Queueing Systems manuscript No.
(will be inserted by the editor)

05/03/12

Diffusion approximation for an overloaded X model via an averaging principle

Ohad Perry · Ward Whitt

Received: date / Accepted: date

Abstract In previous papers we developed a deterministic fluid approximation for an overloaded Markovian queueing system having two customer classes and two service pools, known in the call-center literature as the X model. The system uses the fixed-queue-ratio-with-thresholds (FQR-T) control, which we proposed as a way for one service system to help another in face of an unexpected overload. Under FQR-T, customers are served by their own service pool until a threshold is exceeded. Then, one-way sharing is activated with customers from one class allowed to be served in both pools. The control aims to keep the two queues at a pre-specified fixed ratio. We supported the fluid approximation by establishing a many-server heavy-traffic functional weak law of large numbers (FWLLN) involving an averaging principle. In this paper we develop a refined diffusion approximation for the same model based on a many-server heavy-traffic functional central limit theorem (FCLT).

1 Introduction

In this paper we establish a many-server heavy-traffic *functional central limit theorem* (FCLT) for an overloaded large-scale Markovian queueing system hav-

Ohad Perry
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA
Tel.: +847-467-443
Fax: +847-491-8005
E-mail: ohad.perry@northwestern.edu

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA
Tel.: +212-854-7255
Fax: +212-854-8103
E-mail: ww2040@columbia.edu

ing two classes and two service pools, known as the X model [7], using the *fixed-queue-ratio with thresholds* (FQR-T) routing, which we proposed in [21].

In particular, we consider a system in which each class has its own designated service pool, but with all agents, in both pools, capable of serving customers from both classes. The control aims to prevent sharing of customers (i.e., sending customers from one class to be served at the other class pool) when both classes are normally loaded, and activate sharing when the system unexpectedly experiences an overloaded, due to an unforseen shift in the arrival rates.

When sharing is taking place, the control aims at keeping a pre-specified fixed ratio between the two queues. This ratio is chosen according to a deterministic ("fluid") optimization problem; see §5.3 in [21], where it is also shown that sharing should not be allowed at both directions simultaneously, i.e., at any time there should be at most one pool working with both classes. In general there are two different ratios: If class 1 is overloaded, then an optimal ratio $r_{1,2}$ should hold between the queues. If class 2 is overloaded, then an optimal ratio $r_{2,1}$ should hold between the queues. In [23] we showed that the FQR-T control achieves the objectives above, asymptotically in the fluid limit. Moreover, the FQR-T control produces a tractable fluid limit. Here we establish a refined stochastic limit.

The FQR-T control here is a modification of the FQR control (without the thresholds), which is a special case of the *queue-and-idleness ratio* (QIR) controls suggested in [11]. These QIR and FQR controls were analyzed in [10], [11] and [12] for critically loaded systems, operating in the *quality and efficiency driven* (QED) many-server heavy-traffic regime; see [8,13]. Heavy-traffic limits for networks having cyclic graphs, such as the X model, were obtained under the condition that the service rates are class or pool dependent; see Theorem 3.1 in [11]. In general, when the service rate depends on both the class and the pool, FQR can perform badly in cyclic networks, creating severe congestion even if each pool is not congested by itself; see §4.1 in [21] and §EC.2 in [22].

We suggested the FQR-T control in [21], and analyzed the X model using a heuristic stationary fluid approximation. In [22] we determined the transient behavior of that same fluid model, based on a stochastic *averaging principle* (AP), but that AP was introduced there as a heuristic engineering principle, supported only by simulation. The purpose of our subsequent papers [23,24] was to establish key mathematical properties of the fluid model, expressed as an *ordinary differential equation* (ODE), and show that the fluid model, heuristically derived in [21,22], arises as the many-server heavy traffic limit of a sequence of X models in the many-server *efficiency driven* (ED) regime. That FWLLN is challenging, because the fluid limit depends critically on the AP. For each n, the system evolves as a 6-dimensional *continuous-time Markov chain* (CTMC), but there is (a somewhat complicated) statistical regularity associated with the many-server heavy-traffic limit. In particular, the limiting fluid approximation is a deterministic function characterized by an ODE (and an initial condition), which is driven by the time-varying instantaneous average

behavior of a family of *fast-time-scale stochastic processes* (FTSP's), which produces the AP. See §1.3 of [24] for a discussion of the literature on AP's; notable contributions in the queueing literature are [4,15]. See [6] for a FCLT involving an AP, building on [15].

We now build on the FWLLN and the AP to describe the distribution of the stochastic fluctuations about the fluid path; i.e., we establish the corresponding FCLT, which is Theorem 4 here. There is technical novelty in properly treating the FTSP's alluded to above. The limit process involves an independent Brownian motion term with deterministic time scaling involving the asymptotic variance of the FTSP; see §4.1 and $\hat{L}_2$, $\hat{I}$, γ_2 and γ_3 in Theorem 4. A key step in establishing the main result – the FCLT in Theorem 4 – is a FCLT for the family of FTSP's, Theorem 6, which is of independent interest. This challenging step proves a FCLT for a sequence of CTMC's having time-varying parameters depending on the fluid limit. The new methods developed here should prove useful for analyzing related problems.

From an engineering perspective, Corollary 1 is especially useful for understanding the performance of the FQR-T control. It describes the stochastic-process limit once the fluid has stabilized (i.e. when the fluid is stationary). With a constant fluid state, the key limit process becomes the well-studied *bivariate Ornstein-Uhlenbeck* (BOU) process, which has a Gaussian distribution for each t; see Corollary 1 below. Consequently, the approximating steady-state distribution during the overload is a Gaussian distribution, with mean values equal to the stationary fluid point in Theorem 2 multiplied by n, and variance and covariance terms in (24) multiplied by $\sqrt{n}$.

