Probability Surveys
Vol. 4 (2007) 193–267
ISSN: 1549-5787
DOI: 10.1214/06-PS091

Martingale proofs of many-server heavy-traffic limits for Markovian queues

Guodong Pang, Rishi Talreja and Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699; e-mail: {gp2224, rt2146, ww2040}@columbia.edu

Abstract: This is an expository review paper illustrating the "martingale method" for proving many-server heavy-traffic stochastic-process limits for queueing models, supporting diffusion-process approximations. Careful treatment is given to an elementary model – the classical infinite-server model $M/M/\infty$, but models with finitely many servers and customer abandonment are also treated. The Markovian stochastic process representing the number of customers in the system is constructed in terms of rate-1 Poisson processes in two ways: (i) through random time changes and (ii) through random thinnings. Associated martingale representations are obtained for these constructions by applying, respectively: (i) optional stopping theorems where the random time changes are the stopping times and (ii) the integration theorem associated with random thinning of a counting process. Convergence to the diffusion process limit for the appropriate sequence of scaled queueing processes is obtained by applying the continuous mapping theorem. A key FCLT and a key FWLLN in this framework are established both with and without applying martingales.

AMS 2000 subject classifications: Primary 60F17, 60K25.
Keywords and phrases: multiple-server queues, many-server heavy-traffic limits for queues, diffusion approximations, martingales, functional central limit theorems.

Received December 2006.

Contents

1. Introduction

The purpose of this paper is to illustrate how to do martingale proofs of many-server heavy-traffic limit theorems for queueing models, as in Krichagina and Puhalskii [37] and Puhalskii and Reiman [53]. Even though the method is remarkably effective, it is somewhat complicated. Thus it is helpful to see how the argument applies to simple models before considering more elaborate models. For the more elementary Markovian models considered here, these many-server heavy-traffic limits were originally established by other methods, but new methods of proof have been developed. For the more elementary models, we show how key steps in the proof - a FCLT ((59)) and a FWLLN (Lemma 4.3) - can be done without martingales as well as with martingales. For the argument without martingales, we follow Mandelbaum and Pats [47, 48], but without their level of generality and without discussing strong approximations.

Even though we focus on elementary Markov models here, we are motivated by the desire to treat more complicated models, such as the non-Markovian $GI/GI/n$ models in Krichagina and Puhalskii [37], Puhalskii and Reiman [53], Reed [55] and Kaspi and Ramanan [34], and network generalizations of these, such as non-Markovian generalizations of the network models considered by Dai and Tezcan [14, 15, 16], Gurvich and Whitt [23, 24, 25] and references cited there. Thus we want to do more than achieve a quick proof for the simple models, which need not rely so much on martingales; we want to illustrate martingale methods that may prove useful for more complicated models.

1.1. The Classical Markovian Infinite-Server Model

We start by focusing on what is a candidate to be the easiest queueing model – the classic $M/M/\infty$ model – which has a Poisson arrival process (the first M), i.i.d. exponential service times (the second M), independent of the arrival process, and infinitely many servers. Let the arrival rate be λ and let the individual mean service time be $1/\mu$.

Afterwards, in §7.1, we treat the associated $M/M/n/\infty + M$ (Erlang A or Palm) model, which has n servers, unlimited waiting room, the first-come first-served (FCFS) service discipline and Markovian customer abandonment (the final $+M$); customers abandon (leave) if they have to spend too long waiting in queue. The successive customer times to abandon are assumed to be i.i.d. exponential random variables, independent of the history up to that time. That assumption is often reasonable with invisible queues, as in telephone call centers. Limits for the associated $M/M/n/\infty$ (Erlang C) model (without customer abandonment) are obtained as a corollary, by simply letting the abandonment rate be zero.

For the initial $M/M/\infty$ model, let $Q(t)$ be the number of customers in the system at time t, which coincides with the number of busy servers at time t. It is well known that $Q \equiv \{Q(t) : t \geq 0\}$ is a birth-and-death stochastic process and that $Q(t) \Rightarrow Q(\infty)$ as $t \to \infty$, where $\Rightarrow$ denotes convergence in distribution, provided that $P(Q(0) < \infty) = 1$, and $Q(\infty)$ has a Poisson distribution with mean $E[Q(\infty)] = \lambda/\mu$.

We are interested in heavy-traffic limits in which the arrival rate is allowed to increase. Accordingly, we consider a sequence of models indexed by n and let the arrival rate in model n be

$$\lambda_n \equiv n\mu, \quad n \geq 1 \; . \tag{1}$$

Let $Q_n(t)$ be the number of customers in the system at time t in model n. By the observation above, $Q_n(\infty)$ has a Poisson distribution with mean $E[Q_n(\infty)] = \lambda_n/\mu = n$. Since the Poisson distribution approaches the normal distribution as the mean increases, we know that

$$\frac{Q_n(\infty) - n}{\sqrt{n}} \Rightarrow N(0,1) \quad \text{as} \quad n \to \infty \ , \tag{2}$$

where $N(0,1)$ denotes a standard normal random variables (with mean 0 and variance 1).

However, we want to establish a limit for the entire stochastic process $\{Q_n(t) : t \ge 0\}$ as $n \to \infty$. For that purpose, we consider the scaled processes $X_n \equiv \{X_n(t) : t \ge 0\}$ defined by

$$X_n(t) \equiv \frac{Q_n(t) - n}{\sqrt{n}}, \quad t \ge 0 \ . \tag{3}$$

To establish the stochastic-process limit in (3), we have to be careful about the initial conditions. We will assume that $X_n(0) \Rightarrow X(0)$ as $n \to \infty$. In addition, we assume that the random initial number of busy servers, $Q_n(0)$, is independent of the arrival process and the service times. Since the service-time distribution is exponential, the remaining service times of those customers initially in service have independent exponential distributions because of the lack-of-memory property of the exponential distribution.

The heavy-traffic limit theorem asserts that the sequence of stochastic processes $\{X_n : n \ge 1\}$ converges in distribution in the function space $D \equiv D([0,\infty),\mathbb{R})$ to the Ornstein-Uhlenbeck (OU) diffusion process as $n \to \infty$, provided that appropriate initial conditions are in force; see Billingsley [8] and Whitt [69] for background on such stochastic-process limits.

Here is the theorem for the basic $M/M/\infty$ model:

Theorem 1.1 (heavy-traffic limit in D for the $M/M/\infty$ model) *Consider the sequence of $M/M/\infty$ models defined above. If*

$$X_n(0) \Rightarrow X(0) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty \ , \tag{4}$$

then

$$X_n \Rightarrow X \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ ,$$

where X is the OU diffusion process with infinitesimal mean $m(x) = -\mu x$ and infinitesimal variance $\sigma^2(x) = 2\mu$. Alternatively, X satisfies the stochastic integral equation

$$X(t) = X(0) + \sqrt{2\mu}B(t) - \mu \int_0^t X(s)\, ds \ , \quad t \ge 0 \ , \tag{5}$$

where $B \equiv \{B(t) : t \ge 0\}$ is a standard Brownian motion. Equivalently, X satisfies the stochastic differential equation (SDE)

$$dX(t) = -\mu X(t)dt + \sqrt{2\mu}dB(t), \quad t \ge 0 \ . \tag{6}$$

Much of this paper is devoted to *four proofs* of Theorem 1.1. It is possible to base the proof on the martingale functional central limit theorem (FCLT), as given in §7.1 of Ethier and Kurtz [19] and here in §8, as we will show in §9.2, but it is not necessary to do so. Instead, it can be based on the classic FCLT for the Poisson process, which is an easy consequence of Donsker's FCLT for sums of i.i.d. random variables, and the continuous mapping theorem. Nevertheless, the martingale structure can still play an important role. With that approach, the martingale structure can be used to establish stochastic boundedness of the scaled queueing processes, which we show implies required fluid limits or functional weak laws of large numbers (FWLLN's) for random-time-change stochastic processes, needed for an application of the continuous mapping theorem with the composition map. Alternatively (the third method of proof), the fluid limit can be established by combining the same continuous mapping with the strong law of large numbers (SLLN) for the Poisson process.

It is also necessary to understand the characterization of the limiting diffusion process via (5) and (6). That is aided by the general theory of stochastic integration, which can be considered part of the martingale theory [19, 58].

1.2. The QED Many-Server Heavy-Traffic Limiting Regime

We will also establish many-server heavy-traffic limits for Markovian models with finitely many servers, where the number n of servers goes to infinity along with the arrival rate in the limit. We will consider the sequence of models in the quality-and-efficiency (QED) many-server heavy-traffic limiting regime, which is defined by the condition

$$\frac{n\mu - \lambda_n}{\sqrt{n}} \to \beta\mu \quad \text{as} \quad n \to \infty \ . \tag{7}$$

This limit in which the arrival rate and number of servers increase together according to (7) is just the right way so that the probability of delay converges to a nondegenerate limit (strictly between 0 and 1); see Halfin and Whitt [26].

We will also allow finite waiting rooms of size m_n, where the waiting rooms grow at rate $\sqrt{n}$ as $n \to \infty$, i.e., so that

$$\frac{m_n}{\sqrt{n}} \to \kappa \geq 0 \quad \text{as} \quad n \to \infty \ . \tag{8}$$

With the spatial scaling by $\sqrt{n}$, as in (3), this scaling in (8) is just right to produce a reflecting upper barrier at κ in the limit process.

In addition, we allow Markovian abandonment, with each waiting customer abandoning at rate θ. We let the individual service rate μ and individual abandonment rate θ be fixed, independent of n. These modifications produce a sequence of $M/M/n/m_n + M$ models (with $+M$ indicating abandonment). The special case of the Erlang A (abandonment), B (blocking or loss) and C (delay) models are obtained, respectively, by (i) letting $m_n = \infty$, (ii) letting $m_n = 0$, in which case the $+M$ plays no role, and (iii) letting $m_n = \infty$ and $\theta = 0$. So all the basic many-server Markovian queueing models are covered.

Here is the corresponding theorem for the $M/M/n/m_n + M$ model.

Theorem 1.2 (heavy-traffic limit in D for the $M/M/n/m_n + M$ model) *Consider the sequence of $M/M/n/m_n + M$ models defined above, with the scaling in* (7) *and* (8). *Let X_n be as defined in* (3). *If $X_n(0) \Rightarrow X(0)$ in $\mathbb{R}$ as $n \to \infty$, then $X_n \Rightarrow X$ in D as $n \to \infty$, where the limit X is the diffusion process with infinitesimal mean $m(x) = -\beta\mu - \mu x$ for $x < 0$ and $m(x) = -\beta\mu - \theta x$ for $x > 0$, infinitesimal variance $\sigma^2(x) = 2\mu$ and reflecting upper barrier at κ. Alternatively, the limit process X is the unique $(-\infty, \kappa]$-valued process satisfying the stochastic integral equation*

$$X(t) = X(0) - \beta\mu t + \sqrt{2\mu}B(t) - \int_0^t [\mu(X(s) \wedge 0) + \theta(X(s) \vee 0)]\, ds - U(t) \quad (9)$$

for $t \geq 0$, where $B \equiv \{B(t) : t \geq 0\}$ is a standard Brownian motion and U is the unique nondecreasing nonnegative process in D such that (9) *holds and*

$$\int_0^\infty 1_{\{X(t)<\kappa\}}\, dU(t) = 0 \ . \quad (10)$$

1.3. Literature Review

A landmark in the application of martingales to queues is the book by Brémaud [12]; contributions over the previous decade are described there. As reflected by the references cited in Krichagina and Puhalskii [37] and Puhalskii and Reiman [53], and as substantiated by Puhalskii in personal communication, Liptser was instrumental in developing the martingale method to prove limit theorems for queueing models, leading to diffusion approximations; e.g., see Kogan, Liptser and Smorodinskii [36] and §10.4 of Liptser and Shiryaev [44].

The specific $M/M/\infty$ result in Theorem 1.1 was first established by Iglehart [29]; see Borovkov [9, 10], Whitt [68], Glynn and Whitt [22], Massey and Whitt [50], Mandelbaum and Pats [47, 48], Mandelbaum, Massey and Reiman [46] and Krichagina and Puhalskii [37] for further discussion and extensions. A closely related textbook treatment appears in Chapter 6 of Robert [57]. Iglehart applied a different argument; in particular, he applied Stone's [62] theorem, which shows for birth-and-death processes that it suffices to have the infinitesimal means and infinitesimal variances converge, plus other regularity conditions; see Garnett et al. [21] and Whitt [71] for recent uses of that same technique. Iglehart directly assumed finitely many servers and let the number of servers increase rapidly with the arrival rate; the number of servers increases so rapidly that it is tantamount to having infinitely many servers. The case of infinitely many servers can be treated by a minor modification of the same argument.

The corresponding QED finite-server result in Theorem 1.2, in which the arrival rate and number of servers increase together according to (7), which is just the right way so that the probability of delay converges to a nondegenerate limit (strictly between 0 and 1), was considered by Halfin and Whitt [26] for the case of infinite waiting room and no abandonment. There have been many subsequent results; e.g., see Mandelbaum and Pats [47, 48], Srikant and Whitt [61], Mandelbaum et al. [46], Puhalskii and Reiman [53], Garnett et al. [21], Borst et al. [11], Jelenković et al. [31], Whitt [72], Mandelbaum and Zeltyn [49] and Reed [55] for further discussion and extensions.

Puhalskii and Reiman [53] apply the martingale argument to establish heavy-traffic limits in the QED regime. They establish many-server heavy-traffic limits for the $GI/PH/n/\infty$ model, having renewal arrival process (the GI), phase-type service-time distributions (the PH), n servers, unlimited waiting space, and the first-come first-served (FCFS) service discipline. They focus on the number of customers in each phase of service, which leads to convergence to a multi-dimensional diffusion process. One of our four proofs is essentially their argument.

Whitt [72] applied the same martingale argument to treat the $G/M/n/m_n + M$ model, having a general stationary point process for an arrival process (the G), i.i.d. exponential service times, n servers, m_n waiting spaces, the FCFS service discipline and customer abandonment with i.i.d exponential times to abandon; see Theorem 3.1 in [72]. Whitt [72] is primarily devoted to a heavy-traffic limit for the $G/H_2^*/n/m_n$ model, having a special H_2^* service-time distribution (a mixture of an exponential and a point mass at 0), by a different argument, but there is a short treatment of the $G/M/n/m_n + M$ model by the martingale argument, following Puhalskii and Reiman [53] quite closely. The martingale proof is briefly outlined in §5 there. The extension to general arrival processes is perhaps the most interesting contribution there, but that generalization can also be achieved in other ways.

For our proof without martingales, we follow Mandelbaum and Pats [47, 48], but without using strong approximations as they do. Subsequent related work has established similar asymptotic results for Markovian service networks with multiple stations, each with multiple servers, and possibly multiple customer classes; e.g., see Armony [1], Armony et al. [2, 3], Atar [4], Atar et al. [5, 6], Dai and Tezcan [14, 15, 16], Gurvich and Whitt [23, 24, 25], Harrison and Zeevi [27], and Tezcan [63]. These complex stochastic networks have important applications to telephone call centers; see Gans et al. [20].

1.4. Organization

The rest of this paper is organized as follows: We start in §2 by constructing the Markovian stochastic process Q representing the number of customers in the system in terms of rate-1 Poisson processes. We do this in two ways: (i) through random time changes and (ii) through random thinnings. We also give the construction in terms of arrival and service times used by Krichagina and Puhalskii [37] to treat the $G/GI/\infty$ model.

Section §3 is devoted to martingales. After reviewing basic martingale notions, we construct martingale representations associated with the different constructions. We justify the first two representations by applying, respectively: (i) optional stopping theorems where the random time changes are the stopping times and (ii) the integration theorem associated with random thinning of a counting process. The two resulting integral representations are very similar. These integral representations are summarized in Theorems 3.4 and 3.5. In §3 we also present two other martingale representations, one based on constructing counting processes associated with a continuous-time Markov chain, exploiting the infinitesimal generator, and the other - for $G/GI/\infty$ - based on the sequential empirical process (see (44)).

Section 4 is devoted to the main steps of the proof of Theorem 1.1 using the first two martingale representations. In §4.1 we show that the integral representation has a unique solution and constitutes a continuous function mapping $\mathbb{R} \times D$ into D. In order to establish measurability, we establish continuity in the Skorohod [60] J_1 topology as well as the topology of uniform convergence over bounded intervals. As a consequence of the continuity, it suffices to prove FCLT's for the scaled martingales in these integral representations. For the first martingale representation, the scaled martingales themselves are random time changes of the scaled Poisson process. In §4.2 we show that a FCLT for the martingales based on this first representation, and thus the scaled queueing processes, can be obtained by applying the classical FCLT for the Poisson process and the continuous mapping theorem with the composition map.

To carry out the continuous-mapping argument above, we also need to establish the required fluid limit or functional weak law of large numbers (FWLLN) for the random time changes, needed in the application of the CMT with the composition map. In §4.3, following Mandelbaum and Pats [47, 48], we show that this fluid-limit step can be established by applying the strong law of large numbers (SLLN) for the Poisson process with the same continuous mapping determined by the integral representation. If we do not use martingales, then we observe that it is easy to extend the stochastic-process limit to general arrival processes satisfying a FCLT.

But we can also use martingales to establish the FCLT and the FWLLN. For the FWLLN, the martingale structure can be exploited via the Lenglart-Rebolledo inequality to prove stochastic boundedness, first for the scaled martingales and then for the scaled queueing processes, which in turn can be used to to establish the required FWLLN for the scaled random time changes - Lemma 4.2.

Since this martingale proof of the fluid limit relies on stochastic boundedness, which is related to tightness, we present background on these two important concepts in §5. For the proof of Theorem 1.1, we conclude there that it suffices to show that the predictable quadratic variations of the square-integrable martingales are stochastically bounded in $\mathbb{R}$ in order to have the sequence of scaled queueing processes $\{X_n\}$ be stochastically bounded in D.

We complete the proof of Theorem 1.1 in §6. In Lemma 5.9 and §6.1 we show that stochastic boundedness of $\{X_n\}$ in D implies the desired fluid limit or FWLLN needed for the scaled random-time-change processes needed in an application of the continuous mapping with the composition function. In §6.2 we complete the proof by showing that the predictable quadratic variation processes of the martingales are indeed stochastically bounded in $\mathbb{R}$. In §6.3 we show that it is possible to remove a moment condition imposed on $Q_n(0)$, the initial random number of customers in the system, in the martingale representation; in particular, we show that it is not necessary to assume that $E[Q_n(0)] < \infty$. Finally, in §6.4 we state the $G/GI/\infty$ limit in Krichagina and Puhalskii [37] and show that the special case $M/M/\infty$ is consistent with Theorem 1.1.

In §7 we discuss stochastic-process limits for other queueing models. In §7.1 we present a martingale proof of the corresponding many-server heavy-traffic limit for the $M/M/n/\infty + M$ model. Corresponding results hold for the model without customer abandonment by setting the abandonment rate to zero. The proof is nearly the same as for the $M/M/\infty$ model. The only significant difference is the use of the optional stopping theorem for multiple stopping times with respect to multiparameter martingales, as in Kurtz [40], §§2.8 and 6.2 of Ethier and Kurtz [19] and §12 of Mandelbaum and Pats [48]. We discuss the extension to cover finite waiting rooms in §7.2 and non-Markovian arrival processes in §7.3.

We state a version of the martingale FCLT from p 339 of Ethier and Kurtz [19] in §8. In a companion paper, Whitt [73], we review the proof, elaborating on the proof in Ethier and Kurtz [19] and presenting alternative arguments, primarily based on Jacod and Shiryayev [30]. We present a more extensive review of tightness there. In §9 we show how the martingale FCLT implies both the FCLT for a Poisson process and the required FCLT for the scaled martingales arising in the second and third martingale representations.

2. Sample-Path Constructions

We start by making direct sample-path constructions of the stochastic process Q representing the number of customers in the system in terms of independent Poisson processes. We show how this can be done in two different ways. Afterwards, we present a different construction based on arrival and service times, which only exploits the fact that the service times are mutually independent and independent of the arrival process and the initial number of customers in the system (with appropriate common distribution assumptions).

2.1. Random Time Change of Unit-Rate Poisson Processes

We first represent arrivals and departures as random time changes of independent unit-rate Poisson processes. For that purpose, let $A \equiv \{A(t) : t \geq 0\}$ and $S \equiv \{S(t) : t \geq 0\}$ be two independent Poisson processes, each with rate (intensity) 1. We use the process A to generate arrivals and the process S to generate service completions, and thus departures. Let the initial number of busy servers be $Q(0)$. We assume that $Q(0)$ is a proper random variable independent of the two Poisson processes A and S.

The arrival process is simple. We obtain the originally-specified arrival process with rate λ by simply scaling time in the rate-1 Poisson process A; i.e., we use

$A_\lambda(t) \equiv A(\lambda t)$ for $t \geq 0$. It is elementary to see that the stochastic process A_λ is a Poisson process with rate λ.

The treatment of service completions is more complicated. Let $D(t)$ be the number of departures (service completions) in the interval $[0, t]$. We construct this process in terms of S by setting

$$D(t) \equiv S\left(\mu \int_0^t Q(s)\, ds\right), \quad t \geq 0 \ , \tag{11}$$

but formula (11) is more complicated than it looks. The complication is that $Q(t)$ appearing as part of the argument inside S necessarily depends on the history $\{Q(s) : 0 \leq s < t\}$, which in turn depends on the history of S, in particular, upon $\{S\left(\mu \int_0^s Q(u)\, du\right) : 0 \leq s < t\}$. Hence formula (11) is recursive; we must show that it is well defined.

Of course, the idea is not so complicated: Formula (11) is a consequence of the fact that the intensity of departures at time s is $\mu Q(s)$, where the number $Q(s)$ of busy servers at time s is multiplied by the individual service rate μ. The function $\mu \int_0^t Q(s)\, ds$ appearing as an argument inside S serves as a random time change; see §II.6 of Brémaud [12], Chapter 6 of Ethier and Kurtz [19] and §7.4 of Daley and Vere-Jones [17].

By the simple conservation of flow, which expresses the content at time t as initial content plus flow in minus flow out, we have the basic expression

$$\begin{aligned} Q(t) &\equiv Q(0) + A(\lambda t) - D(t), \quad t \geq 0 \ , \\ &= Q(0) + A(\lambda t) - S\left(\mu \int_0^t Q(s)\, ds\right), \quad t \geq 0 \ . \end{aligned} \tag{12}$$

Lemma 2.1 (construction) *The stochastic process $\{Q(t) : t \geq 0\}$ is well defined as a random element of the function space D by formula* (12). *Moreover, it is a birth-and-death stochastic process with constant birth rate $\lambda_k = \lambda$ and linear state-dependent death rate $\mu_k = k\mu$.*

Proof. To construct a bonafide random element of D, start by conditioning upon the random variable $Q(0)$ and the two Poisson processes A and S. Then, with these sample paths specified, we recursively construct the sample path of the stochastic process $Q \equiv \{Q(t) : t \geq 0\}$. By induction over the sucessive jump times of the process Q, we show that the sample paths of Q are right-continuous piecewise-constant real-valued functions of t. Since the Poisson processes A and S have only finitely many transitions in any finite interval w.p.1, the same is necessarily true of the constructed process Q. This sample-path construction follows Theorem 4.1 in Chapter 8 on p. 327 of Ethier and Kurtz [19]; the argument also appears in Theorem 9.2 of Mandelbaum et al. [48]. Finally, we can directly verify that the stochastic process Q satisfies the differential definition of a birth-and-death stochastic process. Let $\mathcal{F}_t$ represent the history of the system up to time t. That is the sigma field generated by $\{Q(s) : 0 \leq s \leq t\}$. It is then straightforward that the infinitesimal transition rates are as claimed:

$$\begin{aligned} P(Q(t+h)-Q(t)=+1|Q(t)=k,\mathcal{F}_t) &= \lambda h + o(h) \ , \\ P(Q(t+h)-Q(t)=-1|Q(t)=k,\mathcal{F}_t) &= k\mu h + o(h) \ , \\ P(Q(t+h)-Q(t)=0|Q(t)=k,\mathcal{F}_t) &= 1-\lambda h - k\mu h + o(h) \ , \end{aligned}$$

as $h \downarrow 0$ for each $k \geq 0$, where $o(h)$ denotes a function f such that $f(h)/h \to 0$ as $h \downarrow 0$. Ethier and Kurtz [19] approach this last distributional step by verifying that the uniquely constructed process is the solution of the local-martingale problem for the generator of the Markov process, which in our case is the birth-and-death process. ∎

Some further explanation is perhaps helpful. Our construction above is consistent with the construction of the queue-length process as a Markov process, specifically, a birth-and-death process. There are other possible constructions. A different construction would be the standard approach in discrete-event simulation, with event clocks. Upon the arrival of each customer, we might schedule the arrival time of the next customer, by generating an exponential interarrival time. We might also generate the required service time of the current arrival. With the simulation approach, we have information about the future state that we do not have with the Markov-process construction. The critical distinction between the different constructions involves the information available at each time. The information available is captured by the filtration, which we discuss with the martingale representations in the next section.

The random-time-change approach we have used here is natural when applying strong approximations, as was done by Mandelbaum, Massey and Reiman [46] and Mandelbaum and Pats [47, 48]. They applied the strong approximation for a Poisson process, as did Kurtz [39]. A different use of strong approximations to establish heavy-traffic limits for non-Markovian infinite-server models is contained in Glynn and Whitt [22].

2.2. Random Thinning of Poisson Processes

We now present an alternative construction, which follows §II.5 of Brémaud [12] and Puhalskii and Reiman [53]. For this construction, let A_λ and $S_{\mu,k}$ for $k \geq 1$ be independent Poisson processes with rate λ and μ, respectively. As before, we assume that these Poisson processes are independent of the initial number of busy servers, $Q(0)$. This will not alter the overall system behavior, because the service-time distribution is exponential. By the lack-of-memory property, the remaining service times are distributed as i.i.d. exponential random variables, independent of the elapsed service times at time 0.

We let all arrivals be generated from arrival process A_λ; we let service completions from individual servers be generated from the processes $S_{\mu,k}$. To be specific, let the servers be numbered. With that numbering, at any time we use $S_{\mu,k}$ to generate service completions from the busy server with the k^{th} smallest index among all busy servers. (We do not fix attention on a particular server, because we want the initial indices to refer to the busy servers at all time.)

