TWO-PARAMETER HEAVY-TRAFFIC LIMITS FOR INFINITE-SERVER QUEUES: LONGER VERSION

by

Guodong Pang and Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027-6699
Email: {gp2224, ww2040}@columbia.edu
URL: www.columbia.edu/∼ww2040

October 4, 2009

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 (1991), where the case of discrete service-time distributions was treated, and Krichagina and Puhalskii (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.

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. The Markov property also helps in characterizing the approximate steady-state and transient one-dimensional (marginal) distributions. 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.

Conventional Heavy-Traffic Limits. The classic example is 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 exponential interarrival and service times, the stochastic process $\{Q(t) : t \geq 0\}$ is Markov. However, with non-exponential interarrival and service times, the stochastic process $\{Q(t) : t \geq 0\}$ 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 shows that the queue-length process $\{Q(t) : t \geq 0\}$ is approximately equal to a Markov process under heavy loads. In the conventional heavy-traffic limit, the arrival rate λ is allowed to increase while the number of servers, s, and the service-time distribution with mean $1/\mu$ are held fixed, so that the traffic intensity $\rho \equiv \lambda/s\mu$ approaches the critical value 1 from below. As $\rho \uparrow 1$, we obtain convergence of appropriately scaled queue-length processes to reflected Brownian motion (RBM), which is a Markov process [22, 23, 55]. With $Q_\rho(t)$ denoting the queue length at time t when the traffic intensity is ρ, under regularity conditions controlling the way the arrival processes change with ρ, we have

$$(1-\rho)Q_\rho((1-\rho)^{-2}t) \Rightarrow RBM(t; -\nu, \sigma^2) \quad \text{as} \quad \rho \uparrow 1 \tag{1.1}$$

for positive constants $\nu = s$ and $\sigma^2 = s(c_a^2 + c_s^2)$, where $RBM(t; -\nu, \sigma^2)$ denotes an RBM with drift coefficient $-\nu$ and diffusion coefficient σ^2, $\Rightarrow$ denotes convergence in distribution, and

c_a^2 and c_s^2 are the squared coefficients of variation (SCV's, variance divided by the square of the mean) of an interarrival time and a service time, respectively. The limiting distribution of $RBM(t; -\nu, \sigma^2)$ as $t \to \infty$ is exponential with mean $\sigma^2/2\nu$.

In fact, to obtain the limit in (1.1), we do not need the interarrival times and service times to 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 functional central limit theorem (FCLT) converging to independent Brownian motions (BM's). That allows for weak dependence. Thus, in considerable generality, we have the heavy-traffic approximation

$$\{Q_\rho(t) : t \geq 0\} \approx \{(1-\rho)^{-1} RBM((1-\rho)^2 t; -\nu, \sigma^2) : t \geq 0\}. \tag{1.2}$$

Moreover, the Markov property of the limit extends to conventional heavy-traffic limits for networks of queues. It is significant that the limit process is again a Markov process, in particular, a multidimensional RBM in an orthant, as shown by Harrison and Reiman [19].

Many-Server Heavy-Traffic Limits. Unfortunately, however, 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 Halfin and Whitt [18]. For infinite-server models, we index the stochastic processes by the arrival rate λ.

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 [46, 47], Puhalskii and Reed [45], and Mandelbaum and Momcilovic [**?**].

With infinitely many exponential servers, we again obtain Markov diffusion limits, as first shown by Iglehart [20] for the $M/M/\infty$ model; see Pang et al. [44] for a review. For the $M/M/\infty$ model, with i.i.d. exponential interarrival and service times, the established heavy-traffic limit for $Q(t)$ (now coinciding with the number of busy servers) is

$$\frac{Q_\lambda(t) - (\lambda/\mu)}{\sqrt{\lambda/\mu}} \Rightarrow OU(t; \nu, \sigma^2) \quad \text{as} \quad \lambda \to \infty, \tag{1.3}$$

for appropriate positive constants $\nu = \mu$ and $\sigma^2 = 2\mu$, where $OU(t;\nu,\sigma^2)$ is an Ornstein-Uhlenbeck (OU) diffusion process with drift $-\nu x$ and diffusion coefficient σ^2, which has normal marginal distributions. As a consequence,

$$Q_\lambda(t) \approx \frac{\lambda}{\mu} + \sqrt{\lambda/\mu} OU(t;\mu,2\mu) \quad \text{and} \quad Q_\lambda(\infty) \approx N(\lambda/\mu, \lambda/\mu), \tag{1.4}$$

where $N(m,\sigma^2)$ denotes a normal random variable with mean m and variance σ^2. It is significant that essentially the same limit holds for general arrival processes, provided only that they satisfy a FCLT. With renewal arrival processes, we obtain the same OU limit in (1.3) modified only by having $\sigma^2 = \mu(1+c_a^2)$. A systematic way to extend the limit to general arrival processes is given in §7.3 of [44].

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 [21, 54, 38, 15, 32], 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; see Glynn [14] and Krichagina and Puhalskii [32]). For the non-Markov Gaussian limit, that Gaussian process is fully characterized by specifying its covariance structure. Nevertheless, the Gaussian process is in general not Markov. An inference to be drawn is that $Q(t)$ does not contain the relevant state for describing the evolution of the system, even approximately.

What we want to do, then, is to add more to the system state. We want to consider a stochastic process characterizing the system state for which the associated heavy-traffic limit process is Markov. To do so, we consider the two-parameter stochastic process $\{Q^e(t,y) : t \geq 0, y \geq 0\}$, 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 $\{Q^e(t,y) : t \geq 0, y \geq 0\}$ be used in models and measured in practice. Ways to exploit such ages for control were discussed by Duffield and Whitt [10].

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 $\{Q^e(t,\cdot) : t \geq 0\}$ and $\{Q^r(t,\cdot) : t \geq 0\}$ as function-valued stochastic processes, in particular, random elements of the function space $D_D \equiv D([0,\infty), D([0,\infty),\mathbb{R}))$, 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 §2.3. 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.

The $M/GI/\infty$ and $M_t/GI/\infty$ Models. For understanding, it is very helpful to consider the special case of a Poisson arrival process. The function-valued stochastic process $\{Q^e(t,\cdot) : t \geq 0\}$ is clearly Markov when the arrival process is Poisson. (To know $Q^e(t,\cdot)$ is to know $Q^e(t,y)$ for all $y \geq 0$.) The $M/GI/\infty$ model has a very simple story, even for the generalization to a nonhomogeneous Poisson arrival process (M_t), which is described in [11, 41] (where references to earlier work are given). The key idea, expressed in the proof of Theorem 1 of [11], is a Poisson-random-measure representation: The initial step is to put a point at (t,x) in $[0,\infty)^2$ if there is an arrival at time t with service time x. For the $M_t/GI/\infty$ model, that makes the number of points in subsets of $[0,\infty)^2$ a Poisson random measure with intensity $\lambda(t)f(x)$ at (t,x), where f is the probability density function (p.d.f.) associated with the service-time c.d.f. F. (The c.d.f. F is not actually required to have a p.d.f.) Let C_i, $1 \leq i \leq m$, be m disjoint subsets of $[0,\infty)^2$, and let $N(C_i)$ be the number of arrivals at time t with service times x for $(t,x) \in C_i$. The key fact is that $N(C_i)$, $1 \leq i \leq m$, are independent Poisson random variables, with means equal to the integral of the intensity over C_i. In this context, $Q^r(t,y) = N(C_1)$ and $Q^r(t,0) - Q^r(t,y) = N(C_2)$ (or, $Q^e(t,y) = N(C_1)$ and $Q^e(t,t) - Q^t(t,y) = N(C_2)$) for appropriate disjoint sets C_1 and C_2, as depicted in Figure 1 below, and so are independent Poisson random variables.

When the arrival rate is allowed to grow and appropriate scaling is introduced, the discrete Poisson nature is lost, but the random-measure structure with independence over disjoint subsets is preserved. The limits here produce continuous Brownian analogs of the discrete

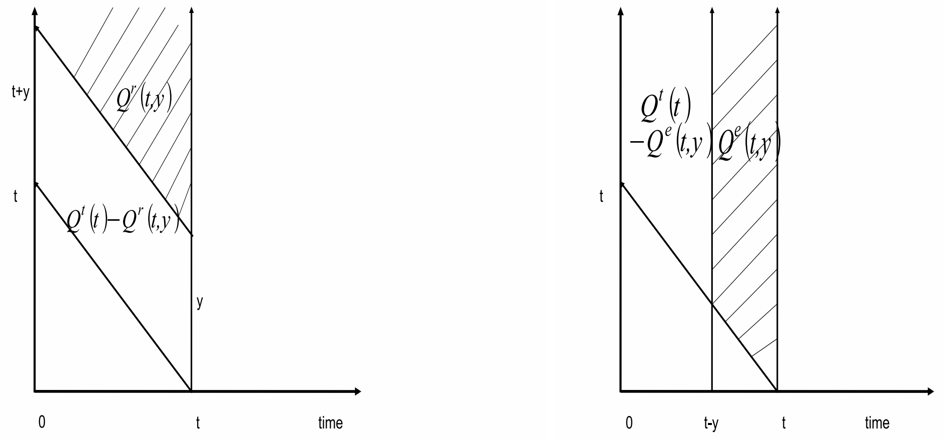

Figure 1: The excess and residual queue-length random variables $Q^e(t,y)$ and $Q^r(t,y)$ counting the number of arrival-time and service-time pairs in designated disjoint subsets of $[0,\infty)^2$.

Poisson results in [11, 41]. The resulting normal approximations are discussed in §9 of [41].

For the $M/GI/\infty$ model, the Markov-process perspective was already applied by Duffield and Whitt [10] to study problems of control and recovery from rare congestion events. The Markov representation there was $(Q^t(t), D(t,y))$, where $Q^t(t) \equiv Q^e(t,t)$ is the total number of customers in the systems at time t and $D(t,y) \equiv Q^e(t,y)/Q^t(t)$ as a function of y $(y \le t)$ is the empirical age distribution of elapsed service times at time t. A rare congestion event was defined as an unusual number $Q^t(t)$ or an unusual age distribution $D(t,y)$.

The $G/GI/\infty$ Model. However, for the more general $GI/GI/\infty$ model, having a non-Poisson renewal arrival process, the stochastic process $\{Q^e(t,\cdot) : t \ge 0\}$ 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; then we would require knowledge of even more of the arrival-process history. 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. (The situation is actually somewhat more complicated, because the constant factor c_a^2 does not appear in all terms in the limit process.) However, there is further simplification when $c_a^2 = 1$; see Corollary 4.1. Thus, a continuous analog of the simple $M/GI/\infty$ story applies in the limit. And it extends to time-varying arrival rates. Thus, we aim to establish heavy-traffic limits for the two-parameter stochastic processes $\{Q^e(t, \cdot) : t \geq 0\}$ and $\{Q^r(t, \cdot) : t \geq 0\}$ in the $G/GI/\infty$ model, including an extension to the non-stationary case G_t.

Proof Strategy. Our proof builds on previous work by Glynn and Whitt [15] and Krichagina and Puhalskii [32]. First, a restricted form of the desired two-parameter stochastic-process limits was already established in Theorem 3 of Glynn and Whitt [15] for the case of service-time distributions with finite support. The result there is only stated for arbitrary fixed second parameter y, but it can be extended. A key idea there was to treat that case by representing the service times as a finite mixture 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 §3 of [15], especially, Proposition 3.1. That step relies on the FCLT for split counting processes, as in §9.5 of [55]. The mixture argument extends quite directly to treat arbitrary discrete distributions. It would then also extend to arbitrary distributions if we can treat continuous service-time c.d.f., because we can regard the general c.d.f. as a mixture of a discrete c.d.f. and a continuous c.d.f. However, the proof in [15] does not seem to extend naturally to continuous service-time c.d.f.'s.

We treat the final case of a continuous service-time c.d.f. here by a different approach. In doing so, we draw heavily on the important paper by Krichagina and Puhalskii [32]. Our limits for continuous service-time c.d.f.'s are extensions of theirs, obtained using the same martingale arguments. The proof in [32] already took a two-parameter approach and 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 is established in the space $D_D \equiv D([0, \infty), D([0, \infty), \mathbb{R}))$ of D-valued functions; see §2.3.

Moreover, Krichagina and Puhalskii [32] already treated general service-time c.d.f.'s, but they do not state limits for two-parameter queueing processes. It might seem that it would be a routine extension to do so, but we show that is *not* so, because the limit process is not a random element of the space D_D for discontinuous service-time c.d.f.'s, as we explain after

Theorem 3.2. Fortunately, however, the argument in [32] can be extended to the two-parameter case if we restrict attention to continuous service-time c.d.f.'s, which is just what we need. In fact, we started with [32] and thought of applying [15] only after discovering this difficulty.

The main work here is our treatment of the case of a continuous service-time c.d.f.. The main result is our FCLT for that case in Theorem 3.2. Sections §§8 - 11 are devoted to its proof. Our main difficulty is extending tightness proofs in [32] from the space D to the space D_D. The rest of the results are relatively routine, so we provide relatively few proof details for them.

Practical Value. In doing this work, our goal has been to obtain useful practical formulas for the approximate distributions of the random variables describing system performance. Two observations set the stage: First, in the heavy-traffic limit, as the arrival rate increases, the sequence of properly scaled arrival processes usually converges to Brownian motion (or a time-transformed version), which has independent increments, so that will be a key initial assumption here. As a consequence, with high arrival rate, the arrival process should have approximately independent increments. Since new arrivals do not interact with customers in the system, because each customer can have his own server, this independent-increments property implies that, for any time t, the system evolution for new arrivals after time t should be approximately independent of the system history up to time t.

Second, to know the pre-limit process $Q^e(t, y)$ for all y up to time t is essentially (aside from customer identity) equivalent to knowing the total number in system at time t plus the arrival times of the customers in the system at that time. Combining these two observations directly shows that the pre-limit stochastic process $\{Q^e(t, \cdot) : t \geq 0\}$ should be approximately equal to a Markov process. Hence, we do not greatly dwell on that issue.

It also shows what is needed: First, to describe the approximate consequence of new arrivals after time t, it suffices to describe the system evolution starting empty. Second, we want to describe the system state at time t, which we can also think of as starting empty in the past. Hence, it suffices to describe the approximate distributions of $Q^e(t, y)$ and $Q^r(t, y)$ for all $t \geq 0$ and $y \geq 0$.

In this paper, we establish heavy-traffic limits that enable us to do just that. Since the limiting random variables are all Gaussian, the approximate distributions of unscaled pre-limit random variables are determined by their means and variances. The approximate mean values are determined by the fluid limits, while the approximate variances are determined by the variances of the limiting random variables. Thus, we meet our objective by providing explicit formulas for both the fluid limits and the variances of the limiting random variables. The fluid limits are given in Theorems 3.1 and 7.1; the variance formulas are given in Theorems 4.2 and 7.3. It is important that our main results - Theorems 3.2 and 7.2 - do indeed yield these important practical consequences.

Other Related Literature. As noted by Krichagina and Puhalskii [32], the relevance of the two-parameter Kiefer process for the infinite-server queue was first observed by Louchard [38]. The results here were briefly outlined in §6.4 in our survey [44]. (The first drafts of this paper were written at that time.) Since then, there has been a flurry of new work: Related fluid limits for measure-valued processes have been obtained by Kaspi and Ramanan [29] for the $GI/GI/s$ model with $s \to \infty$, by Kang and Ramanan [28] and Zhang [61] for the $GI/GI/s$ model with abandonment. However, the first fluid limit for two-parameter processes evidently was the fluid limit in §6 of [56] for the discrete-time version of that more general $G_t(n)/GI/s + GI$ model, having both time-dependent and state-dependent arrivals. Two-parameter processes are also used in [51] to establish limits for waiting times. There is also a substantial body of related limits for two-parameter processes associated with queueing models with non-FCFS service disciplines; see [9, 17, 33, 34] and references therein.

