecreasing homeomorphism $\{\lambda_n\colon n \geq 1\}$ on $[0,T]$ such that $\lim_{n\to\infty} \lambda_n(y) = y$ uniformly in $y \in [0,T]$. We want to show that the following holds:

$$\sup_{0 \leq y \leq T} \big|\hat{G}_n^\epsilon\big(\tau_n+\delta_n, \lambda_n(y)\big) - \hat{G}_n^\epsilon(\tau_n, y)\big| \Rightarrow 0, \quad \text{as } n \to \infty.$$

Now,

$$\begin{aligned}
&\sup_{0\le y\le T}\left|\hat{G}_n^{\epsilon}\bigl(\tau_n+\delta_n,\lambda_n(y)\bigr)-\hat{G}_n^{\epsilon}(\tau_n,y)\right| \\
&\quad=\sup_{0\le y\le T}\left|\int_0^{\tau_n+\delta_n+\lambda_n(y)}\frac{\hat{R}_n(\tau_n+\delta_n+\lambda_n(y)-x,x)}{1-F(x)}\mathbf{1}\bigl(F(x)\le 1-\epsilon\bigr)\,dF(x)\right. \\
&\qquad\left.-\int_0^{\tau_n+y}\frac{\hat{R}_n(\tau_n+y-x,x)}{1-F(x)}\mathbf{1}\bigl(F(x)\le 1-\epsilon\bigr)\,dF(x)\right| \\
&\quad\le\sup_{0\le y\le T}\left|\int_0^{\tau_n+\delta_n+\lambda_n(y)}\frac{\hat{R}_n(\tau_n+\delta_n+\lambda_n(y)-x,x)-\hat{R}_n(\tau_n+y-x,x)}{1-F(x)}\right. \\
&\qquad\left.\times\mathbf{1}\bigl(F(x)\le 1-\epsilon\bigr)\,dF(x)\right| \\
&\qquad+\sup_{0\le y\le T}\left|\int_0^{\tau_n+\delta_n+\lambda_n(y)}\frac{\hat{R}_n(\tau_n+y-x,x)}{1-F(x)}\mathbf{1}\bigl(F(x)\le 1-\epsilon\bigr)\,dF(x)\right. \\
&\qquad\left.-\int_0^{\tau_n+y}\frac{\hat{R}_n(\tau_n+y-x,x)}{1-F(x)}\mathbf{1}\bigl(F(x)\le 1-\epsilon\bigr)\,dF(x)\right| \\
&\quad\Rightarrow 0
\end{aligned}$$

as $n\to\infty$, where the first and the second terms converge to 0 by the stochastic boundedness and weak convergence of $\hat{R}_n$ in $D_D$, and because $\tau_n$ is uniformly bounded, $\lambda_n(y)$ converges to $y$ uniformly in $[0,T]$, and $\delta_n\downarrow 0$ as $n\to\infty$. Hence, the processes $\{\hat{G}_n^{\epsilon}\}$ are tight in $D_D$ and the proof is completed. □

**Lemma 6.4** *Under Assumptions 1 and 2 (F is continuous), the sequence of processes* $\{\hat{H}_n\colon n\ge 1\}\equiv\{\{\hat{H}_n(t,y)\colon t\ge 0,\,y\ge 0\},n\ge 1\}$ *is tight in* $D_D$.

*Proof* As in Lemma 3.7 in [27], we write the process $\hat{H}_n$ as

$$\hat{H}_n(t,y)=\frac{1}{\sqrt{n}}\sum_{i=1}^{A_n(t)}\left(\mathbf{1}\bigl(0\le\eta_i\le t+y-\tau_i^n\bigr)-\int_0^{\eta_i\wedge(t+y-\tau_i^n)^+}\frac{1}{1-F(u)}\,dF(u)\right).$$

We will apply Theorem 6.2 to prove the tightness of $\{\hat{H}_n\colon n\ge 1\}$ in $D_D$. In this case, it is convenient to use criterion (ii) in Theorem 6.2. We will first prove that this criterion holds, and then prove the stochastic boundedness of the sequence of processes $\{\hat{H}_n\colon n\ge 1\}$.