Instead of (11), we define the departure process by

$$D(t) \equiv \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(s-)\geq k\}}\, dS_{\mu,k}(s), \quad t \geq 0 \ , \tag{13}$$

where 1_A is the indicator function of the event A; i.e., $1_A(\omega) = 1$ of $\omega \in A$ and $1_A(\omega) = 0$ otherwise. It is important that we use the left limit $Q(s-)$ in the integrand of (13), so that the intensity of any service completion does not depend on that service completion itself; i.e., the functions $Q(s-)$ and $1_{\{Q(s-)=k\}}$ are left-continuous in s for each sample point. Then, instead of (12), we have

$$Q(t) \quad \equiv \quad Q(0) + A(\lambda t) - \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(s-)\geq k\}}\, dS_{\mu,k}(s), \quad t \geq 0 \ . \tag{14}$$

With this alternative construction, there is an analog of Lemma 2.1 proved in essentially the same way. In this setting, it is even more evident that we can construct a sample path of the stochastic process $\{Q(t) : t \geq 0\}$ by first conditioning on a realization of $Q(0)$, $\{A_\lambda(t) : t \geq 0\}$ and $\{S_{\mu,k}(t) : t \geq 0\}$, $k \geq 1$.

Even though this construction is different that the one in §2.1, it too is consistent with the Markov-process view. Consistent with most applications, we know what has happened up to time t, but not future arrival times and service times.

2.3. Construction from Arrival and Service Times

Now, following Krichagina and Puhalskii [37], we construct the sample paths of the process Q from the arrival and service times. This construction applies to the more general $G/GI/\infty$ system; we will show how the approach applies to the special case $M/M/\infty$; see §§3.7 and 6.4. See [37] for references to related work.

Let $A(t)$ be the cumulative number of arrivals in the interval $[0, t]$ and let τ_i be the time of the i^{th} arrival. Let all the service times be mutually independent random variables, independent of the arrival process and the number, $Q(0)$, of customers in the system at time 0 (before new arrivals). Let $Q(0)$ be independent of the arrival process. Let the $Q(0)$ initial customers have service times from $\{\bar{\eta}_i, i \geq 1\}$ with cumulative distribution function (cdf) F_0. Let the new arrivals have service times from $\{\eta_i, i \geq 1\}$ with cdf F.

Then $D(t)$, the number of customers that leave the system by time t, can be expressed as

$$D(t) = \sum_{i=1}^{Q(0)} \mathbf{1}(\bar{\eta}_i \leq t) + \sum_{i=1}^{A(t)} \mathbf{1}(\tau_i + \eta_i \leq t), \quad t \geq 0 \ . \tag{15}$$

and

$$Q(t) \quad = \quad \sum_{i=1}^{Q(0)} \mathbf{1}(\bar{\eta}_i > t) + \sum_{i=1}^{A(t)} \mathbf{1}(\tau_i + \eta_i > t), \quad t \geq 0 \ . \tag{16}$$

Since this construction does not require A to be Poisson (or even renewal) or the cdf's F and F_0 to be exponential, this approach applies to the non-Markovian $G/GI/\infty$ model. However, the stochastic process Q itself is no longer Markov in the general setting. With Poisson arrivals, we can extend [37] to obtain a Markov process by considering the two-parameter process $\{Q(t,y) : 0 \leq y \leq t, t \geq 0\}$, where $Q(t,y)$ is the number of customers in the system at time t that have elapsed service times greater than or equal to y. (For simplicity in treating the initial conditions, let the initial customers have elapsed service times 0 at time 0. Since $P(A(0) = 0) = 1$, $Q(t,t)$ is (w.p.1) the number of initial customers still in the system at time t.) Then

$$Q(t,y) \quad = \quad \sum_{i=1}^{Q(0)} \mathbf{1}(\bar{\eta}_i > t) + \sum_{i=1}^{A(t-y)} \mathbf{1}(\tau_i + \eta_i > t),\, 0 \leq y \leq t,\, t \geq 0. \quad (17)$$

With renewal (GI) arrivals (and the extra assumption that $P(A(0) = 0) = 1$), we can obtain a Markov process by also appending the elapsed interarrival time. Of course, there is an alternative to (17) if we add remaining times instead of elapsed times, but that information is less likely to be available as time evolves. Heavy-traffic limits for these two-parameter processes follow from the argument of [37], but we leave detailed discussion of these extensions to future work. For other recent constructions and limits in this spirit, see Kaspi and Ramanan [34] and references cited there.

3. Martingale Representations

For each of the sample-path constructions in the previous section, we have associated martingale representations. At a high level, it is easy to summarize how to treat the first two sample-path constructions, drawing only on the first two chapters of Brémaud [12]. We represent the random time changes as stopping times with respect to appropriate filtrations and apply versions of the optional stopping theorem, which states that random time changes of martingales are again martingales, under appropriate regularity conditions; e.g., see Theorem T2 on p. 7 of [12]. The problems we consider tend to produce multiple random time changes, making it desirable to apply the optional stopping theorem for multiparameter random time changes, as in §§2.8 and 6.2 of Ethier and Kurtz [19], but we postpone discussion of that more involved approach until Section 7.1.

For the random thinning, we instead apply the integration theorem for martingales associated with counting processes, as in Theorem T8 on p. 27 of Brémaud [12], which concludes that integrals of predictable processes with respect to martingales associated with counting processes produce new martingales, under regularity conditions.

We also present a third martingale representation, based on martingales associated with generators of Markov processes. That approach applies nicely here, because the queueing processes we consider are birth-and-death processes. Thus we can also apply martingales associated with Markov chains, as on pp 5-6, 294 of [12]. Finally, we present a fourth martingale representation associated with §2.3.

3.1. Martingale Basics

In this subsection we present some preliminary material on continuous-time martingales. There is a large body of literature providing background, including the books by: Brémaud [12], Ethier and Kurtz [19], Liptser and Shiryayev [44], Jacod and Shiryayev [30], Rogers and Williams [58, 59], Karatzas and Shreve [33] and Kallenberg [35]. The early book by Brémaud [12] remains especially useful because of its focus on stochastic point processes and queueing models. More recent lecture notes by Kurtz [41] and van der Vaart [64] are very helpful as well.

Stochastic integrals play a prominent role, but we only need relatively simple cases. Since we will be considering martingales associated with counting processes, we rely on the elementary theory for finite-variation processes, as in §IV.3 of Rogers and Williams [58]. In that setting, the stochastic integrals reduce to ordinary Stieltjes integrals and we exploit integration by parts.

On the other hand, as can be seen from Theorems 1.1 and 1.2, the limiting stochastic processes involve stochastic integrals with respect to Brownian motion, which is substantially more complicated. However, that still leaves us well within the classical theory. We can apply the Ito calculus as in Chapter IV of [58], without having to draw upon the advanced theory in Chapter VI.

For the stochastic-process limits in the martingale setting, it is natural to turn to Ethier and Kurtz [19], Liptser and Shiryayev [44] and Jacod and Shiryayev [30]. We primarily rely on Theorem 7.1 on p. 339 of Ethier and Kurtz [19]. The Shiryayev books are thorough; Liptser and Shiryayev [44] focuses on basic martingale theory, while Jacod and Shiryayev [30] focuses on stochastic-process limits.

We will start by imposing regularity conditions. We assume that all stochastic processes $X \equiv \{X(t) : t \geq 0\}$ under consideration are measurable maps from an underlying probability space $(\Omega, \mathcal{F}, P)$ to the function space $D \equiv D([0,\infty), \mathbb{R})$ endowed with the standard Skorohod J_1 topology and the associated Borel σ-field (generated by the open subsets), which coincides with the usual σ-field generated by the coordinate projection maps; see §§3.3 and 11.5 of [69].

Since we will be working with martingales, a prominent role is played by the **filtration** (histories, family of σ fields) $\mathbf{F} \equiv \{\mathcal{F}_t : t \geq 0\}$ defined on the underlying probability space $(\Omega, \mathcal{F}, P)$. (We have the containments: $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2} \subseteq \mathcal{F}$ for all $t_1 < t_2$.) As is customary, see p. 1 of [44], we assume that all filtrations satisfy the **usual conditions**:

(i) they are **right-continuous**:

$$\mathcal{F}_t = \bigcap_{u:u>t} \mathcal{F}_u \quad \text{for all} \quad t, \quad 0 \leq t < \infty \ , \quad \textit{and}$$

(ii) **complete** ($\mathcal{F}_0$, and thus $\mathcal{F}_t$ contains all P-null sets of $\mathcal{F}$).

We will assume that any stochastic process X under consideration is **adapted** to the filtration; i.e., X is **F**-adapted, which means that $X(t)$ is $\mathcal{F}_t$-measurable for each t. These regularity conditions guarantee desirable measurability properties, such as progressive measurability: The stochastic process X is *progressively measurable* with respect to the filtration $\mathbf{F}$ if, for each $t \ge 0$ and Borel measurable subset A of $\mathbb{R}$, the set $\{(s,\omega) : 0 \le s \le t, \omega \in \Omega, X(s,\omega) \in A\}$ belongs to the product σ field $\mathcal{B}([0,t]) \times \mathcal{F}_t$, where $\mathcal{B}([0,t])$ is the usual Borel σ field on the interval $[0,t]$. See p. 5 of [33] and Section 1.1 of [44]. In turn, progressive measurability implies measurability, regarding $X(t)$ as a map from the product space $\Omega \times [0,\infty)$ to $\mathbb{R}$.

A stochastic process $M \equiv \{M(t) : t \ge 0\}$ is a a **martingale** (submartingale) with respect to a filtration $\mathbf{F} \equiv \{\mathcal{F}_t : t \ge 0\}$ if $M(t)$ is adapted to $\mathcal{F}_t$ and $E[M(t)] < \infty$ for each $t \ge 0$, and

$$E[M(t+s)|\mathcal{F}_t] = (\ge) M(t)$$

with probability 1 (w.p.1) with respect to the underlying probability measure P for each $t \ge 0$ and $s > 0$.

It is often important to have a stronger property than the finite-moment condition $E[M(t)] < \infty$. A stochastic process $X \equiv \{X(t) : t \ge 0\}$ is **uniformly integrable (UI)** if

$$\lim_{n\to\infty} \sup_{t\ge 0} \{E[|X(t)|1_{\{|X(t)|>n\}}]\} = 0 \ ;$$

see p. 286 of [12] and p. 114 of [59]. We remark that UI implies, but is not implied by

$$\sup_{t\ge 0} \{E[|X(t)|]\} < \infty \ . \tag{18}$$

A word of warning is appropriate, because the books are not consistent in their assumptions about *integrability*. A stochastic process $X \equiv \{X(t) : t \ge 0\}$ may be called integrable if $E[|X(t)|] < \infty$ for all t or if the stronger (18) holds. This variation occurs with *square-integrable*, defined below. Similarly, the basic objects may be taken to be martingales, as defined as above, or might instead be UI martingales, as on p. 20 of [44].

The stronger UI property is used in preservation theorems - theorems implying that stopped martingales and stochastic integrals with respect to martingales - remain martingales. In order to get this property when it is not at first present, the technique of **localizing** is applied. We localize by introducing associated stopped processes, where the stopping is done with stopping times. A nonnegative random variable τ is an **F-stopping time** if stopping sometime before t depends only on the history up to time t, i.e., if

$$\{\tau \le t\} \in \mathcal{F}_t \quad \text{for all} \quad t \ge 0 \ .$$

For any class $\mathcal{C}$ of stochastic processes, we define the associated local class $\mathcal{C}_{loc}$ as the class of stochastic processes $\{X(t) : t \ge 0\}$ for which there exists a sequence of stopping times $\{\tau_n : n \geq 1\}$ such that $\tau_n \to \infty$ w.p.1 as $n \to \infty$ and the associated stopped processes $\{X(\tau_n \wedge t) : t \geq 0\}$ belong to class $\mathcal{C}$ for each n, where $a \wedge b \equiv \min\{a, b\}$. We obtain the class of **local martingales** when $\mathcal{C}$ is the class of martingales. Localizing expands the scope, because if we start with martingales, then such stopped martingales remain martingales, so that all martingales are local martingales. Localizing is important, because the stopped processes not only are martingales but can be taken to be UI martingales; see p. 21 of [44] and §IV.12 of [58]. The UI property is needed for the preservation theorems.

As is customary, we will also be exploiting **predictable** stochastic processes, which we will take to mean having left-continuous sample paths. See p. 8 of [12] and §1.2 of [30] for the more general definition and additional discussion. The general idea is that a stochastic process $X \equiv \{X(t) : t \geq 0\}$ is predictable if its value at t is determined by values at times prior to t. But we can obtain simplification by working with stochastic processes with sample paths in D. In the setting of D, we can have a left-continuous process by either (i) considering the left-limit version $\{X(t-) : t \geq 0\}$ of a stochastic process X in D (with $X(0-) \equiv X(0)$) or (ii) considering a stochastic process in D with continuous sample paths. Once we restrict attention to stochastic processes with sample paths in D, we do not need the more general notion, because the left-continuous version is always well defined. If we allowed more general sample paths, that would not be the case.

We will be interested in martingales associated with counting processes, adapted to the appropriate filtration. These counting processes will be nonnegative submartingale processes. Thus we will be applying the following special case of the Doob-Meyer decomposition.

Theorem 3.1 (Doob-Meyer decomposition for nonnegative submartingales) *If Y is a submartingale with nonnegative sample paths, $E[Y(t)] < \infty$ for each t, and Y is adapted to a filtration $\boldsymbol{F} \equiv \{\mathcal{F}_t\}$, then there exists an $\boldsymbol{F}$-predictable process A, called the **compensator** of Y or the dual-predictable projection, such that A has nonnegative nondecreasing sample paths, $E[A(t)] < \infty$ for each t, and $M \equiv Y - A$ is an $\boldsymbol{F}$-martingale. The compensator is unique in the sense that the sample paths of any two versions must be equal w.p.1.*

Proof. See §1.4 of Karatzas and Shreve [33]. The DL condition in [33] is satisfied because of the assumed nonnegativity; see Definition 4.8 and Problem 4.9 on p. 24. For a full account, see §VI.6 of [58]. ■

3.2. Quadratic Variation and Covariation Processes

A central role in the martingale approach to stochastic-process limits is played by the quadratic-variation and quadratic-covariation processes, as can be seen from the martingale FCLT stated here in §8. That in turn depends on the notion of square-integrability. We say that a stochastic process $X \equiv \{X(t) : t \geq 0\}$ is **square integrable** if $E[X(t)^2] < \infty$ for each $t \geq 0$. We thus say that a martingale $M \equiv \{M(t) : t \geq 0\}$ (with respect to some filtration) is square integrable if $E[M(t)^2] < \infty$ for each $t \geq 0$. Again, to expand the scope, we can localize, and focus on the class of locally square integrable martingales. Because we can localize to get the square-integrability, the condition is not very restrictive.

If M is a square-integrable martingale, then $M^2 \equiv \{M(t)^2 : t \geq 0\}$ is necessarily a submartingale with nonnegative sample paths, and thus satisfies the conditions of Theorem 3.1. The **predictable quadratic variation** (PQV) of a square-integrable martingale M, denoted by $\langle M \rangle \equiv \{\langle M \rangle(t) : t \geq 0\}$ the (**angle-bracket process**), is the compensator of the submartingale M^2; i.e., the stochastic process $\langle M \rangle$ is the unique nondecreasing nonnegative predictable process such that $E[\langle M \rangle(t)] < \infty$ for each t and $M^2 - \langle M \rangle$ is a martingale with respect to the reference filtration. (Again, uniqueness holds to the extent that any two versions have the same sample paths w.p.1.) Not only does square integrability extend by localizing, but Theorem 3.1 has a local version; see p. 375 of [58]. As a consequence the PQV is well defined for any locally square-integrable martingale.

Given two locally square-integrable martingales M_1 and M_2, the **predictable quadratic covariation** can be defined as

$$\langle M_1, M_2 \rangle \equiv \frac{1}{4} \left(\langle M_1 + M_2 \rangle - \langle M_1 - M_2 \rangle \right) \ ;$$

see p. 48 of [44]. It can be characterized as the unique (up to equality of sample paths w.p.1) nondecreasing nonnegative predictable process such that $E[\langle M_1, M_2 \rangle(t)] < \infty$ for each t and $M_1 M_2 - \langle M_1, M_2 \rangle$ is a martingale.

We will also be interested in another quadratic variation of a square-integrable martingale M, the so-called **optional quadratic variation** $[M]$ (OQV), the **square-bracket process)**. The square-bracket process is actually more general than the angle-bracket process, because the square-bracket process is well defined for any local martingale, as opposed to only all locally square-integrable martingales; see §§IV. 18, 26 and VI. 36, 37 of [58]. The following is Theorem 37.8 on p. 389 of [58]. For a stochastic process M, let $\Delta M(t) \equiv M(t) - M(t-)$, the jump at t, for $t \geq 0$.

Theorem 3.2 (optional quadratic covariation for local martingales) *Let M_1 and M_2 be local martingales with $M_1(0) = M_2(0) = 0$. Then there exists a unique process, denoted by $[M_1, M_2]$, with sample paths of finite variation over bounded intervals and $[M_1, M_2](0) = 0$ such that*

$$\begin{aligned} &(i) M_1 M_2 - [M_1, M_2] \quad \textit{is a local martingale} \\ &(ii) \Delta [M_1, M_2] = (\Delta M_1)(\Delta M_2) \ . \end{aligned}$$

For one local martingale M, the optional quadratic variation is then defined as $[M] \equiv [M, M]$.

Note that, for a locally square-integrable martingale M, both $M^2 - \langle M \rangle$ and $M^2 - [M]$ are local martingales, but $\langle M \rangle$ is predictable while $[M]$ is not. Indeed, subtracting, we see that $[M] - \langle M \rangle$ is a local martingale, so that $\langle M \rangle$ is the compensator of both M^2 and $[M]$.

There is also an alternative definition of the OQV; see Theorem 5.57 in §5.8 of [64].

Theorem 3.3 (alternative definition) *If M_1 and M_2 are local martingales with $M_1(0) = M_2(0) = 0$, then*

$$[M_1, M_2](t) \equiv \lim_{n \to \infty} \sum_{i=1}^{\infty} (M_1(t_{n,i}) - M_1(t_{n,i-1}))(M_2(t_{n,i}) - M_2(t_{n,i-1})) \ ,$$

where $t_{n,i} \equiv t \wedge (i2^{-n})$ and the mode of convergence for the limit as $n \to \infty$ is understood to be in probability. The limit is independent of the way that the time points $t_{n,i}$ are selected within the interval $[0,t]$, provided that $t_{n,i} > t_{n,i-1}$ and that the maximum difference $t_{n,i} - t_{n,i-1}$ for points inside the interval $[0,t]$ goes to 0 as $n \to \infty$.

Unfortunately, these two quadratic-variation processes $\langle M \rangle$ and $[M]$ associated with a locally square-integrable martingale M, and the more general covariation processes $\langle M_1, M_2 \rangle$ and $[M_1, M_2]$, are somewhat elusive, since the definitions are indirect; it remains to exhibit these processes. We will exploit our sample-path constructions in terms of Poisson processes above to identify appropriate quadratic variation and covariation processes in following subsections.

Fortunately, however, the story about structure is relatively simple in the two cases of interest to us: (i) when the martingale is a compensated counting process, and (ii) when the martingale has continuous sample paths. The story in the second case is easy to tell: When M is continuous, $\langle M \rangle = [M]$, and this (predictable and optional) quadratic variation process itself is continuous; see §VI.34 of [58]. This case applies to Brownian motion and our limit processes. For standard Brownian motion, $\langle M \rangle(t) = [M](t) = t$, $t \geq 0$.

3.3. Counting Processes

The martingales we consider for our pre-limit processes will be compensated counting processes. By a counting process (or point process), we mean a stochastic process $N \equiv \{N(t) : t \geq 0\}$ with nondecreasing nonnegative-integer-valued sample paths in D and $N(0) = 0$. We say that N is a unit-jump counting process if all jumps are of size 1. We say that N is non-explosive if $N(t) < \infty$ w.p.1 for each $t < \infty$. Equivalently, if $\{T_n : n \geq 1\}$ is the associated sequence of points, where

$$N(t) \equiv \max\{n \geq 0 : T_n \leq t\}, \quad t \geq 0 \ ,$$

with $T_0 \equiv 0$, then N is a unit-jump counting process if $T_{n+1} > T_n$ for all $n \geq 0$; while N is non-explosive if $T_n \to \infty$ w.p.1 as $n \to \infty$; see p. 18 of Brémaud (1981).

As discussed by Brémaud [12], the compensator of a non-explosive unit-jump counting process is typically (under regularity conditions!) a stochastic process with sample paths that are absolutely continuous with respect to Lebesgue measure, so that the compensator A can represented as an integral

$$A(t) = \int_0^t X(s)\,ds, \quad t \geq 0 \ ,$$

where $X \equiv \{X(t) : t \geq 0\}$ is adapted to the filtration $\mathbf{F}$. When the compensator has such an integral representation, the integrand X is called the **stochastic intensity** of the counting process N.

We will apply the following extension of Theorem 3.1.

Lemma 3.1 (PQV for unit-jump counting processes) *If N is a non-explosive unit-jump counting process adapted to $\boldsymbol{F}$ with $E[N(t)] < \infty$ for all t, and if the compensator A of N provided by Theorem* 3.1 *is continuous, then the martingale $M \equiv N - A$ is a square-integrable martingale with respect to $\boldsymbol{F}$ with quadratic variation processes:*

$$\langle M \rangle = A \quad \textit{and} \quad [M] = N \ .$$

We first observe that the conditions of Lemma 3.1 imply that N is a nonnegative $\mathbf{F}$-submartingale, so that we can apply Theorem 3.1. We use the extra conditions to get more. We prove Lemma 3.1 in the Appendix.

3.4. First Martingale Representation

We start with the first sample-path construction in §2.1. As a regularity condition, here we assume that $E[Q(0)] < \infty$. We will show how to remove that condition later in §6.3. It could also be removed immediately if we chose to localize.

Here is a quick summary of how martingales enter the picture: The Poisson processes $A(t)$ and $S(t)$ underlying the first representation of the queueing model in §2.1 as well as the new processes $A(\lambda t)$ and $S\left(\mu \int_0^t Q(s)\,ds\right)$ there have nondecreasing nonnegative sample paths. Consequently, they are submartingales with respect to appropriate filtrations (histories, families of σ-fields). Thus, by subtracting the compensators, we obtain martingales. Then the martingale M so constructed turns out to be square integrable, admitting a martingale representation $M^2 - \langle M \rangle$, where $\langle M \rangle$ is the predictable quadratic variation, which in our context will coincide with the compensator.

In constructing this representation, we want to be careful about the filtration (histories, family of σ fields). Here we will want to use the filtration $\mathbf{F} \equiv \{\mathcal{F}_t : t \geq 0\}$ defined by

$$\mathcal{F}_t \equiv \sigma\left(Q(0), A(\lambda s), S\left(\mu \int_0^s Q(u)\,du\right) : 0 \leq s \leq t\right), \quad t \geq 0 \ , \tag{19}$$

augmented by including all null sets.

The following processes will be proved to be **F**-martingales:

$$\begin{aligned} M_1(t) &\equiv A(\lambda t) - \lambda t, \\ M_2(t) &\equiv S\left(\mu \int_0^t Q(s)\, ds\right) - \mu \int_0^t Q(s)\, ds, \quad t \ge 0 \ , \end{aligned} \tag{20}$$

where here A refers to the arrival process. Hence, instead of (12), we have the alternate martingale representation

$$Q(t) = Q(0) + M_1(t) - M_2(t) + \lambda t - \mu \int_0^t Q(s)\, ds \ , \quad t \ge 0 \ . \tag{21}$$

In applications of Theorem 3.1 and Lemma 3.1, it remains to show that the conditions are satisfied and to identify the compensator. The following lemma fills in that step for a random time change of a rate-1 Poisson process by applying the optional stopping theorem. At this step, it is natural, as in §12 of Mandelbaum and Pats [48], to apply the optional stopping theorem for martingales indexed by directed sets (Theorem 2.8.7 on p. 87 of Ethier and Kurtz [19]) associated with multiparameter random time changes (§6.2 on p. 311 of [19]), but here we can use a more elementary approach. For supporting theory at this point, see Theorem 17.24 and Proposition 7.9 of Kallenberg [35] and §7.4 of Daley and Vere-Jones [17]. We use the mutiparameter random time change in §7.1.

Let $\circ$ be the composition map applied to functions, i.e., $(x \circ y)(t) \equiv x(y(t))$.