For the $G/GI/\infty$ model we consider and generalizations, there also have been FCLT results: A FCLT for the $G/GI/s$ model was announced in [29], but has not appeared by 9/29/09. Decreusefond and Moyal [8] established a FCLT for the $M/GI/\infty$ model. The limit in [8] is expressed in terms of cylindrical Brownian motions on a Hilbert space $L^2(dF)$ with F being the service-time distribution. The limit is expressed as an infinite-dimensional stochastic integral with respect to independent standard BM's. Following Decreusefond and Moyal [8], in contemporaneous work, Reed and Talreja [48] used a very different approach, based on distribution-valued processes, which is a more general framework than measure-valued processes. Moreover, they showed that there is a close connection to early work by Kallianpur and Perez-Abreu [26, 27], Mitoma [42] and others. With this approach, they are able to apply traditional martingale methods with the continuous mapping theorem to establish their limit. Moreover, they are able to characterize the limit as a generalized Ornstein-Uhlenbeck (OU) process. This OU structure is appealing, because it extends the seminal result by Iglehart [20] for the $M/M/\infty$ model. Indeed, this generalization could be anticipated, because in [54] we had already obtained an m-dimensional OU limit for the $GI/PH_m/\infty$ model. A problem with the distribution-valued framework in [48] is that there are far fewer continuous functions on this space, so that the limits have fewer applications with the continuous mapping theorem.

Organization of this paper. We start with preliminaries in §2. In §3 we state our main results, focusing only on new arrivals (ignoring any customers initially in the system) and a continuous service-time c.d.f. In §4 we characterize the limit processes. In §5 we treat the initial conditions, and treat all customers in the system. In §6 we show how the limit reduces to the previous results for the $M/M/\infty$ and $G/M/\infty$ models. In §7 we treat the general service-time c.d.f.'s. In §§8-11 we prove the main theorem: Theorem 3.2. In §9 we prove the continuity of the representation of some key processes in the space D_D. In §10 we continue the proof by establishing tightness of the key processes. In §11 we complete the proof by establishing convergence of the finite-dimensional distributions. We draw conclusions in §12. We present supporting technical details in the Appendix, 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 §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 τ_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, 55]. 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

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

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

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

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 λ, 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 [11, 41, 40].

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 λ^{-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 Chapter 4 of [55]. 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 §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 $\{\eta_i : i \geq 1\}$

with a *continuous* c.d.f. F, independent of n and the arrival processes. (We extend to general c.d.f.'s in §7.) ∎

As in [32], 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 except that it need be continuous. It need not have a finite mean. However, for subsequent results we will need to assume in addition that F has a finite mean μ^{-1} and even a finite second moment with SCV c_s^2. The continuity of F implies that our limit processes will have continuous paths, as will be seen in Theorems 3.2 and 5.1.

Krichagina and Puhalskii [32] observed that it is fruitful to view the service times through the two-parameter *sequential empirical process*

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

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

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

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

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 η_i is uniformly distributed on $[0,1]$, so that $F(x) = x$, $0 \le x \le 1$. Extending previous results by Bickel and Wichura [1], Krichagina and Puhalskii [32] showed that

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

where $U(t,x)$ is the *standard Kiefer process*; see Csörgö and Révész [7], Gaenssler and Stute [13], van der Vaart and Wellner [52] and Appendix A. 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

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

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

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

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

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

From (2.9), we see the connection to the sequential empirical process $\bar{K}_n$ in (2.3). Indeed, the key observation (following [32]) is that we can rewrite the random sums in (2.9) as integrals with respect to the random field $\bar{K}_n$ by

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

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). However, (again following [32]) we will interpret their limits as stochastic integrals, defined in the sense of mean-square convergence. That mean-square-convergence definition is consistent with the two-dimensional Stieltjes integral used for the prelimit processes. We will exploit the mean-square characterization to prove that the finite-dimensional distributions of the scaled processes $\hat{X}_{n,2}$ in (2.13) below converge to those of the limiting process in §11. We will primarily focus on Q_n^r.

Lemma 2.1. (representation of Q_n^r) *The process* Q_n^r *defined in* (2.9) *and* (2.10) *can be represented as*

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

where

$$\hat{X}_{n,1}(t,y) \equiv \int_0^t F^c(t+y-s)\,d\hat{A}_n(s), \tag{2.12}$$

$$\begin{aligned}\hat{X}_{n,2}(t,y) &\equiv \int_0^t\int_0^\infty \mathbf{1}(s+x>t+y)\,d\hat{R}_n(s,x) \\ &= -\int_0^t\int_0^\infty \mathbf{1}(s+x\le t+y)\,d\hat{R}_n(s,x),\end{aligned} \tag{2.13}$$

$$\begin{aligned}\hat{R}_n(t,y) &\equiv \hat{K}_n(\bar{A}_n(t),y) = \frac{1}{\sqrt{n}}\sum_{i=1}^{A_n(t)}(\mathbf{1}(\eta_i\le y) - F(y)) \\ &= \sqrt{n}\bar{K}_n(\bar{A}_n(t),y) - \hat{A}_n(t)F(y) - \sqrt{n}\bar{a}(t)F(y),\end{aligned} \tag{2.14}$$

with the first two integrals in (2.12) *and* (2.13) *both defined as Stieltjes integrals for functions of bounded variation as integrators.*

Proof. Apply (2.4) to get the first relation in (2.14). (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 (2.14) and

$$\begin{aligned}\bar{K}_n(\bar{A}_n(t),x) &= \frac{1}{n}\sum_{i=1}^{A_n(t)}\mathbf{1}(\eta_i\le x) \\ &= \frac{1}{\sqrt{n}}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{A_n(t)}(\mathbf{1}(\eta_i\le x) - F(x))\Big] + \frac{1}{\sqrt{n}}\sqrt{n}(\bar{A}_n(t) - \bar{a}(t))F(x) + \bar{a}(t)F(x) \\ &= \frac{1}{\sqrt{n}}\hat{R}_n(t,x) + \frac{1}{\sqrt{n}}\hat{A}_n(t)F(x) + \bar{a}(t)F(x).\end{aligned} \tag{2.15}$$

Combine (2.10) and (2.15) 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.,

$$F_n^e(t,y) \equiv Q_n^e(t,y)/Q_n^t(t), \quad t \ge 0, \quad 0 \le y \le t, \tag{2.16}$$

and

$$F_n^{r,c}(t,y) \equiv 1 - F_n^r(t,y) \equiv Q_n^r(t,y)/Q_n^t(t), \quad t \ge 0, \quad y \ge 0. \tag{2.17}$$

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

$$I_n(t) \equiv \sum_{i=1}^{A_n(t)} \eta_i, \quad t \ge 0. \tag{2.18}$$

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

$$W_n^r(t,y) \equiv \int_y^\infty Q_n^r(t,x)\,dx, \quad t \ge 0, \quad y \ge 0. \tag{2.19}$$

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, 55] 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 [49] 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 summarize the tightness criteria in the space D_D in Appendix C, which will be applied in §10 to prove tightness of these processes. We refer to Talreja and Whitt [?] 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 [50] and Neuhaus [43], 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 $Q^r_n(t,y)$. 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^o 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.

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^r_n and W^r_n, along with the closely related processes. We give the proofs in §§8-11.

Define the LLN-scaled processes $\bar{Q}^r_n \equiv \{\bar{Q}^r_n(t,y), t\ge 0, y\ge 0\}$ by

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

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

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

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

By Lemma 2.1,

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

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 μ^{-1}), defined by

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

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,*

$$\big(\bar{A}_n, \bar{K}_n, \bar{Q}^r_n, \bar{Q}^t_n, \bar{Q}^e_n, \bar{F}^e_n, \bar{F}^{r,c}_n, \bar{D}_n\big) \Rightarrow \big(\bar{a}, \bar{k}, \bar{q}^r, \bar{q}^t, \bar{q}^e, \bar{f}^e, \bar{f}^{r,c}, \bar{d}\big) \tag{3.5}$$

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

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

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

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

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

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

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

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

for F_e in (3.4)*.*

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

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

as $t \to \infty$.

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

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

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),

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

Moreover,

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

and

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

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.

Theorem 3.2. (FCLT)

(*a*) *Under Assumptions 1 and 2,*

$$(\hat{A}_n, \hat{K}_n, \hat{Q}^r_n, \hat{Q}^t_n, \hat{Q}^e_n, \hat{F}^{r,c}_n, \hat{F}^e_n, \hat{D}_n) \Rightarrow (\hat{A}, \hat{K}, \hat{Q}^r, \hat{Q}^t, \hat{Q}^e, \hat{F}^{r,c}, \hat{F}^e, \hat{D}) \tag{3.13}$$

in $D \times D^2_D \times D \times D^3_D \times D$ *as* $n \to \infty$, *where* $\hat{A}$ *is the limit in* (2.1), $\hat{K}(t,x)$ *is the limit in* (2.6), *which is independent of* $\hat{A}$,

$$\begin{aligned} \hat{Q}^r(t,y) &\equiv \hat{X}_1(t,y) + \hat{X}_2(t,y), \quad t \geq 0, \quad y \geq 0, \\ \hat{X}_1(t,y) &\equiv \int_0^t F^c(t+y-s)d\hat{A}(s), \\ &\qquad = F^c(y)\hat{A}(t) - \int_0^t \hat{A}(s)dF(t+y-s), \\ \hat{X}_2(t,y) &\equiv \int_0^t \int_0^\infty \mathbf{1}(s+x > t+y)\, d\hat{K}(\bar{a}(s),x), \\ &\qquad = -\int_0^t \int_0^\infty \mathbf{1}(s+x \leq t+y)\, d\hat{K}(\bar{a}(s),x), \end{aligned} \tag{3.14}$$

$\hat{Q}^t(t) \equiv \hat{Q}^r(t,0)$, $\hat{Q}^e(t,y) \equiv \hat{Q}^t(t) - \hat{Q}^r(t-y,y)$, $\hat{F}^{r,c}(t,y) \equiv \bar{q}^t(t)^{-1}(\hat{Q}^r(t,y) - \hat{Q}^t(t)f^{r,c}(t,y))$, $\hat{F}^e(t,y) \equiv \bar{q}^t(t)^{-1}(\hat{Q}^e(t,y) - \hat{Q}^t(t)f^e(t,y))$, and $\hat{D} = \hat{A} - \hat{Q}^t$. Moreover, all these limit processes are continuous and $\hat{Q}^e(t,y)$ can be represented as

$$\begin{aligned}
\hat{Q}^e(t,y) &\equiv \hat{X}^e_1(t,y) + \hat{X}^e_2(t,y), \quad t \geq 0, \quad 0 \leq y \leq t, \qquad (3.15)\\
\hat{X}^e_1(t,y) &\equiv \int_{t-y}^t F^c(t-s)d\hat{A}(s),\\
&= \hat{A}(t) - F^c(y)\hat{A}(t-y) - \int_{t-y}^t \hat{A}(s)dF(t-s),\\
\hat{X}^e_2(t,y) &\equiv \int_{t-y}^t \int_0^t \mathbf{1}(s+x>t)\,d\hat{K}(\bar{a}(s),x),\\
&= -\int_{t-y}^t \int_0^t \mathbf{1}(s+x \leq t)\,d\hat{K}(\bar{a}(s),x).
\end{aligned}$$

(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 μ^{-1}, then $(\hat{W}^r_n, \hat{W}^t_n) \Rightarrow (\hat{W}^r, \hat{W}^t)$ in $D_D \times D$ as $n \to \infty$ jointly with the limits in (3.13)*, where*

$$\hat{W}^r(t,y) \equiv \int_y^\infty \hat{Q}^r(t,x)\,dx, \quad \text{and} \quad \hat{W}^t(t) \equiv \hat{W}^r(t,0) = \int_0^\infty \hat{Q}^r(t,x)\,dx. \qquad (3.16)$$

(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

$$\hat{I}(t) \equiv \sqrt{\lambda c_s^2} B_s(t) + \mu^{-1}\hat{A} \quad \text{and} \quad \hat{C}(t) \equiv \hat{I}(t) - \hat{W}^t(t), \quad t \geq 0, \qquad (3.17)$$

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

The two integrals in (3.14) are stochastic integrals. The first integral for $\hat{X}_1$ is a standard Ito integral if $\hat{A}$ is a (time-changed) Brownian motion; otherwise, the expression for $\hat{X}_1$ is interpreted as the form after integration by parts. The relevant version of integration by parts for $\hat{X}_{n,1}$ and $\hat{X}_1$ is given in Bremaud [4], p.336. For $\hat{X}_{n,1}$, it yields

$$\hat{X}_{n,1}(t,y) = F^c(y)\hat{A}_n(t) - \int_0^t \hat{A}_n(s-)\,dF(t+y-s), \qquad (3.18)$$

and similarly for $\hat{X}_1$. The left limit $\hat{A}_n(s-)$ in (3.18) 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 as in the proof of Theorem 4.2, or in the mean-square sense in §11 following [32].

Note that the two limit processes $\hat{X}_1$ and $\hat{X}_2$ are independent since $\hat{A}$ and $\hat{K}$ are independent. The asymptotic variability of the arrival process is captured by $\hat{A}$, which appears only in $\hat{X}_1$, while the asymptotic variability of the service process is captured by $\hat{K}$, which appears only in $\hat{X}_2$. Thus, in some sense, there is additivity of stochastic effects, as pointed out in [38, 32], but this might be misinterpreted. Notice that *both* $\hat{X}_1$ and $\hat{X}_2$ 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 $\hat{X}_1$, so that there is a genuine asymptotic insensitivity to the arrival process beyond its rate in $\hat{X}_2$.

We remark that if the service-time c.d.f. F is discontinuous, the limit process $\hat{X}_2$ 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 [32]. 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 Chapter 15 of [55].

We now establish additional results in the standard case. Let $\overset{\mathrm{d}}{=}$ mean equal in distribution. 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 μ^{-1}. Under Assumptions 1 and 2, the limits in* (3.5) *hold with*

$$\begin{aligned}
\bar{q}^r(t, y) &\equiv \lambda \int_0^t F^c(t + y - s)\, ds = \lambda \int_0^t F^c(y + s)\, ds \\
&\quad \to (\lambda/\mu)\, F_e^c(y) \quad \textit{as} \quad t \to \infty, \\
\bar{q}^e(t, y) &\equiv \lambda \int_{t-y}^t F^c(t - s)\, ds = \lambda \int_0^y F^c(s)\, ds = (\lambda/\mu)\, F_e(y), \quad \textit{for} \quad t \geq 0, \\
\bar{f}^e(t, y) &\equiv \bar{q}^e(t, y)/\bar{q}^t(t) \to F_e(y) \quad \textit{as} \quad t \to \infty, \\
\bar{f}^{r,c}(t, y) &\equiv \bar{q}^r(t, y)/\bar{q}^t(t) \to F_e^c(y) \quad \textit{as} \quad t \to \infty. \qquad (3.19)
\end{aligned}$$

4. Characterizing the Limit Processes

We now want to show that the basic 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. (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,

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

for $0 \le t_1 < t_2$ and $0 \le x_1 < x_2$, 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 Figure 2.