The FCLT extension is essential for truly understanding the system performance under overloads, because the actual performance is not nearly deterministic, as described by the fluid approximation, unless the scale is extremely large. This phenomenon is well illustrated by the example here in §11. For that example, the standard deviations of the queue lengths are about equal to (half of) the mean queue lengths when the number of servers in each pool is 25 (100).

Here is how the paper is organized: After preliminaries in §2, we briefly state the FWLLN and the associated WLLN for the stationary distributions in §3. We state the FCLT and our other main results in §4. We prove the FCLT in §5 except for Lemma 6, establishing joint convergence of the driving processes. We give the proof of Lemma 6 in §6 except for two supporting results. The key supporting result is a FCLT for the FTSP with time-varying parameter state function in Theorem 6. We prove Theorem 6 in §7. Our proof of Lemma 14 to prove Theorem 6 exploits the martingale FCLT for triangular arrays. We state these supporting martingale results in §8. We then prove five remaining lemmas in §9. A key technical step in the proofs is approximating the given process with time-varying parameters over appropriate subintervals by associated *frozen* processes, where the parameters are fixed (frozen) at designated values. Those approximation steps are justified in §10 by using coupling constructions. In particular, we prove Lemmas 8 and 12 there. Finally, we evaluate the quality of the approximations by making comparisons with simulations in §11.

2 Preliminaries

2.1 Notation

Let $\mathbb{R}$, $\mathbb{Z}$ and $\mathbb{N}$ denote the real numbers, integers and nonnegative integers, respectively. Let $\equiv$ denote equality by definition. For a subinterval I of $[0,\infty)$, let $\mathcal{D} \equiv \mathcal{D}(I) \equiv \mathcal{D}(I,\mathbb{R})$ be the space of all right-continuous $\mathbb{R}$-valued functions on I with limits from the left everywhere, endowed with the familiar Skorohod J_1 topology [32]. Let $\mathcal{C}$ be the subset of continuous functions in $\mathcal{D}$. Let a subscript k appended to one of these spaces denote the set of all k-dimensional vectors with components from the space, endowed with the corresponding product topology, e.g., $\mathbb{R}_k$ and $\mathcal{D}_k$.

Let d_{J_1} denote a metric on $\mathcal{D}_k(I)$ inducing the convergence. Since we will be considering continuous limits, the topology is equivalent to uniform convergence on compact subintervals of I. Let e be the identity function in $\mathcal{D} \equiv \mathcal{D}_1$; i.e., $e(t) \equiv t$, $t \in I$. Let $\circ$ be the composition function, i.e., $(x \circ y)(t) \equiv x(y(t))$. Let $\Rightarrow$ denote convergence in distribution [32].

We use the familiar big-O and small-o notation for deterministic functions: For two real functions f and g, we write

$$f(x) = O(g(x)) \quad \text{whenever} \quad 0 < \limsup_{x\to\infty} |f(x)/g(x)| < \infty,$$
$$f(x) = o(g(x)) \quad \text{whenever} \quad \limsup_{x\to\infty} |f(x)/g(x)| = 0.$$

(Note that our definition of $O(g(x))$ deviates from the standard definition which allows for the lim sup in the right-hand side to be equal to 0.) For a function $x : [0,\infty) \to \mathbb{R}$ and $0 < t < \infty$, let $\|x\|_t \equiv \sup_{0\le s\le t} |x(s)|$.

For a stochastic process $Y \equiv \{Y(t) : t \ge 0\}$ and a deterministic function $f : [0,\infty) \to [0,\infty)$, we say that Y is $o_P(f(t))$ if $\|Y\|_t / f(t) \Rightarrow 0 \quad \text{as} \quad t \to \infty$.

For a sequence of stochastic processes or random variables, $\{Y^n : n \ge 1\}$, we denote its fluid-scaled version by $\bar{Y}^n \equiv Y^n/n$. We let $\breve{Y}^n \equiv Y^n/\sqrt{n}$ be the $\sqrt{n}$-scaled processes without the centering about the fluid limit, and $\hat{Y}^n$ denote the diffusion-scaled processes centered about the fluid limit, as in (15) below.

2.2 A Sequence of Overloaded Markovian X Models

We consider a sequence of overloaded Markovian X models, indexed by superscript n. There are two customer classes and two service pools. We are looking at these models during the overload incident, after the arrival rates have changed. The arrival rates are considered fixed, but the system is typically not yet in its new steady-state during the overload (assuming that the overload would persist). For each n and $i = 1, 2$, there is a class-i Poisson arrival process with rate λ_i^n. Customers have limited patience, and may abandon when waiting in queue. The times to abandon are i.i.d. exponential variables

with rate θ_i for each class-i customer in queue. Service pool j has m_j^n homogeneous agents (servers). Service times of class-i customers by pool-j agents are mutually independent and exponentially distributed with rate $\mu_{i,j}$, $i,j = 1,2$. The abandonment and service rates are independent of n.

Since we are considering an overload incident, we will scale to achieve an efficiency-driven (ED) many-server heavy-traffic regime.

Assumption 1 (*many-server heavy-traffic scaling*) *For* $\lambda_i, m_i > 0$, $i = 1,2$,

$$\frac{\lambda_i^n - n\lambda_i}{\sqrt{n}} \to 0 \quad \textit{and} \quad \frac{m_i^n - nm_i}{\sqrt{n}} \to 0 \quad \textit{as} \quad n \to \infty.$$

We could instead obtain a modified, more general, FCLT if there were nondegenerate limits in Assumption 1, but we consider our choice natural, because the system operates in an overload regime. (The modified limit includes a deterministic term ct in the diffusion limit, but there is no difference in the variability of the limit process, as can be seen from (30). For the FWLLN, it is sufficient that $\lambda_i^n/n \to \lambda_i$ and $m_i^n/n \to m_i$ as $n \to \infty$, $i = 1,2$.)