Lemma 3.2 (random time change of a rate-1 Poisson process) *Suppose that S is a rate-1 Poisson process adapted to a filtration $\boldsymbol{F} \equiv \{\mathcal{F}_t : t \ge 0\}$ and $I \equiv \{I(t) : t \ge 0\}$ is a stochastic process with continuous nondecreasing nonnegative sample paths, where $I(t)$ is an $\boldsymbol{F}$-stopping time for each $t \ge 0$. In addition, suppose that the following moment conditions hold:*

$$E[I(t)] < \infty \quad \textit{and} \quad E[S(I(t))] < \infty \quad \textit{for all} \quad t \ge 0. \tag{22}$$

Then $S \circ I \equiv \{S(I(t)) : t \ge 0\}$ is a non-explosive unit-jump counting process such that $M \equiv S \circ I - I \equiv \{S(I(t)) - I(t) : t \ge 0\}$ is a square-integrable martingale with respect to the filtration $\boldsymbol{F}_I \equiv \{\mathcal{F}_{I(t)} : t \ge 0\}$, having quadratic variation processes

$$\langle M\rangle(t) = I(t) \quad \textit{and} \quad [M](t) = S(I(t)), \quad t \ge 0 \ . \tag{23}$$

Proof. Since the sample paths of I are continuous, it is evident that $S \circ I$ is a unit-jump counting process. By condition (22), it is non-explosive. In order to apply the optional stopping theorem, we now localize by letting

$$I^m(t) \equiv I(t) \wedge m, \quad t \ge 0 \ .$$

Since $I(t)$ is an **F**-stopping time, $I^m(t)$ is a bounded **F**-stopping time for each $m \ge 1$. The optional stopping theorem then implies that $M^m \equiv S \circ I^m - I^m \equiv$

$\{S(I^m(t)) - I^m(t) : t \geq 0\}$ is an $\mathbf{F}_I$-martingale; e.g., see p. 7 of [12] or p. 61 of [19]. As a consequence I^m is the compensator of $S \circ I^m$. Since we have the moment conditions in (22), we can let $m \uparrow \infty$ and apply the monotone convergence theorem with conditioning, as on p. 280 of Brémaud, to deduce that $M \equiv S \circ I - I \equiv \{S(I(t)) - I(t) : t \geq 0\}$ is a martingale. Specifically, given that

$$E[S(I^m(t+s)) - I^m(t+s)|\mathcal{F}_{I(s)}] = S(I^m(s)) - I^m(s) \quad \text{w.p.1}$$

for all m,

$$E[S(I^m(t+s))|\mathcal{F}_{I(s)}] \to E[S(I(t+s))|\mathcal{F}_{I(s)}] \quad \text{w.p.1} \quad \text{as} \quad m \to \infty \ ,$$

$$E[I^m(t+s)|\mathcal{F}_{I(s)}] \to E[I(t+s)|\mathcal{F}_{I(s)}] \quad \text{w.p.1} \quad \text{as} \quad m \to \infty \ ,$$

$S(I^m(s))) \uparrow S(I(s))$ and $I^m(s) \uparrow I(s)$ as $m \to \infty$, we have

$$E[S(I(t+s)) - I(t+s)|\mathcal{F}_{I(s)}] = S(I(s)) - I(s) \quad \text{w.p.1}$$

Lemma 3.1 implies the square-integrability and identifies the quadratic variation processes. ■

To apply Lemma 3.2 to our $M/M/\infty$ queueing problem, we need to verify the finite-moment conditions in (22). For that purpose, we use a crude inequality:

Lemma 3.3 (crude inequality) *Given the representation* (12),

$$Q(t) \leq Q(0) + A(\lambda t), \quad t \geq 0 \ , \tag{24}$$

so that

$$\int_0^t Q(s)\, ds \leq t(Q(0) + A(\lambda t)), \quad t \geq 0 \ . \tag{25}$$

Now we want to show that our processes in (20) actually are martingales with respect to the filtration in (19). To do so, we will apply Lemma 3.2. However, to apply Lemma 3.2, we will first alter the filtration. In order to focus on the service completions in the easiest way, we initially condition on the entire arrival process, and consider the filtration $\mathbf{F}^1 \equiv \{\mathcal{F}_t^1 : t \geq 0\}$ defined by

$$\mathcal{F}_t^1 \equiv \sigma\left(Q(0), \{A(u) : u \geq 0\}, S\left(s\right) : 0 \leq s \leq t\right), \quad t \geq 0 \ , \tag{26}$$

augmented by including all null sets. Then, as in the statement of Lemma 3.2, we consider the associated filtration $\mathbf{F}_I^1 \equiv \{\mathcal{F}_{I(t)}^1 : t \geq 0\}$. Finally, we are able to obtain the desired martingale result with respect to the desired filtration $\mathbf{F}$ in (19).

Lemma 3.4 (verifying the conditions.) *Suppose that $E[Q(0)] < \infty$ in the setting of* §2.1 *with $\{Q(t) : t \geq 0\}$ defined in* (12),

$$I(t) \equiv \mu \int_0^t Q(s)\, ds, \quad t \geq 0 \ , \tag{27}$$

and the filtration being $\boldsymbol{F}^1$ in (26). *Then the conditions of Lemma* 3.2 *are satisfied, so that $S \circ I - I$ is a square-integrable $\boldsymbol{F}_I^1$-martingale with $\boldsymbol{F}_I^1$-compensator I in* (27). *As a consequence, $S \circ I - I$ is also a square-integrable $\boldsymbol{F}$-martingale with $\boldsymbol{F}$-compensator I in* (27) *for filtration $\boldsymbol{F}$ in* (19).

Proof. First, we can apply the crude inequality in (25) to establish the required moment conditions: Since $E[Q(0)] < \infty$,

$$E\left[\mu \int_0^t Q(s)\, ds\right] \le \mu t(E[Q(0)] + E[A(\lambda t)]) = \mu t E[Q(0)] + \mu\lambda t^2 < \infty, \quad t \ge 0 \ ,$$

and

$$S\left(\mu \int_0^t Q(s)\, ds\right) \le S\left(\mu t(Q(0) + A(\lambda t)\right), \quad t \ge 0 \ ,$$

so that

$$\begin{aligned} E\left[S\left(\mu \int_0^t Q(s)\, ds\right)\right] &\le E\left[S\left(\mu t(Q(0) + A(\lambda t))\right)\right], \\ &= E\left[E\left[S\left(\mu t(Q(0) + A(\lambda t))\right) | Q(0) + A(\lambda t)\right]\right], \quad t \ge 0 \ , \\ &= \mu t\left(E[Q(0)] + \lambda t\right) < \infty, \quad t \ge 0 \ . \end{aligned}$$

Then, by virtue of (12) and the recursive construction in Lemma 2.1, for each $t \ge 0$, $I(t)$ in (27) is a stopping time relative to $\mathcal{F}_x^1$ for all $x \ge 0$, i.e.,

$$\{I(t) \le x\} \in \mathcal{F}_x^1 \quad \text{for all} \quad x \ge 0 \quad \text{and} \quad t \ge 0 \ .$$

(This step is a bit tricky: To know $I(t)$, we need to know $Q(s), 0 \le s < t$, but, by (2.2), that depends on $I(s), 0 \le s < t$. Hence, to know whether or not $\{I(t) \le x\}$ holds, it suffices to know $S(u) : 0 \le u \le x$.) Since $\{S(t) - t : t \ge 0\}$ is a martingale with respect to $\mathbf{F}^1$ and the moment conditions in (22) are satisfied, we can apply Lemma 3.2 to deduce that $\{S(I(t)) - I(t) : t \ge 0\}$ is a square-integrable martingale with respect to the filtration $\mathbf{F}_I^1 \equiv \{\mathcal{F}_{I(t)}^1 : t \ge 0\}$ augmented by including all null sets and that I in (27) is both the compensator and the predictable quadratic variation. Finally, since the process representing the arrivals after time t, i.e., the stochastic process $\{A(t+s) - A(t) : s \ge 0\}$, is independent of $Q(s)$, $0 \le s \le t$, by virtue of the recursive construction in Lemma 2.1 (and the assumption that A is a Poisson process), we can replace the filtration $\mathbf{F}_I^1$ by the smaller filtration $\mathbf{F}$ in (19). That completes the proof. ■

We now introduce corresponding processes associated with the **sequence of models** indexed by n. We have

$$Q_n(t) = Q_n(0) + M_{n,1}^*(t) - M_{n,2}^*(t) + \lambda_n t - \mu \int_0^t Q_n(s)\, ds \ , \quad t \ge 0 \ . \qquad (28)$$

where

$$\begin{aligned} M_{n,1}^*(t) &\equiv A(\lambda_n t) - \lambda_n t, \\ M_{n,2}^*(t) &\equiv S\left(\mu \int_0^t Q_n(s)\, ds\right) - \mu \int_0^t Q_n(s)\, ds, \end{aligned} \qquad (29)$$

The filtrations change with n in the obvious way

We now introduce the scaling, just as in (3). Let the scaled martingales be

$$M_{n,1}(t) \equiv \frac{M^*_{n,1}(t)}{\sqrt{n}} \quad \text{and} \quad M_{n,2}(t) \equiv \frac{M^*_{n,2}(t)}{\sqrt{n}}, \quad t \ge 0 \ . \tag{30}$$

Then, from (28)–(30), we get

$$\begin{aligned} X_n(t) &\equiv \frac{Q_n(t)-n}{\sqrt{n}} \\ &= \frac{Q_n(0)-n}{\sqrt{n}} + \frac{M^*_{n,1}(t)}{\sqrt{n}} - \frac{M^*_{n,2}(t)}{\sqrt{n}} + \frac{\lambda_n t - n\mu t}{\sqrt{n}} \\ &\quad -\mu \int_0^t \left(\frac{Q_n(s)-n}{\sqrt{n}}\right) ds \ , \\ &= \frac{Q_n(0)-n}{\sqrt{n}} + \frac{M^*_{n,1}(t)}{\sqrt{n}} - \frac{M^*_{n,2}(t)}{\sqrt{n}} - \mu \int_0^t \left(\frac{Q_n(s)-n}{\sqrt{n}}\right) ds \ , \\ &= X_n(0) + M_{n,1}(t) - M_{n,2}(t) - \mu \int_0^t X_n(s)\, ds \ , \quad t \ge 0 \ . \end{aligned} \tag{31}$$

Now we summarize this martingale representation for the scaled processes, depending upon the index n. Here is the implication of the analysis above:

Theorem 3.4 (first martingale representation for the scaled processes) *If $E[Q_n(0)] < \infty$ for each $n \ge 1$, then the scaled processes X_n in* (3) *have the martingale representation*

$$X_n(t) \equiv X_n(0) + M_{n,1}(t) - M_{n,2}(t) - \mu \int_0^t X_n(s)\, ds \ , \quad t \ge 0 \ , \tag{32}$$

where $M_{n,i}$ are given in (29) *and* (30). *These processes $M_{n,i}$ are square-integrable martingales with respect to the filtrations $\boldsymbol{F}_n \equiv \{\mathcal{F}_{n,t} : t \ge 0\}$ defined by*

$$\mathcal{F}_{n,t} \equiv \sigma\left(Q_n(0), A(\lambda_n s), S\left(\mu \int_0^s Q_n(u)\, du\right) : 0 \le s \le t\right), \quad t \ge 0 \ ,$$

augmented by including all null sets. Their associated predictable quadratic variations are $\langle M_{n,1}\rangle(t) = \lambda_n t/n$, $t \ge 0$, and

$$\langle M_{n,2}\rangle(t) = \frac{\mu}{n} \int_0^t Q_n(s)\, ds, \quad t \ge 0 \ , \tag{33}$$

where $E[\langle M_{n,2}\rangle(t)] < \infty$ for all $t \ge 0$ and $n \ge 1$. The associated optional quadratic variations are $[M_{n,1}](t) = A(\lambda_n t)/n$, $t \ge 0$ and

$$[M_{n,2}](t) = \frac{S\left(\mu \int_0^t Q_n(s)\, ds\right)}{n}, \quad t \ge 0 \ .$$

Note that X_n appears on both sides of the integral representation (32), but $X_n(t)$ appears on the left, while $X_n(s)$ for $0 \le s \le t$ appears on the right. In §4.1 we show how to work with this integral representation.

3.5. Second Martingale Representation

We can also start with the second sample-path construction and obtain another integral representation of the form (32).

Now we start with the martingales:

$$\begin{aligned}
M_A(t) &\equiv A_\lambda(t) - \lambda t, \\
M_{S_{\mu,k}}(t) &\equiv S_{\mu,k}(t) - \mu t, \\
M_S(t) &\equiv \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(s-)\geq k\}}\, dS_{\mu,k}(s) - \sum_{k=1}^{\infty} \int_0^t \mu 1_{\{Q(s-)\geq k\}}\, ds, \\
&= \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(s-)\geq k\}}\, dS_{\mu,k}(s) - \int_0^t \sum_{k=1}^{\infty} \mu 1_{\{Q(s-)\geq k\}}\, ds, \\
&= \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(s-)\geq k\}}\, dS_{\mu,k}(s) - \mu \int_0^t Q(s-)\, ds, \qquad (34)
\end{aligned}$$

so that, instead of (12), we have the alternate representation

$$Q(t) = Q(0) + M_A(t) - M_S(t) + \lambda t - \mu \int_0^t Q(s-)\, ds, \;, \quad t \geq 0 \;, \qquad (35)$$

where M_A and M_S are square-integrable martingales with respect to the filtration $\mathbf{F} \equiv \{\mathcal{F}_t : t \geq 0\}$ defined by

$$\mathcal{F}_t \equiv \sigma\left(Q(0), A_\lambda(s), S_{\mu,k}(s), \quad k \geq 1 : 0 \leq s \leq t\right), \quad t \geq 0 \;, \qquad (36)$$

again augmented by the null sets. Notice that this martingale representation is very similar to the martingale representation in (21). The martingales in (21) and (35) are different and the filtrations in (19) and (36) are different, but the predictable quadratic variations are the same and the form of the integral representation is the same. Thus, there is an analog of Theorem 3.4 in this setting.

We now provide theoretical support for the claims above. First, we put ourselves in the setting of Lemma 3.1.

Lemma 3.5 (a second integrable counting process with unit jumps) *If* $E[Q(0)] < \infty$*, then, in addition to being adapted to the filtration* $\{\mathcal{F}_t\}$ *in* (36)*, the stochastic process* Y *defined by*

$$Y(t) \equiv \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(s-)\geq k\}}\, dS_k(s), \quad t \geq 0, \qquad (37)$$

is a unit-jump counting process such that $E[Y(t)] < \infty$ *for all* $t \geq 0$.

Proof. It is immediate that Y is a counting process with unit jumps, but there is some question about integrability To establish integrability, we apply the crude inequality (24) to get

$$Y(t) \le \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(0)+A(\lambda t)\ge k\}}\, dS_k(s) \le \sum_{k=1}^{\infty} 1_{\{Q(0)+A(\lambda t)\ge k\}} S_k(t) \ ,$$

so that

$$\begin{aligned} E[Y(t)] &\le \sum_{k=1}^{\infty} P(Q(0)+A(\lambda t)\ge k) E[S_k(t)] \\ &\le \sum_{k=1}^{\infty} P(Q(0)+A(\lambda t)\ge k)\mu t \\ &\le \mu t E[Q(0)+A(\lambda t)] = \mu t (E[Q(0)]+\lambda t) < \infty \ . \quad \blacksquare \end{aligned}$$

Given Lemmas 3.1 and 3.5, it only remains to identify the compensator of the counting process Y, which we call $\tilde{Y}$ since A is used to refer to the arrival process. For that purpose, we can apply the integration theorem, as on p. 10 of Brémaud [12]. But our process Y in (37) is actually a sum involving infinitely many Poisson processes, so we need to be careful.

Lemma 3.6 (identifying the compensator of Y) *The compensator of Y in* (37) *is given by*

$$\tilde{Y}(t) \equiv \sum_{k=1}^{\infty} \int_0^t 1_{\{Q(s-)\ge k\}} \mu\, ds = \mu \int_0^t Q(s-)\, ds, \quad t \ge 0 \ ; \tag{38}$$

i.e., $Y - \tilde{Y}$ is an $\boldsymbol{F}$-martingale for the filtration in (36).

Proof. As indicated above, we can apply the integration theorem on p. 10 of Brémaud, but we have to be careful because Y involves infinitely many Poisson processes. Hence we first consider the first n terms in the sum. With that restriction, since the integrand is an indicator function for each k, we consider the integral of a bounded predictable process with respect to the martingale $\{\sum_{k=1}^{n}(S_{\mu,k}(t)-\mu t) : t \ge 0\}$, which is a martingale of "integrable bounded variation," as required (and defined) by Brémaud. As a consequence, $\{M_S^n(t) : t \ge 0\}$ is an $\mathbf{F}$-martingale, where

$$M_S^n(t) \ \equiv \ \sum_{k=1}^{n} \int_0^t 1_{\{Q(s-)\ge k\}}\, dS_{\mu,k}(s) - \sum_{k=1}^{n} \int_0^t \mu 1_{\{Q(s-)\ge k\}}\, ds \ . \tag{39}$$

However, given that $E[Y(t)] < \infty$, we can apply the monotone convergence theorem to each of the two terms in (39) in order to take the limit as $n \to \infty$ to deduce that $E[\tilde{Y}(t)] < \infty$ and M_S itself, as defined in (34), is an $\mathbf{F}$-martingale, which implies that the compensator of Y in (37) is indeed given by (38). $\blacksquare$

By this route we obtain another integral representation for the scaled processes of exactly the same form as in Theorem 3.4. As before in (28)-(31), we introduce the sequence of models indexed by n. The martingales and filtrations are slightly different, but in the end the predictable quadratic variation processes are essentially the same.

Theorem 3.5 (second martingale representation for the scaled processes) *If $E[Q_n(0)] < \infty$ for each $n \geq 1$, then the scaled processes X_n in* (3) *have the martingale representation*

$$X_n(t) \equiv X_n(0) + M_{n,1}(t) - M_{n,2}(t) - \mu \int_0^t X_n(s)\, ds \ , \quad t \geq 0 \ , \tag{40}$$

where $M_{n,i}$ are given in (30)*, but instead of* (29)*, we have*

$$\begin{aligned} M^*_{n,1}(t) &\equiv A_{\lambda_n}(t) - \lambda_n t, \\ M^*_{n,2}(t) &\equiv \sum_{k=1}^{\infty} \int_0^t 1_{\{Q_n(s-) \geq k\}}\, dS_{\mu,k}(s) - \mu \int_0^t Q_n(s-)\, ds \ . \end{aligned}$$

These processes $M_{n,i}$ are square-integrable martingales with respect to the filtrations $\boldsymbol{F}_n \equiv \{\mathcal{F}_{n,t} : t \geq 0\}$ defined by

$$\mathcal{F}_{n,t} \equiv \sigma\left(Q_n(0), A_{\lambda_n}(s), S_{\mu,k}(s), \quad k \geq 1 : 0 \leq s \leq t\right), \quad t \geq 0 \ ,$$

augmented by including all null sets. Their associated predictable quadratic variations are $\langle M_{n,1}\rangle(t) = \lambda_n t/n$, $t \geq 0$, and

$$\langle M_{n,2}\rangle(t) = \frac{\mu}{n} \int_0^t Q_n(s)\, ds, \quad t \geq 0 \ , \tag{41}$$

where $E[\langle M_{n,2}\rangle(t)] < \infty$ for all $t \geq 0$ and $n \geq 1$ and $\langle M_{n,1}\rangle(t) = (\lambda_n t/n) = \mu t$. The associated optional quadratic variations are $[M_{n,1}](t) = A_{\lambda_n}(t)/n$, $t \geq 0$, and

$$[M_{n,2}](t) = \frac{\sum_{k=1}^{\infty} \int_0^t 1_{\{Q_n(s-) \geq k\}}\, dS_{\mu,k}(s)}{n}, \quad t \geq 0 \ . \tag{42}$$

3.6. Third Martingale Representation

We can also obtain a martingale representation for the stochastic process Q by exploiting the fact that the stochastic process Q is a birth-and-death process. We have the basic representation

$$Q(t) = Q(0) + A(t) - D(t), \quad t \geq 0 \ ,$$

where A is the arrival process and D is the departure process, as in (12). Since Q is a birth-and-death process, we can apply the Lévy and Dynkin formulas, as on p. 294 of Brémaud [12] to obtain martingales associated with various counting processes associated with Q, including the counting processes A and D. Of course, A is easy, but the Dynkin formula immediately yields the desired martingale for D:

$$M_D(t) \equiv D(t) - \mu \int_0^t Q(s)\, ds, \quad t \geq 0 \ ,$$

where the compensator of M_D is just as in the first two martingale representation, i.e., as in (20), (27). (34) and (38); see pp 6 and 294 of [12]. We thus again obtain the martingale representation of the form (21) and (35). Here, however, the filtration can be taken to be

$$\mathcal{F}_t \equiv \sigma\left(Q(s) : 0 \leq s \leq t\right), \quad t \geq 0 \ .$$

The proof of Theorem 1.1 is then the same as for the second representation, which will be by an application of the martingale FCLT in §8.

3.7. Fourth Martingale Representation

In this section, we present the martingale representation for the construction in terms of arrival and service times in §2.3, but without any proofs. Consider a sequence of $G/GI/\infty$ queues indexed by n and let $Q_n(0)$, Q_n, A_n, and D_n be the corresponding quantities in the n^{th} queueing system, just as defined in §2.3. For any cdf F, let the associated complementary cdf (ccdf) be $F^c \equiv 1 - F$.

Given representation (17), the Krichagina and Puhalskii [37] insight is to write the process Q_n as

$$\begin{aligned} Q_n(t) \quad &= \quad \sum_{i=1}^{Q_n(0)} \left(\mathbf{1}(\tilde{\eta}_i > t) - F_0^c(t)\right) + Q_n(0) F_0^c(t) \\ &\qquad + n \int_0^t \int_0^\infty \mathbf{1}(s + x > t) dK_n\left(\frac{A_n(s)}{n}, x\right), \end{aligned} \tag{43}$$

where

$$K_n(t,x) \equiv \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}(\eta_i \leq x), \quad t \geq 0, \quad x \geq 0 \ , \tag{44}$$

is a sequential empirical process (a random field, having two parameters), so that

$$K_n\left(\frac{A_n(t)}{n}, x\right) = \frac{1}{n} \sum_{i=1}^{A_n(t)} \mathbf{1}(\eta_i \leq x), \quad t \geq 0, \quad x \geq 0. \tag{45}$$

The division by n in (44) provides a law-of-large-numbers (LLN) or fluid-limit scaling. To proceed, we define associated queueing processes with LLN scaling. In particular, define the normalized processes $\bar{Q}_n \equiv \{\bar{Q}_n(t), t \geq 0\}$, $\bar{A}_n \equiv \{\bar{A}_n(t), t \geq 0\}$ and $\bar{D}_n \equiv \{\bar{D}_n(t), t \geq 0\}$ as

$$\bar{Q}_n(t) \equiv \frac{1}{n} Q_n(t), \quad \bar{A}_n(t) \equiv \frac{1}{n} A_n(t), \quad \bar{D}_n(t) \equiv \frac{1}{n} D_n(t), \quad t \geq 0 \; . \tag{46}$$

For our general arrival process, we assume that $\bar{A}_n(t) \to a(t) \equiv \mu t$ w.p.1 as $n \to \infty$. For the $M/M/\infty$ special case, that follows from (1).

Next write equation (45) as

$$\begin{aligned} K_n\left(\frac{A_n(t)}{n}, x\right) &= \frac{1}{\sqrt{n}}\Big[\frac{1}{\sqrt{n}} \sum_{i=1}^{A_n(t)} (\mathbf{1}(\eta_i \leq x) - F(x))\Big] + \frac{1}{n} A_n(t) F(x) \\ &= \frac{1}{\sqrt{n}}\Big[\frac{1}{\sqrt{n}} \sum_{i=1}^{A_n(t)} (\mathbf{1}(\eta_i \leq x) - F(x))\Big] \\ &\quad + \frac{1}{\sqrt{n}}\Big[\sqrt{n}(\bar{A}_n(t) - a(t))\Big] F(x) + a(t) F(x). \end{aligned}$$

Now we introduce stochastic processes with central-limit-theorem (CLT) scaling. In particular, let

$$V_n(t,x) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{A_n(t)} (\mathbf{1}(\eta_i \leq x) - F(x)),$$

and

$$\hat{A}_n(t) \equiv \sqrt{n}(\bar{A}_n(t) - a(t)). \tag{47}$$

Then

$$K_n(\bar{A}_n(t), x) = \frac{1}{\sqrt{n}} V_n(t,x) + \frac{1}{\sqrt{n}} \hat{A}_n(t) F(x) + a(t) F(x),$$

so that the process Q_n in (43) can be written as

$$\begin{aligned} Q_n(t) &= \sum_{i=1}^{Q_n(0)} (\mathbf{1}(\bar{\eta}_i > t) - F_0^c(t)) + Q_n(0) F_0^c(t) \\ &\quad + \sqrt{n} \int_0^t \int_0^\infty \mathbf{1}(s + x > t) dV_n(s,x) \\ &\quad + \sqrt{n} \int_0^t \int_0^\infty \mathbf{1}(s + x > t) d\hat{A}_n(s) dF(x) \\ &\quad + n \int_0^t \int_0^\infty \mathbf{1}(s + x > t) da(s) dF(x) \\ &= \sum_{i=1}^{Q_n(0)} (\mathbf{1}(\bar{\eta}_i > t) - F_0^c(t)) + Q_n(0) F_0^c(t) \qquad (48) \\ &\quad + \sqrt{n} \int_0^t \int_0^\infty \mathbf{1}(s + x > t) dV_n(s,x) \\ &\quad + \sqrt{n} \int_0^t F^c(t-s) d\hat{A}_n(s) + n \int_0^t F^c(t-s) da(s) \end{aligned}$$

$$
\begin{aligned}
= &\sum_{i=1}^{Q_n(0)} (\mathbf{1}(\bar{\eta}_i > t) - F_0^c(t)) + Q_n(0)F_0^c(t) + n\int_0^t F^c(t-s)da(s) \\
&+\sqrt{n}(M_{n,1}(t) - M_{n,2}(t)) \ ,
\end{aligned}
$$

where

$$
M_{n,1}(t) \equiv \int_0^t F^c(t-s)d\hat{A}_n(s) \tag{49}
$$

and

$$
\begin{aligned}
M_{n,2}(t) &\equiv -\int_0^t \int_0^\infty \mathbf{1}(s+x>t)dV_n(s,x) \\
&= \int_0^t \int_0^\infty \mathbf{1}(s+x \le t)dV_n(s,x).
\end{aligned} \tag{50}
$$

In contrast to previous representations, note that, except for $Q_n(0)$ which can be regarded as known, $Q_n(s)$ for $s < t$ does not appear on the righthand side of representation (48). Instead of the integral representations in Theorems 3.4 and 3.5, here we have a direct expression of $Q_n(t)$ in terms of other model elements, but we will see that some of these model elements in turn do have integral representations.

By equations (46) and (48), we have

$$
\begin{aligned}
\bar{Q}_n(t) = &\frac{1}{n}\sum_{i=1}^{Q_n(0)} (\mathbf{1}(\bar{\eta}_i > t) - F_0^c(t)) + \bar{Q}_n(0)F_0^c(t) + \int_0^t F^c(t-s)da(s) \\
&+\frac{1}{\sqrt{n}}(M_{n,1}(t) - M_{n,2}(t)), \quad t \ge 0.
\end{aligned} \tag{51}
$$

From equation (51), we can prove the following FWLLN. We remark that we could allow more general limit functions a for the LLN-scaled arrival process.