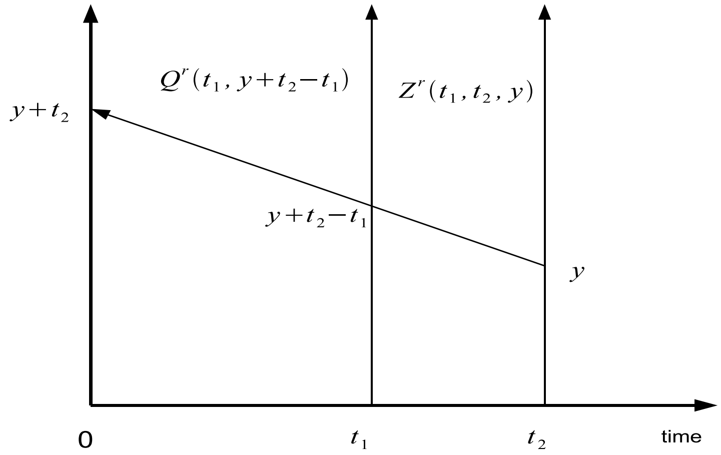

Figure 2: The basic decomposition for $Q^r(t,y)$.

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

positions

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

where $y \geq 0$, $Z^r \equiv Z_1 + Z_2$, and

$$\begin{aligned}
Z_1(t_1,t_2,y) &\equiv \int_{t_1}^{t_2} F^c(t+y-s)d\hat{A}(s),\\
Z_2(t_1,t_2,y) &\equiv \int_{t_1}^{t_2}\int_0^{\infty} \mathbf{1}(s+x>t+y)\, d\hat{K}(\bar{a}(s),x).
\end{aligned}$$

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 three processes $\{\hat{X}_1(t,\cdot) : t \geq 0\}$, $\{\hat{X}_2(t,\cdot) : t \geq 0\}$ and $\{\hat{Q}^r(t,\cdot) : t \geq 0\}$ all have independent increments, and are thus Markov processes* (*with respect to the argument t*)*.*

Proof. The decomposition for $\hat{X}_1(t,y)$, $\hat{X}_2(t,y)$ and $\hat{Q}^r(t,y)$ in (4.2) is by direct construction, as in Figure 2. The independent-increments property is inherited from $\hat{K}$ and $\hat{A}$. ■

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 $\{A_n(t) : t \geq 0\}$ be defined by

$$A_n(t) \equiv \tilde{A}(n\bar{a}(t)), \quad t \geq 0,$$

where $\tilde{A} \equiv \{\tilde{A}(t) : t \geq 0\}$ 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, 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 continuous Gaussian processes. If $\hat{A}(t) = \sqrt{c_a^2}B_a(\bar{a}(t))$ for $t \geq 0$, where $\bar{a}(t) = \int_0^t \lambda(s)ds$ and B_a is* *a standard BM, then for each fixed $t \geq 0$ and $y \geq 0$,*

$$\begin{aligned}
&\hat{X}_1(t,y) \stackrel{\mathrm{d}}{=} N(0, c_a^2\sigma_1^2(t,y)), \quad \hat{X}_2(t,y) \stackrel{\mathrm{d}}{=} N(0, \sigma_2^2(t,y)), \\
&\hat{X}_1^e(t,y) \stackrel{\mathrm{d}}{=} N(0, c_a^2\sigma_{1,e}^2(t,y)), \quad \hat{X}_2^e(t,y) \stackrel{\mathrm{d}}{=} N(0, \sigma_{2,e}^2(t,y)), \\
&\hat{W}^r(t,y) \stackrel{\mathrm{d}}{=} N(0, \sigma_w^2(t,y)),
\end{aligned}$$

where

$$\begin{aligned}
\sigma_1^2(t,y) &\equiv \int_0^t F^c(t+y-s)^2\,\lambda(s)\,ds, \quad t \geq 0, \quad y \geq 0, \\
\sigma_2^2(t,y) &\equiv \int_0^t F(t+y-s)F^c(t+y-s)\,\lambda(s)\,ds, \quad t \geq 0, \quad y \geq 0, \\
\sigma_{1,e}^2(t,y) &\equiv \int_{t-y}^t F^c(t-s)^2\,\lambda(s)\,ds, \quad t \geq 0, \quad 0 \leq y \leq t, \\
\sigma_{2,e}^2(t,y) &\equiv \int_{t-y}^t F(t-s)F^c(t-s)\,\lambda(s)\,ds, \quad t \geq 0, \quad 0 \leq y \leq t, \\
\sigma_w^2(t,y) &\equiv c_a^2 \int_y^\infty \int_y^\infty \int_0^t F^c(t+x-s)F^c(t+z-s)\lambda(s)dsdxdz \\
&\quad + \int_y^\infty \int_y^\infty \int_0^t F(t+x\wedge z-s)F^c(t+x\vee z-s)\lambda(s)dsdxdz, \quad t \geq 0, \quad y \geq 0.
\end{aligned}$$

Moreover, $\hat{X}_1$ and $\hat{X}_2$ are independent, and have covariances

$$\begin{aligned}
Cov(\hat{X}_1(t,y), \hat{X}_1(t',y')) &= c_a^2 \int_0^t F^c(t+y-s)F^c(t'+y'-s)\,\lambda(s)\,ds, \\
Cov(\hat{X}_2(t,y), \hat{X}_2(t',y')) &= \int_0^t F(t+y-s)F^c(t'+y'-s)\,\lambda(s)\,ds,
\end{aligned}$$

for $0 \leq t \leq t'$, $0 \leq y \leq y'$. $\hat{X}_1^e$ and $\hat{X}_2^e$ are also independent, and have covariances

$$\begin{aligned}
Cov(\hat{X}_1^e(t,y), \hat{X}_1^e(t',y')) &= c_a^2 \int_{(t-y)\vee(t'-y')}^t F^c(t-s)F^c(t'-s)\,\lambda(s)\,ds, \\
Cov(\hat{X}_2^e(t,y), \hat{X}_2^e(t',y')) &= \int_{(t-y)\vee(t'-y')}^t F(t-s)F^c(t'-s)\,\lambda(s)\,ds,
\end{aligned}$$

for $0 \leq t \leq t'$, $0 \leq y \leq y'$, $y \leq t$ and $y' \leq t'$.

Proof. We only discuss $\hat{X}_2$ in detail and the variance formulas for the other processes can be obtained similarly. We apply basic properties of the two-parameter Brownian sheet, as reviewed in the Appendix A. In particular, we use the standard representation of the Kiefer process U in terms of the two-parameter Brownian sheet W, i.e.,

$$U(x,y) = W(x,y) - yW(x,1), \quad x \geq 0, \quad 0 \leq y \leq 1. \tag{4.3}$$

We will use the stochastic integral of the first type with respect to two-parameter Brownian sheet. Having $\hat{X}_2$ 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

$$\begin{aligned}
E[\hat{X}_2(t,y)^2] &= E\Big[\Big(\int_0^t\int_0^\infty \mathbf{1}(s+x>t+y)dU(\bar{a}(s),F(x))\Big)^2\Big]\\
&= E\Big[\Big(\int_0^t\int_0^\infty \mathbf{1}(s+x>t+y)d(W(\bar{a}(s),F(x))-F(x)W(\bar{a}(s),1))\Big)^2\Big]\\
&= \int_0^t\int_0^\infty \mathbf{1}(s+x>t+y)dF(x)d\bar{a}(s)+\int_0^t F^c(t+y-s)^2d\bar{a}(s)\\
&\quad -2\int_0^t\int_0^\infty \mathbf{1}(s+x>t+y)F(t+y-s)dF(x)d\bar{a}(s)\\
&= \int_0^t F(t+y-s)F^c(t+y-s)d\bar{a}(s),
\end{aligned}$$

where the second equality uses the identity (4.3) 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. The covariance can be obtained in the similar way.

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

$$E[\hat{W}^r(t,y)^2]=E\Big[\Big(\int_y^\infty \hat{Q}^r(t,x)dx\Big)^2\Big]=E\Big[\Big(\int_y^\infty \hat{X}_1(t,y)dx\Big)^2\Big]+E\Big[\Big(\int_y^\infty \hat{X}_2(t,y)dx\Big)^2\Big].$$

Then by an analogous argument to $E[\hat{X}_2(t,y)^2]$, 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 [32] by understanding the integral in $\hat{X}_2$ as a mean square limit (§11). However, our approach here by applying properties of stochastic integrals with respect to two-parameter Brownian sheets of the first type simplifies the proof. Paralleling the result in Lemma 5.1 [32], we can easily check that

$$\begin{aligned}
&E[(\hat{X}_2(t,y)-\hat{X}_2(t',y'))^2]\\
=&\int_0^t (F(t'+y'-u)-F(t+y-u))(1+F(t+y-u)-F(t'+y'-u))d\bar{a}(u)
\end{aligned}$$

for $0\le t\le t'$, $0\le y\le y'$.

Corollary 4.1. (*the special case* $c_a^2=1$) *If, in addition to the assumptions of Theorem* 4.2, $c_a^2=1$, *then*

$$Var(\hat{Q}^r(t,y))=\sigma_1^2(t,y)+\sigma_2^2(t,y)=\int_0^t F^c(t+y-u)\,\lambda(s)\,ds, \qquad (4.4)$$

for $t \geq 0$ and $y \geq 0$. The limit $\hat{A}$ and all the other limits are the same as if the unscaled arrival processes $\{A_n(t) : t \geq 0\}$ 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 $t \geq 0$ and $y \geq 0$, $\hat{Q}^r(t,y)$ is distributed the same as the limit of

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

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

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

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 [16]. 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 $t \geq 0$ and $y \geq 0$. (In passing, we remark that $\mathcal{Q}_n^r(t,y) \overset{\mathrm{d}}{=} Q_n^r(t,y)$ 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 $n \to \infty$:

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

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

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

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, [55]) and Theorems 3.1 and 3.2, we obtain the weak convergence of $\hat{\mathcal{Q}}_n^r(t,y)$:

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

where

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

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

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

Since $\hat{\mathcal{Q}}^r(t,y)$ and $\hat{Q}^r(t,y)$ are both Gaussian with the same mean and variance, $\hat{\mathcal{Q}}^r(t,y)$ and $\hat{Q}^r(t,y)$ are equal in distribution. When the arrival process is M_t, $Q_n^r(t,y)$ 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 [16]. 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*

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

and

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

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 additon, $c_a^2 = 1$, then

$$\begin{aligned}
Var(\hat{Q}^r(t,y)) &= \lambda\int_0^t F^c(y+s)ds \to \lambda\int_y^\infty F^c(s)ds = \frac{\lambda}{\mu}F_e^c(y), \quad as \quad t\to\infty, \quad y\geq 0, \\
Var(\hat{Q}^e(t,y)) &= \lambda\int_0^y F^c(s)ds = \frac{\lambda}{\mu}F_e(y), \quad t,y\geq 0,
\end{aligned}$$

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

5. Initial Conditions

So far, we have concentrated on new arrivals. Now we turn our attention to customers in the system initially, before any new arrivals come. Like the generality of the service-time c.d.f., the initial conditions present technical difficulties. Our assumptions will be similar to those made for the initial conditions in [32] and similar to those for the new arrivals in §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. In other words, we can think of the system starting in the distant past with just new arrivals, and we will be able to approximate the two-parameter processes by the Markov limit processes. However, there is some justification for what we do below.

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 (2.19).

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

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

where $\bar{q}^{i,r}$ is the FWLLN limit of $\bar{Q}_n^{i,r}$ to be proved. 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 $(\hat{A}_n, \hat{Q}^{i,r}_n) \Rightarrow (\hat{A}, \hat{Q}^{i,r})$ 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 $\{\eta^i_j : j \geq 1\}$ of i.i.d. nonnegative random variables with a *continuous* 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, $Q^{i,t}_n$, 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

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

Paralleling Lemma 2.1, we have the representation result.

Lemma 5.1. (representation of $Q^{i,r}_n$) *The process $Q^{i,r}_n$ can be represented as*

$$Q^{i,r}_n(y) = \sum_{j=1}^{Q^{i,t}_n} \left(\mathbf{1}(\eta^i_j > y) - F^c_i(y)\right) + Q^{i,t}_n F^c_i(y), \quad y \geq 0. \tag{5.3}$$

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

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

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 $Q^{T,r}_n(t, y)$. The key representation is

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

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

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

in D_D as $n \to \infty$, where $\hat{X}_1$ and $\hat{X}_2$ are given in (3.14).

6. Exponential Service Times

We now discuss how the results simplify when the service-time c.d.f.'s are exponential.

6.1. The Total Queue Length

From previous work [20, 3, 54], we know that the one-parameter limit process $\hat{Q}^{T,t}$ is an OU process when the arrival-process limit $\hat{A}$ is BM and the service-time distributions F and F_i are a common exponential distribution. We show how that conclusion for the queue-length process limit $\hat{Q}^{T,t}$ can be deduced from our results.

Assumptions for the Standard Case with Exponential Service. Assume that $F(x) = F_i(x) = 1 - e^{-\mu x}$ for $x \geq 0$, $\bar{a}(t) = \lambda t$ for $t \geq 0$, $\hat{A} = \sqrt{\lambda c_a^2}B$ where B is standard BM. ■

For this exponential case, the FWLLN in Corollary 5.1 gives the limit

$$\begin{aligned}
\bar{q}^{T,r}(t,y) &= F^c_i(t+y)\bar{q}^{i,t} + \int_0^t F^c(t+y-s)\, d\bar{a}(s), \\
&= e^{-\mu(t+y)}\bar{q}^{i,t} + \lambda \int_0^t e^{-\mu(t+y-s)}\, ds \\
&= e^{-\mu y}\left(\frac{\lambda}{\mu} + \left(\bar{q}^{i,t} - \frac{\lambda}{\mu}\right)e^{-\mu t}\right) \to \frac{\lambda}{\mu}e^{-\mu y} \quad \text{as} \quad t \to \infty. \qquad (6.1)
\end{aligned}$$

The limiting behavior as $t \to \infty$ is consistent with Corollary 3.1. For example, we see that the mean steady-state number λ/μ is approached by $\bar{q}^{T,t}(t) \equiv \bar{q}^{T,r}(t,0)$ as $t \to \infty$. This consistency is expected because, under Assumptions 3 and 4, the initial conditions will have no impact on the limiting behavior as $t \to \infty$.

We now turn to the limit in the FCLT. We consider each of the four terms appearing in the limit of the FCLT for all customers in Corollary 5.1. In addition to $\hat{X}_1$ and $\hat{X}_2$, which describe the limit for new arrivals, let the limits for the initial customers be

$$\begin{aligned}\hat{X}_3(t,y) &\equiv F_i^c(t+y)\hat{Q}^{i,t} = e^{-\mu(t+y)}\hat{Q}^{i,t} = -\mu\int_0^t \hat{X}_3(s,y)ds + e^{-\mu y}\hat{Q}^{i,t} \\ \hat{X}_4(t,y) &\equiv \sqrt{\bar{q}^{i,t}}B^0(F_i(t+y)) = \sqrt{\bar{q}^{i,t}}B^0(1-e^{-\mu(t+y)}). \end{aligned} \tag{6.2}$$

By the fact that Brownian bridge B^0 is the unique solution to a one-dimensional SDE (§5.6.B, [30]), we can rewrite $\hat{X}_4$ as

$$\begin{aligned}\hat{X}_4(t,y) &= \sqrt{\bar{q}^{i,t}}\left(-\int_0^{1-e^{-\mu(t+y)}} \frac{B^0(u)}{1-u}du + B_b\left(1-e^{-\mu(t+y)}\right)\right) \\ &= \sqrt{\bar{q}^{i,t}}\left(-\mu\int_0^t B^0(1-e^{-\mu(s+y)})ds + B_b(1-e^{-\mu(t+y)})\right) \\ &= -\mu\int_0^t \hat{X}_4(s,y)ds + \sqrt{\bar{q}^{i,t}}B_b(1-e^{-\mu(t+y)}), \end{aligned} \tag{6.3}$$

where B_b is a standard Brownian motion, independent of B in the assumptions for the standard case with exponential service.