Let

$$\rho_i \equiv \frac{\lambda_i}{m_i\mu_{i,i}} \quad \text{and} \quad q_i^a \equiv \frac{(\lambda_i - \mu_{i,i}m_i)^+}{\theta_i}, \quad i = 1,2,$$

where, for $y \in \mathbb{R}$, $y^+ \equiv \max\{0, y\}$. Then ρ_i is the traffic intensity for pool i and q_i^a is the stationary class-i fluid-limit queue, when both pools operate independently. We say that pool i is overloaded if $\rho_i > 1$. However, with sharing allowed, pool i can be overloaded even if $\rho_i < 1$ provided that enough class j customers are routed to be served there, $j \neq i$. The next assumption makes precise our notion of system overload.

Assumption 2 (*system overload, with class* 1 *more overloaded*)
The rates in the system are such that
(I) $\theta_1 q_1^a > \mu_{1,2}m_2(1-\rho_2)^+$ *and* (II) $q_1^a > r_{1,2}q_2^a$.

Clearly, $\rho_1 > 1$ by Condition (I), so that class 1 is overloaded. However, Condition (I) also ensures that pool 2 is overloaded if sharing is taking place. That is so because, even if $\rho_2 < 1$, there is not enough extra service capacity in pool 2 to take care of all the class-1 customers that pool 1 cannot serve. Condition (II) in the assumption implies that even if pool 2 is overloaded by itself (i.e., if $\rho_2 > 1$), then class 1 is the one that should receive help from pool 2.

2.3 The FQR-T Control

We now describe the FQR-T control for each system n. The purpose of the FQR-T control is: (i) to prevent sharing under normal loads, (ii) to activate sharing as soon as an overload incident begins, and (iii) to keep close to the desired ratio between the two queues, making sure that sharing takes place in the needed direction only. The control is based on two positive thresholds, $k_{1,2}^n$

and $k^n_{2,1}$, and the two ratio parameters discussed above, $r_{1,2}$ and $r_{2,1}$, which satisfy $r_{1,2} \geq r_{2,1}$; see Proposition EC.2 and Equation (EC.11) in [21].

Let $Q^n_i(t)$ be the number of customers in the class-i queue and let $Z^n_{i,j}(t)$ be the number of class-i customers being served in service pool j, at time t, $i, j = 1, 2$ (in the n^{th} system). The FQR-T routing is based on the queue-difference stochastic processes

$$\begin{aligned} D^n_{1,2}(t) &\equiv Q^n_1(t) - k^n_{1,2} - r_{1,2}Q^n_2(t), \quad \text{and} \\ D^n_{2,1}(t) &\equiv r_{2,1}Q^n_2(t) - k^n_{2,1} - Q^n_1(t), \quad t \geq 0. \end{aligned} \tag{1}$$

As long as $D^n_{1,2}(t) \leq 0$ and $D^n_{2,1}(t) \leq 0$, no sharing of customers is allowed, i.e., a server in pool j takes only class j customers, $j = 1, 2$. It follows from [8] that thresholds of order larger than $O(\sqrt{n})$ will prevent sharing in such circumstances, asymptotically as $n \to \infty$. Once one of the queue-difference processes in (1) becomes strictly positive (so that one of the thresholds is crossed) sharing is initiated. It follows from the Corollary 2.1 in [33], that thresholds of size $o(n)$ will detect an overload relatively quickly (instantly, asymptotically as $n \to \infty$). We thus choose the thresholds according to the following assumption.

Assumption 3 (*scaling of the thresholds*) *For $k_{1,2}, k_{2,1} > 0$ and a sequence of positive numbers $\{c_n : n \geq 1\}$, where $c_n/n \to 0$ and $c_n/\sqrt{n} \to \infty$ as $n \to \infty$,*

$$k^n_{1,2}/c_n \to k_{1,2} \text{ and } k^n_{2,1}/c_n \to k_{2,1} \quad \text{as} \quad n \to \infty.$$

Finally, only one-way sharing is allowed at any time. For example, a newly available pool-2 agent at time t serves a class-1 customer if $D^n_{1,2}(t) > 0$, provided no class-2 customers are served in pool 1 at that same time t; otherwise he serves a class-2 customer.

2.4 Dimension Reduction

For the X model operating under FQR-T, the six-dimensional process

$$X^n_6 \equiv (Q^n_1, Q^n_2, Z^n_{1,1}, Z^n_{1,2}, Z^n_{2,1}, Z^n_{2,2}) \tag{2}$$

is a CTMC for each $n \geq 1$. However, there is an important dimension reduction established in §6 of [24]. It was shown, under the assumptions above and with appropriate initial conditions, that asymptotically the two service pools remain fully occupied with no pool-1 servers serving class 2; i.e., for each $T > 0$,

$$P(Z^n_{1,1}(t) = m^n_1, Z^n_{2,1}(t) = 0, Z^n_{1,2} + Z^n_{2,2} = m^n_2, 0 \leq t \leq T) \to 1 \quad \text{as} \quad n \to \infty.$$

Thus, the system is characterized by an essentially three-dimensional process

$$X^{n,*}_6 \equiv (Q^n_1, Q^n_2, m^n_1, Z^n_{1,2}, 0, m^n_2 - Z^n_{1,2}), \tag{3}$$

having the vector of essential components

$$X^n \equiv (Q^n_1, Q^n_2, Z^n_{1,2}), \tag{4}$$

whose evolution is directly specified, and will be specified here in Theorem 1. Theorem 1 concludes that $\bar{X}_6^{n,*}$ and $\bar{X}_6^n$ are asymptotically equivalent, so that $\bar{X}^n$ is sufficient to characterize the FWLLN and, in turn, to prove the FCLT. That implies that $\bar{X}_6^n \Rightarrow x_6$ in $\mathcal{D}_6$ if and only if $\bar{X}^n \Rightarrow x$ in $\mathcal{D}_3$ as $n \to \infty$, with $x(t) \in \mathbb{S} \equiv [0,\infty)^2 \times [0, m_2]$, for $t \geq 0$; see Theorem 1 below. We thus restrict attention to the space $\mathbb{S}$.