Let $\mathbf{H}_n=\{\mathcal{H}_n(t)\colon t\in[0,T]\}$ be a filtration defined by

$$\begin{aligned}
\mathcal{H}_n(t)&=\sigma\bigl\{\hat{H}_n(s,\cdot)\colon 0\le s\le t\bigr\}\vee\mathcal{N} \\
&=\sigma\bigl\{\eta_i\le s+x-\tau_i^n\colon 1\le i\le A_n(t),x\ge 0,0\le s\le t\bigr\} \\
&\quad\vee\bigl\{A_n(s)\colon 0\le s\le t\bigr\}\vee\mathcal{N},
\end{aligned}$$

where $\mathcal{N}$ includes all the null sets. The filtration $\mathbf{H}_n$ satisfies the usual conditions (see p. 254 in [27]).

Let $\delta > 0$ and $\{\kappa_n\colon n \geq 1\}$ be a uniformly bounded sequence, where for each $n$, $\kappa_n$ is a stopping time with respect to the filtration $\mathbf{H}_n$. It suffices to show that

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \sup_{\kappa_n} E\big[d_{J_1}\big(\hat{H}_n(\kappa_n+\delta,\cdot), \hat{H}_n(\kappa_n,\cdot)\big)^2\big] = 0. \tag{6.9}$$

Consider any sequence of nondecreasing homeomorphism $\{\lambda_n\colon n \geq 1\}$ on $[0,T]$ such that $\lim_{n\to\infty} \lambda_n(y) = y$ uniformly in $y \in [0,T]$. We want to show that the following holds:

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \sup_{\kappa_n} E\Big[\Big(\sup_{0\leq y\leq T} \big|\hat{H}_n\big(\kappa_n+\delta, \lambda_n(y)\big) - \hat{H}_n(\kappa_n, y)\big|\Big)^2\Big] = 0. \tag{6.10}$$

Define the processes $\hat{H}_{n,i} \equiv \{\hat{H}_{n,i}(t,y)\colon t, y \geq 0\}$ by

$$\hat{H}_{n,i}(t,y) \equiv \mathbf{1}\big(0 \leq \eta_i \leq t+y-\tau_i^n\big) - \int_0^{\eta_i \wedge (t+y-\tau_i^n)^+} \frac{1}{1-F(u)}\, dF(u).$$

As in Lemma 3.5 in [27], one can check that for each fixed $y$ and for each $i$, the process $\{\hat{H}_{n,i}(t,y)\colon t \geq 0\}$ is a square integrable martingale with respect to the filtration $\mathbf{H}_n$ and it has predictable quadratic variation

$$\big\langle \hat{H}_{n,i}(\cdot,y)\big\rangle(t) = \langle \hat{H}_{n,i}\rangle(t,y) = \int_0^{\eta_i \wedge (t+y-\tau_i^n)^+} \frac{1}{1-F(u)}\, dF(u), \quad \text{for } t \geq 0,$$

and that the $\mathbf{H}_n$ martingales $\hat{H}_{n,i}(\cdot,y)$ and $\hat{H}_{n,j}(\cdot,y)$ for each fixed $y$ are orthogonal for $i \neq j$.

Thus, for each fixed $y$ and constant $K > 0$, the process $\hat{H}_n^{(K)} = \{\hat{H}_n^{(K)}(t,y)\colon t \geq 0\}$ defined by

$$\hat{H}_n^{(K)}(t,y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n(\bar{A}_n(t)\wedge K)} \Big(\mathbf{1}\big(0 \leq \eta_i \leq t+y-\tau_i^n\big) - \int_0^{\eta_i \wedge (t+y-\tau_i^n)^+} \frac{1}{1-F(u)}\, dF(u)\Big),$$

is an $\mathbf{H}_n$ square integrable martingale with predictable quadratic variation

$$\big\langle \hat{H}_n^{(K)}(\cdot,y)\big\rangle(t) = \big\langle \hat{H}_n^{(K)}\big\rangle(t,y) = \frac{1}{n} \sum_{i=1}^{n(\bar{A}_n(t)\wedge K)} \int_0^{\eta_i \wedge (t+y-\tau_i^n)^+} \frac{1}{1-F(u)}\, dF(u),$$

for $t \geq 0$. By the SLLN,

$$\frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \int_0^{\eta_i} \frac{1}{1-F(u)}\, dF(u) \to t, \quad \text{a.s. as } n \to \infty. \tag{6.11}$$

So for each fixed $y$, the sequence of quadratic variations $\{\langle \hat{H}_n^{(K)}(\cdot, y)\rangle\colon\ n \geq 1\}$ is $C$-tight by the continuity of $\bar{a}$. (Recall that a sequence $\{Y_n\}$ is said to be C-tight if it is tight and the limit of any convergent subsequence must have continuous sample paths.) It follows by Theorem 3.6 in [50] that the sequence $\{\hat{H}_n^{(K)}(\cdot, y)\colon\ n \geq 1\}$ is $C$-tight for each fixed $y$.