By Ito's formula, we can write $\hat{X}_1$ as

$$\hat{X}_1(t,y) = -\mu\int_0^t \hat{X}_1(s,y)ds + e^{-\mu y}\sqrt{\lambda c_a^2}B(t), \tag{6.4}$$

and by Proposition A.1 for the Kiefer process, we obtain

$$\begin{aligned}U(\lambda t, 1-e^{-\mu x}) &= -\int_0^{1-e^{-\mu x}} \frac{U(\lambda t, y)}{1-y}dy + W(\lambda t, 1-e^{-\mu x}) \\ &= -\mu\int_0^x U(\lambda t, 1-e^{-\mu y})dy + W(\lambda t, 1-e^{-\mu x}),\end{aligned}$$

which implies that

$$\hat{X}_2(t,y) = -\mu\int_0^t \hat{X}_2(s,y)ds + \int_0^t\int_0^\infty \mathbf{1}(s+x \le t+y)dW(\lambda s, 1-e^{-\mu x}). \tag{6.5}$$

Combining the above (6.2), (6.3), (6.4), (6.5) and the FCLT in Corollary 5.1, we can write the limit $\hat{Q}^{T,r}$ as

$$\begin{aligned}\hat{Q}^{T,r}(t,y) &= \hat{X}_1(t,y) + \hat{X}_2(t,y) + \hat{X}_3(t,y) + \hat{X}_4(t,y) \\ &= e^{-\mu y}\hat{Q}^{i,t} - \mu\int_0^t \hat{Q}^{T,r}(s,y)ds + \hat{B}(t,y), \end{aligned} \tag{6.6}$$

where

$$\begin{aligned}\hat{B}(t,y) &= \sqrt{\bar{q}^{i,t}}B_b(1-e^{-\mu(t+y)}) + e^{-\mu y}\sqrt{\lambda c_a^2}B(t) \\ &\quad + \int_0^t\int_0^\infty \mathbf{1}(s+x\le t+y)dW(\lambda s, 1-e^{-\mu x}),\end{aligned}$$

with B, B_b and W being mutually independent. The process $\{\hat{B}(t,y) : t, y \geq 0\}$ is a well defined Gaussian process with mean 0 and covariance for $t < t', y < y'$, (see Appendix D for the calculation)

$$E[\hat{B}(t,y)\hat{B}(t',y')] = (\lambda + \lambda c_a^2 e^{-\mu(y+y')})t + \bar{q}^{i,t} - \frac{\lambda}{\mu}e^{-\mu y} + \left(\frac{\lambda}{\mu} - \bar{q}^{i,t}\right) e^{-\mu(t+y)}.$$

In particular, since the total number of customers in the system $\hat{Q}^{T,t}(t) \equiv \hat{Q}^{T,r}(t,0)$, when $\bar{q}^{i,t} = 1$ and $\lambda = \mu$, we obtain

$$\hat{Q}^{T,t}(t) = \hat{Q}^{i,t} - \mu \int_0^t \hat{Q}^{T,t}(s)ds + \tilde{B}(t), \quad t \geq 0, \tag{6.7}$$

where $\tilde{B}(t) = \hat{B}(t,0)$. It is easy to check that the process $\tilde{B}$ is a Brownian motion with mean 0 and covariance $E[\tilde{B}(t)\tilde{B}(t')] = \lambda(1+c_a^2)(t \wedge t')$. Thus, $\hat{Q}^{T,t}(t)$ in (6.7) is an $OU(t;\mu,\lambda(1+c_a^2))$ process, consistent with the heavy-traffic limit for the queue-length processes of the infinite server queues with exponential service times (Theorem 1, [54]). Moreover, if $c_a^2 = 1$, $\hat{Q}^{T,t}(t)$ in (6.7) is an $OU(t;\mu,2\mu)$ process, consistent with the limit for $M/M/\infty$ (Theorem 1.1, [44]).

6.2. Explicit Variance Formulas

In this subsection, we give the explicit variance formulas in Theorem 4.2 for the standard case with exponential service times. For each $t \geq 0$ and $y \geq 0$,

$$\begin{aligned}
\sigma_1^2(t,y) &= \frac{\lambda}{2\mu}e^{-2\mu y}(1-e^{-2\mu t}), \\
\sigma_2^2(t,y) &= \frac{\lambda}{\mu}e^{-\mu y}(1-e^{-\mu t}) - \frac{\lambda}{2\mu}e^{-2\mu y}(1-e^{-2\mu t}), \\
Var(W^r(t,y)) = \sigma_w^2(t,y) &= \frac{\lambda}{\mu^2}e^{-\mu y}(1-e^{-\mu t}) + \frac{\lambda}{2\mu^3}(c_a^2-1)e^{-2\mu y}(1-e^{-2\mu t}) \\
&\to \frac{\lambda}{\mu^2}e^{-\mu y} + \frac{\lambda}{2\mu^3}(c_a^2-1)e^{-2\mu y}, \quad \text{as} \quad t \to \infty.
\end{aligned}$$

So we obtain the variances for the limiting queue-length processes $\hat{Q}^r(t,y)$ and $\hat{Q}^t(t)$:

$$\begin{aligned}
Var(\hat{Q}^r(t,y)) &= \frac{\lambda}{\mu}e^{-\mu y}(1-e^{-\mu t}) + \frac{\lambda}{2\mu}(c_a^2-1)e^{-2\mu y}(1-e^{-2\mu t}) \\
&\to \frac{\lambda}{\mu}e^{-\mu y} + \frac{\lambda}{2\mu}(c_a^2-1)e^{-2\mu y} \quad \text{as} \quad t \to \infty,
\end{aligned}$$

and

$$Var(\hat{Q}^t(\infty)) = Var(\hat{Q}^r(\infty,0)) = \frac{\lambda}{\mu} + \frac{\lambda}{2\mu}(c_a^2-1) = \frac{\lambda}{2\mu}(c_a^2+1).$$

Similar, somewhat more complicated, formulas can be obtained when the service-time c.d.f. is a mixture of exponential c.d.f.'s. That special case is discussed in [54].

7. General Service-Time Distributions

In this section, we treat general service-time c.d.f.'s. Instead of Assumption 2, we assume the following:

Assumption 5: a general c.d.f. We assume that the service times of new arrivals come from a sequence of i.i.d. nonnegative random variables $\{\eta_i : i \geq 1\}$ with a general c.d.f. F, which is independent of n and the arrival processes. We assume that $F(0) < 1$. ■

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., $p_d \equiv \sum_{x \geq 0} \Delta F(x) \leq 1$ and $p_c = 1 - p_d \leq 1$, 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 $\{\bar{x}_1, \bar{x}_2, ...\}$ be such that $\Delta F(\bar{x}_i) \geq \Delta F(\bar{x}_{i+1})$. Define two proper c.d.f.'s F_c and F_d for a continuous random variable η^c and a discrete random variable η^d, respectively, by

$$F_c(x) \equiv P(\eta^c \leq x) \equiv \frac{1}{p_c}\Big(F(x) - \sum_{y \leq x} \Delta F(y)\Big), \quad x \geq 0,$$

and

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

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

Let $A^c_n(t)$, $A^d_n(t)$ and $A^d_{n,i}(t)$ count the number of arrivals by time t with continuous service time, with a discrete service time, and with a deterministic service time $\bar{x}_i$, $i = 1, 2, ...$, respectively. Clearly, $A^d_n(t) = \sum_{i=1}^{\infty} A^d_{n,i}(t)$ and $A_n(t) = A^d_n(t) + A^c_n(t)$ for $t \geq 0$. Define the LLN-scaled processes $\bar{A}^c_n \equiv n^{-1} A^c_n$, $\bar{A}^d_n \equiv n^{-1} A^d_n$, and $\bar{A}^d_{n,i} \equiv n^{-1} A^d_{n,i}$.

Under Assumptions 1 and 5, for a general service-time c.d.f., we can decompose the system into two subsystems, one with arrivals processes A^c_n and service-time distribution F_c and the other with arrival processes A^d_n and discrete service times $\{\bar{x}_i : i \geq 1\}$ with distribution F_d. We analyze the first subsystem using the approach in this paper to obtain the limits for processes in Theorems 3.1 and 3.2, and analyze the second subsystem using the approach in [15], and then we put them together to obtain the limits for the whole system.

First, we obtain a generalization of the FWLLN in Theorem 3.1. Just as with Theorem 3.1, this will be a consequence of the following FCLT in Theorem 7.2 below.

Theorem 7.1. (FWLLN) *Under Assumptions 1 and 5, the conclusions* (3.5) *in Theorem* 3.1 *remain valid. Moreover,*

$$(\bar{A}_n^c, \bar{A}_n^d, \{\bar{A}_{n,i}^d : i \geq 1\}, \bar{Q}_n^r, \bar{Q}_n^e) \Rightarrow (\bar{a}^c, \bar{a}^d, \{\bar{a}_i^d : i \geq 1\}, \bar{q}^r, \bar{q}^e), \tag{7.1}$$

in $D^2 \times D^\infty \times D_D^2$ *as* $n \to \infty$, *where* $\bar{a}^c \equiv p_c \bar{a}$, $\bar{a}^d \equiv p_d \bar{a}$, $\bar{a}_i^d \equiv p_{d,i} \bar{a}^d$, *for* $i \geq 1$, *and the limit processes* $\bar{q}^r(t, y)$ *and* $\bar{q}^e$ *are*

$$\begin{aligned} \bar{q}^r(t,y) &\equiv \int_0^t F_c^c(t+y-s)d\bar{a}^c(s) + \sum_{i=1}^\infty (\bar{a}_i^d(t) - \bar{a}_i^d(t-(\bar{x}_i - y)^+)), \\ &= \int_0^t F^c(t+y-s)d\bar{a}(s), \quad t \geq 0, \quad y \geq 0, \end{aligned} \tag{7.2}$$

and

$$\begin{aligned} \bar{q}^e(t,y) &\equiv \int_{t-y}^t F_c^c(t-s)d\bar{a}^c(s) + \sum_{i=1}^\infty (\bar{a}_i^d(t) - \bar{a}_i^d(t-(\bar{x}_i \wedge y))), \\ &= \int_{t-y}^t F^c(t-s)d\bar{a}(s), \quad t \geq 0, \quad 0 \leq y \leq t. \end{aligned} \tag{7.3}$$

To see how the two terms in (7.2) reduce to the one term in (3.6), note that

$$\begin{aligned} &\int_0^t F_c^c(t+y-s)d\bar{a}^c(s) + \sum_{i=1}^\infty (\bar{a}_i^d(t) - \bar{a}_i^d(t-(\bar{x}_i - y)^+)) \\ = &\int_0^t \Big(F^c(t+y-s) - \sum_{u>t+y-s} \Delta F(u)\Big)d\bar{a}(s) + \sum_{i=1}^\infty \Delta F(\bar{x}_i)(\bar{a}(t) - \bar{a}(t-(\bar{x}_i - y)^+)) \\ = &\int_0^t \Big(F^c(t+y-s) - \sum_{u>t+y-s} \Delta F(u)\Big)d\bar{a}(s) + \int_0^t \Big(\sum_{u>t+y-s} \Delta F(u)\Big)d\bar{a}(s) \\ = &\int_0^t F^c(t+y-s)d\bar{a}(s). \end{aligned}$$

Similarly, the two terms in (7.3) reduce to the one term in (3.7).

For the FCLT, 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

$$\hat{A}_n^c(t) \equiv n^{1/2}(\bar{A}_n^c(t) - \bar{a}^c(t)), \quad \hat{A}_n^d(t) \equiv n^{1/2}(\bar{A}_n^d(t) - \bar{a}^d(t)), \quad \hat{A}_{n,i}^d(t) \equiv n^{1/2}(\bar{A}_{n,i}^d(t) - \bar{a}_i^d(t)),$$

for $t \geq 0$ and $i \geq 1$. By applying the FCLT for split counting processes (Theorem 9.5.1, [55]), we obtain a generalization of the FCLT in Theorem 3.2. Let $\circ$ be the composition function, i.e., $(x \circ y)(t) \equiv x(y(t))$, $t \geq 0$. Let $\overset{\mathrm{d}}{=}$ mean equal in distribution (as random elements).

Theorem 7.2. (FCLT) *Under Assumptions 1 and 5, the conclusions of Theorem* 3.2 *remain valid, provided we modify the limit $\hat{Q}^r$. In addition,*

$$(\hat{A}^c_n, \hat{A}^d_n, \{\hat{A}^d_{n,i} : i \geq 1\}, \hat{Q}^r_n, \hat{Q}^e_n) \Rightarrow (\hat{A}^c, \hat{A}^d, \{\hat{A}^d_i : i \geq 1\}, \hat{Q}^r, \hat{Q}^e), \tag{7.4}$$

in $D^3 \times D^\infty \times D^2_D$ as $n \to \infty$ jointly with the limits in Theorem 3.2*, where*

$$\begin{aligned}
\hat{A}^c &= p_c\hat{A} + S^c \circ \bar{a}, \quad \hat{A}^d = p_d\hat{A} + S^d \circ \bar{a}, \quad \hat{A}^d_i = p_d p_{d,i}\hat{A} + S^d_i \circ \bar{a},\\
S^c &= -S^d, \quad S^c \overset{\mathrm{d}}{=} \sqrt{p_c(1-p_c)}B, \quad S^d \overset{\mathrm{d}}{=} \sqrt{p_d(1-p_d)}B,\\
S^d_i &\overset{\mathrm{d}}{=} \sqrt{p_d p_{d,i}(1 - p_d p_{d,i})}B, \quad i \geq 1,
\end{aligned} \tag{7.5}$$

where B is a standard BM, independent of $\hat{A}$, and the process $(S^c, S^d, \{S^d_i : i \geq 1\})$ 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 $i \geq 1$, $\mathbf{C}_{i,c} = \mathbf{C}_{c,i} = -p_c p_d p_{d,i}$, $\mathbf{C}_{i,d} = \mathbf{C}_{d,i} = -p^2_d p_{d,i}$ and $\mathbf{C}_{i,j} = -p^2_d p_{d,i} p_{d,j}$ for $i \neq j$. The new representations for $\hat{Q}^r$ and $\hat{Q}^e$ are*

$$\begin{aligned}
\hat{Q}^r(t,y) &= \hat{X}^c_1(t,y) + \hat{X}^c_2(t,y) + \hat{X}^d(t,y), \quad t \geq 0, \quad y \geq 0, \\
\hat{Q}^e(t,y) &= \hat{X}^{c,e}_1(t,y) + \hat{X}^{c,e}_2(t,y) + \hat{X}^{d,e}(t,y), \quad t \geq 0, \quad 0 \leq y \leq t,
\end{aligned} \tag{7.6}$$

where

$$\begin{aligned}
\hat{X}^c_1(t,y) &\equiv \int_0^t F^c_c(t+y-s)d\hat{A}^c(s), \quad \hat{X}^{c,e}_1(t,y) \equiv \int_{t-y}^t F^c_c(t-s)d\hat{A}^c(s),\\
\hat{X}^c_2(t,y) &\equiv \int_0^t \int_0^\infty \mathbf{1}(s+x > t+y)\, d\hat{K}^c(\bar{a}^c(s), x),\\
\hat{X}^{c,e}_2(t,y) &\equiv \int_{t-y}^t \int_0^\infty \mathbf{1}(s+x > t)\, d\hat{K}^c(\bar{a}^c(s), x),\\
\hat{X}^d(t,y) &\equiv \sum_{i=1}^\infty (\hat{A}^d_i(t) - \hat{A}^d_i(t - (\bar{x}_i - y)^+)),\\
\hat{X}^{d,e}(t,y) &\equiv \sum_{i=1}^\infty (\hat{A}^d_i(t) - \hat{A}^d_i(t - (\bar{x}_i \wedge y))),
\end{aligned}$$

with $\hat{K}^c(\bar{a}^c(s), x) = U(\bar{a}^c(s), F_c(x))$. The other processes in Theorem 3.2 remain the same after we replace $\hat{Q}^r$ by (7.6). If, in addition, $\hat{A} = \sqrt{c^2_a}B_a \circ \bar{a}$, as when A_n is nonhomogeneous Poisson, then $\hat{A}^d$ and $\hat{A}^c$ are independent.