2.5 The Fast-Time-Scale Process

Given that the system is overloaded with class 1 needing help from pool 2, as determined by Assumptions 1 and 2, the FQR-T control is driven by the process $D_{1,2}^n$ in (1). Since the queue lengths are asymptotically of order $O(n)$, the queue-difference process $D_{1,2}^n$ has transitions at rate $O(n)$. However, Theorem 4.5 in [24] shows that, under regularity conditions, the sequence $\{D_{1,2}^n(t) : n \geq 1\}$ is stochastically bounded in $\mathbb{R}$, so that the difference process should be analyzed without any spatial scaling. On the other hand, Theorem 4.4 in [22] also shows that this sequence is *not* $\mathcal{D}$-tight. Thus, these difference processes do not converge to nondegenerate limits in $\mathcal{D}$ as $n \to \infty$ without spatial scaling. Nevertheless, both the FWLLN and FCLT depend heavily on the asymptotic behavior of functionals of that driving queue-difference process and on the analysis of a related family of *fast time scale process* (FTSP's).

Fix $t_0 \geq 0$ and consider the *time expanded queue-difference process*

$$\{D_e^n(\mathit{\Gamma}^n, s) : s \geq 0\} \equiv \{D_{1,2}^n(t_0 + s/n) : s \geq 0\}, \tag{5}$$

where $\mathit{\Gamma}^n$ is a random vector in $\mathbb{R}_3$, representing a possible state of X^n, and we condition on $X^n(t_0) = \mathit{\Gamma}^n$. Theorem 4.4 in [24] shows, under the assumptions of the FWLLN in Theorem 1 below, that

$$\{D_e^n(\mathit{\Gamma}^n, s) : s \geq 0\} \Rightarrow \{D(\gamma, s) : s \geq 0\} \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \tag{6}$$

if $\mathit{\Gamma}^n/n \Rightarrow \gamma \in \mathbb{S}$ and $D_e^n(\mathit{\Gamma}^n, 0) \Rightarrow D(\gamma, 0)$ in $\mathbb{R}$ as $n \to \infty$. The limit process $D(\gamma, \cdot)$ is the FTSP, an irreducible pure-jump (time homogeneous) Markov process having transition rates that are the limit of the instantaneous rates of $D_{1,2}^n(t_0)$ at time t_0 (given the state of the CTMC $X_6^n(t_0)$), divided by n. Since the distribution of the FTSP is determined by γ, we obtain a different FTSP $D(\gamma, \cdot)$ for each $\gamma \in \mathbb{S}$, and thus for each $t \geq 0$. The name "FTSP" becomes clear when observing that it arises as the limit in (5) achieved by "slowing" time in the neighborhood of each time point t_0 in $D_{1,2}^n(t_0)$.

As explained in §2.3, the purpose of the FQR-T control during overload periods (with class 1 receiving help) is to keep the two queues approximately fixed at the target ratio r. In this paper we will be concerned with the region of the state space in which $q_1 = rq_2$ and the FTSP is positive recurrent. In particular, for $\gamma \equiv (q_1, q_2, z_{1,2})$ we let

$$\mathbb{S}^b \equiv \{\gamma \in \mathbb{S} : q_1 = rq_2\}$$

denote the 'boundary' set of points in $\mathbb{S}$ which is part of the state space to which the control drives the process. We then let $\mathbb{A}$ denote the set of all $\gamma \in \mathbb{S}^b$, such that $D(\gamma, \cdot)$ is positive recurrent, with $D(\gamma, \infty)$ denoting a random variable distributed as the stationary distribution of the FTSP $D(\gamma, \cdot)$. For each $\gamma \in \mathbb{S}^b$, let

$$\pi_{1,2}(\gamma) \equiv P(D(\gamma, \infty) > 0). \tag{7}$$

By Lemma 3.1 in [24], $\pi_{1,2}(\gamma)$ is well defined for all $\gamma \in \mathbb{S}$, but $D(\gamma, \cdot)$ is positive recurrent if and only if $0 < \pi_{1,2}(\gamma) < 1$ **and** $\gamma \in \mathbb{S}^b$. By Theorem 6.1 of [23],

$$\mathbb{A} = \{\gamma \in \mathbb{S}^b : 0 < \pi_{1,2}(\gamma) < 1\} = \{\gamma \in \mathbb{S}^b : \delta_+(\gamma) < 0 \quad \text{and} \quad \delta_-(\gamma) > 0\}, \tag{8}$$

where $\delta_+(\gamma)$ and $\delta_-(\gamma)$, respectively, are the constant drift rates in the positive region $\{s : D(\gamma, s) > 0\}$ and the non-positive region $\{s : D(\gamma, s) \leq 0\}$.

Both the FWLLN and the FCLT depend critically on distributional and topological characteristics of the FTSP's. A simplification is achieved by representing the FTSP as a *quasi-birth-and-death* (QBD) process, which can be done by assuming that $r_{1,2}$ is rational. The QBD representation is not straightforward, thus we refer to §6.2 in [23] for more details on the QBD representation of the FTSP, and to [18] for the general theory of QBD processes. See also Theorem 6.1 and Equation (7.2) in [23] for how the QBD representation simplifies the characterization of $\mathbb{A}$, as well as §11 in [23], where an efficient algorithm for computing the fluid limit numerically is developed, based on that QBD representation. For our purposes here, it only matters that the FTSP can be analyzed as a QBD, provided that the queue ratios are rational number. We thus make the following assumption.

Assumption 4 (*queue ratios parameters*) *The queue ratios $r_{1,2}$ and $r_{2,1}$ are positive rational numbers.*

Since we are considering the case when sharing is taking place with class-1 customers receiving help, we essentially need only consider $r_{1,2}$, which we henceforth denote by r, i.e., $r \equiv r_{1,2}$.