Now, to prove (6.10), we have

$$
\begin{aligned}
&E\Big[\Big(\sup_{0\leq y\leq T}\big|\hat{H}_n\big(\kappa_n+\delta,\lambda_n(y)\big)-\hat{H}_n(\kappa_n,y)\big|\Big)^2\Big]\\
&\quad\leq 2E\Big[\sup_{0\leq y\leq T}\big|\hat{H}_n\big(\kappa_n+\delta,\lambda_n(y)\big)-\hat{H}_n\big(\kappa_n,\lambda_n(y)\big)\big|^2\Big]\\
&\qquad+2E\Big[\sup_{0\leq y\leq T}\big|\hat{H}_n\big(\kappa_n,\lambda_n(y)\big)-\hat{H}_n(\kappa_n,y)\big|^2\Big]\\
&\quad=2\lim_{K\to\infty}E\Big[\sup_{0\leq y\leq T}\big|\hat{H}_n^{(K)}\big(\kappa_n+\delta,\lambda_n(y)\big)-\hat{H}_n^{(K)}\big(\kappa_n,\lambda_n(y)\big)\big|^2\Big]\\
&\qquad+2\lim_{K\to\infty}E\Big[\sup_{0\leq y\leq T}\big|\hat{H}_n^{(K)}\big(\kappa_n,\lambda_n(y)\big)-\hat{H}_n^{(K)}(\kappa_n,y)\big|^2\Big],
\end{aligned}
$$

where the equality holds by the dominated convergence and by stochastic boundedness of $A_n$. The first term converges to 0 as $n\to\infty$ and $\delta\downarrow 0$ by the assumptions on $\kappa_n$ and $\lambda_n$ and $C$-tightness of $\{\hat{H}_n^{(K)}\colon\ n\geq 1\}$. We conclude that the second term converges to 0 by observing

$$
\begin{aligned}
&\hat{H}_n^{(K)}\big(\kappa_n,\lambda_n(y)\big)-\hat{H}_n^{(K)}(\kappa_n,y)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^{A_n(\kappa_n)\wedge K}\Bigg(\mathbf{1}\big(0\leq\eta_i\leq\kappa_n+\lambda_n(y)-\tau_i^n\big)-\mathbf{1}\big(0\leq\eta_i\leq\kappa_n+y-\tau_i^n\big)\\
&\qquad-\Bigg(\int_0^{\eta_i\wedge(\kappa_n+\lambda_n(y)-\tau_i^n)^+}\frac{1}{1-F(u)}\,dF(u)\\
&\qquad-\int_0^{\eta_i\wedge(\kappa_n+y-\tau_i^n)^+}\frac{1}{1-F(u)}\,dF(u)\Bigg)\Bigg).
\end{aligned}
$$

Thus we obtain (6.10).

Now we prove the stochastic boundedness of $\{\hat{H}_n\colon\ n\geq 1\}$ in $D_D$. We observe that for each $n$, each sample path of the process $\hat{H}_n$ is bounded by that of the process $\tilde{H}_n$ defined by

$$
\begin{aligned}
\tilde{H}_n(t,y)=\frac{1}{\sqrt{n}}\sum_{i=1}^{A_n(t+y)}\Bigg(&\mathbf{1}\big(0\leq\eta_i\leq t+y-\tau_i^n\big)\\
&-\int_0^{\eta_i\wedge(t+y-\tau_i^n)^+}\frac{1}{1-F(u)}\,dF(u)\Bigg).
\end{aligned}
$$

The stochastic boundedness of $\{\tilde{H}_n\colon\ n \geq 1\}$ in $D_D$ follows directly from the proof of Lemma 3.7 in [27]. Therefore, $\{\hat{H}_n\colon\ n \geq 1\}$ is stochastically bounded, so that tightness of $\{\hat{H}_n\colon\ n \geq 1\}$ in $D_D$ is proved. □