Theorem 3.6 (FWLLN for the fourth martingale representation) *If there is convergence $(\bar{Q}_n(0), \bar{A}_n) \Rightarrow (q(0), a)$ in $\mathbb{R} \times D$ as $n \to \infty$, where $a(t) \equiv \mu t$, $t \ge 0$, then $\bar{Q}_n \Rightarrow q$, where*

$$
q(t) = q(0)F_0^c(t) + \int_0^t F^c(t-s)da(s), \quad t \ge 0 \ . \tag{52}
$$

For the $M/M/\infty$ special case,

$$
q(t) = q(0)e^{-\mu t} + \mu\int_0^t e^{-\mu(t-s)}ds = 1 - (1-q(0))e^{-\mu t}, \quad t \ge 0.
$$

If, in addition, $q(0) = 1$, then $q(t) = 1$ for $t \ge 0$.

Let the scaled process X_n be defined by

$$X_n(t) = \sqrt{n}(\bar{Q}_n(t) - q(t)), \quad t \geq 0. \tag{53}$$

If $q(t) = 1$ for $t \geq 0$, then (53) coincides with (3). By equations (53), (51) and (52), we obtain the following theorem for the scaled processes.

Theorem 3.7 (fourth martingale representation for the scaled processes) *The scaled process X_n in* (53) *has the representation*

$$\begin{aligned} X_n(t) \quad = \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^{Q_n(0)} (\mathbf{1}(\bar{\eta}_i > t) - F_0^c(t)) + \sqrt{n}(\bar{Q}_n(0) - q(0))F_0^c(t) \\ & + M_{n,1}(t) - M_{n,2}(t), \quad t \geq 0, \end{aligned} \tag{54}$$

where $M_{n,1}$ and $M_{n,2}$ are defined as in (49) *and* (50)*, respectively.*

The situation is more complicated here, because the processes $M_{n,1}$ and $M_{n,2}$ in (49) and (50) are not naturally martingales for the $G/GI/\infty$ model or even the $M/M/\infty$ special case, with respect to the obvious filtration, but they can be analyzed by martingale methods. In particular, associated martingales can be exploited to establish stochastic-process limits. In particular, the proof of the FCLT for the processes X_n in (53) - see §6.4 - exploits semimartingale decompositions of the following two-parameter process $U_n \equiv \{U_n(t,x), t \geq 0, 0 \leq x \leq 1\}$ (and related martingale properties):

$$U_n(t,x) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}(\zeta_i \leq x) - x) \ , \tag{55}$$

where the ζ_i are independent and uniformly distributed on $[0,1]$.

Extending Bickel and Wichura [7], Krichagina and Puhalskii [37] proved that the sequence of processes $\{U_n, n \geq 1\}$ converges in distribution to the Kiefer process U in $D([0,\infty), D([0,1]))$. For properties of Kiefer processes, we refer to Csörgó M. and P. Révéz [13] and Khoshnevisan [35]. The importance of the Kiefer process for infinite-server queues was evidently first observed by Louchard [45].

The process U_n has the following semimartingale decomposition (See Chapter IX of Jacod and Shiryaev [30]):

$$U_n(t,x) = -\int_0^x \frac{U_n(t,y)}{1-y} dy + M_{n,0}(t,x),$$

where

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

is a square-integrable martingale relative to the filtration $F^n = \bigvee_{i \leq \lfloor nt \rfloor} \mathcal{F}^i(x)$ and $\mathcal{F}^i(x) = \sigma(\mathbf{1}(\zeta_i \leq y), 0 \leq y \leq x) \vee \mathcal{N}$ for all $x \in [0,1]$.

Hence $V_n(t,x) = U_n(a_n(t), F(x))$ can be written as

$$V_n(t,x) = -\int_0^x \frac{V_{n,-}(t,y)}{1-F_-(y)} dF(y) + L_n(t,x) \ ,$$

where

$$L_n(t,x) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{A_n(t)} (\mathbf{1}(\eta_i \le x) - \int_0^{x\wedge\eta_i} \frac{1}{1-F_-(y)} dF(y),$$

$F_-(y)$ is the left-continuous version of F, $F_-(0) \equiv 0$ and $V_{n,-}$ is the left-continuous version of V_n in the second argument. Therefore, $M_{n,2}$ can be written as

$$M_{n,2}(t) = G_n(t) + H_n(t) \ ,$$

where

$$\begin{aligned} G_n(t) &\equiv \int_0^t \int_0^\infty \mathbf{1}(s+x \le t) d\Big(-\int_0^x \frac{V_{n,-}(s,y)}{1-F_-(y)} dF(y)\Big) \\ &= -\int_0^t \frac{V_{n,-}(t-x,x)}{1-F_-(x)} dF(x) \end{aligned}$$

and

$$H_n(t) \equiv \int_0^t \int_0^\infty \mathbf{1}(s+x \le t) dL_n(s,x).$$

In closing this subsection, we remark that an associated representation holds for the two-parameter process $Q(t,y)$ in (17). Let the associated scaled two-parameter process be defined by

$$X_n(t,y) \equiv \sqrt{n}(\bar{Q}_n(t,y) - q(t,y)), \quad t \ge 0, \tag{56}$$

where $\bar{Q}_n(t,y) \equiv Q_n(t,y)/n$, $\bar{Q}_n \Rightarrow q$ as $n \to \infty$ and

$$q(t,y) = q(0)F_0^c(t) + \int_0^{t-y} F^c(t-s)\, da(s) \ . \tag{57}$$

Corollary 3.1 (associated representation for the scaled two-parameter processes) *Paralleling* (54)*, the scaled process in* (56) *has the representation*

$$\begin{aligned} X_n(t,y) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{Q_n(0)} (\mathbf{1}(\bar{\eta}_i > t) - F_0^c(t)) + \sqrt{n}(\bar{Q}_n(0) - q(0))F_0^c(t) \\ &\quad + M_{n,1}(t,y) - M_{n,2}(t,y), \quad t \ge 0, \end{aligned} \tag{58}$$

where, paralleling (49) *and* (50)*,*

$$M_{n,1}(t,y) = \int_0^{t-y} F^c(t-s) d\hat{A}_n(s)$$

and

$$M_{n,2}(t,y) = \int_0^{t-y} \int_0^\infty \mathbf{1}(s+x \le t) dV_n(s,x).$$

4. Main Steps in the Proof of Theorem 1.1

In this section we indicate the main steps in the proof of Theorem 1.1 starting from one of the first three martingale representations in the previous section. First, in §4.1 we show that the integral representation appearing in both Theorems 3.4 and 3.5 has a unique solution, so that it constitutes a continuous function from D to D. Next, in §4.2 we show how the limit can be obtained from the functional central limit theorem for the Poisson process and the continuous mapping theorem, but a fluid limit (Lemma 4.2 or Lemma 4.3) remains to be verified. In §4.3 we show how the proof can be completed without martingales by directly establishing that associated fluid limit. In §§5-6 we show how martingales can achieve the same result. In §6.4 we indicate how to complete the proof with the fourth martingale representation.

4.1. Continuity of the Integral Representation

We apply the **continuous-mapping theorem (CMT)** with the integral representation in (32) and (40) in order to establish the desired convergence; for background on the CMT, see §3.4 of [69]. In subsequent sections we will show that the scaled martingales converge weakly to independent Brownian motions, i.e.,

$$(M_{n,1}, M_{n,2}) \Rightarrow (\sqrt{\mu}B_1, \sqrt{\mu}B_2) \quad \text{in} \quad D^2 \equiv D \times D \quad \text{as} \quad n \to \infty \;, \tag{59}$$

where B_1 and B_2 are two independent standard Brownian motions, from which an application of the CMT with subtraction yields

$$M_{n,1} - M_{n,2} \Rightarrow \sqrt{\mu}B_1 - \sqrt{\mu}B_2 \overset{\mathrm{d}}{=} \sqrt{2\mu}B \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \;, \tag{60}$$

where B is a single standard Brownian motion.

We then apply the CMT with the function $f : D \times \mathbb{R} \to D$ taking (y, b) into x determined by the integral representation

$$x(t) = b + y(t) - \mu \int_0^t x(s)\, ds \;, \quad t \geq 0 \;. \tag{61}$$

In the pre-limit, the function y in (61) is played by $M_{n,1} - M_{n,2} \equiv \{M_{n,1}(t) - M_{n,2}(t) : t \geq 0\}$ in (60), while b is played by $X_n(0)$. In the limit, the function y in (61) is played by the limit $\sqrt{2\mu}B$ in (60), while b is played by $X(0)$. (The constant b does not play an essential role in (61); it is sometimes convenient when we want to focus on the solution x as a function of the initial conditions.)

For our application, the limiting stochastic process in (60) has continuous sample paths. Moreover, the function f in (61) maps continuous functions into continuous functions, as we show below. Hence, it suffices to show that the map $f : D \times \mathbb{R} \to D$ is measurable and continuous at continuous limits. Since the limit is necessarily continuous as well, the required continuity follows from continuity when the function space D appearing in both the domain and the range is endowed with the topology of uniform convergence on bounded intervals. However, if we only establish such continuity, then that leaves open the issue of measurability. It is significant that the σ field on D generated by the topology of uniform convergence on bounded intervals is not the desired customary σ field on D, which is generated by the coordinate projections or by any of the Skorohod topologies; see §11.5 of [69] and §18 of Billingsley [8]. We prove measurability with respect to the appropriate σ field on D (generated by the J_1 topology) by proving continuity when the function space D appearing in both the domain and the range is endowed with the Skorohod J_1 topology. That implies the required measurability. At the same time, of course, it provides continuity in that setting.

We now establish the basic continuity result. We establish a slightly more general form than needed here in order to be able to treat other cases. In particular, we introduce a Lipschitz function $h : \mathbb{R} \to \mathbb{R}$; i.e., we assume that there exists a constant $c > 0$ such that

$$|h(s_1) - h(s_2)| \leq c|s_1 - s_2| \quad \text{for all} \quad s_1, s_2 \in \mathbb{R} \ . \tag{62}$$

We apply the more general form to treat the Erlang A model in §7.1. Theorem 7.3 in §7.2 involves an even more general version in which $h : D \to D$.

Theorem 4.1 (*continuity of the integral representation*) *Consider the integral representation*

$$x(t) = b + y(t) + \int_0^t h(x(s))\, ds \ , \quad t \geq 0 \ , \tag{63}$$

where $h : \mathbb{R} \to \mathbb{R}$ satisfies $h(0) = 0$ and is a Lipschitz function as defined in (62). *The integral representation in* (63) *has a unique solution x, so that the integral representation constitutes a function $f : D \times \mathbb{R} \to D$ mapping (y, b) into $x \equiv f(y, b)$. In addition, the function f is continuous provided that the function space D* (*in both the domain and range*) *is endowed with either:* (*i*) *the topology of uniform convergence over bounded intervals or* (*ii*) *the Skorohod J_1 topology. Moreover, if y is continuous, then so is x.*

Proof. If y is a piecewise-constant function, then we can directly construct the solution x of the integral representation by doing an inductive construction, just as in Lemma 2.1. Since any element y of D can be represented as the limit of piecewise-constant functions, where the convergence is uniform over bounded intervals, using endpoints that are continuity points of y, we can then extend the function f to arbitrary elements of D, exploiting continuity in the topology of uniform convergence over bounded intervals, shown below. Uniqueness follows from the fact that the only function x in D satisfying the inequality

$$|x(t)| \leq c \int_0^t |x(s)|\, ds, \quad t \geq 0 \ ,$$

is the zero function, which is a consequence of Gronwall's inequality, which we re-state in Lemma 4.1 below in the form needed here.

For the remainder of the proof, we apply Gronwall's inequality again. We introduce the norm

$$||x||_T \equiv \sup_{0 \le t \le T} |x(t)| \ .$$

First consider the case of the topology of uniform convergence over bounded intervals. We need to show that, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $||x_1 - x_2||_T < \epsilon$ when $|b_1 - b_2| + ||y_1 - y_2||_T < \delta$, where (y_i, x_i) are two pairs of functions satisfying the relation (63). From (63), we have

$$\begin{aligned} |x_1(t) - x_2(t)| &\le |b_1 - b_2| + |y_1(t) - y_2(t)| + \int_0^t |h(x_1(s)) - h(x_2(s))| \, ds \ , \\ &\le |b_1 - b_2| + |y_1(t) - y_2(t)| + c \int_0^t |x_1(s) - x_2(s)| \, ds \ . \qquad (64) \end{aligned}$$

Suppose that $|b_1 - b_2| + ||y_1 - y_2||_T \le \delta$. By Gronwall's inequality,

$$|x_1(t) - x_2(t)| \le \delta e^{ct} \quad \text{and} \quad ||x_1 - x_2||_T \le \delta e^{cT} \ .$$

Hence it suffices to let $\delta = \epsilon e^{-cT}$.

We now turn to the Skorohod J_1 topology; see §§3.3 and 11.5 and Chapter 12 of [69] for background. To treat this non-uniform topology, we will use the fact that the function x is necessarily bounded. That is proved later in Lemma 5.5. We want to show that $x_n \to x$ in $D([0, \infty), \mathbb{R}, J_1)$ when $b_n \to b$ in $\mathbb{R}$ and $y_n \to y$ in $D([0, \infty), \mathbb{R}, J_1)$. For y given, let the interval right endpoint T be a continuity point of y. Then there exist increasing homeomorphisms λ_n of the interval $[0, T]$ such that $\|y_n - y \circ \lambda_n\|_T \to 0$ and $\|\lambda_n - e\|_T \to 0$ as $n \to \infty$. Moreover, it suffices to consider homeomorphisms λ_n that are absolutely continuous with respect to Lebesgue measure on $[0, T]$ having derivatives $\dot{\lambda}_n$ satisfying $\|\dot{\lambda}_n - 1\|_T \to 0$ as $n \to \infty$. The fact that the topology is actually unchanged is a consequence of Billingsley's equivalent complete metric d_0 on pp 112–114 of Billingsley [8]. Hence, for y given, let $M \equiv \sup_{0 \le t \le T} \{|x(t)|\}$. Since h in (62) is Lipschitz, we have

$$\sup_{0 \le t \le T} \{|h(x(t))|\} \le h(0) + \sup_{0 \le t \le T} \{|h(x(t)) - h(0)|\} \le h(0) + cM = cM \ .$$

Thus we have

$$\begin{aligned} |x_n(t) - x(\lambda_n(t))| &\le |b_n - b| + \|y_n - y \circ \lambda_n\|_T \\ &\qquad + \left| \int_0^t h(x_n(u)) \, du - \int_0^{\lambda_n(t)} h(x(u)) \, du \right| \\ &\le |b_n - b| + \|y_n - y \circ \lambda_n\|_T \\ &\qquad + \left| \int_0^t h(x_n(u)) \, du - \int_0^t h(x(\lambda_n(u))) \dot{\lambda}_n(u) \, du \right| \end{aligned}$$

$$\begin{aligned}
&\leq \quad |b_n - b| + \|y_n - y \circ \lambda_n\|_T + \|\dot{\lambda}_n - 1\|_T \int_0^T |h(x(u))|\, du \\
&\qquad + \int_0^t |h(x_n(u)) - h(x(\lambda_n(u)))|\, du \\
&\leq \quad |b_n - b| + \|y_n - y \circ \lambda_n\|_T + \|\dot{\lambda}_n - 1\|_T (cMT) \\
&\qquad + c\int_0^t |x_n(u) - x(\lambda_n(u))|\, du \ .
\end{aligned}$$

Choose n_0 such that $\|\dot{\lambda}_n - 1\|_T < \delta/(2cMT)$ and $|b_n - b| + \|y_n - y \circ \lambda_n\|_T < \delta/2$. Then Gronwall's inequality implies that

$$|x_n(t) - x(\lambda_n(t))| \leq \delta e^{ct} \quad \text{for all} \quad t, \quad 0 \leq t \leq T \ ,$$

so that

$$\|x_n - x \circ \lambda_n\|_T \leq \delta e^{cT} \ .$$

Hence, for $\epsilon > 0$ given, choose $\delta < \epsilon e^{-cT}$ to have $\|x_n - x \circ \lambda_n\|_T \leq \epsilon$ for $n \geq n_0$. If necessary, choose n larger to make $\|\lambda_n - e\|_T < \epsilon$ and $\|\dot{\lambda}_n - 1\|_T < \epsilon$ as well. Finally, for the inheritance of continuity, note that

$$x(t+s) - x(t) = y(t+s) - y(t) + \int_t^{t+s} h(x(u))\, du \ ,$$

so that

$$|x(t+s) - x(t)| \leq |y(t+s) - y(t)| + \int_t^{t+s} |h(x(u))|\, du \ .$$

Since x is bounded over $[0, T]$, x is continuous if y is continuous. ■

In our case we can simply let $h(s) = \mu s$, but we will need the more complicated function h in (63) and (62) in §7.1. To be self-contained, we now state a version of Gronwall's inequality; see p. 498 of [19]. See §11 of [46] for other versions of Gronwall's inequality.

Lemma 4.1 (*version of Gronwall's inequality*) *Suppose that* $g : [0, \infty) \to [0, \infty)$ *is a Borel-measurable function such that*

$$0 \leq g(t) \leq \epsilon + M \int_0^t g(s)\, ds, \quad 0 \leq t \leq T \ ,$$

for some positive finite ϵ *and* M. *Then*

$$g(t) \leq \epsilon e^{Mt}, \quad 0 \leq t \leq T \ .$$

It thus remains to establish the limit in (59). Our proof based on the first martingale representation in Theorem 3.4 relies on a FCLT for the Poisson process and the CMT with the composition map. The application of the CMT with the composition map requires a fluid limit, which requires further argument. That is contained in subsequent sections.

4.2. Poisson FCLT Plus the CMT

As a consequence of the last section, it suffices to show that the scaled martingales converge, as in (59). From the martingale perspective, it is natural to achieve that goal by directly applying the martingale FCLT, as in §7.1 of Ethier and Kurtz [19], and as reviewed here in §8, and that works. In particular, the desired limit (59) follows from Theorems 3.4 and 8.1 (ii) (or Theorems 3.5 and 8.1 (ii)) plus Lemma 4.2 below. Lemma 4.2 shows that the scaled predictable quadratic variation processes in (33) and (41) converge, as required in condition (133) of Theorem 8.1 here; see §9.2.

However, starting with the first martingale representation in Theorem 3.4, we do not need to apply the martingale FCLT. Instead, we can justify the martingale limit in (59) by yet another application of the CMT, using the composition map associated with the random time changes, in addition to a functional central limit theorem (FCLT) for scaled Poisson processes. Our approach also requires establishing a limit for the sequence of scaled predictable quadratic variations associated with the martingales, so the main steps of the argument become the same as when applying the martingale FCLT.

The FCLT for Poisson processes is a classical result. It is a special case of the FCLT for a renewal process, appearing as Theorem 17.3 in Billingsley [8]. It and generalizations are also discussed extensively in [69]; see §§6.3, 7.3, 7.4, 13.7 and 13.8. The FCLT for a Poisson process can also be obtained via a strong approximation, as was done by Kurtz [39], Mandelbaum and Pats [47, 48] and Mandelbaum, Massey and Reiman [46]. Finally, the FCLT for a Poisson process itself can be obtained as an easy application of the martingale FCLT, as we show in §8.

We start with the scaled Poisson processes

$$M_{A,n}(t) \equiv \frac{A(nt)-nt}{\sqrt{n}} \quad \text{and} \quad M_{S,n}(t) \equiv \frac{S(nt)-nt}{\sqrt{n}}, \quad t \ge 0 \; . \tag{65}$$

We employ the following basic FCLT: Since A and S are independent rate-1 Poisson processes, we have

Theorem 4.2 (*FCLT for independent Poisson processes*) *If A and S are independent rate-1 Poisson processes, then*

$$(M_{A,n}, M_{S,n}) \Rightarrow (B_1, B_2) \quad in \quad D^2 \equiv D \times D \quad as \quad n \to \infty \; , \tag{66}$$

where $M_{A,n}$ and $M_{S,n}$ are the scaled processes in (65), while B_1 and B_2 are independent standard Brownian motions.

We can prove the desired limit in (59) for both martingale representations, but we will only give the details for the first martingale representation in Theorem 3.4. In order to get the desired limit in (59), we introduce a deterministic and a random time change. For that purpose, let $e : [0,\infty) \to [0,\infty)$ be the identity function in D, defined by $e(t) \equiv t$ for $t \ge 0$. Then let

$$\Phi_{A,n}(t) \quad \equiv \quad \frac{\lambda_n t}{n} = \mu t \equiv (\mu e)(t),$$

$$\Phi_{S,n}(t) \quad \equiv \quad \frac{\mu}{n}\int_0^t Q_n(s)\, ds, \quad t \ge 0 \ . \tag{67}$$

We will establish the following fluid limit, which can be regarded as a functional weak law of large numbers (FWLLN). Here below, and frequently later, we have convergence in distribution to a deterministic limit; that is **equivalent to convergence in probability**; see p. 27 of [8].

Lemma 4.2 (*desired fluid limit*) *Under the conditions of Theorem* 1.1*,*

$$\Phi_{S,n} \Rightarrow \mu e \quad in \quad D \quad as \quad n \to \infty \ , \tag{68}$$

where $\Phi_{S,n}$ *is defined in* (67)*.*

For that purpose, it suffices to establish another more basic fluid limit. Consider the stochastic process

$$\Psi_{S,n}(t) \equiv \frac{Q_n(t)}{n} \ , \quad t \ge 0 \ . \tag{69}$$

Let ω be the function that is identically 1 for all t.

Lemma 4.3 (*basic fluid limit*) *Under the conditions of Theorem* 1.1*,*

$$\Psi_{S,n} \Rightarrow \omega \quad in \quad D \quad as \quad n \to \infty \ , \tag{70}$$

where $\Psi_{S,n}$ *is defined in* (69) *and* $\omega(t) = 1$*,* $t \ge 0$*.*

Proof of Lemma 4.2. The desired fluid limit in Lemma 4.2 follows from the basic fluid limit in Lemma 4.3 by applying the CMT with the function $h : D \to D$ defined by

$$h(x)(t) \equiv \mu \int_0^t x(s)\, ds \ , \quad t \ge 0 \ . \quad \blacksquare \tag{71}$$

We thus have the following result

Lemma 4.4 (all but the fluid limit) *If the limit in* (70) *holds, then*

$$(M_{n,1}, M_{n,2}) \Rightarrow (\sqrt{\mu} B_1, \sqrt{\mu} B_2) \quad in \quad D^2 \tag{72}$$

as required to complete the proof of Theorem 1.1*.*

Proof. From the limit in (66), the desired fluid limit in (68) and Theorem 11.4.5 of [69], it follows that

$$(M_{A,n}, \mu e, M_{S,n}, \Phi_{S,n}) \Rightarrow (B_1, \mu e, B_2, \mu e) \quad \text{in} \quad D^4 \equiv D \times \cdots \times D \tag{73}$$

as $n \to \infty$. From the CMT with the composition map, as in §§3.4 and 13.2 of [69] - in particular, with Theorem 13.2.1 - we obtain the desired limit in (59):

$$(M_{n,1}, M_{n,2}) \equiv (M_{A,n} \circ \mu e, M_{S,n} \circ \Phi_{S,n}) \Rightarrow (B_1 \circ \mu e, B_2 \circ \mu e) \quad \text{in} \quad D^2 \tag{74}$$

as $n \to \infty$. By basic properties of Brownian motion,

$$(B_1 \circ \mu e, B_2 \circ \mu e) \stackrel{\mathrm{d}}{=} (\sqrt{\mu} B_1, \sqrt{\mu} B_2) \quad \text{in} \quad D^2 \ . \quad \blacksquare$$

It thus remains to establish the key fluid limit in Lemma 4.3. In the next section we show how to do that directly, without martingales, by applying the continuous mapping provided by Theorem 4.1 in the fluid scale or, equivalently, by applying Gronwall's inequality again. We would stop there if we only wanted to analyze the $M/M/\infty$ model, but in order to illustrate other methods used in Krichagina and Puhalskii [37] and Puhalskii and Reiman [53], we also apply martingale methods. Thus, in the subsequent four sections we show how to establish that fluid limit using martingales. Here is an **outline of the remaining martingale argument:**

(1) To prove the needed Lemma 4.3, it suffices to demonstrate that $\{X_n\}$ is stochastically bounded in D. (Combine Lemma 5.9 and §6.1.)

(2) However, $\{X_n\}$ is stochastically bounded in D if the sequences of martingales $\{M_{n,1}\}$ and $\{M_{n,2}\}$ are stochastically bounded in D. (Combine Theorem 3.4 and Lemma 5.5.)

(3) But then the sequences of martingales $\{M_{n,1}\}$ and $\{M_{n,2}\}$ are stochastically bounded in D if the associated sequences of predictable quadratic variations $\{\langle M_{n,1}\rangle(t)\}$ and $\{\langle M_{n,2}\rangle(t)\}$ are stochastically bounded in $\mathbb{R}$ for each $t > 0$ (Apply Lemma 5.8. One of these is trivial because it is deterministic.)

(4) Finally, we establish stochastic boundedness of $\{\langle M_{n,2}\rangle(t)\}$ (the one nontrivial case) through a crude bound in §6.2.

This alternate route to the fluid limit is much longer, but all the steps might be considered well known. We remark that the fluid limit seems to be required by any of the proofs, including by the direct application of the martingale FCLT.

4.3. Fluid Limit Without Martingales

In this section we prove Lemma 4.3 without using martingales. We do so by establishing a stochastic-process limit in the fluid scale which is similar to the corresponding stochastic-process limit with the more refined scaling. This is a standard line of reasoning for heavy-traffic stochastic-process limits; e.g., see the proofs of Theorems 9.3.4, 10.2.3 and 14.7.4 of Whitt [69]. The specific argument here follows §6 of Mandelbaum and Pats [47]. With this approach, even though we exploit the martingale representations, we do not need to mention martingales at all. We are only applying the continuous mapping theorem.