Moreover, in general, the limit processes $\hat{Q}^r$ and $\hat{Q}^e$ can also be expressed as the sum of the *following three mutually independent processes*

$$\begin{aligned}
\hat{Q}^r(t,y) &= \hat{X}_1(t,y)+\hat{X}_2^c(t,y)+\hat{X}_3(t,y), \quad t,y\geq 0, \\
\hat{Q}^e(t,y) &= \hat{X}_1^e(t,y)+\hat{X}_2^{c,e}(t,y)+\hat{X}_3^e(t,y), \quad t\geq 0, \quad 0\leq y\leq t,
\end{aligned} \tag{7.7}$$

where $\hat{X}_1(t,y)$ and $\hat{X}_1^e(t,y)$ are defined in (3.14) and (3.15), respectively, and

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

Proof. Only a few details require discussion. First, we need to explain (7.7). By (7.5), we can write $\hat{X}_1^c(t,y)$ and $\hat{X}^d(t,y)$ in (7.6) as

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

and

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

Thus, $\hat{X}_1^c(t,y)+\hat{X}^d(t,y)=\hat{X}_1(t,y)+\hat{X}_3(t,y)$ 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.

Next, we remark that in [15], the convergence to the limit $\hat{X}^d(t,y)$ is proved in the space D for each fixed $y\geq 0$, 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 $\hat{X}_n^d$ is

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

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

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

is continuous in the Skorohod J_1 topology and then apply the continuous mapping theorem. Suppose $z_n \to z$ in D as $n \to \infty$, we need to show that $d_{D_D}(\phi(z_n), \phi(z)) \to 0$ as $n \to \infty$. Fix $T > 0$. By the convergence of $z_n \to z$, there exist increasing homeomorphisms λ_n over the interval $[0,T]$ such that $||z_n - z \circ \lambda_n||_T \to 0$ and $||\lambda_n - e||_T \to 0$ as $n \to \infty$, where $e(t) = t$ for all $t \geq 0$. It suffices to show that

$$\sup_{t,y \leq T} \Big| \sum_{i=1}^{\infty} (z_n(t) - z_n(t - (\bar{x}_i - y)^+) - (z(\lambda_n(t)) - z(\lambda_n(t) - (\bar{x}_i - \lambda_n(y))^+) \Big| \to 0$$

as $n \to \infty$. This follows easily from the assumptions on λ_n and the convergence of $z_n \to z$.

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

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

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

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

Since the limit $\hat{A}$ is continuous, it suffices to show the uniform continuity of the mapping ψ on compact intervals, which follows from a direct argument. ■

Since $\hat{X}^d(t,y)$ and $\hat{X}^{d,e}(t,y)$ only involve $\hat{A}_i^d$, if the limit process $\hat{A}^d$ has independent increments (inherited from the process $\hat{A}$), then $\hat{X}^d(t,y)$ and $\hat{X}^{d,e}(t,y)$ also have independent increments and are thus Markov processes (with respect to argument t). Thus, Theorem 4.1 extends to the process $\hat{Q}^r$ in (7.6). Moreover, we obtain the following generalization of the Gaussian property in Theorem 4.2.

Theorem 7.3. (Gaussian property) *If, in addition to the assumptions of Theorem* 7.2*, then the limit process $\hat{A}$ is Gaussian, the limit process in* (7.6) *and* (7.7) *is also a continuous Gaussian process. 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 $t, y \geq 0$,*

$$\hat{Q}^r(t,y) \overset{\mathrm{d}}{=} N(0, \sigma_{q,r}^2(t,y)), \quad \hat{Q}^e(t,y) \overset{\mathrm{d}}{=} N(0, \sigma_{q,e}^2(t,y)), \quad \hat{W}^r(t,y) \overset{\mathrm{d}}{=} N(0, \sigma_w^2(t,y)), \quad (7.8)$$

where

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

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 (7.7) to calculate them and the mutual independence between the three terms in the expression of $\hat{Q}^r$ gives $\sigma^2_{q,r}(t,y) = \sigma^2_1(t,y) + \sigma^2_{2,c}(t,y) + \sigma^2_3(t,y)$, where $\sigma^2_1(t,y) = E[(\hat{X}_1(t,y))^2]$, $\sigma^2_{2,c}(t,y) = E[(\hat{X}^c_2(t,y))^2]$ and $\sigma^2_3(t,y) = E[(\hat{X}_3(t,y))^2]$. We have

$$\sigma^2_1(t,y) = c_a^2 \int_0^t F^c(t+y-s)^2 d\bar{a}(s),$$

and

$$\begin{aligned}
\sigma^2_3(t,y) &= p_d p_c \int_0^t F^c_c(t+y-s)^2 d\bar{a}(s) + \sum_{i=1}^\infty \Big(p_d p_{d,i}(1-p_d p_{d,i})(\bar{a}(t)-\bar{a}(t-(\bar{x}_i-y)^+))\Big)\\
&\quad - 2p_d^2 \sum_{i<j} p_{d,i}p_{d,j}\big(\bar{a}(t)-\bar{a}(t-((\bar{x}_i\wedge\bar{x}_j)-y)^+)\big)\\
&\quad - 2\sum_{i=1}^\infty p_c p_d p_{d,i} \int_0^t F^c_c(t+y-s) d(\bar{a}(s)-\bar{a}(s-(\bar{x}_i-y)^+)),
\end{aligned}$$

and

$$\sigma^2_{2,c}(t,y) \;=\; \int_0^t F_c(t+y-s)F^c_c(t+y-s) d\bar{a}^c(s).$$

Notice that

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

Moreover, $F^c = p_c F^c_c + p_d F^c_d$ and $FF^c = (1-p_cF^c_c - p_dF^c_d)(p_cF^c_c + p_dF^c_d) = p_cF^c_c(1-p_cF^c_c) + p_dF^c_d(1-p_dF^c_d) - 2p_cF^c_c p_dF^c_d$. Then, simple algebra calculation gives the final expression for $\sigma^2_{q,r}(t,y)$. Similar argument applies to the calculation of $\sigma^2_{q,e}(t,y)$ and $\sigma^2_w(t,y)$. ∎

We consider a simple example to check the variance formulas in Theorem 7.3, where the arrival processes are Poisson and the service time distribution is a mixture of an exponential service time of mean μ^{-1} with probability p and a deterministic service time $\bar{x}$ ($\bar{x} > 0$) with probability $1-p$. So $F = pF_c + (1-p)F_d$, where $F_c(x) = 1 - e^{-\mu x}$ and $F_d(x) = \mathbf{1}(x \geq \bar{x})$ for $x \geq 0$. By Poisson thinning, $\hat{A} = \hat{A}^c + \hat{A}^d = B_a \circ \lambda e$, where $\hat{A}^c \overset{\mathrm{d}}{=} B_a \circ \lambda p e$, $\hat{A}^d \overset{\mathrm{d}}{=} B_a \circ \lambda(1-p)e$, $\hat{A}^c$ is independent of $\hat{A}^d$, and $e(t) = t$ for all $t \geq 0$. Then the three terms in (7.6) are mutually independent, so we obtain $Var(\hat{Q}^r(t,y))$ easily from the variances of the three terms

$$\begin{aligned} & Var(\hat{Q}^r(t,y)) = Var(\hat{X}_1^c(t,y)) + Var(\hat{X}_2^c(t,y)) + Var(\hat{X}^d(t,y)) \\ = \;& p\frac{\lambda}{2\mu}e^{-2\mu y}(1-e^{-2\mu t}) + p\Big(\frac{\lambda}{\mu}e^{-\mu y}(1-e^{-\mu t}) - \frac{\lambda}{2\mu}e^{-2\mu y}(1-e^{-2\mu t})\Big) + (1-p)\lambda(\bar{x}-y)^+ \\ = \;& p\frac{\lambda}{\mu}e^{-\mu y}(1-e^{-\mu t}) + (1-p)\lambda(\bar{x}-y)^+. \end{aligned} \tag{7.9}$$

This is consistent with our intuition that the first term accounts for the variance from the exponential service time, while the second term accounts for the variance from the deterministic service time. It is easy to see that the variance formula given in Theorem 7.3 will give us the same result. Note that the split arrival process in (7.5) gives $\hat{A}^c \overset{\mathrm{d}}{=} pB_a + S^c \circ \lambda e$ and $\hat{A}^d \overset{\mathrm{d}}{=} (1-p)B_a + S^d \circ \lambda e$, where $S^c \overset{\mathrm{d}}{=} \sqrt{p(1-p)}B$, $S^d \overset{\mathrm{d}}{=} \sqrt{p(1-p)}B$, B is a standard BM, independent of B_a, and $Cov(S^c, S^d) = -p(1-p)$. Then, by applying the formulas in the proof of Theorem 7.3, we obtain

$$\begin{aligned} \sigma_1^2(t,y) &= \lambda\int_0^t \big(1 - p(1-e^{-\mu(y+s)}) - (1-p)\mathbf{1}(y+s \geq \bar{x})\big)^2 ds \\ &= \lambda(1-p)^2(\bar{x}-y)^+ + p^2\frac{\lambda}{2\mu}e^{-2\mu y}(1-e^{-2\mu t}) + 2p(1-p)\frac{\lambda}{\mu}e^{-\mu y}(1-e^{-\mu(\bar{x}-y)^+}), \\ \sigma_{2,c}^2(t,y) &= p\Big(\frac{\lambda}{\mu}e^{-\mu y}(1-e^{-\mu t}) - \frac{\lambda}{2\mu}e^{-2\mu y}(1-e^{-2\mu t})\Big), \\ \sigma_3^2(t,y) &= p(1-p)\frac{\lambda}{2\mu}e^{-2\mu y}(1-e^{-2\mu t}) + \lambda p(1-p)(\bar{x}-y)^+ \\ &\quad - 2p(1-p)\frac{\lambda}{\mu}e^{-\mu y}(1-e^{-\mu(\bar{x}-y)^+}). \end{aligned}$$

Summing up the three terms gives us the same result as in (7.9), which confirms that our variance formulas are correct.

8. Proof of the FCLT

Our goal now is to prove the FCLT in Theorem 3.2. 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 $\hat{Q}^r_n$ 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 $\hat{Q}^r_n$. 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 $\hat{Q}^r_n$ 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 (3.18). However, $\hat{X}_{n,2}$ presents a problem. Unfortunately, the general results of weak convergence of stochastic integrals and differential equations in [35, 39, 36] does not seem to apply. Thus, instead, we will follow Krichagina and Puhalskii [32] and prove the convergence in the classical way, by proving tightness and convergence of the finite-dimensional distributions, exploiting more involved arguments, including the semi-martingale decomposition used in [32].

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 §13.2 of [55]; see [**?**] 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 [55]. 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

$$(\hat{A}_n, \bar{A}_n, \hat{K}_n, \hat{R}_n) \Rightarrow (\hat{A}, \bar{a}, \hat{K}, \hat{R}) \quad \text{in} \quad D^2 \times D^2_D, \tag{8.1}$$

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

$$(\hat{A}_n, \bar{A}_n, \hat{K}_n, \hat{R}_n, \hat{X}_{n,1}, \hat{X}_{n,2}) \Rightarrow (\hat{A}, \bar{a}, \hat{K}, \hat{R}, \hat{X}_1, \hat{X}_2) \tag{8.2}$$

in $D^2 \times D^4_D$: We show convergence of the finite-dimensional distributions and tightness. We get tightness for $\{(\hat{A}_n, \bar{A}_n, \hat{K}_n, \hat{R}_n) : n \geq 1\}$ from the convergence in (8.1). 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 [55]. Since we can write $\hat{X}_{n,1}$ as (3.18), 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 (3.18) 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 §9. We then establish tightness for $\{(\hat{X}_{n,1}, \hat{X}_{n,2}) : n \geq 1\}$ in §10 and the required convergence of the finite-dimensional distributions associated with $\{(\hat{X}_{n,1}, \hat{X}_{n,2}) : n \geq 1\}$ in §11. Given the limit in (8.2), 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 [55]. The component limits require finite second moments.

9. Continuity of the Representation for $\hat{X}_{n,1}$ in D_D

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

$$\begin{aligned} \phi(x)(t,y) &\equiv F^c(y)x(t) - \int_0^t x(s-)dF(t+y-s), \\ &= F^c(y)x(t) - \int_y^{t+y} x((t+y-s)-)dF(s), \end{aligned} \tag{9.1}$$

for $x \in D$ and $t, y \geq 0$. By (3.18) and (3.14), 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 9.1. *The mapping* ϕ *defined in* (9.1) *is continuous in* D_D.

Proof. Suppose $x_n \to x$ in D, we need to show that

$$d_{D_D}(\phi(x_n), \phi(x)) \to 0, \quad \text{as} \quad n \to \infty. \tag{9.2}$$

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 λ_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

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

Now,

$$\begin{aligned}
& |\phi(x_n)(t,y) - \phi(x)(\lambda_n(t),y)| \\
= \; & \Big| F^c(y)x_n(t) - \int_0^t x_n(s-)dF(t+y-s) \\
& \quad - F^c(y)x(\lambda_n(t)) + \int_0^{\lambda_n(t)} x(s-)dF(\lambda_n(t)+y-s) \Big| \\
\leq \; & F^c(y)\big|x_n(t) - x(\lambda_n(t))\big| \\
& + \Big| \int_0^t x_n(s-)dF(t+y-s) - \int_0^{\lambda_n(t)} x(s-)dF(\lambda_n(t)+y-s) \Big| \\
= \; & F^c(y)\big|x_n(t) - x(\lambda_n(t))\big| \\
& + \Big| \int_0^t x_n(s-)dF(t+y-s) - \int_0^t x(\lambda_n(s)-)dF(\lambda_n(t)+y-\lambda_n(s)) \Big| \\
\leq \; & F^c(y)\big|x_n(t) - x(\lambda_n(t))\big| + \Big| \int_0^t (x_n(s-) - x(\lambda_n(s)-))dF(t+y-s) \Big| \\
& + \Big| \int_0^t x(\lambda_n(s)-)d(F(\lambda_n(t)+y-\lambda_n(s)) - F(t+y-s)) \Big| \\
\leq \; & F^c(y)\big|x_n(t) - x(\lambda_n(t))\big| + ||x_n - x \circ \lambda_n||_T |F(y) - F(t+y)| \\
& + M|F(\lambda_n(t)+y) - F(t+y)| \\
\leq \; & 3||x_n - x \circ \lambda_n||_T + M|F(\lambda_n(t)+y) - F(t+y)|.
\end{aligned}$$

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 that (9.2) holds, i.e., the mapping $\phi : D \to D_D$ is continuous. ■

10. 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 $\{\hat{X}_{n,1} : n \geq 1\}$ and $\{\hat{X}_{n,2} : n \geq 1\}$ in D_D. By Assumption 1, the sequence of processes $\{\hat{A}_n : n \geq 1\}$ is tight. The tightness of $\{\hat{X}_{n,1}\}$ follows from the continuity of the mapping ϕ in D_D. It remains to show the tightness of $\{\hat{X}_{n,2}\}$ and then we obtain tightness of the sequences of processes $\{\bar{Q}^r_n : n \geq 1\}$ and $\{\hat{Q}^r_n : n \geq 1\}$ using the fact that tightness of product spaces is equivalent to the tightness on each of the component spaces.