### 6.3 Convergence of the finite-dimensional distributions

In this section, we complete the proof of the convergence $(\hat{X}_{n,1}, \hat{X}_{n,2}) \Rightarrow (\hat{X}_1, \hat{X}_2)$ in $D_D \times D_D$ by proving that the finite-dimensional distributions of $(\hat{X}_{n,1}, \hat{X}_{n,2})$ converge to those of $(\hat{X}_1, \hat{X}_2)$ since we have proved the tightness of $\{(\hat{X}_{n,1}, \hat{X}_{n,2})\colon\ n \geq 1\}$ in Sect. 6.2. We will mostly have to deal with $\hat{X}_{n,2}$, since we have already shown convergence of $\hat{X}_{n,1}$. Our argument for $\hat{X}_{n,2}$ will also enable us to establish joint convergence of the two finite-dimensional distributions.

**Lemma 6.5** *Under Assumptions 1 and 2 (F is continuous), the finite-dimensional distributions of* $(\hat{X}_{n,1}, \hat{X}_{n,2})$ *converge to those of* $(\hat{X}_1, \hat{X}_2)$ *as* $n \to \infty$.

*Proof* First of all, we understand the integrals $\hat{X}_{n,2}$ in (2.14) and $\hat{X}_2$ ($\equiv \hat{X}_2^{c,r}$) in (3.16) as mean-square integrals, so that they can be represented as

$$\hat{X}_{n,2}(t,y) = \text{l.i.m.}_{k\to\infty} \hat{X}_{n,2,k}(t,y), \quad \text{and} \quad \hat{X}_2(t,y) = \text{l.i.m.}_{k\to\infty} \hat{X}_{2,k}(t,y),$$

where l.i.m. means limit in mean square, that is,

$$\lim_{k\to\infty} E\big[\big(\hat{X}_{n,2}(t,y) - \hat{X}_{n,2,k}(t,y)\big)^2\big] = 0 \quad \text{and}$$

$$\lim_{k\to\infty} E\big[\big(\hat{X}_2(t,y) - \hat{X}_{2,k}(t,y)\big)^2\big] = 0,$$

$$\begin{aligned}\hat{X}_{n,2,k}(t,y) &\equiv -\int_0^t \int_0^\infty \mathbf{1}_{k,t}^y(s,x)\, d\hat{U}_n\big(\bar{A}_n(s), F(x)\big) \\ &= -\sum_{i=1}^k \big[\Delta_{\hat{U}_n}\big(\bar{A}_n\big(s_{i-1}^k\big), \bar{A}_n\big(s_i^k\big), 0, F\big(t+y-s_i^k\big)\big)\big],\end{aligned}$$

and

$$\begin{aligned}\hat{X}_{2,k}(t,y) &\equiv -\int_0^t \int_0^\infty \mathbf{1}_{k,t}^y(s,x)\, dU\big(\bar{a}(s), F(x)\big) \\ &= -\sum_{i=1}^k \big[\Delta_U\big(\bar{a}\big(s_{i-1}^k\big), \bar{a}\big(s_i^k\big), 0, F\big(t+y-s_i^k\big)\big)\big],\end{aligned}$$

where $\mathbf{1}_{k,t}^y$ is defined by

$$\mathbf{1}_{k,t}^y(s,x) = \mathbf{1}(s=0)\mathbf{1}(x \leq t+y) + \sum_{i=1}^k \mathbf{1}\big(s \in \big(s_{i-1}^k, s_i^k\big]\big)\mathbf{1}\big(x \leq t+y-s_i^k\big), \tag{6.12}$$

with the points $0 = s_0^k < s_1^k < \cdots < s_k^k = t$ chosen so that $\max_{1\le i\le k} |s_{i-1}^k - s_i^k| \to 0$ as $k \to \infty$, and $\Delta_{\hat{U}_n}$ and $\Delta_U$ are defined as $\Delta_{\hat{K}}$ in (4.1).