By essentially the same reasoning as in §3.4, we obtain a fluid-scale analog of (31) and (32):

$$\bar{X}_n(t) \quad \equiv \quad \frac{Q_n(t) - n}{n}$$

$$\begin{aligned} &= \frac{Q_n(0)-n}{n} + \frac{M^*_{n,1}(t)}{n} - \frac{M^*_{n,2}(t)}{n} + \frac{\lambda_n t - n\mu t}{n} \\ &\quad -\mu \int_0^t \left(\frac{Q_n(s)-n}{n} \right) ds \ , \\ &= \bar{X}_n(0) + \bar{M}_{n,1}(t) - \bar{M}_{n,2}(t) - \mu \int_0^t \bar{X}_n(s)\, ds \ , \quad t \ge 0 \ , \end{aligned} \tag{75}$$

where

$$\bar{M}_{n,1}(t) \equiv \frac{M^*_{n,1}(t)}{n} \quad \text{and} \quad \bar{M}_{n,2}(t) \equiv \frac{M^*_{n,2}(t)}{n}, \quad t \ge 0 \ , \tag{76}$$

with $M^*_{n,i}(t)$ defined in (29).

Notice that the limit

$$\bar{X}_n \Rightarrow \eta \quad \text{in} \quad D^k \quad \text{as} \quad n \to \infty \ , \tag{77}$$

where

$$\eta(t) \equiv 0, \quad t \ge 0 \ , \tag{78}$$

is equivalent to the desired conclusion of Lemma 4.3. Hence we will prove the fluid limit in (77).

The assumed limit in (4) implies that $\bar{X}_n(0) \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$. We can apply Theorem 4.1 or directly Gronwall's inequality in Lemma 4.1 to deduce the desired limit (77) if we can establish the following lemma.

Lemma 4.5 (*fluid limit for the martingales*) *Under the conditions of Theorem 1.1,*

$$\bar{M}_{n,i} \Rightarrow \eta \quad \textit{in} \quad D \quad \textit{w.p.1} \quad \textit{as} \quad n \to \infty \ , \tag{79}$$

for $i = 1, 2$, *where* $\bar{M}_{n,i}$ *is defined in* (76) *and* η *is defined in* (78).

Proof of Lemma 4.5. We can apply the SLLN for the Poisson process, which is equivalent to the more general functional strong law of large numbers (FSLLN); see §3.2 of [70]. (Alternatively, we could apply the FWLLN, which is a corollary to the FCLT.) First, the SLLN for the Poisson process states that

$$\frac{A(t)}{t} \to 1 \quad \text{and} \quad \frac{S(t)}{t} \to 1 \quad \text{w.p.1} \quad \text{as} \quad t \to \infty \ ,$$

which implies the corresponding FSLLN's

$$\sup_{0 \le t \le T} \{ \frac{A(nt)}{n} - t \} \to 0 \quad \text{and} \quad \sup_{0 \le t \le T} \{ \frac{S(nt)}{n} - t \} \to 0 \quad \text{w.p.1} \tag{80}$$

as $n \to \infty$ for each T with $0 < T < \infty$. We thus can treat $\bar{M}_{n,1}$ directly. To treat $\bar{M}_{n,2}$, we combine (80) with the crude inequality in (25) and the representation in (75)–(76) in order to obtain the desired limit (79). To elaborate, the crude inequality in (25) implies that, for any $T_1 > 0$, there exists T_2 such that

$$P\left(\frac{\mu}{n} \int_0^{T_1} Q_n(s)\, ds > T_2 \right) \to 0 \quad \text{as} \quad n \to \infty \ .$$

That provides the key, because

$$P\left(\|\bar{M}_{n,2}\|_{T_1} > \epsilon\right) \le P\left(\frac{\mu}{n}\int_0^{T_1} Q_n(s)\,ds > T_2\right) + P\left(\|\bar{S}_n\|_{T_2} > \epsilon/2\right) \ ,$$

where

$$\bar{S}_n(t) \equiv \frac{S(nt) - nt}{n}, \quad t \ge 0 \ . \ \blacksquare$$

5. Tightness and Stochastic Boundedness

5.1. Tightness

As indicated at the end of §4.2, we can also use a stochastic-boundedness argument in order to establish the desired fluid limit. Since stochastic boundedness is closely related to tightness, we start by reviewing tightness concepts. In the next section we apply the tightness notions to stochastic boundedness. The next three sections contain extra material not really needed for the current proofs. Additional material on tightness criteria appears in Whitt [73].

We work in the setting of a complete separable metric space (CSMS), also known as a Polish space; see §§13 and 19 of Billingsley [8], §§3.8-3.10 of Ethier and Kurtz [19] and §§11.1 and 11.2 of [69]. (The space $D^k \equiv D([0,\infty),\mathbb{R})^k$ is made a CSMS in a standard way and the space of probability measures on D^k becomes a CSMS as well.) Key concepts are: closed, compact, tight, relatively compact and sequentially compact. We assume knowledge of metric spaces and compactness in metric spaces.

Definition 5.1 (tightness) *A set A of probability measures on a metric space S is **tight** if, for all $\epsilon > 0$, there exists a compact subset K of S such that*

$$P(K) > 1 - \epsilon \quad \textit{for all} \quad P \in A \ .$$

A set of random elements of the metric space S is tight if the associated set of their probability laws on S is tight. Consequently, a sequence $\{X_n : n \ge 1\}$ of random elements of the metric space S is tight if, for all $\epsilon > 0$, there exists a compact subset K of S such that

$$P(X_n \in K) > 1 - \epsilon \quad \textit{for all} \quad n \ge 1 \ .$$

Since a continuous image of a compact subset is compact, we have the following lemma.

Lemma 5.1 (continuous functions of random elements) *Suppose that $\{X_n : n \ge 1\}$ is a tight sequence of random elements of the metric space S. If $f : S \to S'$ is a continuous function mapping the metric space S into another metric space S', then $\{f(X_n) : n \ge 1\}$ is a tight sequence of random elements of the metric space S'.*

Proof. As before, let $\circ$ be used for composition: $(f \circ g)(x) \equiv f(g(x))$. For any function $f : S \rightarrow S'$ and any subset A of S, $A \subseteq f^{-1} \circ f(A)$. Let $\epsilon > 0$ be given. Since $\{X_n : n \geq 1\}$ is a tight sequence of random elements of the metric space S, there exists a compact subset K of S such that

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

Then $f(K)$ will serve as the desired compact set in S', because

$$P(f(X_n) \in f(K)) = P(X_n \in (f^{-1} \circ f)(K)) \geq P(X_n \in K) > 1 - \epsilon$$

for all $n \geq 1$. ∎

We next observe that on products of separable metric spaces tightness is characterized by tightness of the components; see §11.4 of [69].

Lemma 5.2 (tightness on product spaces) *Suppose that $\{(X_{n,1}, \ldots, X_{n,k}) : n \geq 1\}$ is a sequence of random elements of the product space $S_1 \times \cdots \times S_k$, where each coordinate space S_i is a separable metric space. The sequence $\{(X_{n,1}, \ldots, X_{n,k}) : n \geq 1\}$ is tight if and only if the sequence $\{X_{n,i} : n \geq 1\}$ is tight for each i, $1 \leq i \leq k$.*

Proof. The implication from the random vector to the components follows from Lemma 5.1 because the component $X_{n,i}$ is the image of the projection map $\pi_i : S_1 \times \cdots \times S_k \rightarrow S_i$ taking $(x_1, \ldots, x_k)$ into x_i, and the projection map is continuous. Going the other way, we use the fact that

$$A_1 \times \cdots \times A_k = \bigcap_{i=1}^{k} \pi_i^{-1}(A_i) = \bigcap_{i=1}^{k} \pi_i^{-1} \circ \pi_i(A_1 \times \cdots \times A_k)$$

for all subsets $A_i \subseteq S_i$. Thus, for each i and any $\epsilon > 0$, we can choose K_i such that $P(X_{n,i} \notin K_i) < \epsilon/k$ for all $n \geq 1$. We then let $K_1 \times \cdots \times K_k$ be the desired compact for the random vector. We have

$$\begin{aligned} P\left((X_{n,1}, \ldots, X_{n,k}) \notin K_1 \times \cdots \times K_k\right) &= P\left(\bigcup_{i=1}^{k} \{X_{n,i} \notin K_i\}\right) \\ &\leq \sum_{i=1}^{k} P\left(X_{n,i} \notin K_i\right) \leq \epsilon \ . \quad \blacksquare \end{aligned}$$

Tightness goes a long way toward establishing convergence because of Prohorov's theorem. It involves the notions of sequential compactness and relative compactness.

Definition 5.2 (relative compactness and sequential compactness) *A subset A of a metric space S is **relatively compact** if every sequence $\{x_n : n \geq 1\}$ from A has a subsequence that converges to a limit in S (which necessarily belongs to the closure $\bar{A}$ of A).*

We can now state Prohorov's theorem; see §11.6 of [69]. It relates compactness of sets of measures to compact subsets of the underlying sample space S on which the probability measures are defined.

Theorem 5.1 (Prohorov's theorem) *A subset of probability measures on a CSMS is tight if and only if it is relatively compact.*

We have the following elementary corollaries:

Corollary 5.1 (convergence implies tightness) *If $X_n \Rightarrow X$ as $n \to \infty$ for random elements of a CSMS, then the sequence $\{X_n : n \geq 1\}$ is tight.*

Corollary 5.2 (individual probability measures) *Every individual probability measure on a CSMS is tight.*

As a consequence of Prohorov's Theorem, we have the following method for establishing convergence of random elements:

Corollary 5.3 (convergence in distribution via tightness) *Let $\{X_n : n \geq 1\}$ be a sequence of random elements of a CSMS S. We have*

$$X_n \Rightarrow X \quad \textit{in} \quad S \quad \textit{as} \quad n \to \infty$$

if and only if (i) the sequence $\{X_n : n \geq 1\}$ is tight and (ii) the limit of every convergent subsequence of $\{X_n : n \geq 1\}$ is the same fixed random element X (has a common probability law).

In other words, once we have established tightness, it only remains to show that the limits of all converging subsequences must be the same. With tightness, we only need to uniquely determine the limit. When proving Donsker's theorem, it is natural to uniquely determine the limit through the finite-dimensional distributions. Convergence of all the finite-dimensional distributions is not enough to imply convergence on D, but it does uniquely determine the distribution of the limit; see pp 20 and 121 of Billingsley [8] and Example 11.6.1 in [69].

This approach is applied to prove the martingale FCLT stated in §8; see [73]. In the martingale setting it is natural instead to use the martingale characterization of Brownian motion, originally established by Lévy [42] and proved by Ito's formula by Kunita and Watanabe [38]; see p. 156 of Karatzas and Shreve [33], and various extensions, such as to continuous processes with independent Gaussian increments, as in Theorem 1.1 on p. 338 of Ethier and Kurtz [19]. A thorough study of martingale characterizations appears in Chapter 4 of Liptser and Shiryayev [44] and in Chapters VIII and IX of Jacod and Shiryayev [30].

We have not discussed conditions to have tightness; they are reviewed in [73].

5.2. Stochastic Boundedness

We start by defining stochastic boundedness and relating it to tightness. We then discuss situations in which stochastic boundedness is preserved. Afterwards, we give conditions for a sequence of martingales to be stochastically bounded in D involving the stochastic boundedness of appropriate sequences of $\mathbb{R}$-valued random variables. Finally, we show that the FWLLN follows from stochastic boundedness.

5.2.1. Connection to Tightness

For random elements of $\mathbb{R}$ and $\mathbb{R}^k$, stochastic boundedness and tightness are equivalent, but tightness is stronger than stochastic boundedness for random elements of the functions spaces C and D (and the associated product spaces C^k and D^k).

Definition 5.3 (stochastic boundedness for random vectors) *A sequence $\{X_n : n \geq 1\}$ of random vectors taking values in $\mathbb{R}^k$ is **stochastically bounded (SB)** if the sequence is tight, as defined in Definition* 5.1*.*

The notions of tightness and stochastic boundedness thus agree for random elements of $\mathbb{R}^k$, but these notions differ for stochastic processes. For a function $x \in D^k \equiv D([0,\infty),\mathbb{R})^k$, let

$$\|x\|_T \equiv \sup_{0\leq t\leq T} \{|x(t)|\} ,$$

where $|b|$ is a norm of $b \equiv (b_1, b_2, \ldots, b_k)$ in $\mathbb{R}^k$ inducing the Euclidean topology, such as the maximum norm: $|b| \equiv \max\{|b_1|, |b_2|, \ldots, |b_k|\}$. (Recall that all norms on Euclidean space $\mathbb{R}^k$ are equivalent.)

Definition 5.4 (stochastic boundedness for random elements of D^k) *A sequence $\{X_n : n \geq 1\}$ of random elements of D^k is **stochastically bounded in** D^k if the sequence of real-valued random variables $\{\|X_n\|_T : n \geq 1\}$ is stochastically bounded in $\mathbb{R}$ for each $T > 0$, using Definition* 5.3*.*

For random elements of D^k, tightness is a strictly stronger concept than stochastic boundedness. Tightness of $\{X_n\}$ in D^k implies stochastic boundedness, but not conversely; see §15 of Billingsely [8]. However, stochastic boundedness is sufficient for us, because it alone implies the desired fluid limit.

5.2.2. Preservation

We have the following analog of Lemma 5.2, which characterizes tightness for sequences of random vectors in terms of tightness of the associated sequences of components.

Lemma 5.3 (stochastic boundedness on D^k via components) *A sequence*

$$\{(X_{n,1}, \ldots, X_{n,k}) : n \geq 1\} \quad in \quad D^k \equiv D \times \cdots \times D$$

is stochastically bounded in D^k if and only if the sequence $\{X_{n,i} : n \geq 1\}$ is stochastically bounded in $D \equiv D^1$ for each i, $1 \leq i \leq k$.

Proof. Assume that we are using the maximum norm on product spaces. We can apply Lemma 5.2 after noticing that

$$\|(x_1, \ldots, x_k)\|_T = \max\{\|x_i\|_T : 1 \le i \le k\}$$

for each element $(x_1, \ldots, x_k)$ of D^k. Since other norms are equivalent, the result applies more generally. ∎

Lemma 5.4 (stochastic boundedness in D^k for sums) *Suppose that*

$$Y_n(t) \equiv X_{n,1}(t) + \cdots + X_{n,k}(t), \quad t \ge 0,$$

for each $n \ge 1$, where $\{(X_{n,1}, \ldots, X_{n,k}) : n \ge 1\}$ is a sequence of random elements of the product space $D^k \equiv D \times \cdots \times D$. If $\{X_{n,i} : n \ge 1\}$ is stochastically bounded in D for each i, $1 \le i \le k$, then the sequence $\{Y_n : n \ge 1\}$ is stochastically bounded in D.

Note that the converse is not true: We could have $k = 2$ with $X_{n,2}(t) = -X_{n,1}(t)$ for all n and t. In that case we have $Y_n(t) = 0$ for all $X_{n,1}(t)$.

We now provide conditions for the stochastic boundedness of integral representations such as (32).

Lemma 5.5 (stochastic boundedness for integral representations) *Suppose that*

$$X_n(t) \equiv X_n(0) + Y_{n,1}(t) + \cdots + Y_{n,k}(t) + \int_0^t h(X_n(s))\, ds, \quad t \ge 0 \ ,$$

where h is a Lipschitz function as in (62) and $(X_n(0), (Y_{n,1}, \ldots, Y_{n,k}))$ is a random element of $\mathbb{R} \times D^k$ for each $n \ge 1$. If the sequences $\{X_n(0) : n \ge 1\}$ and $\{Y_{n,i} : n \ge 1\}$ are stochastically bounded (in $\mathbb{R}$ and D, respectively,) for $1 \le i \le k$, then the sequence $\{X_n : n \ge 1\}$ is stochastically bounded in D.

Proof. For the proof here, it is perhaps easiest to imitate the proof of Theorem 4.1. In particular, we will construct the following bound

$$||X_n||_T \le Ke^{cT},$$

assuming that $|X_n(0)| + ||Y_{n,1}||_T + \cdots + ||Y_{n,k}||_T + T|h(0)| \le K$. As before, we apply Gronwall's inequality, Lemma 4.1. Inserting absolute values into all the terms of the integral representation, we have

$$\begin{aligned} |X_n(t)| &\le |X_n(0)| + |Y_{n,1}(t)| + \cdots + |Y_{n,k}(t)| + \int_0^t |h(X_n(s))|\, ds, \\ &\le |X_n(0)| + \|Y_{n,1}\|_T + \cdots + \|Y_{n,k}\|_T + Th(0) + c\int_0^t |X_n(s)|\, ds \ , \end{aligned}$$

for $t \ge 0$, where c is the Lipschitz modulus. Suppose that $|X_n(0)| + ||Y_{n,1}||_T + \cdots + ||Y_{n,k}||_T + T|h(0)| \le K$. By Gronwall's inequality,

$$|X_n(t)| \le Ke^{ct} \quad \text{and} \quad ||X_n||_T \le Ke^{cT} \ . \ ∎$$

5.2.3. Stochastic Boundedness for Martingales

We now provide ways to get stochastic boundedness for sequences of martingales in D from associated sequences of random variables. Our first result exploits the classical submartingale-maximum inequality; e.g., see p. 13 of Karatzas and Shreve [33]. We say that a function $f : \mathbb{R} \to \mathbb{R}$ is *even* if $f(-x) = f(x)$ for all $x \in \mathbb{R}$.

Lemma 5.6 (SB from the maximum inequality) *Suppose that, for each $n \geq 1$, $M_n \equiv \{M_n(t) : t \geq 0\}$ is a martingale (with respect to a specified filtration) with sample paths in D. Also suppose that, for each $T > 0$, there exists an even nonnegative convex function $f : \mathbb{R} \to \mathbb{R}$ with first derivative $f'(t) > 0$ for $t > 0$ (e.g., $f(t) \equiv t^2$), there exists a positive constant $K \equiv K(T, f)$, and there exists an integer $n_0 \equiv n_0(T, f, K)$, such that*

$$E[f(M_n(T))] \leq K \quad \textit{for all} \quad n \geq n_0 \ .$$

Then the sequence of stochastic processes $\{M_n : n \geq 1\}$ is stochastically bounded in D.

Proof. Since any set of finitely many random elements of D is automatically tight, Theorem 1.3 of Billingsley [8], it suffices to consider $n \geq n_0$. Since f is continuous and $f'(t) > 0$ for $t > 0$, $t > c$ if and only if $f(t) > f(c)$ for $t > 0$. Since f is even,

$$E[f(M_n(t))] = E[f(|M_n(t)|)] \leq E[f(|M_n(T)|)] = E[f(M_n(T))] \leq K$$

for all t, $0 \leq t \leq T$. Since these moments are finite and f is convex, the stochastic process $\{f(M_n(t)) : 0 \leq t \leq T\}$ is a submartingale for each $n \geq 1$, so that we can apply the submartingale-maximum inequality to get

$$P(\|M_n\|_T > c) = P(\|f \circ M_n\|_T > f(c)) \leq \frac{E[f(M_n(T))]}{f(c)} \leq \frac{K}{f(c)}$$

for all $n \geq n_0$. Since $f(c) \to \infty$ as $c \to \infty$, we have the desired conclusion. ∎

We now establish another sufficient condition for stochastic boundedness of square-integrable martingales by applying the Lenglart-Rebolledo inequality; see p. 66 of Liptser and Shiryayev [44] or p. 30 of Karatzas and Shreve [33].

Lemma 5.7 (Lenglart-Rebolledo inequality) *Suppose that $M \equiv \{M(t) : t \geq 0\}$ is a square-integrable martingale (with respect to a specified filtration) with predictable quadratic variation $\langle M \rangle \equiv \{\langle M \rangle(t) : t \geq 0\}$, i.e., such that $M^2 - \langle M \rangle \equiv \{M(t)^2 - \langle M \rangle(t) : t \geq 0\}$ is a martingale by the Doob-Meyer decomposition. Then, for all $c > 0$ and $d > 0$,*

$$P\left(\sup_{0 \leq t \leq T} \{|M(t)|\} > c\right) \leq \frac{d}{c^2} + P\left(\langle M \rangle(T) > d\right) \ . \tag{81}$$

As a consequence we have the following criterion for stochastic boundedness of a sequence of square-integrable martingales.

Lemma 5.8 (SB criterion for square-integrable martingales) *Suppose that, for each $n \geq 1$, $M_n \equiv \{M_n(t) : t \geq 0\}$ is a square-integrable martingale (with respect to a specified filtration) with predictable quadratic variation $\langle M_n \rangle \equiv \{\langle M_n \rangle(t) : t \geq 0\}$, i.e., such that $M_n^2 - \langle M_n \rangle \equiv \{M_n(t)^2 - \langle M_n \rangle(t) : t \geq 0\}$ is a martingale by the Doob-Meyer decomposition. If the sequence of random variables $\{\langle M_n \rangle(T) : n \geq 1\}$ is stochastically bounded in $\mathbb{R}$ for each $T > 0$, then the sequence of stochastic processes $\{M_n : n \geq 1\}$ is stochastically bounded in D.*

Proof. For $\epsilon > 0$ given, apply the assumed stochastic boundedness of the sequence $\{\langle M_n \rangle(T) : n \geq 1\}$ to obtain a constant d such that

$$P\left(\langle M_n \rangle(T) > d\right) < \epsilon/2 \quad \text{for all} \quad n \geq 1 \ .$$

Then for that determined d, choose c such that $d/c^2 < \epsilon/2$. By the Lenglart-Rebolledo inequality (81), these two inequalities imply that

$$P\left(\sup_{0 \leq t \leq T} \{|M_n(t)|\} > c\right) < \epsilon \ . \quad \blacksquare$$

5.2.4. FWLLN from Stochastic Boundedness

We will want to apply stochastic boundedness in D to imply the desired fluid limit in Lemmas 4.2 and 4.3. The fluid limit corresponds to a **functional weak law of large numbers (FWLLN)**.

Lemma 5.9 (FWLLN from stochastic boundedness in D^k) *Let $\{X_n : n \geq 1\}$ be a sequence of random elements of D^k. Let $\{a_n : n \geq 1\}$ be a sequence of positive real numbers such that $a_n \to \infty$ as $n \to \infty$. If the sequence $\{X_n : n \geq 1\}$ is stochastically bounded in D^k , then*

$$\frac{X_n}{a_n} \Rightarrow \eta \quad in \quad D^k \quad as \quad n \to \infty \ , \tag{82}$$

where $\eta(t) \equiv (0, 0, \ldots, 0)$, $t \geq 0$.

Proof. As specified in Definition 5.4, stochastic boundedness of the sequence $\{X_n : n \geq 1\}$ in D^k corresponds to stochastic boundedness of the associated sequence $\{\|X_n\|_T : n \geq 1\}$ in $\mathbb{R}$ for each $T > 0$. By Definition 5.3, stochastic boundedness in $\mathbb{R}$ is equivalent to tightness. It is easy to verify directly that we then have tightness (or, equivalently, stochastic boundedness) for the associated sequence $\{\|X_n\|_T / a_n : n \geq 1\}$ in $\mathbb{R}$. By Prohorov's theorem, Theorem 5.1, tightness on $\mathbb{R}$ (or any CSMS) is equivalent to relative compactness. Hence consider a convergent subsequence $\{\|X_{n_k}\|_T / a_{n_k} : k \geq 1\}$ of the sequence $\{\|X_n\|_T / a_n : n \geq 1\}$ in $\mathbb{R}$: $\|X_{n_k}\|_T / a_{n_k} \to L$ as $k \to \infty$. It suffices to show that $P(L = 0) = 1$; then all convergent subsequences will have the same limit, which implies convergence to that limit. For that purpose, consider the associated subsequence $\{\|X_{n_k}\|_T : k \geq 1\}$ in $\mathbb{R}$. It too is tight. So by Prohorov's theorem again, it too is relatively compact. Thus there exists a convergent sub-subsequence: $\|X_{n_{k_l}}\|_T \Rightarrow L'$ in $\mathbb{R}$. It follows immediately that

$$\frac{\|X_{n_{k_l}}\|_T}{a_{n_{k_l}}} \Rightarrow 0 \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad l \to \infty \ . \tag{83}$$

This can be regarded as a consequence of the **generalized continuous mapping theorem (GCMT)**, Theorem 3.4.4 of [69], which involves a sequence of continuous functions: Consider the functions $f_n : \mathbb{R} \to \mathbb{R}$ defined by $f_n(x) \equiv x/a_n$, $n \geq 1$, and the limiting zero function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) \equiv 0$ for all x. It is easy to see that $f_n(x_n) \to f(x) \equiv 0$ whenever $x_n \to x$ in $\mathbb{R}$. Thus the GCMT implies the limit in (83). Consequently, the limit L we found for the subsequence $\{\|X_{n_k}\|_T/a_{n_k} : k \geq 1\}$ must actually be 0. Since that must be true for all convergent subsequences, we must have the claimed convergence in (82). ■

6. Completing the Proof of Theorem 1.1

In the next three subsections we complete the proof of Theorem 1.1 by the martingale argument, as outlined at the end of §4.2. In §6.1 we show that the required fluid limit in Lemma 4.3 follows from the stochastic boundedness of the the sequence of stochastic processes $\{X_n : n \geq 1\}$. In §6.2 we finish the proof of the fluid limit by proving that the associated predictable quadratic variation processes are stochastically bounded. Finally, in §6.3 we show how to remove the condition on the initial conditions.

6.1. Fluid Limit from Stochastic Boundedness in D

We now show how to apply stochastic boundedness to imply the desired fluid limit in Lemma 4.3. Here we simply apply Lemma 5.9 with $a_n \equiv \sqrt{n}$ for $n \geq 1$ to our particular sequence of stochastic processes $\{X_n : n \geq 1\}$ in (3) or (32).

Lemma 6.1 (application to queueing processes) *Let X_n be the random elements of D defined in* (3) *or* (32). *If the sequence $\{X_n : n \geq 1\}$ is stochastically bounded in D, then*

$$\|n^{-1}Q_n - \omega\|_T \Rightarrow 0 \quad \textit{as} \quad n \to \infty \quad \textit{for all} \quad T > 0 \ , \tag{84}$$

where $\omega(t) \equiv 1$, $t \geq 0$. Equivalently,

$$\sup_{0 \leq t \leq T} \{|(Q_n(t)/n) - 1|\} \Rightarrow 0 \quad \textit{as} \quad n \to \infty \quad \textit{for all} \quad T > 0 \ . \tag{85}$$

Proof. As a consequence of Lemma 5.9, From the stochastic boundedness of $\{X_n\}$ in D, we will obtain

$$\frac{X_n}{\sqrt{n}} \Rightarrow \eta \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ ,$$

where η is the zero function defined above. This limit is equivalent to (84) and (85). ■

In §4.2 we already observed that Lemma 4.3 implies the desired Lemma 4.2. So we have completed the required proof.