3 The Fluid Limit

We now review the FWLLN for the process $\bar{X}_6^n$ in (2) and the WLLN for the associated sequence of stationary random variables $\bar{X}_6^n(\infty)$, established in [24]. For these, we assume that the fluid $x(t)$ is in the set $\mathbb{A}$, where the FTSP is positive recurrent. We conclude by reviewing a result stating that the fluid model eventually remains in $\mathbb{A}$.

3.1 The FWLLN

We now describe the fluid limit, i.e., the limit of $\bar{X}_6^n$ for X_6^n in (2). The FWLLN requires an assumption about the initial conditions. In [24] we considered a (more general) version of the following.

Assumption 5 *Assume that*

$$P(Z^n_{2,1}(0) = 0, Q^n_i(0) > a_n, i = 1, 2) = 1 \quad \textit{for all } n \geq 1,$$
$$\bar{X}^n(0) \Rightarrow x(0) \in \mathbb{A} \quad \textit{and} \quad D^n_{1,2}(0) \Rightarrow L \quad \textit{as } n \to \infty,$$

where L is a finite random variable, $x(0)$ is deterministic and $\{a_n : n \geq 1\}$ is a sequence of numbers satisfying $a_n/c_n \to a$, $0 < a \leq \infty$, for c_n in Assumption 3.

We note that in [24] $x(0)$ was not necessarily in $\mathbb{A}$. The following theorem is a version of the main result - Theorem 4.1 - in [24], adapted to our needs here.

Theorem 1 (FWLLN) *Under Assumptions 1-5,*

$$\bar{X}^n_6 \Rightarrow x_6 \quad \textit{in} \quad \mathcal{D}_6([0, \infty)) \quad \textit{as} \quad n \to \infty,$$

for X^n_6 in (2), where $x_6 \equiv (q_i, z_{i,j}; i, j = 1, 2)$, is a deterministic element of $\mathcal{C}_6$, with $z_{1,1} = m_1 e$, $z_{2,1} = 0e$ and $z_{2,2} = m_2 e - z_{1,2}$ and $x \equiv (q_1, q_2, z_{1,2})$ being the unique solution to the three-dimensional ODE

$$\begin{aligned}
\dot{q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(x(t))\left[z_{1,2}(t)\mu_{1,2} + z_{2,2}(t)\mu_{2,2}\right] - \theta_1 q_1(t) \\
\dot{q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(x(t)))\left[z_{2,2}(t)\mu_{2,2} + z_{1,2}(t)\mu_{1,2}\right] - \theta_2 q_2(t) \\
\dot{z}_{1,2}(t) &\equiv \pi_{1,2}(x(t))z_{2,2}(t)\mu_{2,2} - (1 - \pi_{1,2}(x(t)))z_{1,2}(t)\mu_{1,2},
\end{aligned} \tag{9}$$

for $\pi_{1,2}(x(t)) \equiv P(D(x(t), \infty) > 0)$ in (7). Moreover, there exists δ, $0 < \delta \leq \infty$, such that $x(t) \in \mathbb{A}$, so that $0 < \pi_{1,2}(x(t)) < 1$ and $q_1(t) = rq_2(t)$, for all $t \in [0, \delta)$.

Just as the routing of customers at each time $t \geq 0$ in the prelimit is determined by whether $D^n_{1,2}(t) > 0$ or ≤ 0, so also the instantaneous future evolution of the fluid limit $x(t)$ at time $t \geq 0$, is determined by whether the FTSP corresponding to $x(t)$, $D(x(t), \cdot)$, is positive or nonpositive. However, that evolution is determined by the *long-run average behavior* of the FTSP corresponding to time t, i.e., by $\pi_{1,2}(x(t))$, giving rise to the term "averaging principle". Loosely speaking, $D^n_{1,2}(t)$ achieves a local steady state (the steady state of the FTSP) instantaneously as $n \to \infty$, at each time $t \geq 0$.

Observe that Theorem 1 concludes that if $x(0) \in \mathbb{A}$, then $x(t) \in \mathbb{A}$ for all t over some interval $[0, \delta)$ (that part of the theorem follows from Theorem 4.5 in [24]), so that we have SSC in the sense that the original six-dimensional process is a deterministic function of a two-dimensional process. More importantly for the FCLT, we also have that $Q^n_1(t) - k^n_{1,2} - rQ^n_2(t) = o(\sqrt{n})$ for $t \in (t_1, t_2)$ if $x(t) \in \mathbb{A}$ over $[t_1, t_2)$, so the SSC to two dimensions holds in diffusion scale as well; see Lemma 2 below.

3.2 The Stationary Fluid Limit

Our main theorem here will be establishing the FCLT about the fluid trajectory, given that the trajectory is in $\mathbb{A}$. An important consequence will be the BOU limit when the fluid limit is stationary. Since the fluid limit of $\bar{X}^n$ in (4) is the unique solution to the ODE (9), there is an immediate equivalence between stationarity of the fluid limit and stationarity of the dynamical system in (9), and we do not distinguish between the two.

Definition 1 (fluid stationarity) A point $x^* \in \mathbb{S}$ is a stationary point of the unique solution $x \equiv \{x(t) : t \geq 0\}$ to the ODE (9) if $x(0) = x^*$ implies $x = x^*e$. If $x = x^*e$, then x is said to be stationary.

Since the ODE is autonomous (i.e., time invariant), we can replace time 0 with any $t > 0$ in the definition 1. That is, if $x(T) = x^*$ for some $T > 0$, then $x(t) = x^*$ for all $t > T$. Time invariance also implies that $x(t)$ is stationary at time t ($x(t) = x^*$) if and only if $\dot{x}(t) \equiv (\dot{q}_1(t), \dot{q}_2(t), \dot{z}_{1,2}(t)) = (0, 0, 0)$; see §8 of [23].