We prove the convergence of the finite-dimensional distributions of $\hat{X}_{n,2}$ to those of $\hat{X}_2$ by taking advantage of the convergence of $\hat{U}_n \Rightarrow U$ as $n \to \infty$ in $D([0,\infty), D([0,1],\mathbb{R}))$ (see (2.5)), for which we define another process $\{\tilde{X}_{n,2,k}(t,y)\colon t, y \ge 0\}$ in $D_D$ for each $n$ by replacing the $\bar{A}_n$ terms in $\Delta_{\hat{U}_n}$ of $\hat{X}_{n,2,k}$ by $\bar{a}$ as follows,

$$\begin{aligned}\tilde{X}_{n,2,k}(t,y) &\equiv -\int_0^t \int_0^\infty \mathbf{1}_{k,t}^y(s,x)\, d\hat{U}_n\big(\bar{a}(s), F(x)\big) \\ &= -\sum_{i=1}^k \big[\Delta_{\hat{U}_n}\big(\bar{a}\big(s_{i-1}^k\big), \bar{a}\big(s_i^k\big), 0, F\big(t+y-s_i^k\big)\big)\big].\end{aligned}$$

Hence, we easily obtain the convergence of the finite-dimensional distributions of $\tilde{X}_{n,2,k}$ to those of $\hat{X}_{2,k}$ as $n \to \infty$, since $\bar{a}$ and $F$ are both continuous by Assumptions 1 and 2, and the finite-dimensional distributions of $\hat{U}_n$ converge to those of $U$ as $n \to \infty$ and $U$ is continuous.

Moreover, since $\hat{K}_n$ ($\hat{U}_n$) and $A_n$ are independent by Assumptions 1 and 2, $\tilde{X}_{n,2,k}$ and $\hat{X}_{n,1}$ are independent, and since the limit processes $\hat{X}_{2,k}$ and $\hat{X}_1$ are also independent, we obtain the joint convergence of the finite-dimensional distributions of $(\hat{X}_{n,1}, \tilde{X}_{n,2,k})$ to those of $(\hat{X}_1, \hat{X}_{2,k})$ as $n \to \infty$.

Now it suffices to show that the difference between $\hat{X}_{n,2,k}$ and $\tilde{X}_{n,2,k}$ is asymptotically negligible in probability as $n \to \infty$, and the difference between $\hat{X}_{n,2}$ and $\hat{X}_{n,2,k}$ is asymptotically negligible in probability as $n \to \infty$ and $k \to \infty$, i.e.,

$$\lim_{n\to\infty} P\Big(\sup_{0\le t\le T, y\ge 0} \big|\hat{X}_{n,2,k}(t,y) - \tilde{X}_{n,2,k}(t,y)\big| > \epsilon\Big) = 0, \quad T > 0,\ \epsilon > 0. \tag{6.13}$$

and

$$\lim_{k\to\infty} \limsup_{n\to\infty} P\big(\big|\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)\big| > \epsilon\big) = 0, \quad t, y \ge 0,\ \epsilon > 0. \tag{6.14}$$

We obtain (6.13) easily from Assumption 1 and (2.5) since $\bar{a}$ and $U$ are continuous. Now we proceed to prove (6.14). We will follow a martingale approach argument similar to the one used in Lemma 5.3 of [27], which relies on their technical Lemma 5.2. Fortunately, for our two-parameter processes, the conditions of Lemma 5.2 [27] are satisfied by fixing the second argument. We have for $t, y \ge 0$ and $\Upsilon > 0$,

$$\begin{aligned}&P\big(\big|\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)\big| > \epsilon\big) \\ &\quad \le P\big(\bar{A}_n(t) > \Upsilon\big) + P\big(\big|\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)\big| > \epsilon, \bar{A}_n(t) \le \Upsilon\big).\end{aligned} \tag{6.15}$$

On $\{\bar{A}_n(t) \le \Upsilon\}$,

$$\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)$$

$$= \int_0^t \int_0^\infty \big(\mathbf{1}_{k,t}^y(s,x) - \mathbf{1}(s+x \le t+y)\big)\, d\hat{U}_n\big(\bar{A}_n(s), F(x)\big)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{A_n(t)\wedge(n\Upsilon)} \beta_i\big(\tau_i^n, \eta_i\big)(t,y),$$

where

$$\beta_i\big(\tau_i^n, \eta_i\big)(t,y) = \sum_{j=1}^{k} \mathbf{1}\big(s_{j-1}^k < \tau_i^n \le s_j^k\big)\big(\mathbf{1}\big(t+y-s_j^k < \eta_i \le t+y-\tau_i^n\big) - \big(F\big(t+y-\tau_i^n\big) - F\big(t+y-s_j^k\big)\big)\big).$$