Recap: How this all goes together. The consequence of the last three sections is as follows: In order to complete the martingale argument to establish the required fluid limit in Lemmas 4.2 and 4.3, it suffices to establish the stochastic boundedness of two sequences of random variables: $\{\langle M_{n,1}\rangle(T) : n \geq 1\}$ and $\{\langle M_{n,2}\rangle(T) : n \geq 1\}$ for any $T > 0$.

Here is an explanation: For each $n \geq 1$, the associated stochastic processes $\{\langle M_{n,1}\rangle(t) : t \geq 0\}$ and $\{\langle M_{n,2}\rangle(t) : t \geq 0\}$ are the predictable quadratic variations (compensators) of the scaled martingales $M_{n,1}$ and $M_{n,2}$ in (29), (30) and (59). First, by Lemma 5.8, the stochastic boundedness of these two sequences of PQV random variables implies stochastic boundedness in D of the two sequences of scaled martingales $\{M_{n,1} : n \geq 1\}$ and $\{M_{n,2} : n \geq 1\}$ in (30). That, in turn, under condition (4), implies the stochastic boundedness in D of the sequence of scaled queue-length processes $\{X_n : n \geq 1\}$ in (32) and (40) by Lemma 5.5. Finally, by Lemma 5.9 the stochastic boundedness of $\{X_n\}$ in D implies the required fluid limit in (84). The fluid limit in (84) was just what was needed in Lemmas 4.2 and 4.3.

6.2. Stochastic Boundedness of the Quadratic Variations

We have observed at the end of the last section that it only remains to establish the stochastic boundedness of two sequences of PQV random variables: $\{\langle M_{n,1}\rangle(t) : n \geq 1\}$ and $\{\langle M_{n,2}\rangle(t) : n \geq 1\}$ for any $t > 0$. For each $n \geq 1$, the associated stochastic processes $\{\langle M_{n,1}\rangle(t) : t \geq 0\}$ and $\{\langle M_{n,2}\rangle(t) : t \geq 0\}$ are the predictable quadratic variations (compensators) of the scaled martingales $M_{n,1}$ and $M_{n,2}$ in (33). We note that the stochastic boundedness of $\{\langle M_{n,1}\rangle(t) : t \geq 0\}$ is trivial because it is deterministic.

Lemma 6.2 (stochastic boundedness of $\{\langle M_{n,2}\rangle\}$ *Under the assumptions in Theorems* 1.1 *and* 3.4, *the sequence* $\{\langle M_{n,2}\rangle(t)\} : n \geq 1\})$ *is stochastically bounded for each* $t > 0$, *where*

$$\langle M_{n,2}\rangle(t) \equiv \frac{\mu}{n}\int_0^t Q_n(s)\, ds, \quad t \geq 0 \ .$$

Proof. It suffices to apply the crude inequality in Lemma 3.3:

$$\frac{\mu}{n}\int_0^t Q_n(s)\,ds \le \frac{\mu t}{n}(Q_n(0) + A(\lambda nt)) \ .$$

By Lemma 5.4, it suffices to show that the two sequences $\{Q_n(0)/n\}$ and $\{A(\lambda nt)/n\}$ are stochastically bounded. By condition (4) and the CMT, we have the WLLN's

$$\frac{Q_n(0)-n}{n} \Rightarrow 0 \quad \text{as} \quad n\to\infty \quad \text{so that} \quad \frac{Q_n(0)}{n} \Rightarrow 1 \quad \text{as} \quad n\to\infty \ ,$$

but that convergence implies stochastic boundedness; see Corollary 5.1. Turning to the other term, since $\lambda_n = n\mu$, we have

$$\frac{A(\lambda_n t)}{n} \to \mu t \quad \text{as} \quad n\to\infty \quad \text{w.p.1}$$

by the SLLN for Poisson processes. But convergence w.p.1 implies convergence in distribution, which in turn implies stochastic boundedness; again apply Corollary 5.1. Hence we have completed the proof. ■

6.3. The Initial Conditions

The results in §§3–6.1 have used the finite moment condition $E[Q_n(0)] < \infty$, but we want to establish Theorem 1.1 without this condition. In particular, note that this moment condition appears prominently in Lemmas 3.4 and 3.5 and in the resulting final martingale representations for the scaled processes, e.g., as stated in Theorems 3.4 and 3.5.

However, for the desired Theorem 1.1, we do not need to directly impose this moment condition. We do need the assumed convergence of the scaled initial conditions in (4), but we can circumvent this moment condition by defining bounded initial conditions that converge to the same limit. For that purpose, we can work with

$$\hat{Q}_n(0) \equiv Q_n(0) \wedge 2n, \quad n \ge 0 \ ,$$

and

$$\hat{X}_n(0) \equiv \frac{\hat{Q}_n(0)-n}{\sqrt{n}}, \quad n \ge 1 \ .$$

Then, for each $n \ge 1$, we use $\hat{Q}_n(0)$ and $\hat{X}_n(0)$ instead of $Q_n(0)$ and $X_n(0)$. We then obtain Theorem 1.1 for these modified processes: $\hat{X}_n \Rightarrow X \quad \text{in} \quad D \quad \text{as}$ $n\to\infty$. However, $P(X_n \ne \hat{X}_n) \to 0 \quad \text{as} \quad n\to\infty$. Hence, we have $X_n \Rightarrow X$ as well.

6.4. Limit from the Fourth Martingale Representation

In this section we state the FCLT limit for the general $G/GI/\infty$ queue stemming from the fourth martingale representation in §3.7, but we omit proofs and simply refer to Krichagina and Puhalskii [37]. (We do not call the FCLT a diffusion limit because the limit is not a diffusion process.) Even in the $M/M/\infty$ special case, the limit process has a different representation from the representation in Theorem 1.1. We will show that the two representations are actually equivalent.

Theorem 6.1 (FCLT from the fourth martingale representation) *Let X_n be defined in* (53) *and let $\hat{A}_n$ be defined in* (46) *and* (47)*. If $X_n(0) \Rightarrow X(0)$ in $\mathbb{R}$ and $\hat{A}_n \Rightarrow Z$ in D, where Z is a process with continuous sample paths and $Z(0) = 0$, then $X_n \Rightarrow X$ in D, where*

$$X(t) \equiv F_0^c(t)X(0) + \sqrt{q(0)}W^0(F_0(t)) + M_1(t) - M_2(t), \qquad t \geq 0, \tag{86}$$

where $W^0 \equiv \{W^0(t) : 0 \leq t \leq 1\}$ is a Brownian bridge, $U \equiv \{U(t,x), t \geq 0, 0 \leq x \leq 1\}$ is a Kiefer process, $X(0)$, W^0, Z and U are independent,

$$M_1(t) = \int_0^t F^c(t-s)dZ(s), \qquad t \geq 0,$$

and

$$M_2(t) = \int_0^t \int_0^t \mathbf{1}(s + x \leq t)dU(a(s), F(x)), \qquad t \geq 0.$$

For $M/M/\infty$ queues, $Z = \sqrt{\mu}B$, where B is a standard Brownian motion and the limit process is

$$\begin{aligned} X(t) &= e^{-\mu t}X(0) + \sqrt{q(0)}W^0(1 - e^{-\mu t}) + \sqrt{\mu}\int_0^t e^{-\mu(t-s)}dB(s) \\ &\quad - \int_0^t \int_0^t \mathbf{1}(s + x \leq t)dU(\mu s, 1 - e^{-\mu x}), \qquad t \geq 0. \end{aligned} \tag{87}$$

Remark 6.1 *A corresponding limit holds for the two-parameter processes $X_n \equiv \{X_n(t,y)\}$ in Corollary 3.1 by a minor variation of the same argument.*

Connection to Theorem 1.1. We now show that the two characterizations of the limit X in (87) and (5) are equivalent for the $M/M/\infty$ special case. For that purpose, express the process X in (87) as $X = Z_1 + Z_2 + Z_3 + Z_4$, where

$$Z_1(t) = e^{-\mu t}X(0), \quad Z_2(t) = \sqrt{q(0)}W^0(1 - e^{-\mu t}),$$

$$Z_3(t) = \sqrt{\mu}\int_0^t e^{-\mu(t-s)}dB(s),$$

$$Z_4(t) = -\int_0^t \int_0^t \mathbf{1}(s + x \leq t)dU(\mu s, 1 - e^{-\mu x}), \quad t \geq 0 \ .$$

Clearly, $Z_1(t)$ can be written as

$$Z_1(t) = -\mu \int_0^t Z_1(s)ds + X(0). \tag{88}$$

By the solution to the linear SDE, §5.6 in [33], we have

$$Z_3(t) = -\mu \int_0^t Z_3(s)ds + \sqrt{\mu}B(t), \tag{89}$$

where B is again standard Brownian motion.

Recall - see §5.6.B, [33] - that the Brownian bridge W^0 is the unique strong solution to the one-dimensional SDE

$$dY(t) = -\frac{Y(t)}{1-t}dt + dB_2(t), \qquad Y(0) = 0, \qquad 0 \le t \le 1,$$

where B_2 is second independent standard Brownian motion. So we can write

$$W^0(x) = -\int_0^x \frac{W^0(y)}{1-y}dy + B_2(x) \tag{90}$$

and it follows that

$$\begin{aligned} Z_2(t) &= \sqrt{q(0)}\Big(-\int_0^{1-e^{-\mu t}} \frac{W^0(y)}{1-y}dy + B_2(1-e^{-\mu t})\Big) \\ &= \sqrt{q(0)}\Big(-\mu\int_0^t W^0(1-e^{-\mu s})ds + B_2(1-e^{-\mu t})\Big) \\ &= -\mu\int_0^t Z_2(s)ds + \sqrt{q(0)}B_2(1-e^{-\mu t}). \end{aligned} \tag{91}$$

Paralleling (90), the Kiefer process U is related to the Brownian sheet $W \equiv \{W(t,x) : t \ge 0, 0 \le x \le 1\}$ by

$$U(t,x) = -\int_0^x \frac{U(t,y)}{1-y}dy + W(t,x), \qquad t \ge 0, \qquad 0 \le x \le 1\ . \tag{92}$$

Hence, for $t \ge 0, x \ge 0$, we have

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

Next, by similar reasoning, it can be shown that

$$Z_4(t) = -\mu\int_0^t Z_4(s)ds + \int_0^t\int_0^t \mathbf{1}(s+x \le t)dW(\mu s, 1-e^{-\mu x}). \tag{93}$$

By (88), (89), (91) and (93), we have

$$\begin{aligned} X(t) &= X_0 - \mu\int_0^t X(s)ds + \sqrt{\mu}B(t) + \sqrt{q(0)}B_2(1-e^{-\mu t}) \\ &\quad + \int_0^t\int_0^t \mathbf{1}(s+x \le t)dW(\mu s, 1-e^{-\mu x}), \end{aligned} \tag{94}$$

where B and B_2 are independent standard Brownian motions. Let $\hat{B}$ be the sum of the last three components in (94), i.e.,

$$\hat{B}(t) \equiv \sqrt{\mu}B(t) + \sqrt{q(0)}B_2(1 - e^{-\mu t}) + \int_0^t \int_0^t \mathbf{1}(s + x \le t) dW(\mu s, 1 - e^{-\mu x}).$$

It is evident that the process $\hat{B}$ is a continuous Gaussian process with mean 0 and, for $s < t$,

$$E[\hat{B}(t) - \hat{B}(s)]^2 = 2\mu(t - s) - (1 - q(0))(e^{-\mu s} - e^{-\mu t}),$$

and

$$E[\hat{B}(t)\hat{B}(s)] = 2\mu s - (1 - q(0))(1 - e^{-\mu s}).$$

However, in the $M/M/\infty$ case, we have $q(0) = 1$, so that $E[\hat{B}(t)\hat{B}(s)] = 2\mu(s \wedge t)$, which implies that the process $\{\hat{B}(t) : t \ge 0\}$ is distributed the same as $\{\sqrt{2\mu}B(t) : t \ge 0\}$, where B is a standard Brownian motion. Therefore, we have shown that the $M/M/\infty$ case of Theorem 6.1 is consistent with Theorem 1.1.

7. Other Models

In this section we discuss how to treat other models closely related to our initial $M/M/\infty$ model. We consider the Erlang-A model in §7.1; we also consider limits for the waiting time there. We consider finite waiting rooms in §7.2. Finally, we indicate how to treat general non-Poisson arrival processes in §7.3.

7.1. Erlang A Model

In this section we prove the corresponding many-server heavy-traffic limit for the $M/M/n/\infty + M$ (or Erlang-A or Palm) model in the QED regime. As before, the arrival rate is λ and the individual service rate is μ. Now there are n servers and unlimited waiting room with the FCFS service discipline. Customer times to abandon are i.i.d. exponential random variables with a mean of $1/\theta$. Thus individual customers waiting in queue abandon at a constant rate θ. The Erlang-C model arises when there is no abandonment, which occurs when $\theta = 0$. The Erlang C model is covered as a special case of the result below.

Let $Q(t)$ denote the number of customers in the system at time t, either waiting or being served. It is well known that the stochastic process $Q \equiv \{Q(t) : t \ge 0\}$ is a birth-and-death stochastic process with constant birth rate $\lambda_k = \lambda$ and state-dependent death rate $\mu_k = (k \wedge n)\mu + (k - n)^+\theta$, $k \ge 0$, where $a \wedge b \equiv \min\{a, b\}$, $a \vee b \equiv \max\{a, b\}$ and $(a)^+ \equiv a \vee 0$ for real numbers a and b.

As in Theorem 1.1, the many-server heavy-traffic limit involves a sequence of Erlang-A queueing models. As before, we let this sequence be indexed by n, but now this n coincides with the number of servers. Thus now we are letting the number of servers be finite, but then letting that number grow. At the same time, we let the arrival rate increase. As before, we let the arrival rate in model

n be λ_n, but now we stipulate that λ_n grows with n. At the same time, we hold the individual service rate μ and abandonment rate θ fixed. Let $\rho_n \equiv \lambda_n/n\mu$ be the **traffic intensity** in model n. We stipulate that

$$(1-\rho_n)\sqrt{n} \to \beta \quad \text{as} \quad n \to \infty \ , \tag{95}$$

where β a (finite) real number. That is equivalent to assuming that

$$\frac{n\mu - \lambda_n}{\sqrt{n}} \to \beta\mu \quad \text{as} \quad n \to \infty \ , \tag{96}$$

as in (7). Conditions (95) and (96) are known to characterize the QED many-server heavy-traffic regime; see Halfin and Whitt [26] and Puhalskii and Reiman [53]. The many-server heavy-traffic limit theorem for the Erlang-A model was proved by Garnett, Mandelbaum and Reiman [21], exploiting Stone [62]. See Zeltyn and Mandelbaum [74] and Mandelbaum and Zeltyn [49] for extensions and elaboration. For related results for single-server models with customer abandonment, see Ward and Glynn [65, 66, 67].

Here is the QED many-server heavy-traffic limit theorem for the $M/M/n/\infty+M$ model:

Theorem 7.1 (heavy-traffic limit in D for the $M/M/n+M$ model) *Consider the sequence of $M/M/n/\infty+M$ models defined above, with the scaling in* (95). *Let X_n be as defined in* (3). *If $X_n(0) \Rightarrow X(0)$ in $\mathbb{R}$ as $n \to \infty$, then $X_n \Rightarrow X$ in D as $n \to \infty$, where X is the diffusion process with infinitesimal mean $m(x) = -\beta\mu - \mu x$ for $x < 0$ and $m(x) = -\beta\mu - \theta x$ for $x > 0$, and infinitesimal variance $\sigma^2(x) = 2\mu$. Alternatively, the limit process X satisfies the stochastic integral equation*

$$X(t) = X(0) - \mu\beta t + \sqrt{2\mu}B(t) - \int_0^t \left[\mu(X(s)\wedge 0) + \theta(X(s)\vee 0)\right] ds$$

for $t \geq 0$, where B is a standard Brownian motion. Equivalently, X satisfies the SDE

$$dX(t) = -\mu\beta - \mu(X(t)\wedge 0)dt - \theta(X(t)\vee 0)dt + \sqrt{2\mu}dB(t), \quad t \geq 0 \ .$$

Proof. In the rest of this section we very quickly summarize the proof; the argument mostly differs little from what we did before. Indeed, if we use the second martingale representation as in §§2.2 and 3.5, then there is very little difference. However, if we use the first martingale representation, as in §§2.1 and 3.4, then there is a difference, because now we want to use the optional stopping theorem for multiparameter random time changes, as in §§2.8 and 6.2 of Ethier and Kurtz [19]. That approach follows Kurtz [40], which draws on Helms [28]. That approach has been applied in §12 of Mandelbaum and Pats [48]. To illustrate this alternate approach, we use the random time change approach here.

Just as in §2.1, we can construct the stochastic process Q in terms of rate-1 Poisson processes. In addition to the two Poisson processes A and S introduced before, now we have an extra rate-1 Poisson process R used to generate abandonments. Instead of (12), here we have representation

$$\begin{aligned} Q(t) &\equiv Q(0) + A(\lambda t) - D(t) - L(t), \quad t \geq 0 \ , \\ &= Q(0) + A(\lambda t) - S\left(\mu \int_0^t (Q(s) \wedge n)\, ds\right) \\ &\qquad -R\left(\theta \int_0^t (Q(s) - n)^+ \, ds\right) \ , \end{aligned} \tag{97}$$

for $t \geq 0$, where $D(t)$ is the number of departures (service completions) in the time interval $[0,t]$, while $L(t)$ is the number of customers lost because of abandonment in the time interval $[0,t]$. Since there are only n servers, the instantaneous overall service rate at time s is $\mu(Q(s) \wedge n)$, while the instantaneous overall abandonment rate (which is only from waiting customers, not yet in service) at time s is $\theta(Q(s) - n)^+$.

Paralleling (14), we have the martingale representation

$$\begin{aligned} Q(t) &= Q(0) + M_1(t) - M_2(t) - M_3(t) + \lambda t \\ &\qquad -\mu \int_0^t (Q(s) \wedge n)\, ds - \theta \int_0^t (Q(s) - n)^+ \, ds \end{aligned} \tag{98}$$

for $t \geq 0$, where

$$\begin{aligned} M_1(t) &\equiv A(\lambda t) - \lambda t, \\ M_2(t) &\equiv S\left(\mu \int_0^t (Q(s) \wedge n)\, ds\right) - \mu \int_0^t (Q(s) \wedge n)\, ds \ , \\ M_3(t) &\equiv R\left(\theta \int_0^t (Q(s) - n)^+ \, ds\right) - \theta \int_0^t (Q(s) - n)^+ \, ds \end{aligned} \tag{99}$$

for $t \geq 0$ and the filtration is $\mathbf{F} \equiv \{\mathcal{F}_t : t \geq 0\}$ defined by

$$\begin{aligned} \mathcal{F}_t \equiv \sigma\Bigg(Q(0), A(\lambda s), S\left(\mu \int_0^s (Q(u) \wedge n)\, du\right), \\ R\left(\theta \int_0^s (Q(u) - n)^+ \, du\right) : 0 \leq s \leq t\Bigg) \ , \end{aligned} \tag{100}$$

for $t \geq 0$, augmented by including all null sets.

We now want to justify the claims in (98)–(100). Just as before, we can apply Lemmas 3.1 and 3.3 to justify this martingale representation, but now we need to replace Lemmas 3.2 and 3.4 by corresponding lemmas involving the optional stopping theorem with multiparameter random time changes, as in §§2.8 and 6.2 of [19]. We now sketch the approach: We start with the three-parameter filtration

$$\begin{aligned} \mathbf{H} &\equiv \mathcal{H}(t_1, t_2, t_3) \\ &\equiv \sigma\left(Q(0), A(s_1), S(s_2), R(s_3) : 0 \leq s_1 \leq t_1, 0 \leq s_2 \leq t_2, 0 \leq s_3 \leq t_3\right) \end{aligned} \tag{101}$$

augmented by adding all null sets. Next introduce the three nondecreasing nonnegative stochastic processes

$$\begin{aligned} I_1(t) &\equiv \lambda t, \\ I_2(t) &\equiv \mu \int_0^t (Q(s) \wedge n)\, ds, \\ I_3(t) &\equiv \theta \int_0^t (Q(s) - n)^+\, ds, \quad t \ge 0\ . \end{aligned}$$

Then observe that the vector $(I_1(t), I_2(t), I_3(t))$ is an **H**-stopping time. (This is essentially by the same arguments as we used in Lemma 3.4. But here we directly gain control of the arrival process, because the event $\{I_1(t) \le x_1\}$ coincides with the requirement that $\{\lambda t \le x_1\}$, which ensures that we always have enough of the arrival process to construct $Q(s)$, $0 \le s \le t$.)

Moreover, since A, S and R are assumed to be independent rate-1 Poisson processes, the stochastic process

$$\bar{M} \equiv (\bar{M}_1, \bar{M}_2, \bar{M}_3) \equiv \{(\bar{M}_1(s_1), \bar{M}_2(s_2), \bar{M}_3(s_3) : s_1 \ge 0, s_2 \ge 0, s_3 \ge 0\}$$

where

$$\begin{aligned} \bar{M}_1(t) &\equiv A(t) - t, \\ \bar{M}_2(t) &\equiv S(t) - t, \\ \bar{M}_3(t) &\equiv R(t) - t, \quad t \ge 0, \end{aligned}$$

is an **H**-multiparameter martingale. As a consequence of the optional stopping theorem, Theorem 8.7 on p. 87 of [19],

$$(\bar{M}_1 \circ I_1, \bar{M}_2 \circ I_2, \bar{M}_3 \circ I_3) \equiv \{\bar{M}_1(I_1(t)), \bar{M}_2(I_2(t)), \bar{M}_3(I_3(t)) : t \ge 0\}$$

is a martingale with respect to the filtration $\mathbf{F} \equiv \{\mathcal{F}_t : t \ge 0\}$ in (100), because

$$\begin{aligned} & \mathcal{H}(I_1(t), I_2(t), I_3(t)) \\ = \; & \sigma\left(Q(0), A(s_1), S(s_2), R(s_3) : 0 \le s_1 \le I_1(t), 0 \le s_2 \le I_2(t), 0 \le s_3 \le I_3(t)\right) \\ = \; & \sigma\Bigg(Q(0), A\left(\lambda s\right), S\left(\mu \int_0^s (Q(u) \wedge n)\, du\right), \\ & \qquad R\left(\mu \int_0^t (Q(s) - n)^+\, du\right) : 0 \le s \le t\Bigg) = \mathcal{F}_t \quad \text{for all} \quad t \ge 0\ . \end{aligned}$$

As in §3.4, we use the crude inequality in Lemma 3.3 to guarantee that the moment conditions are satisfied:

$$E[I_j(t)] < \infty \quad \text{and} \quad E[M_j(t)] < \infty \quad \text{for} \quad j = 1, 2, 3\ .$$

Just as before, we then consider the sequence of models indexed by n. Just as in (28)-(31), we define associated scaled processes:

$$M_{n,1}(t) \equiv n^{-1/2}\left[A(\lambda_n t) - \lambda_n t\right], \tag{102}$$

$$
\begin{aligned}
M_{n,2}(t) &\equiv n^{-1/2}\left[S\left(\mu\int_0^t (Q_n(s)\wedge n)\,ds\right) - \mu\int_0^t (Q_n(s)\wedge n)\,ds\right] \ , \\
M_{n,3}(t) &\equiv n^{-1/2}\left[R\left(\theta\int_0^t (Q_n(s)-n)^+\,ds\right) - \theta\int_0^t (Q_n(s)-n)^+\,ds\right],
\end{aligned}
$$

for $t \ge 0$.

We thus obtain the following analog of Theorem 3.4:

Theorem 7.2 (first martingale representation for the scaled processes in the Erlang A model) *If $E[Q_n(0)] < \infty$, then the scaled processes have the martingale representation*

$$
\begin{aligned}
X_n(t) &\equiv X_n(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) + \frac{(\lambda_n - \mu n)t}{\sqrt{n}} \\
&\quad - \int_0^t \left[\mu(X_n(s)\wedge 0) + \theta X_n(s)^+\right]\,ds, \quad t \ge 0 \ , \qquad (103)
\end{aligned}
$$

where $M_{n,i}$ are the scaled martingales in (102)*. These processes $M_{n,i}$ are square-integrable martingales with respect to the filtrations $\boldsymbol{F}_n \equiv \{\mathcal{F}_{n,t} : t \ge 0\}$ defined by*

$$
\begin{aligned}
\mathcal{F}_{n,t} &\equiv \sigma\Bigg(Q_n(0), A(\lambda_n s), S\left(\mu\int_0^s (Q_n(u)\wedge n)\,du\right), \\
&\qquad R\left(\theta\int_0^s (Q_n(u)-n)^+\,du\right) : 0 \le s \le t\Bigg) \ , \quad t \ge 0 \ , \qquad (104)
\end{aligned}
$$

augmented by including all null sets. Their associated predictable quadratic variations are

$$
\begin{aligned}
\langle M_{n,1}\rangle(t) &= \frac{\lambda_n t}{n} \ , \\
\langle M_{n,2}\rangle(t) &= \frac{\mu}{n}\int_0^t (Q_n(s)\wedge n)\,ds, \\
\langle M_{n,3}\rangle(t) &= \frac{\theta}{n}\int_0^t (Q_n(s)-n)^+\,ds, \quad t \ge 0 \ , \qquad (105)
\end{aligned}
$$

where $E[\langle M_{n,i}\rangle(t)] < \infty$ for all i, $t \ge 0$ and $n \ge 1$.

The representation in (103) satisfies the Lipschitz condition in Theorem 4.1, so that the integral representation is again a continuous mapping. In particular, here we have

$$
\begin{aligned}
X_n(t) &\equiv X_n(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) + \frac{(\lambda_n - \mu n)t}{\sqrt{n}} \\
&\quad + \int_0^t h(X_n(s))\,ds \ , \quad t \ge 0 \ , \qquad (106)
\end{aligned}
$$

where $h : \mathbb{R} \to \mathbb{R}$ is the function

$$
h(s) = -\mu(s\wedge 0) - \theta(s)^+, \quad s \in \mathbb{R} \ , \qquad (107)
$$

so that h is Lipschitz as required for Theorem 4.1:

$$|h(s_1) - h(s_2)| \leq (\mu \vee \theta)|s_1 - s_2| \quad \text{for all} \quad s_1, s_2 \in \mathbb{R} \ .$$

Hence the proof can proceed exactly as before. Note that we have convergence

$$\frac{(\lambda_n - \mu n)}{\sqrt{n}} \to -\mu\beta \quad \text{as} \quad n \to \infty$$

for the deterministic term in (106) by virtue of the QED scaling assumption in (95).