Theorem 10.1. *Under Assumptions 1 and 2, the sequence of processes* $\{\hat{X}_{n,1} : n \geq 1\}$*,* $\{\hat{X}_{n,2} : n \geq 1\}$*,* $\{\bar{Q}^r_n : n \geq 1\}$ *and* $\{\hat{Q}^r_n : n \geq 1\}$ *are individually and jointly tight.*

We will also need to generalize the tightness criteria in Lemma VI.3.32 in [25] 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 [25] with inequalities for the modulus of continuity for functions in the space D_D.

Lemma 10.1. *Suppose that a sequence of processes* $\{X_n : n \geq 1\}$ *in the space* D_D *can be decomposed into two sequences* $\{Y^q_n : n \geq 1\}$ *and* $\{Z^q_n : n \geq 1\}$ *for some parameter* $q \in \mathbb{N}$*, i.e.,* $X_n = Y^q_n + Z^q_n$ *for each* $n \geq 1$*, and that* (i) *the sequence* $\{Y^q_n : n \geq 1\}$ *is tight in the space* D_D *and* (ii) *for all* $T > 0$ *and* $\delta > 0$*,* $\lim_{q\to\infty} \limsup_{n\to\infty} P(\sup_{t,y\leq T} |Z^q_n(t,y)| > \delta) = 0$*. Then, the sequence* $\{X_n : n \geq 1\}$ *is tight in the space* D_D*.*

In order to prove the tightness of $\{\hat{X}_{n,2} : n \geq 1\}$ defined in (2.13), we will closely follow the approach in [32] but we have to adjust to the tightness criteria in D_D (see Appendix C for the tightness criteria in D_D). We first give a decomposition of the process $\hat{X}_{n,2}$ for each n. By Proposition A.2 (following [25] and [32]), $\hat{R}_n(t,y)$ in (2.14) can be written 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

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

We remark that we need not consider the left-hand limit of $\hat{R}_n$ in the second argument, as was done in [32], since the service-time c.d.f F is assumed to be continuous, while F is allowed to be discontinuous in [32]. Hence, $\hat{X}_{n,2}$ can be written as

$$\hat{X}_{n,2}(t,y) = \hat{G}_n(t,y) + \hat{H}_n(t,y), \quad \text{for} \quad t \geq 0 \quad \text{and} \quad y \geq 0, \tag{10.1}$$

where

$$\begin{aligned} \hat{G}_n(t,y) &\equiv \int_0^t \int_0^\infty \mathbf{1}(s+x \leq t+y) d\Big(-\int_0^x \frac{\hat{R}_n(s,v)}{1-F(v)} dF(v) \Big) \\ &= -\int_0^{t+y} \frac{\hat{R}_n(t+y-x,x)}{1-F(x)} dF(x), \end{aligned} \tag{10.2}$$

and

$$\hat{H}_n(t,y) \equiv \int_0^t \int_0^\infty \mathbf{1}(s+x \leq t+y) d\hat{L}_n(s,x). \tag{10.3}$$

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 10.2. *Under Assumptions 1 and 2, the sequence of processes* $\{\hat{G}_n : n \geq 1\} \equiv \{\{\hat{G}_n(t,y) : t \geq 0, y \geq 0\}, n \geq 1\}$ *is tight in* D_D.

Proof. We will apply Lemma 10.1. We define the sequence of processes $\{\hat{G}_n^\epsilon : n \geq 1\}$, for some $\epsilon \in (0,1)$, by

$$\hat{G}_n^\epsilon(t,y) \equiv -\int_0^{t+y} \frac{\hat{R}_n(t+y-x,x)}{1-F(x)} \mathbf{1}(F(x) \leq 1-\epsilon) dF(x), \quad t,y \geq 0. \tag{10.4}$$

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

$$\lim_{\epsilon \downarrow 0} \limsup_n P\Big(\sup_{t,y \leq T} \Big| \int_0^{t+y} \frac{\hat{R}_n(t+y-x,x)}{1-F(x)} \mathbf{1}(F(x) > 1-\epsilon) dF(x) \Big| > \delta \Big) = 0, \tag{10.5}$$

for each $\delta > 0$ and $T > 0$, and thus will conclude that the sequence $\{\hat{G}_n\}$ is tight in D_D by Lemma 10.1. It is easy to see that (10.5) follows easily from (3.23) in [32]. So we only need to prove the tightness of the sequence of processes $\{\hat{G}_n^\epsilon : n \geq 1\}$.

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, [55]), we obtain

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

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 stochastic bounded, so the integral representation of $\hat{G}_n^\epsilon$ in terms of $\hat{R}_n$ in (10.4) implies that $\{\hat{G}_n^\epsilon : n \geq 1\}$ is also stochastically bounded in D_D. We apply Theorem C.1 to prove the tightness of $\{\hat{G}_n^\epsilon : n \geq 1\}$ in D_D. In this case, it is convenient to use the sufficient criterion in the remark right after Theorem C.1.

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

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

where $\mathcal{N}$ includes all the null sets. Note that the filtration $\mathbf{G}_n$ satisfies the usual conditions (Chapter 1, [30] and proof of Lemma 3.1 in [32]).

Let $\delta_n \downarrow 0$ and $\{\tau_n : n \geq 1\}$ be a uniformly bounded sequence, where for each n, τ_n is a stopping times with respect to the filtration $\mathbf{G}_n$. Then, it suffices to show that

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

Consider any sequence of nondecreasing homeomorphism $\{\lambda_n : n \geq 1\}$ 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:

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

Now,

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

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 τ_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. We present an alternative direct argument via the lemma below. ■

The following lemma provides an alternative direct way to complete the last step of the proof of Lemma 10.2 above, instead of applying Lemma 10.1. For our alternative argument, we apply the following lemma together with Prohorov's theorem, p. 387 of [55]. Prohorov's theorem tells us that a set Π of probability measures on a complete separable metric space S is tight if and only if it is relatively compact, i.e., if its closure is sequentially compact (and thus just compact) in the complete separable metric space $\mathcal{P}(S)$ of probability measures on the space S, with the usual topology associated with weak convergence, which is generated by the Prohorov metric; see p. 77 of [55]. Thus to characterize tightness, we can focus on relative compactness in the set of probability measures. A sequence of probability measures is relatively compact if each subsequence has a further subsequence converging to a limit (not necessarily an element of the sequence). The following lemma is applied in the space of probability measures on D_D. Condition (ii) below is implied by property (10.5). Property (10.5) directly implies that the usual distance associated with convergence in probability converges uniformly to 0, but that distance is greater than or equal to the Prohorov distance; see p. 375 of [55].

Lemma 10.3. *Suppose that* $\{x_n : n \geq 1\}$ *and* $\{x_n^\epsilon : n \geq 1\}$ *are sequences in a complete separable metric space* (S, d) *for each* $\epsilon > 0$*. If (i)* $\{x_n^\epsilon : n \geq 1\}$ *is relatively compact for each* $\epsilon > 0$ *and (ii)* $\sup_{n \geq 1}\{d(x_n, x_n^\epsilon)\} \rightarrow 0$ *as* $\epsilon \rightarrow 0$*, then* $\{x_n : n \geq 1\}$ *is relatively compact.*

Proof. We want to show that the sequence $\{x_n\}$ is relatively compact; i.e., we want to show that each subsequence itself has a convergent subsequence (a subsubsequence). Start by choosing a subsequence of $\{x_n\}$. We now want to find a convergent subsequence of that subsequence. We will construct one such directly.

To do so, choose a sequence $\{\epsilon_n\}$ decreasing to 0. For ϵ_1, start with the initial subsequence, i.e., a subsequence of $\{x_n^{\epsilon_1}\}$ with the same indices as the convergent subsequence of $\{x_n\}$. Since $\{x_n^{\epsilon_1}\}$ is relatively compact, there exists a subsequence of that initial subsequence for ϵ_1 for which there is convergence to some limit L_1. Let that convergent subsequence (the indices) be the initial sequence for ϵ_2. That is, for ϵ_2, start with a subsequence having the indices of the final convergent subsequence for ϵ_1. Since $\{x_n^{\epsilon_2}\}$ is relatively compact, there exists a subsequence of that initial subsequence for which there is convergence to some limit L_2. That will still yield convergence for ϵ_1, because a subsequence of a converging sequence is again converging. We can continue in that way recursively. Hence, at stage n, we have a common subsequence (indices) working for $\epsilon_i, 1 \leq i \leq n$, for which we get convergence of $x_n^{\epsilon_i}$ to a limit L_i for each i. We now want to deduce, first, that $\{L_n\}$ must converge to some L as $n \rightarrow \infty$ and that, second, also that some subsequence of x_n must converge to this same L. We now apply condition (ii) to deduce that the sequence $\{L_n\}$ must be Cauchy. By completeness, we deduce that there is an L such that L_n converges to L.

It remains to construct the subsequence of $\{x_n\}$. To do so, we use the diagonal sequence, i.e., we let the index of the k^{th} term be the k^{th} term (index) of the convergent subsequence constructed for $\{x_n^{\epsilon_k}\}$. Observe that, with these indices, $\{x_n^{\epsilon_k}\}$ converges to L_k as $n \rightarrow \infty$ for each k. Finally, we can apply condition (ii) again to deduce that this diagonal subsequence of x_n also converges to L. Hence, the sequence $\{x_n\}$ itself must be relatively compact. ■

Lemma 10.4. *Under Assumptions 1 and 2, the sequence of processes* $\{\hat{H}_n : n \geq 1\} \equiv \{\{\hat{H}_n(t, y) : t \geq 0, y \geq 0\}, n \geq 1\}$ *is tight in* D_D*.*

Proof. As in Lemma 3.7 in [32], we write the process $\hat{H}_n$ as

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

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

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

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

where $\mathcal{N}$ includes all the null sets. The filtration $\mathbf{H}_n$ satisfies the usual conditions (see p. 254 in [32]).

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

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \sup_{\kappa_n} E[d_{J_1}(\hat{H}_n(\kappa_n+\delta,\cdot), \hat{H}_n(\kappa_n,\cdot))^2] = 0. \tag{10.6}$$

Consider any sequence of nondecreasing homeomorphism $\{\lambda_n : n \ge 1\}$ 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:

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \sup_{\kappa_n} E\Big[\Big(\sup_{0 \le y \le T} |\hat{H}_n(\kappa_n+\delta, \lambda_n(y)) - \hat{H}_n(\kappa_n, y)| \Big)^2\Big] = 0. \tag{10.7}$$

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

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

As in Lemma 3.5 in [32], one can check that for each fixed y and for each i, the process $\{\hat{H}_{n,i}(t,y) : t \ge 0\}$ is a square integrable martingale with respect to the filtration $\mathbf{H}_n$ and it has predictable quadratic variation

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

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 $\hat{H}_n^{(K)} = \{\hat{H}_n^{(K)}(t,y) : t \ge 0\}$ defined by

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

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

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

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

$$\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor}\int_0^{\eta_i}\frac{1}{1-F(u)}dF(u) \to t, \quad a.s. \quad \text{as} \quad n\to\infty. \tag{10.8}$$

So for each fixed y, the sequence of quadratic variations $\{\langle \hat{H}_n^{(K)}(\cdot, y)\rangle : n \geq 1\}$ is C-tight by the continuity of $\bar{a}$ in Assumption 1 (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 [57] that the sequence $\{\hat{H}_n^{(K)}(\cdot, y) : n \geq 1\}$ is C-tight for each fixed y.

Now, to prove (10.7), we have

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

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 κ_n and λ_n and C-tightness of $\{\hat{H}_n^{(K)} : n \geq 1\}$. We obtain the convergence to 0 for the second term by observing that

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

Thus we obtain (10.7).

Now we prove the stochastic boundedness of $\{\hat{H}_n : n \geq 1\}$ 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

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

The stochastic boundedness of $\{\tilde{H}_n : n \geq 1\}$ in D_D follows directly from the proof of Lemma 3.7 in [32]. Thus, $\{\hat{H}_n : n \geq 1\}$ is stochastically bounded, so that tightness of $\{\hat{H}_n : n \geq 1\}$ in D_D is proved. ■

11. 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 $\{(\hat{X}_{n,1}, \hat{X}_{n,2}) : n \geq 1\}$ in §10. 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 11.1. *Under Assumptions 1 and 2, 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 (2.13) and $\hat{X}_2$ in (3.14) 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,

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

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

and

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

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

$$\mathbf{1}^y_{k,t}(s,x) = \mathbf{1}(s=0)\mathbf{1}(x \le t+y) + \sum_{i=1}^{k} \mathbf{1}(s \in (s^k_{i-1}, s^k_i])\mathbf{1}(x \le t+y-s^k_i), \tag{11.1}$$

with the points $0 = s^k_0 < s^k_1 < ... < s^k_k = t$ chosen so that $\max_{1\le i\le k} |s^k_{i-1} - s^k_i| \to 0$ as $k \to \infty$, and $\Delta_{\hat{U}_n}$ and Δ_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 $\{\tilde{X}_{n,2,k}(t,y) : t, y \ge 0\}$ 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,

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

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 is asymptotically negligible in probability as $n \to \infty$ and $k \to \infty$, i.e.,

$$\lim_{n\to\infty} P\Big(\sup_{0\le t\le T, y\ge 0} |\hat{X}_{n,2,k}(t,y) - \tilde{X}_{n,2,k}(t,y)| > \epsilon\Big) = 0, \quad T > 0, \quad \epsilon > 0. \tag{11.2}$$

and

$$\lim_{k\to\infty} \limsup_{n\to\infty} P(|\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)| > \epsilon) = 0, \quad t, y \ge 0, \quad \epsilon > 0. \tag{11.3}$$

We obtain (11.2) easily from Assumption 1 and (2.5) since $\bar{a}$ and U are continuous. Now we proceed to prove (11.3). We will follow a martingale approach argument similar to the one used in Lemma 5.3 of [32], which relies on their technical Lemma 5.2. Fortunately, for our two-parameter processes, the conditions of Lemma 5.2 [32] are satisfied by fixing the second argument. We can write

$$\begin{aligned} & P(|\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)| > \epsilon) \\ \leq \quad & P(\bar{A}_n(t) > \delta) + P(|\hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y)| > \epsilon, \bar{A}_n(t) \leq \delta) \qquad (11.4) \end{aligned}$$

for $t, y \geq 0$ and $\delta > 0$.