Define the process $Z_n^{(\Upsilon)} \equiv \{Z_n^{(\Upsilon)}(t,y)\colon\ t,y \ge 0\}$ by

$$Z_n^{(\Upsilon)}(t,y) \equiv \sum_{i=1}^{A_n(t)\wedge(n\Upsilon)} \beta_i\big(\tau_i^n, \eta_i\big)(t,y), \quad t,y \ge 0.$$

As in Lemma 5.2 in [27], one can check that for each fixed $y > 0$, the process $Z_n^{(\Upsilon)}(\cdot,y) = \{Z_n^{(\Upsilon)}(t,y)\colon\ t \ge 0\}$ is a square integrable martingale with respect to the filtration $\mathbf{F}_n = \{\mathcal{F}_n(t), t \ge 0\}$, where

$$\mathcal{F}_n(t) = \sigma\big\{\eta_i \le s+x\colon\ 1 \le i \le A_n(t), x \ge 0, 0 \le s \le t\big\} \vee \big\{A_n(s)\colon\ 0 \le s \le t\big\} \vee \mathcal{N},$$

and the quadratic variation of $Z_n^{(\Upsilon)}(\cdot,y)$ is

$$\begin{aligned}
\big\langle Z_n^{(\Upsilon)}(\cdot,y)\big\rangle(t) &= \big\langle Z_n^{(\Upsilon)}\big\rangle(t,y) \\
&= \sum_{i=1}^{A^n(t)\wedge(n\Upsilon)} E\big[\beta_i\big(\tau_i^n,\eta_i\big)(t,y)^2\big] \\
&= \sum_{i=1}^{A^n(t)\wedge(n\Upsilon)} \sum_{j=1}^{k} \big[\mathbf{1}\big(s_{j-1} < \tau_i^n \le s_j^k\big)\big(F\big(t+y-\tau_i^n\big) - F\big(t+y-s_j^k\big)\big) \\
&\quad \cdot \big(1 - \big(F\big(t+y-\tau_i^n\big) - F\big(t+y-s_j^k\big)\big)\big)\big] \\
&\le \sum_{i=1}^{A^n(t)\wedge(n\Upsilon)} \sum_{j=1}^{k} \big[\mathbf{1}\big(s_{j-1} < \tau_i^n \le s_j^k\big)\big(F\big(t+y-\tau_i^n\big) - F\big(t+y-s_j^k\big)\big)\big] \\
&= \sum_{j=1}^{k} \big(F\big(t+y-s_{j-1}^k\big) - F\big(t+y-s_j^k\big)\big)\big(A_n\big(s_j^k\big) - A_n\big(s_{j-1}^k\big)\big) \\
&\le \sup_{1\le j\le k} \big\{A_n\big(s_j^k\big) - A_n\big(s_{j-1}^k\big)\big\},
\end{aligned}$$

where the last inequality follows from the fact that the sum of the coefficients before the $A_n(s_j^k) - A_n(s_{j-1}^k)$ terms is less than 1. So for fixed $y \geq 0$, and on $\{\bar{A}^n(t) \leq \Upsilon\}$,

$$\begin{aligned}
&\lim_{k\to\infty} \limsup_{n\to\infty} E\big[\big(\hat{X}_{n,2}(t,y) - \hat{X}_{n,2,k}(t,y)\big)^2\big] \\
&\quad = \lim_{k\to\infty} \limsup_{n\to\infty} E\left[\left\langle \frac{1}{\sqrt{n}} Z_n^{(\Upsilon)}(\cdot, y)\right\rangle(t)\right] \\
&\quad \leq \lim_{k\to\infty} \limsup_{n\to\infty} E\Big[\sup_{1\leq j\leq k} \big\{\bar{A}_n\big(s_j^k\big) - \bar{A}_n\big(s_{j-1}^k\big)\big\}\Big] = 0,
\end{aligned}$$

where the convergence to 0 holds because of the continuity of $\bar{a}$, Assumption 1 and $\max_{1\leq j\leq k}(s_j^k - s_{j-1}^k) \to 0$ as $k \to \infty$.