The analog of Theorem 4.2 is the corresponding FCLT for three independent rate-1 Poisson processes, now including R as well as A and S. We now have three random-time-change processes: $\Phi_{A,n}$, $\Phi_{S,n}$ and $\Phi_{R,n}$, which here take the form:

$$\begin{aligned}
\Phi_{A,n}(t) &\equiv \langle M_{n,1}\rangle(t) = \frac{\lambda_n t}{n}, \\
\Phi_{S,n}(t) &\equiv \langle M_{n,2}\rangle(t) = \frac{\mu}{n}\int_0^t (Q(s) \wedge n)\, ds, \\
\Phi_{R,n}(t) &\equiv \langle M_{n,3}\rangle(t) = \frac{\theta}{n}\int_0^t (Q(s) - n)^+\, ds \ , \qquad (108)
\end{aligned}$$

drawing upon (105). By the same line of reasoning as before, we obtain the deterministic limits

$$\Phi_{A,n}(t) \Rightarrow \mu e, \quad \Phi_{S,n}(t) \Rightarrow \mu e \quad \text{and} \quad \Phi_{R,n}(t) \Rightarrow \eta \ , \qquad (109)$$

where, as before, e is the identity map $e(t) \equiv t, \quad t \geq 0$, and η is the zero function $\eta(t) \equiv 0$, $t \geq 0$. In particular, we again get the sequence $\{X_n : n \geq 1\}$ stochastically bounded in D by the reasoning in §§5.2 and 6.2. Then, by Lemma 5.9 and §6.1, we get the FWLLN corresponding to Lemma 4.3. Finally, from Lemma 4.3, we can prove (109), just as we proved Lemma 4.2, using analogs of the continuous map in (71) for the random time changes in (108). The new functions for applications of the CMT are

$$h_1(x)(t) \equiv \mu \int_0^t (x(s) \wedge 1)\, ds \quad \text{and} \quad h_2(x)(t) \equiv \theta \int_0^t (x(s) - 1)^+\, ds$$

for $t \geq 0$.

Paralleling (73), here we have

$$(M_{A,n}, \Phi_{A,n}, M_{S,n}, \Phi_{S,n}, M_{R,n}, \Phi_{R,n}) \Rightarrow (B_1, \mu e, B_2, \mu e, B_3, \eta) \quad \text{in} \quad D^6$$

as $n \to \infty$. Hence we can apply the CMT with composition just as before. Paralleling (74), here we obtain first

$$(M_{A,n} \circ \Phi_{A,n}, M_{S,n} \circ \Phi_{S,n}, M_{R,n} \circ \Phi_{R,n}) \Rightarrow (B_1 \circ \mu e, B_2 \circ \mu e, B_3 \circ \eta)$$

in D^3 as $n \to \infty$, and then

$$(M_{A,n} \circ \Phi_{A,n} - M_{S,n} \circ \Phi_{S,n} - M_{R,n} \circ \Phi_{R,n}) \Rightarrow (B_1 \circ \mu e - B_2 \circ \mu e - B_3 \circ \eta)$$

in D. However,

$$B_1 \circ \mu e - B_2 \circ \mu e - B_3 \circ \eta \overset{\mathrm{d}}{=} B_1 \circ \mu e - B_2 \circ \mu e - \eta \overset{\mathrm{d}}{=} \sqrt{2\mu} B \ . \tag{110}$$

Finally, the CMT with the integral representation (103) and Theorem 4.1 completes the proof. Note that the limiting Brownian motion associated with R does not appear, because $\Phi_{R,n}$ is asymptotically negligible. That is why the infinitesimal variance is the same as before. ■

7.2. Finite Waiting Rooms

We can also obtain stochastic-process limits for the number of customers in the system in associated $M/M/n/0$ (Erlang-B), $M/M/n/m_n$ and $M/M/n/m_n+M$ models, which have finite waiting rooms. For the Erlang-B model, there is no waiting room at all; for the other models there is a waiting room on size m_n in model n, where m_n is allowed to grow with n so that $m_n/\sqrt{n} \to \kappa \geq 0$ as $n \to \infty$ as in (8). The QED many-server heavy-traffic limit was stated as Theorem 1.2.

The proof can be much the same as in §7.1. The idea is to introduce the finite waiting room via a reflection map, as in §§3.5, 5.2, 13.5, 14.2, 14.3 and 14.8 of Whitt [69], corresponding to an upper barrier at κ, but the reflection map here is more complicated than for single-server queues and networks of such queues, because it is not applied to a free process. We use an extension of Theorem 4.1 constructing a mapping from $D \times \mathbb{R}$ into D^2, taking model data into the content function and the upper-barrier regulator function.

Theorem 7.3 (*a continuous integral representation with reflection*) *Consider the modified integral representation*

$$x(t) = b + y(t) + \int_0^t h(x(s))\, ds - u(t) \ , \quad t \geq 0 \ , \tag{111}$$

where $x(t) \leq \kappa$, $h : \mathbb{R} \to \mathbb{R}$ satisfies $h(0) = 0$ and is a Lipschitz function as in (62)*, and u is a nondecreasing nonnegative function in D such that* (111) *holds and*

$$\int_0^\infty 1_{\{x(t) < \kappa\}}\, du(t) = 0 \ .$$

The modified integral representation in (111) *has a unique solution (x, u), so that it constitutes a bonafide function $(f_1, f_2) : D \times \mathbb{R} \to D \times D$ mapping (y, b) into $x \equiv f_1(y, b)$ and $u \equiv f_2(y, b)$. In addition, the function (f_1, f_2) is continuous provided that the product topology is used for product spaces and the function space D* (*in both the domain and range*) *is endowed with either:* (*i*) *the topology of uniform convergence over bounded intervals or* (*ii*) *the Skorohod J_1 topology. Moreover, if y is continuous, then so are x and u.*

Proof. We only show the key step, for which we follow the argument in §3 of Mandelbaum and Pats [48] and §4 of Reed and Ward [56]; see these sources for additional details and references. The idea is to combine classical results for the conventional one-dimensional reflection map, as in §§5.2 and 13.5 of [69] with a modification of Theorem 4.1. Let $(\phi_\kappa, \psi_\kappa)$ be the one-sided reflection map with upper barrier at κ, so that $\phi_\kappa(y) = y - \psi_\kappa(y)$, with $\phi_\kappa(y)$ being the content function and $\psi_\kappa(y)$ being the nondecreasing regulator function; see §§5.2 and 13.5 of [69]. We observe that the map in (111) can be expressed as $x = \phi_\kappa(w)$ and $u = \psi_\kappa(w)$, where

$$w(t) \equiv \xi(b, y)(t) \equiv b + y(t) + \int_0^t h(\phi_\kappa(w(s)))\, ds, \quad t \geq 0 \ . \tag{112}$$

This lets us represent the desired map as the composition of the maps $(\phi_\kappa, \psi_\kappa)$ and ξ. The argument to treat ξ is essentially the same as in the proof of Theorem 4.1, but we need to make a slight adjustment; we could apply it directly if we had $h : D \to D$ in Theorem 4.1. Recall that ϕ_κ is Lipschitz continuous on $D([0, t]$ for each t with the uniform norm, $\|\phi_\kappa(y_1) - \phi_\kappa(y_2)\|_t \leq 2\|y_1 - y_2\|_t$, with modulus 2 independent of t. Hence, paralleling (64), we have

$$\|w_1 - w_2\|_t \leq |b_1 - b_2| + \|y_1 - y_2\|_t + 2c \int_0^t \|w_1 - w_2\|_s\, ds$$

for each $t > 0$. Hence we can apply Gronwall's inequality in Lemma 4.1 to establish (Lipschitz) continuity of the map ξ on $D([0, T]) \times \mathbb{R}$. Combining this with the known (Lipschitz) continuity of the reflection map $(\phi_\kappa, \psi_\kappa)$, we have the desired continuity for the overall map in the uniform topology. We can extend to the J_1 topology as in the proof of Theorem 4.1. ■

Now that we understand how we are treating the finite waiting rooms, the QED many-server heavy-traffic limit theorem is as stated in Theorem 1.2. This modification alters the limiting diffusion process in Theorem 7.1 only by the addition of a reflecting upper barrier at κ for the sequence of models with waiting rooms of size m_n, where $\kappa = 0$ for the Erlang-B model. When $\kappa = 0$, X is a reflected OU (ROU) process. Properties of the ROU process are contained in Ward and Glynn [66]. Proofs for the two cases $\kappa > 0$ and $\kappa = 0$ by other methods are contained in §4.5 of Whitt [72] and Theorem 4.1 of Srikant and Whitt [61]. General references on reflection maps are Lions and Sznitman [43] and Dupuis and Ishii [18].

Proof of Theorem 1.2. We briefly sketch the proof. Instead of (97), here we have representation

$$\begin{aligned} Q_n(t) &\equiv Q_n(0) + A(\lambda_n t) - D_n(t) - L_n(t) - U_n(t), \quad t \geq 0 \ , \\ &= Q_n(0) + A(\lambda_n t) - S\left(\mu \int_0^t (Q_n(s) \wedge n)\, ds\right) \\ &\quad -R\left(\theta \int_0^t (Q_n(s) - n)^+\, ds\right) - U_n(t) \ , \end{aligned} \tag{113}$$

for $t \geq 0$, where $U_n(t)$ is the number of arrivals in the time interval $[0,t]$ when the system is full in model n, i.e., when $Q_n(t) = n + m_n$. In particular,

$$U_n(t) \equiv \int_0^t 1_{\{Q_n(s)=n+m_n\}} \, dA(\lambda_n s), \quad t \geq 0 \; . \tag{114}$$

To connect to Theorem 7.3, it is significant that U_n can also be represented as the unique nondecreasing nonnegative process such that $Q_n(t) \leq n + m_n$, (113) holds and

$$\int_0^\infty 1_{\{Q_n(t)<\kappa\}} \, dU_n(t) = 0 \; . \tag{115}$$

We now construct a martingale representation, just as in (98)–(102). The following is the natural extension of Theorem 7.2:

Theorem 7.4 (first martingale representation for the scaled processes in the $M/M/n/m_n{+}M$ model) *If $\kappa < \infty$, then the scaled processes have the martingale representation*

$$\begin{aligned} X_n(t) \;\; &\equiv \;\; X_n(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) + \frac{(\lambda_n - \mu n)t}{\sqrt{n}} \\ &\quad - \int_0^t \left[\mu(X_n(s) \wedge 0) + \theta X_n(s)^+\right] \, ds - V_n(t), \quad t \geq 0 \; , \end{aligned} \tag{116}$$

$M_{n,i}$ are the scaled martingales in (102) *and*

$$V_n(t) \equiv \frac{U_n(t)}{\sqrt{n}}, \quad t \geq 0 \; , \tag{117}$$

for U_n in (113)*–*(115)*. The scaled processes $M_{n,i}$ are square-integrable martingales with respect to the filtrations $\boldsymbol{F}_n \equiv \{\mathcal{F}_{n,t} : t \geq 0\}$ defined by*

$$\begin{aligned} \mathcal{F}_{n,t} \;\; &\equiv \;\; \sigma\left(Q_n(0), A(\lambda_n s), S\left(\mu \int_0^s (Q_n(u) \wedge n) \, du\right) , \right. \\ &\quad \left. R\left(\theta \int_0^s (Q_n(u) - n)^+ \, du\right) : 0 \leq s \leq t\right) \; , \quad t \geq 0 \; , \end{aligned} \tag{118}$$

augmented by including all null sets. Their associated predictable quadratic variations are

$$\begin{aligned} \langle M_{n,1}\rangle(t) \;\; &= \;\; \frac{\lambda_n t}{n} \; , \\ \langle M_{n,2}\rangle(t) \;\; &= \;\; \frac{\mu}{n} \int_0^t (Q_n(s) \wedge n) \, ds, \\ \langle M_{n,3}\rangle(t) \;\; &= \;\; \frac{\theta}{n} \int_0^t (Q_n(s) - n)^+ \, ds, \quad t \geq 0 \; , \end{aligned} \tag{119}$$

where $E[\langle M_{n,i}\rangle(t)] < \infty$ for all i, $t \geq 0$ and $n \geq 1$.

By combining Theorems 7.3 and 7.4, we obtain the joint convergence

$$(X_n, V_n) \Rightarrow (X, U) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty \ ,$$

for X_n and V_n in (116) and (117), where the vector (X, U) is characterized by (9) and (10). That implies Theorem 1.2 stated in §1.

Remark 7.1 (correction) The argument in this section follows Whitt [72], but provides more detail. We note that the upper-barrier regulator processes are incorrectly expressed in formulas (5.2) and (5.8) of [72].

7.3. General Non-Markovian Arrival Processes

In this section, following §5 of Whitt [72], we show how to extend the many-server heavy-traffic limit from $M/M/n/m_n + M$ models to $G/M/n/m_n + M$ models, where the arrival processes are allowed to be general stochastic point processes satisfying a FCLT. They could be renewal processes (GI) or even more general arrival processes. The limit of the arrival-process FCLT need not have continuous sample paths. (As noted at the end of §4.3, this separate argument is not needed if we do not use martingales.)

Let $\bar{A}_n \equiv \{\bar{A}_n(t) : t \geq 0\}$ be the general arrival process in model n and let

$$A_n(t) \equiv \frac{\bar{A}_n(t) - \lambda_n t}{\sqrt{n}}, \quad t \geq 0 \ , \tag{120}$$

be the associated scaled arrival process. We assume that

$$A_n \Rightarrow A \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ . \tag{121}$$

We also assume that, conditional on the entire arrival process, model n evolves as the Markovian queueing process with i.i.d. exponential service times and i.i.d. exponential times until abandonment.

Thus, instead of Theorem 7.4, we have

Theorem 7.5 (first martingale representation for the scaled processes in the $G/M/n/m_n{+}M$ model) *Consider the family of $G/M/n/m_n{+}M$ models defined above, evolving as a Markovian queue conditional on the arrival process. If $m_n < \infty$, then the scaled processes have the martingale representation*

$$\begin{aligned} X_n(t) \quad &\equiv \quad X_n(0) + A_n(t) - M_{n,2}(t) - M_{n,3}(t) + \frac{(\lambda_n - \mu n)t}{\sqrt{n}} \\ &\quad - \int_0^t \left[\mu(X_n(s) \wedge 0) + \theta X_n(s)^+\right] ds - V_n(t), \quad t \geq 0 \ , \end{aligned} \tag{122}$$

where A_n is the scaled arrival process in (120)*, $M_{n,i}$ are the scaled martingales in* (102) *and $V_n(t) \equiv U_n(t)/\sqrt{n}$, $t \geq 0$, for U_n in* (113)-(115)*. The scaled processes*

$M_{n,i}$ are square-integrable martingales with respect to the filtrations $\boldsymbol{F}_n \equiv \{\mathcal{F}_{n,t} : t \ge 0\}$ defined by

$$\begin{aligned}\mathcal{F}_{n,t} \quad &\equiv \quad \sigma\left(Q_n(0), \{A_n(u) : u \ge 0\}, S\left(\mu \int_0^s (Q_n(u) \wedge n)\, du\right),\right. \\ &\left. R\left(\theta \int_0^s (Q_n(u) - n)^+\, du\right) : 0 \le s \le t\right) \ , \quad t \ge 0 \ , \qquad (123)\end{aligned}$$

augmented by including all null sets. The associated predictable quadratic variations of the two martingales are

$$\begin{aligned}\langle M_{n,2}\rangle(t) \quad &= \quad \frac{\mu}{n} \int_0^t (Q_n(s) \wedge n)\, ds, \\ \langle M_{n,3}\rangle(t) \quad &= \quad \frac{\theta}{n} \int_0^t (Q_n(s) - n)^+\, ds, \quad t \ge 0 \ , \qquad (124)\end{aligned}$$

where $E[\langle M_{n,i}\rangle(t)] < \infty$ for all i, $t \ge 0$ and $n \ge 1$.

Here is the corresponding theorem for the $G/M/n/m_n + M$ model.

Theorem 7.6 (heavy-traffic limit in D for the $G/M/n/m_n + M$ model) *Consider the sequence of $G/M/n/m_n + M$ models defined above, with the scaling in (120), (7) and (8). Let X_n be as defined in (3). If*

$$X_n(0) \Rightarrow X(0) \quad in \quad \mathbb{R} \quad and \quad A_n \Rightarrow A \quad in \quad D \quad as \quad n \to \infty \ ,$$

then $X_n \Rightarrow X$ in D as $n \to \infty$, where the limit process X satisfies the stochastic integral equation

$$\begin{aligned}X(t) \quad &= \quad X(0) + A(t) - \beta\mu t + \sqrt{\mu} B(t) \\ &- \int_0^t [\mu(X(s) \wedge 0) + \theta(X(s) \vee 0)]\, ds - U(t) \qquad (125)\end{aligned}$$

for $t \ge 0$ with $B \equiv \{B(t) : t \ge 0\}$ being a standard Brownian motion and U being the unique nondecreasing nonnegative process in D such that $X(t) \le \kappa$ for all t, (125) holds and

$$\int_0^\infty 1_{\{X(t)<\kappa\}}\, dU(t) = 0 \ .$$

Proof. We start with the FCLT for the arrival process assumed in (121) and (120). We condition on possible realizations of the arrival process: For each $n \ge 1$, let ζ_n be a possible realization of the scaled arrival process A_n and let ζ be a possible realization of the limit process A. Let $X_n^{\zeta_n}$ be the scaled process X_n conditional on $A_n = \zeta_n$, and let X^ζ be the limit process X conditional on $A = \zeta$.

Since X_n and A_n are random elements of D, these quantities $X_n^{\zeta_n}$ and X^ζ are well defined via regular conditional probabilities; e.g., see §8 of Chapter V, pp 146-150, of Parthasarathy [51]. In particular, we can regard $P(X_n \in \cdot | A_n = z)$

as a probability measure on D for each $z \in D$ and we can regard $P(X_n \in B|A_n = z)$ as a measurable function of z in D for each Borel set B in D, where

$$P(X_n \in B) = \int_B P(X_n \in B|A_n = z)\, dP(A_n = z) \ .$$

And similarly for the pair (X, A).

A minor modification of the previous proof of Theorem 1.2 establishes that

$$X_n^{\zeta_n} \Rightarrow X^{\zeta} \quad \text{in} \quad D \quad \text{whenever} \quad \zeta_n \to \zeta \quad \text{in} \quad D \ ;$$

i.e., for each continuous bounded real-valued function f on D,

$$E[f(X_n^{\zeta_n})] \to E[f(X^{\zeta})] \quad \text{as} \quad n \to \infty \tag{126}$$

whenever $\zeta_n \to \zeta$ in D. Now fix a continuous bounded real-valued function f and let

$$h_n(\zeta_n) \equiv E[f(X_n^{\zeta_n})] \quad \text{and} \quad h(\zeta) \equiv E[f(X^{\zeta})] \ . \tag{127}$$

Since we have regular conditional probabilities, we can regard the functions h_n and h as measurable functions from D to $\mathbb{R}$ (depending on f).

We are now ready to apply the generalized continuous mapping theorem, Theorem 3.4.4 of [69]. Since h_n and h are measurable functions such that $h_n(\zeta_n) \to h(\zeta)$ whenever $\zeta_n \to \zeta$ and since $A_n \Rightarrow A$ in D, we have $h_n(A_n) \Rightarrow h(A)$ as $n \to \infty$. Since the function f used in (126) and (127) is bounded, these random variables are bounded. Hence convergence in distribution implies convergence of moments. Hence, for that function f, we have

$$E[f(X_n)] \equiv E[h_n(A_n)] \to E[h(A)] \equiv E[f(X)] \quad \text{as} \quad n \to \infty \ .$$

Since this convergence holds for all continuous bounded real-valued functions f on D, we have shown that $X_n \Rightarrow X$, as claimed. ■

8. The Martingale FCLT

We now turn to the martingale FCLT. For our queueing stochastic-process limits, it is of interest because it provides one way to prove the FCLT for a Poisson process in Theorem 4.2 and because we can base our entire proof of Theorem 1.1 on the martingale FCLT. However, the gain in the proof of Theorem 1.1 is not so great.

We now state a version of the martingale FCLT for a sequence of local martingales $\{M_n : n \geq 1\}$ in D^k, based on Theorem 7.1 on p. 339 of Ethier and Kurtz [19], hereafter referred to as EK. Another important reference is Jacod and Shiryayev [30], hereafter referred to as JS. See Section VIII.3 of JS for related results; see other sections of JS for generalizations.

We will state a special case of Theorem 7.1 of EK in which the limit process is multi-dimensional Brownian motion. However, the framework always produces limits with continuous sample paths and independent Gaussian increments. Most applications involve convergence to Brownian motion. Other situations are covered by JS, from which we see that proving convergence to discontinuous processes is more complicated.

The key part of each condition below is the convergence of the quadratic covariation processes. Condition (i) involves the optional quadratic-covariation (square-bracket) processes $[M_{n,i}, M_{n,j}]$, while condition (ii) involves the predictable quadratic-covariation (angle-bracket) processes $\langle M_{n,i}, M_{n,j}\rangle$. Recall from §3.2 that the square-bracket process is more general, being well defined for any local martingale (and thus any martingale), whereas the associated angle-bracket process is well defined only for any locally square-integrable martingale (and thus any square-integrable martingale).

Thus the key conditions below are the assumed convergence of the quadratic-variation processes in conditions (130) and (133). The other conditions (129), (131) and (133) are technical regularity conditions. There is some variation in the literature concerning the extra technical regularity conditions; e.g., see Rebolledo [54] and JS.

Let J be the **maximum-jump function**, defined for any $x \in D$ and $T > 0$ by

$$J(x,T) \equiv \sup\left\{|x(t) - x(t-)| : 0 < t \le T\right\} . \tag{128}$$

Theorem 8.1 (multidimensional martingale FCLT) *For $n \ge 1$, let $M_n \equiv (M_{n,1}, \ldots, M_{n,k})$ be a local martingale in D^k with respect to a filtration $\boldsymbol{F}_n \equiv \{\mathcal{F}_{n,t} : t \ge 0\}$ satisfying $M_n(0) = (0, \ldots, 0)$. Let $C \equiv (c_{i,j})$ be a $k \times k$ covariance matrix, i.e., a nonnegative-definite symmetric matrix of real numbers.*

Assume that one of the following two conditions holds:

(i) The expected value of the maximum jump in M_n is asymptotically negligible; i.e., for each $T > 0$,

$$\lim_{n\to\infty} \left\{E\left[J(M_n, T)\right]\right\} = 0 \tag{129}$$

and, for each pair (i,j) with $1 \le i \le k$ and $1 \le j \le k$, and each $t > 0$,

$$\left[M_{n,i}, M_{n,j}\right](t) \Rightarrow c_{i,j}t \quad in \quad \mathbb{R} \quad as \quad n \to \infty \ . \tag{130}$$

(ii) The local martingale M_n is locally square-integrable, so that the predictable quadratic-covariation processes $\langle M_{n,i}, M_{n,j}\rangle$ can be defined. The expected value of the maximum jump in $\langle M_{n,i}, M_{n,j}\rangle$ and the maximum squared jump of M_n are asymptotically negligible; i.e., for each $T > 0$ and (i,j) with $1 \le i \le k$ and $1 \le j \le k$,

$$\lim_{n\to\infty} \left\{E\left[J\left(\langle M_{n,i}, M_{n,j}\rangle, T\right)\right]\right\} = 0 \ , \tag{131}$$

$$\lim_{n\to\infty} \left\{E\left[J\left(M_n, T\right)^2\right]\right\} = 0 \ , \tag{132}$$

and

$$\langle M_{n,i}, M_{n,j}\rangle(t) \Rightarrow c_{i,j}t \quad in \quad \mathbb{R} \quad as \quad n \rightarrow \infty \tag{133}$$

for each $t > 0$ and for each (i, j).

Conclusion:

If indeed one of the the conditions (i) or (ii) above holds, then

$$M_n \Rightarrow M \quad in \quad D^k \quad as \quad n \rightarrow \infty\ ,$$

*where M is a **k-dimensional** $(0, C)$**-Brownian motion**, having mean vector and covariance matrix*

$$E[M(t)] = (0, \dots, 0) \quad and \quad E[M(t)M(t)^{tr}] = Ct, \quad t \geq 0\ ,$$

where, for a matrix A, A^{tr} is the transpose.

Of course, a common simple case arises when C is a diagonal matrix; then the k component marginal one-dimensional Brownian motions are independent. When $C = I$, the identity matrix, M is a standard k-dimensional Brownian motion, with independent one-dimensional standard Brownian motions as marginals.

At a high level, Theorem 8.1 says that, under regularity conditions, convergence of martingales in D is implied by convergence of the associated quadratic covariation processes. At first glance, the result seems even stronger, because we need convergence of only the one-dimensional quadratic covariation processes for a single time argument. However, that is misleading, because the stronger weak convergence of these quadratic covariation processes in D^{k^2} is actually equivalent to the weaker required convergence in $\mathbb{R}$ for each t, i, j in conditions (130) and (133); see [73].

9. Applications of the Martingale FCLT

In this section we make two applications of the preceding martingale FCLT. First, we apply it to prove the FCLT for the scaled Poisson process, Theorem 4.2. Then we apply it to provide a third proof of Theorem 1.1. In the same way we could obtain alternate proofs of Theorems 1.2 and 7.1.