There are several issues regarding stationarity, which we addressed in [23]. In advance, neither existence of a stationary point to the fluid limit nor uniqueness are immediate. Even if there exists a unique stationary point, it needs to be identified. Moreover, it must be shown that the fluid limit converges to a stationary point as $t \to \infty$. (There are still other issues regarding stability of the dynamical system in (9), and we refer to §8.3 in [23] for a discussion.) Finally, the fluid limit of $\bar{X}_6^n$ in (2) is characterized by the fluid limit of the three-dimensional $\bar{X}^n$ in (4), but that does not directly imply any relation between the stationary fluid limit and the stationary stochastic prelimit.

We now present the most relevant results for the FCLT regarding fluid stationarity.

Theorem 2 (*fluid stationarity*) *Under Assumptions 1-5, the following hold:*

(*i*) *For each n,* $\bar{X}_6^n(t) \Rightarrow \bar{X}_6^n(\infty)$ *in* $\mathbb{R}$ *as* $t \to \infty$*, with* $\bar{X}_6^n(\infty)$ *being the unique stationary distribution of the CTMC, and* $\bar{X}_6^n(\infty) \Rightarrow x_6^*$ *in* $\mathbb{R}$ *as* $n \to \infty$ *for*

$$x_6^* \equiv (q_1^*, q_2^*, m_1, z_{1,2}^*, 0, m_2 - z_{1,2}^*), \tag{10}$$

where

$$z_{1,2}^* = \frac{\theta_2(\lambda_1 - m_1\mu_{1,1}) - r\theta_1(\lambda_2 - m_2\mu_{2,2})}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} \wedge m_2,$$

$$q_1^* = \frac{\lambda_1 - m_1\mu_{1,1} - \mu_{1,2}z_{1,2}^*}{\theta_1} \quad \textit{and} \quad q_2^* = \frac{\lambda_2 - \mu_{2,2}(m_2 - z_{1,2}^*)}{\theta_2}.$$

(*ii*) $x^* \equiv (q_1^*, q_2^*, z_{1,2}^*)$ *is the unique stationary point of* x*, the unique solution to the ODE* (9)*.*

(*iii*) $\pi_{1,2}(x^*) \equiv P(D(x^*, \infty) > 0) = \pi_{1,2}^*$*, where* $D(x^*, \infty)$ *is a random variable with the stationary distribution of the FTSP* $D(x^*, \cdot)$ *and*

$$\pi_{1,2}^* \equiv \frac{\mu_{1,2}z_{1,2}^*}{\mu_{1,2}z_{1,2}^* + (m_2 - z_{1,2}^*)\mu_{2,2}}. \tag{11}$$

(*iv*) $x(t) \to x^*$ *as* $t \to \infty$ *exponentially fast.*

Proof Parts (*i*), (*ii*) and (*iii*), and (*iv*), respectively, are covered by Theorem 4.2 in [24], §8 of [23] and Theorem 9.2 in [23]. Explicit exponential bounds on the rate of convergence to stationarity in (*iv*) are given in [23]. We now elaborate on (*ii*) and (*iii*). First, if $x^* \notin \mathbb{A}$, then the fact that x^* is a stationary point of x follows immediately from the fact that $\pi_{1,2}(x^*) = 0$ or $= 1$. In that case, it is also easy to see that $\pi^*_{1,2}$ in (11) is equal to $\pi_{1,2}(x^*)$; see Corollary 8.1 in [23]. It is the unique stationary point by Theorem 8.1 in [23]. The more challenging case, in which $x^* \in \mathbb{A}$ and the existence of a stationary point is nontrivial, is proved in Theorem 8.2 in [23]. □

3.3 Eventually Remaining in the Set where the FTSP is Positive Recurrent

The FCLT will be stated under the assumption that the associated fluid limit lies in the set $\mathbb{A}$. Thus we now explain why this makes sense and introduce an additional assumption.

Note that x^*_6 in (10) is completely characterized by x^*, which involves only the rates in the system, and does not require any knowledge of the transient fluid limit or the initial condition. (In particular, SSC to three dimensions holds for the WLLN of the stationary distributions.) Simple algebra shows that if $0 < z^*_{1,2} < m_2$, then $q^*_1 = rq^*_2$. Together with (8) and (11) we see that $x^* \in \mathbb{A}$ if and only if $0 < z^*_{1,2} < m_2$. It follows from Assumption 2 and (10) that $z^*_{1,2} > 0$ (see also Corollary 8.2 in [23]), so that, under Assumption 2,

$$x^* \in \mathbb{A} \quad \text{if and only if} \quad z^*_{1,2} < m_2. \tag{12}$$

The next theorem, which follows from Theorem 10.2 in [23], shows that there is not much loss in assuming that the limit x lies entirely in $\mathbb{A}$ whenever $x^* \in \mathbb{A}$.

Theorem 3 *If* $x^* \in \mathbb{A}$ *then there exists* $T_A < \infty$ *such that* $x(t) \in \mathbb{A}$ *for all* $t \geq T_A$.

Since we are interested in the case $x^* \in \mathbb{A}$, which is the main case, as is clear from (12), we make the following assumption

Assumption 6 *For all* $t \geq 0$, $x(t) \in \mathbb{A}$.

Assumption 6 is not essential for our results; we make it only for simplicity of the exposition. Without this assumption, the FCLT can be proved over a finite interval over which $x \in \mathbb{A}$. In applications, the fluid limit is likely to hit $\mathbb{A}$ immediately after the overload begins, and remain in $\mathbb{A}$ thereafter; see §11.3 in [23].

4 The Main Results

In preparation for the FCLT, we indicate how the limit is affected by the FTSP in §4.1. We then state the main FCLT and important corollaries in §4.2 and §4.3. We conclude in §4.4 by indicating how the results simplify in the special case $r \equiv r_{1,2} = 1$, where FQR reduces to serving the longer queue.