Hence, (6.15) becomes

$$\begin{aligned}
&P\big(\big|\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)\big| > \epsilon\big) \\
&\quad \leq P\big(\bar{A}_n(t) > \Upsilon\big) + \frac{1}{\epsilon^2} E\left[\left\langle \frac{1}{\sqrt{n}} Z_n^{(\Upsilon)}(\cdot, y)\right\rangle(t)\right] \\
&\quad \leq P\big(\bar{A}_n(t) > \Upsilon\big) + \frac{1}{\epsilon^2} E\Big[\sup_{1\leq j\leq k} \big\{\bar{A}_n\big(s_j^k\big) - \bar{A}_n\big(s_{j-1}^k\big)\big\}\Big].
\end{aligned}$$

Therefore, by the stochastic boundedness of $\bar{A}_n$, (6.14) is proved. That concludes the demonstration that the finite-dimensional distributions of $(\hat{X}_{n,1}, \hat{X}_{n,2})$ converge to those of $(\hat{X}_1, \hat{X}_2)$ as $n \to \infty$. □

**Acknowledgements** This research was supported by NSF grants DMI-0457095 and CMMI-0948190.

## References

1. Bickel, P.J., Wichura, M.J.: Convergence criteria for multiparameter stochastic processes and some applications. Ann. Math. Stat. **42**, 1656–1670 (1971)
2. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968) (2nd edn., 1999)
3. Borovkov, A.A.: On limit laws for service processes in multi-channel systems. Sib. Math J. **8**, 746–763 (1967) (in Russian)
4. Brémaud, P.: Point Processes and Queues: Martingale Dynamics. Springer, Berlin (1981)
5. Cairoli, R.: Sur une equation differentielle stochastique. C. R. Acad. Sci. Paris Ser. A **274**, 1739–1742 (1972)
6. Cairoli, R., Walsh, J.B.: Stochastic integrals in the plane. Acta Math. **134**, 111–183 (1975)
7. Csörgö, M., Révész, P.: Strong Approximations in Probability and Statistics. Wiley, New York (1981)
8. Decreusefond, L., Moyal, P.: A functional central limit theorem for the $M/GI/\infty$ queue. Ann. Appl. Probab. **18**(6), 2156–2178 (2008)
9. Duffield, N.G., Whitt, W.: Control and recovery from rare congestion events in a large multi-server system. Queueing Syst. **26**, 69–104 (1997)
10. Eick, S.G., Massey, W.A., Whitt, W.: The physics of the $M_t/G/\infty$ queue. Oper. Res. **41**, 731–742 (1993)
11. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. Wiley, New York (1986)
12. Gaenssler, P., Stute, W.: Empirical processes: a survey of results for independent and identically distributed random variables. Ann. Probab. **7**, 193–243 (1979)