On $\{\bar{A}_n(t) \leq \delta\}$,

$$\begin{aligned} \hat{X}_{n,2,k}(t,y) - \hat{X}_{n,2}(t,y) &= \int_0^t \int_0^\infty (\mathbf{1}_{k,t}^y(s,x) - \mathbf{1}(s+x \leq t+y)) d\hat{U}_n(\bar{A}_n(s), F(x)) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{A_n(t)\wedge(n\delta)} \beta_i(\tau_i^n, \eta_i)(t,y), \end{aligned}$$

where

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

Define the process $Z_n^{(\delta)} \equiv \{Z_n^{(\delta)}(t,y) : t, y \geq 0\}$ by

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

As in Lemma 5.2 in [32], one can check that for each fixed $y > 0$, the process $Z_n^{(\delta)}(\cdot, y) = \{Z_n^{(\delta)}(t,y) : t \geq 0\}$ is a square integrable martingale with respect to the filtration $\mathbf{F}_n = \{\mathcal{F}_n(t), t \geq 0\}$, where

$$\mathcal{F}_n(t) \quad = \quad \sigma\{\eta_i \leq s+x : 1 \leq i \leq A_n(t), x \geq 0, 0 \leq s \leq t\} \vee \{A_n(s) : 0 \leq s \leq t\} \vee \mathcal{N},$$

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

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

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 $\{\bar{A}^n(t) \leq \delta\}$,

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

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, (11.4) becomes

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

Therefore, by the stochastic boundedness of $\bar{A}_n$, (11.3) 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 $n \to \infty$. ■

12. Conclusions

We have established heavy-traffic limits for a sequence of infinite-server queues with increasing arrival rates, allowing i.i.d. service times with a general c.d.f. We allowed general G and even nonstationary G_t arrival processes; the key condition (Assumption 1) is that the arrival processes satisfy a FCLT, where the limit has continuous sample paths. For the Markov property of $\hat{Q}^r$ in D_D, we required that the limit $\hat{A}$ have independent increments. Of course, the limit process $\hat{A}$ will usually be some form of Brownian motion. The standard case is $\hat{A} = \sqrt{\lambda c_a^2} B_a$, where B_a is a standard BM. For G_t arrival processes, a common case should be the time-transformed analog $\hat{A}(t) = \sqrt{\lambda c_a^2} B_a(\bar{a}(t))$, where $\bar{a}(t) \equiv \int_0^t \lambda(s)\, ds$, where λ is the time-varying arrival-rate function, as in Theorem 4.2 (a) In general, the parameter c_a^2 should be time-varying as well..

In order to capture all the relevant state information, we considered the two-parameter processes $Q_n^e(t,y)$ and $Q_n^r(t,y)$, representing the number of customers in the system at time t having elapsed service times less than or equal to time y, or residual service times strictly greater than y. We regarded these as random elements of the space D_D. From these basic processes we constructed several related processes of interest, such as the workload processes $W_n^r(t, y)$, for which we also established heavy-traffic limits. Our main theorems - Theorems 3.2 and 7.2 - establish a joint heavy-traffic FCLT for twelve processes.

Our analysis indicates that, in order to describe the distribution of future events in a many-server service system, we should keep track of the elapsed service times of customers in the system in addition to the number of customers in the system; i.e., we should monitor the stochastic process

$$Q_n^e \equiv \{Q_n^e(t, \cdot) : t \geq 0\} \equiv \{\{Q_n^e(t, y) : y \geq 0\} : t \geq 0\} \quad \text{in} \quad D_D.$$

We can then approximate $Q_n^e(t, y)$ by $nq^e(t, y) + \sqrt{n}\hat{Q}^e(t, y)$, $y \geq 0$, $t \geq 0$. These two-parameter heavy-traffic limits provide a rich description of the state of the system. In applications, the limit processes $\hat{Q}^e$ and $\hat{Q}^r$ can be used to approximate the steady-state and transient distributions. With the aid of these approximations, methods for control and recovery developed in [10] can be extended to non-Poisson arrival processes.

We expect that the approach here can be be used to establish two-parameter heavy-traffic limits for networks of infinite-server queues, extending [41], and finite-server systems, using methods as in [45, 46, 47]. We expect that the results here can be used to develop effective control and recovery schemes for multi-server systems, extending [10].

13. *

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

14. *

References

[1] Bickel P.J. and M. J. Wichura. 1971. Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* 42, 1656–1670.

[2] Billingsley, P. 1968. *Convergence of Probability Measures*, Wiley (second edition, 1999).

[3] Borovkov, A. A. 1967. On limit laws for service processes in multi-channel systems (in Russian). *Siberian Math J.* 8, 746–763.

[4] Brémaud, P. 1981. *Point Processes and Queues: Martingale Dynamics*, Springer.

[5] Cairoli, R. 1972. Sur une equation differentielle stochastique. *Compte Rendus Acad. Sc.* Paris 274 Ser. A, 1739-1742.

[6] Cairoli, R. and J.B. Walsh. 1975. Stochastic integrals in the plane. *Acta Math.* 134, 111-183.

[7] Csörgö, M. and P. Révész. 1981. *Strong Approximations in Probabilitiy and Statistics*, Wiley, New York.

[8] Decreusefond, L. and P. Moyal. 2008. A functional central limit theorem for the $M/GI/\infty$ queue. *Ann. Appl. Prob.* Vol. 18, No. 6, 2156-2178.

[9] Descreusefond, L., and P. Moyal. 2008. Fluid limit of a heavily loaded edf queue with impatient customers. *Markov Processes and Related Fields* 14,131–158.

[10] Duffield, N. G. and W. Whitt. 1997. Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems* 26, 69–104.

[11] Eick, S. G., W. A. Massey and W. Whitt. 1993. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41, 731–742.

[12] Ethier, S. N. and T. G. Kurtz. 1986. *Markov Processes: Characterization and Convergence.* Wiley.

[13] Gaenssler, P. and W. Stute. 1979. Empirical processes: a survey of results for independent and identically distributed random variables. *Ann. Probab.* 7, 193–243.

[14] Glynn, P. W. 1982. On the Markov property of the $GI/G/\infty$ Gaussian limit. *Adv. Appl. Prob.* 14, 191–194.

[15] Glynn, P. W. and W. Whitt. 1991. A new view of the heavy-traffic limit theorem for the infinite-server queue. *Adv. Appl. Prob.* 23, 188–209.

[16] Goldberg, D. A. and W. Whitt. 2008. The last departure time from an $M_t/GI/\infty$ queue with a terminating arrival process. *Queueing Systems* 58, 77–104.

[17] Gromoll, H. C., Ph. Robert, B. Zwart. Fluid limits for processor sharing queues with impatience. *Math. Oper. Res.* 33 (2008) 375–402.

[18] Halfin, S. and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–587.

[19] Harrison, J. M. and M. I. Reiman. 1981. Reflected Brownian motion in an orthant. *Ann. Probab.* 9, 302–308.

[20] Iglehart, D. L. 1965. Limit diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* 2, 429–441.

[21] Iglehart, D.L. 1973. Weak convergence of compound stochastic processes. *Stoch. Proc. Appl.* 1, 11–31.

[22] Iglehart, D. L. and W. Whitt. 1970. Multichannel queues in heavy traffic, I. *Adv. Appl. Prob.* 2, 150–177.

[23] Iglehart, D. L. and W. Whitt. 1970. Multichannel queues in heavy traffic, II: sequences, networks and batches. *Adv. Appl. Prob.* 2, 355–369..

[24] Ivanoff, G. and E. Merzbach. 2000. *Set-indexed Martingales.* Chapman and Hall/CRC.

[25] Jacod, J. and A. N. Shiryayev. 1987. *Limit Theorems for Stochastic Processes*, Springer.

[26] Kallianpur, G., V. Perez-Abreu. Stochastic evolution equations driven by nuclear-space-valued martingales. *Appl. Math. Optim.* 17 (1988) 237–272.

[27] Kallianpur, G., V. Perez-Abreu. Weak convergence of solutions of stochastic evolution equations on nuclear spaces. In G. Da Prato and L. Tubaor (eds.), *Stochastic Partial*

Differential Equations and Applications, II, volume 1390, *Lecture Notes in Mathematics*, Springer, 119–131.

[28] Kang, W. and K. Ramanan. 2008. Fluid limits of many-server queues with reneging. *In preparation.*

[29] Kaspi, H. and K. Ramanan. 2007. Law of large numbers limits for many-server queues. *working paper.*

[30] Karatzas, I. and S. Shreve. 1991. *Brownian Motion and Stochastic Calculus*. Springer.

[31] Khoshnevisan, D. 2002. *Multiparameter Processes: An Introduction to Random Fields.* Springer.

[32] Krichagina, E. V. and A. A. Puhalskii. 1997. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* 25, 235–280.

[33] Kruk, L., J. Lehoczky, S. Shreve, S. Yeung. Multiple-input heavy-traffic real-time queues, *Ann. Appl. Prob.* 13 (2003) 54–99.

[34] Kruk, L., J. Lehoczky, S. Shreve, S. Yeung. Earliest-deadline-first service in heavy-traffic acyclic networks. *Ann. Appl. Prob.Z* 14 (2004) 1306–1352.

[35] Kurtz, T.G. and P. Protter. 1991. Weak limit theorems for stochastic integrals and stochastic differential equations. *Ann. Probab.* 19, 1035–1070.

[36] Kurtz, T.G. and P. Protter. 1996. Weak convergence of stochastic integrals and differential equations II: Infinite dimensional case. *Lecture Notes in Mathematics.* Vol. 1627, 197-285.

[37] Liptser, R. Sh. and A. N. Shiryayev. 1986. *Theory of Martingales.* Kluwer.

[38] Louchard, G. 1988. Large finite population queuing systems. Part I: the infinite server model. *Stochastic Models* 4, 373–505.

[39] Mamatov, K.M. 1986. Weak convergence of stochastic integrals with respect to semimartingales. *Russ. Math. Surv.* 41 (5), 155–156.

[40] Mandelbaum, A., W. A. Massey and M. I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* 30, 149–201.

[41] Massey, W. A. and W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13, 183–250.

[42] Mitoma, I. Tightness of probabilities on $C([0,1];\mathcal{S}')$ and $D([0,1];\mathcal{S}')$. *Ann. Probab.* 11 (1983) 989–999.

[43] Neuhaus, G. 1971. On weak convergence of stochastic processes with multidimensional time parameter. *Ann. Math. Statist.* 42, 1285–1295.

[44] Pang, G., R. Talreja and W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys.* 4, 193–267.

[45] Puhalskii, A. A. and J.E. Reed. 2008. On many-server queues in heavy traffic. *working paper.*

[46] Reed, J. E. (2007a) The G/GI/N queue in the Halfin-Whitt regime I: infinite-server queue system equations. working paper, The Stern School, NYU.

[47] Reed, J. E. (2007b) The G/GI/N queue in the Halfin-Whitt regime II: idle-time system equations. working paper, The Stern School, NYU.

[48] Reed, J. and R. Talreja. 2009. Distribution-valued heavy-traffic limits for $GI/GI/\infty$ queues. *In preparation.*

[49] Skorohod, A. V. 1956. Limit theorems for stochastic processes. *Prob. Theory Appl.* 1, 261–290.

[50] Straf, M.L. 1971. Weak convergence of stochastic processes with several parameters. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* 2, 187–221.

[51] Talreja, R. and W. Whitt. 2009. Heavy-traffic limits for waiting times in many-server queues with abandonments. To appear in *Annals of Applied Probability.*

[52] van Der Vaart, A. W. and J. Wellner. 1996. *Weak Convergence and Empirical Processes*, Springer.

[53] Walsh, J.B. 1986. Martingales with a multidimensional parameter and stochastic integrals in the plane. *Lectures in Probability and Statistics.* 329-491, Springer.

[54] Whitt, W. 1982. On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Prob.* 14, 171–190.

[55] Whitt, W. 2002. *Stochastic-Process Limits.* Springer.

[56] Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* 54, 37–54

[57] Whitt, W. 2007. Proofs of the martingale FCLT: a review. *Probability Surveys.* Vol. 4, 268-302.

[58] Wong, E. and M. Zakai, 1974. Martingales and stochastic integrals for processes with a multidimensional parameter. *Z.Wahrscheinlichkeitstheorie verw. Gebiete.* 29, 109-122.

[59] Wong, E. and M. Zakai. 1976. Weak martingales and stochastic integrals in the plane. *Ann. Probability.* 4, 570-586.

[60] Wong, E. and M. Zakai. 1977. An extension of stochastic integrals in the plane. *Ann. Probability.* 5, 770-778.

[61] Zhang, J. 2009. Fluid models of multi-server queues with abandonment. *Submitted.*

APPENDIX

In this appendix we provide additional background information. In §A we review basic properties of the Brownian sheet and the Kiefer process. In §B we give background on two-dimensional stochastic integrals. In §C we state a tightness criterion in the space D_D. In §D, we give the detailed calculation of the covariance of $\hat{B}(t,y)$ in (6.6).

A. The Classical Two-parameter Processes

We now provide background on the classic two-parameter processes: (i) the two-parameter Wiener process or Brownian sheet and (ii) the Kiefer process. We start with the Brownian sheet, because the Kiefer process is naturally defined in terms of it.

Just like Brownian motion (the ordinary Wiener process), the Brownian sheet can be understood as the limit of appropriately normalized partial sums. However, now the setting is a two-dimensional array of i.i.d. random variables $\{X_{i,j} : i \geq 1, j \geq 1\}$, which we take to have mean 0 and variance 1. We then form the partial sums

$$S_{k,l} \equiv \sum_{1 \leq i \leq k, 1 \leq j \leq l} \{X_{i,j}\}, \quad k \geq 1, l \geq 1.$$

We also let $S_{0,l} \equiv 0$ for each l and $S_{k,0} \equiv 0$ for each k. Then we scale time by introducing the two-parameter process

$$\hat{S}_n(s,t) \equiv \frac{S_{\lfloor ns \rfloor, \lfloor nt \rfloor}}{\sqrt{n}}, \quad 0 \leq s \leq 1, 0 \leq t \leq 1.$$

Then, analogous to Donsker's theorem, we have $\hat{S}_n \Rightarrow W$ in $D([0,1], D([0,1], \mathbb{R}))$ (and also in $D([0,1] \times [0,1], \mathbb{R})$), where W is the Brownian sheet.

Here is a direct definition: A stochastic process $W = \{W(t) : t \in \mathbb{R}^2_+\}$ is called a Brownian sheet (two-parameter Wiener process) if

1. $W(R)$ with $R = [x_1, x_2) \times [y_1, y_2)$ has a normal distribution with mean 0 and variance $(x_2 - x_1)(y_2 - y_1)$, where

$$W(R) \quad \equiv \quad \Delta_W(x_1, x_2, y_1, y_2) \equiv W(x_2, y_2) - W(x_1, y_2) - W(x_2, y_1) + W(x_1, y_1),$$

2. $W(0, y) = W(x, 0) = 0$ for all $0 \leq x, y < \infty$,

3. W has independent increments, i.e., $W(R_1), W(R_2), ..., W(R_n), n \geq 2$ are independent random variables if $R_1, R_2, ..., R_n$ are disjoint,

4. The sample path function $W(t;\omega)$ is continuous in t with probability one.

Note that the covariance function of a Brownian sheet $W(t)$ is

$$Cov(W(t_1), W(t_2)) = (x_1 \wedge x_2)(y_1 \wedge y_2),$$

where $t_1 \equiv (x_1, y_1)$ and $t_2 \equiv (x_2, y_2)$. For any fixed $0 < x_0 < \infty$, the process $\{x_0^{-1/2} W(x_0, y) : y \geq 0\}$ is an ordinary Brownian motion, as is $\{y_0^{-1/2} W(x, y_0) : x \geq 0\}$. The processes $\{xW(1/x, y) : x > 0, y \geq 0\}$, $\{yW(x, 1/y) : x \geq 0, y > 0\}$ and $\{xyW(1/x, 1/y) : x > 0, y > 0\}$ are also Brownian motion processes.