4.1 The Role of the FTSP's in the Stochastic Limit

Just as the limiting ODE in (9) arising in the FWLLN depends on the FTSP's $D(\gamma,\cdot)$ (through the probability $\pi_{1,2}(x(t))$), so too the stochastic limit process arising in the FCLT refinement depends on these same FTSP's. Since the FTSP $D(\gamma,\cdot)$ depending on the state γ is a positive recurrent QBD under the assumption that $\gamma \in \mathbb{A}$, the stochastic refinement depends on the asymptotic variability of the FTSP. In particular, since the FTSP $D(\gamma,\cdot)$ is a regenerative process (which can be represented as a QBD whenever the ratio r is rational), the associated cumulative process obtained by integrating the indicator functions $1_{\{D(\gamma,s)>0\}}$ obeys a FCLT; i.e.,

$$\hat{C}^n_{QBD}(t;\gamma) \equiv n^{-1/2} \int_0^{nt} \left(1_{\{D(\gamma,s)>0\}} - \pi_{1,2}(\gamma)\right) ds \Rightarrow B(\sigma^2(\gamma)t) \qquad (13)$$

in the functions space $\mathcal{D}$ as $n \to \infty$, where B is a standard Brownian motion (BM) for each $\gamma \in \mathbb{A}$.

The constant $\sigma^2(\gamma)$ appearing inside the BM on the right in (13) is often called the *asymptotic variance* (see [3,9,31]) of the regenerative process $D(\gamma,s)$ (and the function f with $f(D(\gamma,s)) \equiv 1_{\{D(\gamma,s)>0\}}$). For each $\gamma \in \mathbb{A}$, it is defined as the limit

$$\sigma^2(\gamma) \equiv \lim_{t\to\infty} \frac{1}{t} Var\left(\int_0^t 1_{\{D(\gamma,s)>0\}} ds\right).$$

In this paper we will be making extensive use of the *regenerative structure*; see [3,9] for background. In our QBD context, the underlying *regenerative cycles* can be determined by successive visits of $D(\gamma,\cdot)$ to any fixed state, i.e., starting at a transition into the state and ending at the next transition into that state after first leaving that state. (The next transition into the state after leaving is the beginning of the next cycle; the cycles are closed on the left and open on the right.) The asymptotic behavior is determined by the random length of a cycle, $\tau(\gamma)$, and either the random integral over a cycle, $\tilde{Y}(\gamma)$, or the random centered integral over a cycle, $Y(\gamma)$, where

$$\tilde{Y}(\gamma) \equiv \int_0^{\tau(\gamma)} 1_{\{D(\gamma,s)>0\}}\, ds \quad \text{and} \quad Y(\gamma) \equiv \int_0^{\tau(\gamma)} (1_{\{D(\gamma,s)>0\}} - \pi_{1,2}(\gamma))\, ds.$$

The key asymptotic quantities here can be expressed in terms of the means of the first two variables and the variance of $Y(\gamma)$ via

$$\pi_{1,2}(\gamma) = \frac{E[\tilde{Y}(\gamma)]}{E[\tau(\gamma)]} \quad \text{and} \quad \sigma^2(\gamma) = \frac{Var(Y(\gamma))}{E[\tau(\gamma)]}; \qquad (14)$$

see [3,9]. Of course, $Y(\gamma) = \tilde{Y}(\gamma) - \pi_{1,2}(\gamma)\tau(\gamma)$, so that $Var(Y(\gamma))$ can be expressed in terms the means, variances and the covariance of the variables $\tau(\gamma)$ and $\tilde{Y}(\gamma)$, where $0 \le \tilde{Y}(\gamma) \le \tau(\gamma)$ w.p.1. Here we have strong regularity, with the random variable $\tau(\gamma)$ having a finite moment generating function and all these quantities being continuous functions of the state γ, by virtue of Lemma C.5 of [24].

4.2 The FCLT

Let $A_i^n(t)$ count the number of class-i customer arrivals, let $S_{i,j}^n(t)$ count the number of service completions of class-i customers by agents in pool j, an let $U_i^n(t)$ count the number of class-i customers to abandon from queue, all in model n during the time interval $[0,t]$. Let $D_{1,2}^n(t)$ be the queue-difference process in (1) and let $Q_s^n(t) \equiv Q_1^n(t) + Q_2^n(t)$, all at time t. Let $p_1 \equiv r/(1+r)$ and $p_2 \equiv 1 - p_1 = 1/(1+r)$, where $r \equiv r_{1,2}$. For $t \geq 0$ and $i, j = 1, 2$, let the diffusion-scaled processes be

$$\begin{aligned}
\hat{A}_i^n(t) &\equiv \frac{A_i^n(t) - n\lambda_i(t)}{\sqrt{n}}, \qquad \hat{U}_i^n(t) \equiv \frac{U_i^n(t) - n\theta_i \int_0^t q_i(s)\, ds}{\sqrt{n}}, \\
\hat{Z}_{i,j}^n(t) &\equiv \frac{Z_{i,j}^n(t) - nz_{i,j}(t)}{\sqrt{n}}, \quad \hat{S}_{i,j}^n(t) \equiv \frac{S_{i,j}^n(t) - n\mu_{i,j} \int_0^t z_{i,j}(s)\, ds}{\sqrt{n}}, \\
\hat{Q}_1^n(t) &\equiv \frac{Q_1^n(t) - nq_1(t)}{\sqrt{n}}, \qquad \hat{Q}_2^n(t) \equiv \frac{Q_2^n(t) - nq_2(t)}{\sqrt{n}}, \\
\hat{Q}_s^n(t) &\equiv \frac{Q_s^n(t) - nq_s(t)}{\sqrt{n}}, \qquad \hat{D}^n(t) \equiv \frac{D_{1,2}^n(t)}{\sqrt{n}} \\
\hat{I}^n(t) &\equiv \sqrt{n} \int_0^t (1_{\{D_{1,2}^n(s) > 0\}} - \pi_{1,2}(x(s)))ds, \quad t \geq 0,
\end{aligned} \tag{15}$$

where $x \equiv (q_1, q_2, z_{1,2})$ is the customary three-dimensional representation of the fluid limit, $z_{1,1} \equiv m_1 e$, $z_{2,1} = 0e$ $z_{2,2} \equiv m_2 e - z_{1,2}$, $q_s \equiv q_1 + q_2$ and $\pi_{1,2}(x(s)) \equiv P(D(x(s), \infty) > 0)$, with $D(x(s), \infty)$ being a random variable with the steady-state distribution of the FTSP $\{D(x(s), t) : t \geq 0\}$ associated with the fluid limit $x(s)$ at time s.