13. Glynn, P.W.: On the Markov property of the $GI/G/\infty$ Gaussian limit. Adv. Appl. Probab. **14**, 191–194 (1982)
14. Glynn, P.W., Whitt, W.: A new view of the heavy-traffic limit theorem for the infinite-server queue. Adv. Appl. Probab. **23**, 188–209 (1991)
15. Goldberg, D.A., Whitt, W.: The last departure time from an $M_t/GI/\infty$ queue with a terminating arrival process. Queueing Syst. **58**, 77–104 (2008)
16. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**(3), 567–587 (1981)
17. Harrison, J.M., Reiman, M.I.: Reflected Brownian motion in an orthant. Ann. Probab. **9**, 302–308 (1981)
18. Iglehart, D.L.: Limit diffusion approximations for the many server queue and the repairman problem. J. Appl. Probab. **2**, 429–441 (1965)
19. Iglehart, D.L.: Weak convergence of compound stochastic processes. Stoch. Process. Appl. **1**, 11–31 (1973)
20. Iglehart, D.L., Whitt, W.: Multichannel queues in heavy traffic, I. Adv. Appl. Probab. **2**, 150–177 (1970)
21. Iglehart, D.L., Whitt, W.: Multichannel queues in heavy traffic, II: sequences, networks and batches. Adv. Appl. Probab. **2**, 355–369 (1970)
22. Jacod, J., Shiryayev, A.N.: Limit Theorems for Stochastic Processes. Springer, Berlin (1987)
23. Kang, W., Ramanan, K.: Fluid limits of many-server queues with reneging. Ann. Appl. Probab. (2010, to appear)
24. Karatzas, I., Shreve, S.: Brownian Motion and Stochastic Calculus. Springer, Berlin (1991)
25. Kaspi, H., Ramanan, K.: Law of large numbers limits for many-server queues. Ann. Appl. Probab. (2010, to appear)
26. Khoshnevisan, D.: Multiparameter Processes: An Introduction to Random Fields. Springer, Berlin (2002)
27. 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)
28. Kurtz, T.G., Protter, P.: Weak limit theorems for stochastic integrals and stochastic differential equations. Ann. Probab. **19**, 1035–1070 (1991)
29. Kurtz, T.G., Protter, P.: Weak convergence of stochastic integrals and differential equations II: Infinite dimensional case. Lect. Notes Math. **1627**, 197–285 (1996)
30. Louchard, G.: Large finite population queuing systems. Part I: the infinite server model. Stoch. Models **4**, 373–505 (1988)
31. Mamatov, K.M.: Weak convergence of stochastic integrals with respect to semimartingales. Russ. Math. Surv. **41**(5), 155–156 (1986)
32. Mandelbaum, A., Momcilovic, P.: Queues with many servers and impatient customers. EECS Department, University of Michigan (2009)
33. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. Queueing Syst. **30**, 149–201 (1998)
34. Massey, W.A., Whitt, W.: Networks of infinite-server queues with nonstationary Poisson input. Queueing Syst. **13**, 183–250 (1993)
35. Neuhaus, G.: On weak convergence of stochastic processes with multidimensional time parameter. Ann. Math. Stat. **42**, 1285–1295 (1971)
36. Pang, G., Whitt, W.: Two-parameter heavy-traffic limits for infinite-server queues: longer version. Columbia University (2009). Available at: http://www.columbia.edu/~ww2040
37. Pang, G., Talreja, R., Whitt, W.: Martingale proofs of many-server heavy-traffic limits for Markovian queues. Probab. Surv. **4**, 193–267 (2007)
38. Puhalskii, A.A., Reed, J.E.: On many-server queues in heavy traffic. Ann. Appl. Probab. **20**(1), 129–195 (2010)
39. Reed, J.E.: The G/GI/N queue in the Halfin–Whitt regime I: infinite-server queue system equations. Ann. Appl. Probab. **19**(6), 2211–2269 (2009)
40. Reed, J.E.: The G/GI/N queue in the Halfin–Whitt regime II: idle-time system equations. Working paper, The Stern School, NYU (2007)
41. Reed, J., Talreja, R.: Distribution-valued heavy-traffic limits for $GI/GI/\infty$ queues. Preprint (2009)
42. Skorohod, A.V.: Limit theorems for stochastic processes. Probab. Theory Appl. **1**, 261–290 (1956)
43. Straf, M.L.: Weak convergence of stochastic processes with several parameters. Proc. Sixth Berkeley Symp. Math. Stat. Probab. **2**, 187–221 (1971)

44. Talreja, R., Whitt, W.: Heavy-traffic limits for waiting times in many-server queues with abandonments. Ann. Appl. Probab. **19**(6), 2137–2175 (2009)
45. van Der Vaart, A.W., Wellner, J.: Weak Convergence and Empirical Processes. Springer, Berlin (1996)
46. Walsh, J.B.: Martingales with a multidimensional parameter and stochastic integrals in the plane. In: Lectures in Probability and Statistics, pp. 329–491. Springer, Berlin (1986)
47. Whitt, W.: On the heavy-traffic limit theorem for $GI/G/\infty$ queues. Adv. Appl. Probab. **14**, 171–190 (1982)
48. Whitt, W.: Stochastic-Process Limits. Springer, Berlin (2002)
49. Whitt, W.: Fluid models for multiserver queues with abandonments. Oper. Res. **54**, 37–54 (2006)
50. Whitt, W.: Proofs of the martingale FCLT: a review. Probab. Surv. **4**, 268–302 (2007)
51. Wong, E., Zakai, M.: Martingales and stochastic integrals for processes with a multidimensional parameter. Z. Wahrscheinlichkeitstheor. Verw. Geb. **29**, 109–122 (1974)
52. Wong, E., Zakai, M.: An extension of stochastic integrals in the plane. Ann. Probab. **5**, 770–778 (1977)
53. Zhang, J.: Fluid models of multi-server queues with abandonment. Preprint (2009)