To understand the Kiefer process, we should recall the Brownian bridge, often denoted by B_0. It is Brownian motion on $[0,1]$, conditioned to be 0 at time 1. It can be defined in terms of a Brownian motion B by $B_0(t) \equiv B(t) - tB(1)$, $0 \leq t \leq 1$. It can be understood as the limit of the properly scaled usual one-dimensional partial sums $S_n \equiv X_1 + \cdots X_n$, conditional on $S_n = 0$. The Brownian bridge arises as the limit for the properly scaled and centered empirical distribution function of uniform random variables; see §2.2 of [55]. The standard Kiefer process can be understood as the limit of the sequential empirical process in (2.3) for the special case of the uniform distribution, as stated in (2.5). It is easy to see that we must have $\hat{U}_n(t, 1) = 0$ for all t, $0 \leq t \leq 1$.

Analogous to the Brownian bridge, the Kiefer process $\{K(x, y); 0 \leq x < \infty, 0 \leq y \leq 1\}$ can be defined directly in terms of the Brownian sheet by

$$K(x, y) = W(x, y) - yW(x, 1),$$

where $W(x, y)$ is a two parameter Wiener process. The Kiefer process can be understood to be the Brownian sheet conditioned to be 0 when $y = 1$ (tied down at the upper boundary).

The Kiefer process K has the following properties:

1. $E[K(x, y)] = 0$ and the covariance function of $K(x, y)$ is

$$E[K(x_1, y_1)K(x_2, y_2)] = (y_1 \wedge y_2 - y_1 y_2)(x_1 \wedge x_2);$$

2. for any $0 < y_0 < 1$, $\left\{W(x) = \frac{K(x,y_0)}{\sqrt{y_0(1-y_0)}} : x \geq 0\right\}$ is a Brownian motion;

3. for any $x_0 > 0$, $\left\{B(y) = x_0^{-1} K(x_0, y) : 0 \leq y \leq 1\right\}$ is a Brownian Bridge;

4. the sample path functions of $K(x, y)$ are continuous with probability one;

5. $W(x,y) = (y+1)K\left(x, \frac{y}{y+1}\right)$, for $x \geq 0$, and $y \geq 0$;

6. $K(x,y) = (1-y)W\left(x, \frac{y}{1-y}\right)$, for $0 \leq y < 1$, and $x \geq 0$.

Proposition A.1. *A Kiefer process $\{K(t,x); t \geq 0, 0 \leq x \leq 1\}$ is the unique solution to the following stochastic integral equation:*

$$K(t,x) = -\int_0^x \frac{K(t,y)}{1-y}dy + W(t,x), \quad \textit{for} \quad t \geq 0 \quad \textit{and} \quad 0 \leq x \leq 1, \tag{A.1}$$

where $\{W(t,x) : t \geq 0, 0 \leq x \leq 1\}$ is a Brownian sheet.

Proof. For each fixed $t \geq 0$, consider the following linear SDE,

$$dK(t,x) = -\frac{K(t,x)}{1-x}dx + dW(t,x), \quad \text{for} \quad 0 \leq x \leq 1,$$

with $K(t,0) = 0$. Since for each t fixed, $\{t^{-1/2}W(t,x) : 0 \leq x \leq 1\}$ is a standard Brownian motion, we can use the classical one-dimensional linear SDE solution result (§5.6, [30]) to obtain (including uniqueness)

$$K(t,x) = (1-x)\int_0^x \frac{1}{1-y}dW(t,y), \quad \text{for} \quad 0 \leq x \leq 1. \tag{A.2}$$

By Ito's integration theory, for each fixed $t \geq 0$, the process $\{K(t,x) : 0 \leq x \leq 1\}$ is a continuous, square integrable martingale with predictable quadratic variation

$$\langle K(t,\cdot)\rangle(x) = (1-x)^2 t\int_0^x \frac{1}{(1-y)^2}dy = (1-x)^2 t\frac{x}{1-x} = (1-x)xt,$$

for $0 \leq x \leq 1$. By the representation of continuous martingales as a time-changed Brownian motion (Theorem 3.4.6, [30]), we can write $K(t,x)$ as $K(t,x) = B(\langle K(t,\cdot)\rangle(x))$ for $0 \leq x \leq 1$ and a standard Brownian motion B, for each fixed $t \geq 0$. So, for each fixed $t \geq 0$, the process $\{K(t,x) : 0 \leq x \leq 1\}$ is a continuous Gaussian process with mean 0 and covariance

$$E[K(t,x)K(t,y)] = (1-x)(1-y)t\int_0^{x\wedge y} \frac{1}{(1-u)^2}du = (x \wedge y - xy)t.$$

This implies that, for each fixed $t \geq 0$, $\{t^{-1}K(t,x) : 0 \leq x \leq 1\}$ is a Brownian bridge. Moreover, we can also obtain from (A.2) that

$$E[K(t_1,x)K(t_2,y)] = (x \wedge y - xy)(t_1 \wedge t_2).$$

Therefore, the proof is complete. ■

The empirical distribution process U_n in (2.5) has the following semimartingale decomposition. (See Chapter IX, Jacod and Shiryaev [25]. We provide a proof here for completeness since it is not in [25].)

Proposition A.2. *The empirical distribution process* $U_n = (U_n(t,x), t \geq 0, 0 \leq x \leq 1)$ *has the semimartingale decomposition*

$$U_n(t,x) = -\int_0^x \frac{U_n(t,y)}{1-y} dy + U_{n,0}(t,x), \tag{A.3}$$

where

$$U_{n,0}(t,x) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \Big(\mathbf{1}(\zeta_i \leq x) - \int_0^{x \wedge \zeta_i} \frac{1}{1-y} dy \Big), \tag{A.4}$$

is a square integrable martingale with respect to the filtration $\mathbf{F}_n = \{\bigvee_{i \leq \lfloor nt \rfloor} \mathcal{F}^i(x) : t \geq 0\}$ *for each fixed* $x \geq 0$*, and* $\mathcal{F}^i(x) = \sigma(\mathbf{1}(\zeta_i \leq y), 0 \leq y \leq x) \vee \mathcal{N}$ *for all* $x \in [0,1]$*,* ζ_i*'s are i.i.d. with uniform distribution on* $[0,1]$.

Proof. For $t \geq 0$ and $0 \leq x \leq 1$,

$$\begin{aligned}
\int_0^x \frac{U_n(t,y)}{1-y} dy &= \int_0^x \frac{1}{1-y} \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}(\zeta_i \leq y) - y) dy \\
&= \frac{1}{\sqrt{n}} \int_0^x \frac{1}{1-y} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}(\zeta_i \leq y) dy - \frac{1}{\sqrt{n}} \int_0^x \frac{y}{1-y} \lfloor nt \rfloor dy \\
&= \frac{1}{\sqrt{n}} \int_0^x \frac{1}{1-y} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}(\zeta_i \leq y) dy + \frac{\lfloor nt \rfloor}{\sqrt{n}} x - \frac{1}{\sqrt{n}} \int_0^x \frac{1}{1-y} \lfloor nt \rfloor dy \\
&= \frac{\lfloor nt \rfloor}{\sqrt{n}} x + \frac{1}{\sqrt{n}} \int_0^x \frac{1}{1-y} \Big(\sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}(\zeta_i \leq y) - \lfloor nt \rfloor \Big) dy \\
&= \frac{\lfloor nt \rfloor}{\sqrt{n}} x - \frac{1}{\sqrt{n}} \int_0^x \frac{1}{1-y} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}(y \leq \zeta_i) dy \\
&= \frac{\lfloor nt \rfloor}{\sqrt{n}} x - \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \int_0^{x \wedge \zeta_i} \frac{1}{1-y} dy \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}(\zeta_i \leq x) - x) + \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \Big(\mathbf{1}(\zeta_i \leq x) - \int_0^{x \wedge \zeta_i} \frac{1}{1-y} dy \Big).
\end{aligned}$$

Hence, we obtain the decomposition of U_n in (A.3). It is easy to check that, for each fixed x, $U_{n,0}$ in (A.4) is a square integrable martingale with respect to the filtration $\mathbf{F}_n$, and the first term of U_n in (A.3) has finite variation. ∎

B. Two-Parameter Stochastic Integrals

We first review the definition of filtrations for two-parameter processes. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $R(s,t)$ be the rectangle $\{(u,v) : 0 \leq u \leq s, 0 \leq v \leq t\}$. Write

$(u,v) \preceq (s,t)$ if $u \leq s$ and $v \leq t$. A collection $\mathbf{F} = \{\mathcal{F}_{(s,t)} : (s,t) \in [0,T] \times [0,T]\}$ of sub-σ-fields of $\mathcal{F}$ is said to be a filtration if (i) $(u,v) \preceq (s,t)$ implies that $\mathcal{F}_{(u,v)} \subseteq \mathcal{F}_{(s,t)}$ and (ii) $\mathcal{F}_{(0,0)}$ contains all the null sets of $\mathcal{F}$, and (iii) $\mathcal{F}_{(s,t)} = \bigcap \mathcal{F}_{(u,v)}$ for $u > s$ and $v > t$. For a stochastic process X that takes value in D_2, the natural filtration it generates is given by $\mathcal{F}_{(s,t)} = \sigma\{X(u,v) : u \leq s, v \leq t\}$ for each $(s,t) \in [0,T] \times [0,T]$.

In the literature, several types of stochastic integrals with respect to two-parameter processes have been defined. The first type is to generalize the definition of Ito's integral directly, i.e., the integral

$$I_{1,M}(\phi) \equiv \int_{[0,s]\times[0,t]} \phi(u,v) dM_{(u,v)}, \tag{B.1}$$

where $(M(u,v), (u,v) \in \mathbb{R}^2_+)$ is a two-parameter (strong, weak)-martingale, ϕ is a class of previsible and square integrable processes, and $R(s,t)$ is a rectangle with $(s,t) \in \mathbb{R}^2_+$. This type of integral was first defined for two-parameter Brownian sheets by Cairoli [5] (see also [53]) and generalized to n-parameter Brownian sheets by Wong and Zakai [58]. It was generalized to general martingales by Cairoli and Walsh [6]. Even more generalization appears in Wong and Zakai [60].

Two-parameter martingales were defined in Cairoli and Walsh [6], Wong and Zakai [58], [59] and Walsh [53]. For more general set-indexed martingales, see Ivanoff and Merzbach [24] and Khoshnevisan [31]. Since we are only using properties in terms of Brownian sheet in this paper, the following definition of two-parameter martingales will suffice for us. For definitions of weak and strong multi-parameter martingales, we refer to [24] and [31].

A process $\{M(s,t), (s,t) \in [0,T] \times [0,T]\}$ is a martingale if $M(s,t) \in \mathcal{F}_{(s,t)}$ is integrable and for each $(s,t) \preceq (s',t')$, $E[M(s',t')|\mathcal{F}_{(s,t)}] = M(s,t)$.

The following processes are martingales related to two-parameter Brownian sheets W:

$$W, \quad \left\{W(s,t)^2 - st : (s,t) \in \mathbb{R}^2_+\right\}, \quad \left\{\exp\left(\lambda W(s,t) - \frac{1}{2}\lambda^2 st\right) : (s,t) \in \mathbb{R}^2_+\right\}.$$

When M is a two-parameter Brownian sheet, the integral $I_{1,M}(\cdot)$ in (B.1) inherits most properties of Ito's integral, e.g.:

- **Linearity**: $I_{1,M}(\alpha\phi + \beta\psi) = \alpha I_{1,M}(\phi) + \beta I_{1,M}(\psi), \quad \alpha, \beta \in \mathbb{R}$,
- **Isometry**: $E[I_{1,M}(\phi) I_{1,M}(\psi)] = \int_{[0,s]\times[0,t]} \phi(u,v)\psi(u,v) du dv$,
- **Martingale**: $I_{1,M}(\phi)$ is a martingale.

The second and third types of integral are introduced in [58] and [59], respectively, which are not applied in this paper. We refer interested readers to those papers.

C. Tightness in D_D

The following tightness criteria come from Theorem 3.8.6 of Ethier and Kurtz [12], adapted to the space $D([0,T],D)$. For a review of tightness criteria for processes in the space D, see [57].

Theorem C.1. *A sequence of stochastic processes $\{X_n : n \geq 1\}$ in $D([0,T],D)$ is tight if and only if*

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

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

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 $\{\tau_n : n \geq 1\}$ where for each n, τ_n is a stopping time with respect to the natural filtration $\mathbf{F}_n = \{\mathcal{F}_n(t), t \in [0,T]\}$ where $\mathcal{F}_n(t) = \sigma\{X_n(s,\cdot) : 0 \leq s \leq t\}$, there exists a constant $\beta > 0$ such that

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

or

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

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

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

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

Remark. The following condition is sufficient, but not necessary, for condition (*ii*) in Theorem C.1:

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

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

D. The Covariance of $\hat{B}$

We calculate the covariance of $\hat{B}(t,y)$ in (6.6) using the properties of stochastic integrals with respect to the two-parameter Brownian sheet of the first type. For $t < t'$ and $y < y'$,

$$\begin{aligned}
& E[\hat{B}(t,y)\hat{B}(t',y')] \\
= \; & E\Big[\Big(e^{-\mu y}\sqrt{\lambda c_a^2}B(t) + \sqrt{\bar{q}^{i,t}}B_b(1 - e^{-\mu(t+y)}) + \int_0^t \int_0^\infty \mathbf{1}(s + x \leq t + y)dW(\lambda s, 1 - e^{-\mu x})\Big) \\
& \Big(e^{-\mu y'}\sqrt{\lambda c_a^2}B(t') + \sqrt{\bar{q}^{i,t}}B_b(1 - e^{-\mu(t'+y')}) + \int_0^{t'} \int_0^\infty \mathbf{1}(s + x \leq t' + y')dW(\lambda s, 1 - e^{-\mu x})\Big)\Big] \\
= \; & \lambda c_a^2 e^{-\mu(y+y')}E\Big[B(t)B(t')\Big] + \bar{q}^{i,t}E\Big[B_b(1 - e^{-\mu(t+y)})B_b(1 - e^{-\mu(t'+y')})\Big] \\
& + E\Big[\int_0^t \int_0^\infty \mathbf{1}(s + x \leq t + y)dW(\lambda s, 1 - e^{-\mu x})\int_0^{t'} \int_0^\infty \mathbf{1}(s + x \leq t' + y')dW(\lambda s, 1 - e^{-\mu x})\Big] \\
= \; & \lambda c_a^2 e^{-\mu(y+y')}t + \bar{q}^{i,t}(1 - e^{-\mu(t+y)}) + \int_0^t \int_0^\infty \mathbf{1}(s + x \leq t + y)d(1 - e^{-\mu x})d(\lambda s) \\
= \; & \lambda c_a^2 e^{-\mu(y+y')}t + \bar{q}^{i,t}(1 - e^{-\mu(t+y)}) + \lambda \int_0^t (1 - e^{-\mu(t+y-s)})ds \\
= \; & c_a^2 e^{-\mu(y+y')}t + \bar{q}^{i,t}(1 - e^{-\mu(t+y)}) + \lambda t - \frac{\lambda}{\mu}e^{-\mu y} + \frac{\lambda}{\mu}e^{-\mu(t+y)} \\
= \; & (\lambda + \lambda c_a^2 e^{-\mu(y+y')})t + \bar{q}^{i,t} - \frac{\lambda}{\mu}e^{-\mu y} + \left(\frac{\lambda}{\mu} - \bar{q}^{i,t}\right)e^{-\mu(t+y)}.
\end{aligned}$$

The second equality holds because of the mutual independence of B, B_b and W and the third equality holds because of the isometry property of the stochastic integral with respect to the two-parameter Brownian sheet of the first type.