9.1. Proof of the Poisson FCLT

We now apply the martingale FCLT to prove the Poisson FCLT in Theorem 4.2. To do so, it suffices to consider the one-dimensional version in D, since the Poisson processes are mutually independent. Let the martingales $M_n \equiv M_{A,n}$ be as defined in (65), i.e.,

$$M_n \equiv M_{A,n}(t) \equiv \frac{A(nt) - nt}{\sqrt{n}}, \quad t \geq 0\ ,$$

with their internal filtrations. The limits in (129) and (132) hold, because

$$J(M_n, T) \leq \frac{1}{\sqrt{n}} \quad \text{and} \quad J(M_n^2, T) \leq \frac{1}{n}, \quad n \geq 1 \ .$$

For each $n \geq 1$, M_n is square integrable, the optional quadratic variation process is

$$[M_n](t) \equiv [M_n, M_n](t) = \frac{A(nt)}{n}, \quad t \geq 0 \ ,$$

and the predictable quadratic variation process is

$$\langle M_n \rangle(t) \equiv \langle M_n, M_n \rangle(t) = \frac{nt}{n} \equiv t, \quad t \geq 0 \ .$$

Hence both (131) and (133) hold trivially. By the SLLN for a Poisson process, $A(nt)/n \rightarrow t$ w.p.1 as $n \rightarrow \infty$ for each $t > 0$. Hence both conditions (i) and (ii) in Theorem 8.1 are satisfied, with $C = c_{1,1} = 1$. ∎

9.2. Completing the Proof of Theorem 1.1

The bulk of this paper has consisted of a proof of Theorem 1.1 based on the first martingale representation in Theorem 3.4, which in turn is based on the representation of the service-completion counting process as a random time change of a rate-1 Poisson process, as in (12). A second proof in §4.3 established the fluid limit directly.

In this subsection we present a third proof of Theorem 1.1 based on the second martingale representation in Theorem 3.5, which in turn is based on a random thinning of rate-1 Poisson processes, as in (14). This third proof also applies to the third martingale representation in §3.6, which is based on constructing martingales for counting processes associated with the birth-and-death process $\{Q(t) : t \geq 0\}$ via its infinitesimal generator.

Starting with the second martingale representation in Theorem 3.5 (or the third martingale representation in Subsection 3.6), we cannot rely on the Poisson FCLT to obtain the required stochastic process limit

$$(M_{n,1}, M_{n,2}) \Rightarrow (\sqrt{\mu} B_1, \sqrt{\mu} B_2) \quad \text{in} \quad D^2 \quad \text{as} \quad n \rightarrow \infty \ , \tag{134}$$

in (59). However, we can apply the martingale FCLT for this purpose, and we show how to do that now.

As in §9.1, we can apply either condition (i) or (ii) in Theorem 8.1, but it is easier to apply (ii), so we will. The required argument looks more complicated because we have to establish the two-dimensional convergence in (134) in D^2 because the scaled martingales $M_{n,1}$ and $M_{n,2}$ are not independent. Fortunately, however, they are orthogonal, by virtue of the following lemma. That still means that we need to establish the two-dimensional limit in (134), but it is not difficult to do so.

We say that two locally square-integrable martingales with respect to the filtration $\mathbf{F}$, M_1 and M_2, are **orthogonal** if the process $M_1 M_2$ is a local martingale with $M_1(0)M_2(0) = 0$. Since $M_1 M_2 - \langle M_1, M_2 \rangle$ is a local martingale, orthogonality implies that $\langle M_1, M_2 \rangle(t) = 0$ for all t.

Lemma 9.1 (orthogonality of stochastic integrals with respect to orthogonal martingales) *Suppose that M_1 and M_2 are locally square-integrable martingales with respect to the filtration $\boldsymbol{F}$, where $M_1(0) = M_2(0) = 0$, while C_1 and C_2 are locally-bounded $\boldsymbol{F}$-predictable processes. If M_1 and M_2 are orthogonal,* (*which is implied by independence*)*, then the stochastic integrals $\int_0^t C_1(s)\,dM_1(s)$ and $\int_0^t C_2(s)\,dM_2(s)$ are orthogonal, which implies that*

$$\left\langle \int C_1(s)\,dM_1(s), \int C_2(s)\,dM_2(s) \right\rangle (t) = 0, \quad t \ge 0 \ .$$

and

$$\left[\int C_1(s)\,dM_1(s), \int C_2(s)\,dM_2(s) \right] (t) = 0, \quad t \ge 0 \ .$$

Lemma 9.1 follows from the following representation for the quadratic covariation of the stochastic integrals; see §5.9 of van der Vaart [64].

Lemma 9.2 (Quadratic covariation of stochastic integrals with respect to martingales) *Suppose that M_1 and M_2 are locally square-integrable martingales with respect to the filtration $\boldsymbol{F}$, while C_1 and C_2 are locally-bounded $\boldsymbol{F}$-predictable processes. Then*

$$\left\{ \int_0^t C_1(s)\,dM_1(s) \int_0^t C_2(s)\,dM_2(s) \right. \\ \left. - \left[\int C_1(s)\,dM_1(s), \int C_2(s)\,dM_2(s) \right] (t) : t \ge 0 \right\}$$

and

$$\left\{ \int_0^t C_1(s)\,dM_1(s) \int_0^t C_2(s)\,dM_2(s) \right. \\ \left. - \left\langle \int C_1(s)\,dM_1(s), \int C_2(s)\,dM_2(s) \right\rangle (t) : t \ge 0 \right\}$$

are local $\boldsymbol{F}$-martingales, where the quadratic covariation processes are

$$\left[\int C_1(s)\,dM_1(s), \int C_2(s)\,dM_2(s) \right] (t) = \int_0^t C_1(s)C_2(s)\,d[M_1, M_2](s)$$

and

$$\left\langle \int C_1(s)\,dM_1(s), \int C_2(s)\,dM_2(s) \right\rangle (t) = \int_0^t C_1(s)C_2(s)\,d\langle M_1, M_2 \rangle(s), \quad t \ge 0 \ .$$

As a consequence of the orthogonality provided by Lemma 9.1, we have $[M_{n,1}, M_{n,2}](t) = 0$ and $\langle M_{n,1}, M_{n,2}\rangle(t) = 0$ for all t and n for the martingales in (134), which in turn come from Theorem 3.5. Thus the orthogonality trivially implies that

$$[M_{n,1}, M_{n,2}] \Rightarrow 0 \quad \text{and} \quad \langle M_{n,1}, M_{n,2}\rangle(t) \Rightarrow 0 \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty$$

for all $t \geq 0$. We then have

$$\langle M_{n,i}, M_{n,i}\rangle(t) \Rightarrow c_{i,i}t = \mu t \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty$$

for each t and $i = 1, 2$ by (41) in Theorem 3.5 and Lemma 4.2, as in the previous argument used in the first proof of Theorem 1.1 in §§4.1-6.2. As stated above, the bulk of the proof is thus identical. By additional argument, we can also show that

$$[M_{n,i}, M_{n,i}](t) \Rightarrow c_{i,i}t = \mu t \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty \ .$$

starting from (42).

We have just shown that (133) holds. It thus remains to show the other conditions in Theorem 8.1 (ii) are satisfied. First, since we have a scaled unit-jump counting process, condition (132) holds by virtue of the scaling in Theorem 3.5 and (30). Next (131) holds trivially because the predictable quadratic variation processes $\langle M_{n,1}\rangle$ and $\langle M_{n,2}\rangle$ are continuous. Hence this third proof is complete. ∎

In closing this section, we observe that this alternate method of proof also applies to the Erlang-A model in §7.1 and the generalization with finite waiting rooms in Theorem 1.2.

Acknowledgments

The authors thank Jim Dai, Itay Gurvich and anonymous referees for helpful comments on this paper. This work was supported by NSF Grant DMI-0457095.

References

[1] Armony, M. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* 51, 287–329. MR2189596

[2] Armony M., I., Gurvich and C. Maglaras. 2006. Cross-selling in a call center with a heterogeneous customer population. Working Paper, New York University, New York, NY, and Columbia University, New York, NY.

[3] Armony M., I. Gurvich and A. Mandelbaum. 2006. Service level differentiation in call Centers with fully flexible servers. *Management Science*, forthcoming.

[4] Atar R. 2005. Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 15, 2606-2650. MR2187306

[5] Atar R., A. Mandelbaum and M. I. Reiman. 2004a. Scheduling a multi-class queue with many exponential servers. *Ann. Appl. Probab.* 14, 1084-1134. MR2071417

[6] Atar R., A. Mandelbaum and M. I. Reiman. 2004b. Brownian control problems for queueing systems in the Halfin-Whitt regime. *Ann. Appl. Probab.* 14, 1084-1134. MR2071417

[7] Bickel P.J. and M.J. Wichura. 1971. Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* 42, 1656–1670. MR0383482

[8] Billingsley, P. 1968. *Convergence of Probability Measures*, Wiley (second edition, 1999). MR1700749

[9] Borovkov, A. A. 1967. On limit laws for service processes in multi-channel systems (in Russian). *Siberian Math J.* 8, 746–763. MR0222973

[10] Borovkov, A. A. 1984. *Asymptotic Methods in Queueing Theory*, Wiley, New York. MR0745620

[11] Borst, S., A. Mandelbaum and M. I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* 52, 17–34. MR2066238

[12] Brémaud, P. 1981. *Point Processes and Queues: Martingale Dynamics*, Springer. MR0636252

[13] Csörgó M. and P. Révéz. 1981. *Strong Approximations in Probability and Statistics*. Akademiai Kiado.

[14] Dai J. G. and T. Tezcan. 2005. State space collapse in many server diffusion limits of parallel server systems. Working Paper, Georgia Institute of Technology, Atlanta, GA.

[15] Dai J. G. and T. Tezcan. 2006. Dynamic control of N systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. Working Paper, Georgia Institute of Technology, Atlanta, GA.

[16] Dai J. G. and T. Tezcan. 2007. Optimal control of parallel server systems with many servers in heavy traffic. Working Paper, Georgia Institute of Technology, Atlanta, GA.

[17] Daley, D. J. and D. Vere-Jones. 2003. *An Introduction to the Theory of Point Processes*, second ed., Springer. MR0950166

[18] Dupuis, P. and H. Ishii. 1991. On when the solution to the Skorohod problem is Lipschitz continuous with applications. *Stochastics* 35, 31–62. MR1110990

[19] Ethier, S. N. and T. G. Kurtz. 1986. *Markov Processes; Characterization and Convergence*, Wiley. MR0838085

[20] Gans, N., G. Koole and A. Mandelbaum. 2003. Telephone call centers: tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5**(2), 79–141.

[21] Garnett, O., A. Mandelbaum and M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4, 208–227.

[22] 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. MR1091098

[23] Gurvich, I. and W. Whitt. 2007a. Queue-and-idleness-ratio controls in many-server servbice systems. working paper, Columbia University. Available at: `http://www.columbia.edu/~ww2040`

[24] Gurvich, I. and W. Whitt. 2007b. Service-level differentiation in many-server service systems: a solution based on fixed-queue-ratio routing. working paper, Columbia University. Available at: `http://www.columbia.edu/~ww2040`

[25] Gurvich, I. and W. Whitt. 2007c. Scheduling Flexible Servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management*, forthcoming. Available at: `http://www.columbia.edu/~ww2040`

[26] Halfin, S. and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29, 567–588. MR0629195

[27] Harrison, J. M. and A. Zeevi. 2005. Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime. *Oper. Res.* 52, 243–257. MR2066399

[28] Helms, L. L. 1974. Ergodic properties of several interacting Poisson particles. *Advances in Math.* 12, 32–57. MR0345247

[29] Iglehart, D. L. 1965. Limit diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* 2, 429–441. MR0184302

[30] Jacod, J. and A. N. Shiryayev. 1987. *Limit Theorems for Stochastic Processes*, Springer. MR0959133

[31] Jelenković, P., A. Mandelbaum and P. Momčilović. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* 47, 53–69. MR2074672

[32] Kallenberg, O. 2002. *Foundations of Modern Probability*, second edition, Springer. MR1876169

[33] Karatzas, I, and S. Shreve. 1988. *Brownian Motion and Stochastic Calculus*, Springer. MR0917065

[34] Kaspi, H. and K. Ramanan. 2007. Law of large numbers limit for many-server queues. Working paper. The Technion and Carnegie Mellon University.

[35] Khoshnevisan D. 2002. *Multiparameter Processes: An Introduction to Random Fields*, Springer. MR1914748

[36] Kogan, Y., R. Sh. Liptser and A. V. Smorodinskii. 1986. Gaussian diffusion approximations of closed Markov models of computer networks. *Problems. Inform. Transmission* 22, 38–51. MR0838688

[37] 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. MR1458591

[38] Kunita, H. and S. Watanabe. 1967. On square-integrble martingales. *Nagoya Math. J.* 30, 209–245. MR0217856

[39] Kurtz, T. 1978. Strong approximation theorems for density dependent Markov chains. *Stoch. Process Appl.* 6, 223–240. MR0464414

[40] Kurtz, T. 1980. Representations of Markov processes as multiparameter time changes. *Ann. Probability* 8, 682–715. MR0577310

[41] Kurtz, T. 2001. *Lectures on Stochastic Analysis*, Department of Mathematics and Statistics, University of Wisconsin, Madison, WI 53706-1388.

[42] Lévy, P. 1948. *Processus Stochastiques et Mouvement Borownien*, Gauthiers-Villars, Paris.

[43] Lions, P. and A. Sznitman. 1984. Stochastic differential equations with reflecting boundary conditions. *Commun. Pure Appl. Math.* 37, 511–537. MR0745330

[44] Liptser, R. Sh. and A. N. Shiryayev. 1989. *Theory of Martingales*, Kluwer (English translation of 1986 Russian edition).

[45] Louchard, G. 1988. Large finite population queueing systems. Part I: The infinite server model. *Comm. Statist. Stochastic Models* 4(3), 473–505. MR0971602

[46] Mandelbaum, A., W. A. Massey and M. I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* 30, 149–201. MR1663767

[47] Mandelbaum, A. and G. Pats. 1995. State-dependent queues: approximations and applications. In *Stochastic Networks*, F. P. Kelly, R. J. Williams (eds.), Institute for Mathematics and its Applications, Vol. 71, Springer, 239–282. MR1381015

[48] Mandelbaum, A. and G. Pats. 1998. State-dependent stochastic networks. Part I: Approximations and Applications with continuous diffusion limits. *Ann. Appl. Prob.* 8, 569–646. MR1624965

[49] Mandelbaum, A. and S. Zeltyn. 2005. The Erlang-A/Palm queue, with applications to call centers. Working paper, The Technion, Haifa, Israel. Available at: `http://iew3.technion.ac.il/serveng/References/references.html`

[50] Massey, W. A. and W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13, 183–250. MR1218848

[51] Parthasarathy, K. R. 1967. *Probability Measures on Metric Spaces*, Academic Press. MR0226684

[52] Puhalskii, A. A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* 19, 946–954. MR1304631

[53] Puhalskii, A. A. and M. I. Reiman. 2000. The mutliclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Prob.* 32, 564-595. MR1778580

[54] Rebolledo, R. 1980. Central limit theorems for local martingales. *Zeitschrift Wahrscheinlichkeitstheorie verw. Gebiete* 51, 269–286. MR0566321

[55] Reed, J. 2005. The $G/GI/N$ queue in the Halfin-Whitt regime. working paper, Georgia Institute of Technology.

[56] Reed, J. and A. R. Ward. 2007. Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: hazard-rate scaling in heavy traffic. working paper, Georgia Institute of Technology.

[57] Robert, P. 2003. *Stochastic Networks and Queues*, Springer. MR1996883

[58] Rogers, L. C. G. and D. Williams. 1987. *Diffusions, Markov Processes and Martingales, Volume 2: Ito Calculus*, Wiley. MR0921238

[59] Rogers, L. C. G. and D. Williams. 2000. *Diffusions, Markov Processes and Martingales, Volume 1: Foundations*, Cambridge University Press.

MR1796539

[60] Skorohod, A. V. 1956. Limit theorems for stochastic processes. *Prob. Theory Appl.* 1, 261–290. MR0084897

[61] Srikant, R. and W. Whitt. 1996. Simulation run lengths to estimate blocking probabilities. *ACM Trans. Modeling Computer Simulations* 6, 7–52.

[62] Stone, C. 1963. Limit theorems for random walks, birth and death processes and diffusion processes. *Illinois J. Math.* 4, 638–660. MR0158440

[63] Tezcan T. 2006. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. Working Paper, University of Illinois at Urbana-Champaign. Available at: `https://netfiles.uiuc.edu/ttezcan/www/TolgaTezcansubmittedMOR121906.pdf`

[64] van der Vaart, A. W. 2006. *Martingales, Diffusions and Financial Mathematics* Lecture Notes, Available at: `http://www.math.vu.nl/sto/onderwijs/mdfm/`

[65] Ward, A. R. and P. W. Glynn. 2003a. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* 43, 103–128. MR1957808

[66] Ward, A. R. and P. W. Glynn. 2003b. Properties of the reflected Ornstein-Uhlenbeck process. *Queueing Systems* 44, 109–123. MR1993278

[67] Ward, A. R. and P. W. Glynn. 2005. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Systems* 50, 371–400. MR2172907

[68] Whitt, W. 1982. On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Prob.* 14, 171–190. MR0644013

[69] Whitt, W. 2002. *Stochastic-Process Limits*, Springer. MR1876437

[70] Whitt, W. 2002a. *Internet Supplement to Stochastic-Process Limits*, Available at: `http://www.columbia.edu/~ww2040/supplement.html`

[71] Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50, 1449–1461.

[72] Whitt, W. 2005. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* 30, 1–27. MR2125135

[73] Whitt, W. 2007. Proofs of the martingale functional central limit theorem. *Probability Surveys*, forthcoming.

[74] Zeltyn S. and A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems* 51, 361–402. MR2189598

APPENDIX

Appendix A: Proof of Lemma 3.1

Let Δ be the **jump function**, i.e., $\Delta X(t) \equiv X(t) - X(t-)$ and let

$$\sum_{s \le t} (\Delta X(s))^2, \quad t \ge 0 \ ,$$

represent the sum of all squared jumps. Since $(\Delta X(s))^2 \geq 0$, the sum is independent of the order. Hence the sum is necessarily well defined, although possibly infinite. (For any positive integer n, a function in D has at most finitely many jumps with $|x(t) - x(t-)| > 1/n$ over any bounded interval; see Lemma 1 on p. 110 of Billingsley [8]. Hence, the sample paths of each stochastic process in D have only countably many jumps.)

The optional quadratic variation has a very simple structure for locally-bounded-variation processes, i.e., processes with sample paths of bounded variation over every bounded interval; see (18.1) on p. 27 of Rogers and Williams [58].

Lemma A.1 (optional quadratic covariation and variation for a locally-bounded-variation process) *If stochastic processes X and Y almost surely have sample paths of bounded variation over each bounded interval, then*

$$[X,Y](t) = \sum_{s \leq t} (\Delta X(s))(\Delta Y(s)), \quad t \geq 0 \ , \tag{135}$$

and

$$[X](t) \equiv [X,X](t) = \sum_{s \leq t} (\Delta X(s))^2, \quad t \geq 0 \ . \tag{136}$$

Lemma A.2 (optional quadratic variation for a counting process) *If N is a non-explosive unit-jump counting process with compensator A, both adapted to the filtration $\boldsymbol{F}$, then $N - A$ is a locally square-integrable martingale of locally bounded variation with square-bracket process*

$$\begin{aligned} [N-A](t) &= \sum_{s \leq t} (\Delta(N-A)(s))^2 \qquad (137) \\ &= N(t) - 2\int_0^t \Delta A(s)\, dN(s) + \int_0^t \Delta A(s)\, dA(s), \quad t \geq 0 \ . \end{aligned}$$

If, in addition, the compensator A is continuous, then

$$[N-A] = N \ , \tag{138}$$

Proof. For (137), we apply (136) in Lemma A.1; see §5.8 of van der Vaart [64]. For (138), we observe that the last two terms in (137) become 0 when A is continuous. ■

Proof of Lemma 3.1. We will directly construct the compensator of the submartingale $\{M(t)^2 : t \geq 0\}$. Since (i) N is a non-explosive counting process, (ii) $E[N(t)] < \infty$ and $E[A(t)] < \infty$ for all t and (iii) N and A have nondecreasing sample paths, the sample paths of the martingale $M \equiv N - A$ provided by Theorem 3.1 are of bounded variation over bounded intervals. Since $N(0) = 0$, $M(0) = 0$. We can thus start by applying integration by parts, referred to as the product formula on p. 336 of Brémaud [12], to write

$$\begin{aligned} M(t)^2 &= \int_0^t M(s-)\,dM(s) + \int_0^t M(s)\,dM(s) \\ &= 2\int_0^t M(s-)\,dM(s) + \int_0^t (M(s) - M(s-))\,dM(s) \\ &= 2\int_0^t M(s-)\,dM(s) + [M](t) \\ &= 2\int_0^t M(s-)\,dM(s) + N(t), \quad t \ge 0 \ , \end{aligned} \tag{139}$$

The last step follows from (138) in Lemma A.2.

We now want to show that the stochastic integral $\int_0^t M(s-)\,dM(s)$ is a martingale. To get the desired preservation of the martingale structure, we can apply the integration theorem, Theorem T6 on p. 10 of [12], but we need additional regularity conditions. At this point, we localize in order to obtain boundedness.

To obtain such bounded martingales associated with N and $M \equiv N - A$, let the stopping times be defined in the obvious way by

$$\tau_m \equiv \inf\{t \ge 0 : |M(t)| \ge m \quad \text{or} \quad A(t) \ge m\} \ . \tag{140}$$

Then $\{\tau_m \le t\} \in \mathcal{F}_t$, $t \ge 0$. We now define the associated stopped processes: Let

$$M^m(t) \equiv M(t \wedge \tau_m), \quad N^m(t) \equiv N(t \wedge \tau_m) \quad \text{and} \quad A^m(t) \equiv A(t \wedge \tau_m) \tag{141}$$

for all $t \ge 0$ and $m \ge 1$. Then $N^m(t) = M^m(t) + A^m(t)$ for $t \ge 0$ and $\{M^m(t) : t \ge 0\}$ is a martingale with respect to $\{\mathcal{F}_t\}$ having compensator $\{A^m(t) : t \ge 0\}$ for each $m \ge 1$, as claimed. Moreover, all three stochastic processes N^m, M^m and A^m are bounded. The boundedness follows since N has unit jumps and A is continuous.

We then obtain the representation for these stopped processes corresponding to (139); in particular,

$$M^m(t)^2 = 2\int_0^t M^m(s-)\,dM^m(s) + N^m(t), \quad t \ge 0 \ , \tag{142}$$

for each $m \ge 1$. With the extra boundedness provided by the stopping times, we can apply Theorem T6 on p. 10 of [12] to deduce that the integral $\int_0^t M^m(s-)dM^m(s)$ is an $\mathbf{F}$-martingale. First, $M^m(t)$ is a martingale of integrable bounded variation with respect to $\mathbf{F}$, as defined on p. 10 of [12]. Moreover, the process $\{M^m(t-) : t \ge 0\}$ is an $\mathbf{F}$-predictable process such that

$$\int_0^t |M^m(s-)|\,d|M^m|(s) < \infty \ .$$

By Theorem T6 on p. 10 of [12], the integral $\int_0^t M^m(s-)dM^m(s)$ is an $\mathbf{F}$-martingale. Thus $\{M^m(t)^2 - N^m(t) : t \ge 0\}$ is an $\mathbf{F}$-martingale. But $\{N^m(t) -$

$A^m(t) : t \geq 0\}$ is also an $\mathbf{F}$-martingale. Adding, we see that $\{M^m(t)^2 - A^m(t) : t \geq 0\}$ is an $\mathbf{F}$-martingale for each m. Thus, for each m, the predictable quadratic variation of M^m is $\langle M^m \rangle(t) = A^m(t)$, $t \geq 0$.

Now we can let $m \uparrow \infty$ and apply Fatou's Lemma to get

$$\begin{aligned} E[M(t)^2] &= E\Big[\lim_{m\to\infty} M^m(t)^2\Big] \\ &\leq \liminf_{m\to\infty} E\Big[M^m(t)^2\Big] = \liminf_{m\to\infty} E\Big[A^m(t)\Big] = E\Big[A(t)\Big] < \infty. \end{aligned}$$

Therefore, M itself is square integrable. We can now apply the monotone convergence theorem in the conditioning framework, as on p. 280 of [12], to get

$$E[M^m(t+s)^2|\mathcal{F}_t] \to E[M(t+s)^2|\mathcal{F}_t] \quad \text{and} \quad E[A^m(t+s)|\mathcal{F}_t] \to E[A(t+s)|\mathcal{F}_t]$$

as $m \to \infty$ for each $t \geq 0$ and $s > 0$. Then, since

$$E[M^m(t+s)^2 - A^m(t+s)|\mathcal{F}_t] = M^m(t)^2 - A^m(t) \quad \text{for all} \quad m \geq 1 \ ,$$

$M^m(t) \to M(t)$ and $A^m(t) \to A(t)$, we have

$$E[M(t+s)^2 - A(t+s)|\mathcal{F}_t] = M(t)^2 - A(t)$$

as well, so that $M^2 - A$ is indeed a martingale. Of course that implies that $\langle M \rangle = A$, as claimed. We get $[M] = N$ from Lemma A.2, as noted at the beginning of the proof. ■

We remark that there is a parallel to Lemma A.2 for the angle-bracket process, applying to cases in which the compensator is not continuous. In contrast to Lemma 3.1, we now do not assume that $E[N(t)] < \infty$, so we need to localize.

Lemma A.3 (predictable quadratic variation for a counting process) *If N is a non-explosive unit-jump counting process with compensator A, both adapted to the filtration $\boldsymbol{F}$, then $N - A$ is a locally square-integrable martingale of locally bounded variation with angle-bracket process*

$$\begin{aligned} \langle N - A \rangle(t) &= \langle [N - A] \rangle(t) \qquad (143) \\ &= A(t) - \int_0^t \Delta A(s)\, dA(s) = \int_0^t (1 - \Delta A(s))\, dA(s), \quad t \geq 0 \ . \end{aligned}$$

If, in addition, the compensator A is continuous, then

$$\langle N - A \rangle = A \ . \qquad (144)$$

Proof. For (143), we exploit the fact that $\langle N - A \rangle = \langle [N - A] \rangle$; see p. 377 of Rogers and Williams [58] and §5.8 of van der Vaart [64]. The third term on the right in (137) is predictable and thus its own compensator. The compensators of the first two terms in (137) are obtained by replacing N by its compensator A. See Problem 3 on p. 60 of Liptser and Shiryayev [44]. ■