Here is the main result of this paper: the FCLT for the overloaded X model operating under FQR-T. Since the limit is clearly a Markov process with continuous sample paths, it is by definition a diffusion process. Most of the rest of the paper is devoted to its proof.

Theorem 4 (FCLT) *If, in addition to Assumptions 1–6,*

$$\left(\hat{Q}_s^n(0), \hat{Z}_{1,2}^n(0)\right) \Rightarrow \left(\hat{Q}_s(0), \hat{Z}_{1,2}(0)\right) \in \mathbb{R}_2 \quad as \quad n \to \infty,$$

then, for $i, j = 1, 2$,

$$\left(\hat{A}_i^n, \hat{U}_i^n, \hat{S}_{i,j}^n, \hat{D}^n, \hat{I}^n, \hat{Q}_i^n, \hat{Q}_s^n, \hat{Z}_{i,j}^n\right) \Rightarrow \left(\hat{A}_i, \hat{U}_i, \hat{S}_{i,j}, \hat{D}, \hat{I}, \hat{Q}_i, \hat{Q}_s, \hat{Z}_{i,j}\right) \tag{16}$$

in $\mathcal{D}_{17}$, where the processes depending on n on the left are defined in (15) *and the limit process has continuous paths w.p.1. The initial* 10*-dimensional component $(\hat{A}_i, \hat{U}_i, \hat{S}_{i,j}, \hat{D}, \hat{I})$ is a vector of independent Brownian motions, time scaled by increasing continuous deterministic functions (for the first* 8*, the fluid limits in the translation terms of* (15)*), with two null components $\hat{S}_{2,1} \equiv 0e$ and $\hat{D} \equiv 0e$. Five components of the limit are determined by the relations*

$\hat{Q}_i \stackrel{\mathrm{d}}{=} p_i \hat{Q}_s$, $\hat{Z}_{2,1} \equiv \hat{Z}_{1,1} \equiv 0e$ *and* $\hat{Z}_{2,2} \equiv -\hat{Z}_{1,2}$. *Finally,* $(\hat{Q}_s, \hat{Z}_{1,2})$ *is the unique solution of the following two-dimensional stochastic integral equation:*

$$\begin{aligned}
\hat{Q}_s(t) &= \hat{Q}_s(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}_{1,2}(s)\, ds - (p_1\theta_1 + p_2\theta_2) \int_0^t \hat{Q}_s(s)\, ds \\
&\quad + \hat{L}_1(t) - \hat{L}_{1,2}(t) - \hat{S}_{1,2}(t) - \hat{L}_{2,2}(t) - \hat{S}_{2,2}(t), \\
\hat{Z}_{1,2}(t) &= \hat{Z}_{1,2}(0) - \int_0^t \left[(\mu_{2,2} - \mu_{1,2})\pi_{1,2}(x(s)) + \mu_{1,2}\right] \hat{Z}_{1,2}(s)\, ds \\
&\quad - \hat{L}_{1,2}(t) + \hat{L}_{2,2}(t) + \hat{L}_2(t),
\end{aligned} \tag{17}$$

where, for $i = 1, 2$,

$$\begin{aligned}
\hat{L}_1 &\equiv \hat{A}_1 + \hat{A}_2 - \hat{U}_1 - \hat{U}_2 - \hat{S}_{1,1} \stackrel{\mathrm{d}}{=} \{B_1(\gamma_1(t)) : t \geq 0\}, \\
\hat{L}_{i,2} &\equiv \{B_{i,2}(\phi_{i,2}(t)) : t \geq 0\}, \quad \hat{S}_{i,2} \equiv \{B_{i,3}(\gamma_{i,2}(t)) : t \geq 0\}, \\
\hat{L}_2 &\equiv \{B_2(\gamma_2(t)) : t \geq 0\} \quad \textit{and} \quad \hat{I} \equiv \{B_2(\gamma_3(t)) : t \geq 0\},
\end{aligned} \tag{18}$$

with B_1, $B_{1,2}$, $B_{2,2}$, $B_{1,3}$, $B_{2,3}$ *and* B_2 *being six independent standard BM's, while* γ_i, $\gamma_{i,2}$ *and* $\phi_{i,2}$ *are strictly increasing continuous deterministic functions. Specifically,*

$$\begin{aligned}
\gamma_1(t) &\equiv (\lambda_1 + \lambda_2 + m_1\mu_{1,1})t + (p_1\theta_1 + p_2\theta_2) \int_0^t q_s(u)\, du \\
\phi_{1,2}(t) &\equiv \mu_{1,2} \int_0^t (1 - \pi_{1,2}(x(u)))z_{1,2}(u)\, du, \\
\phi_{2,2}(t) &\equiv \mu_{2,2} \int_0^t \pi_{1,2}(x(u))(m_2 - z_{1,2}(u))\, du, \\
\gamma_{1,2}(t) &\equiv \mu_{1,2} \int_0^t \pi_{1,2}(x(u))z_{1,2}(u)\, du \\
\gamma_{2,2}(t) &\equiv \mu_{2,2} \int_0^t (1 - \pi_{1,2}(x(u)))(m_2 - z_{1,2}(u))\, du, \\
\gamma_2(t) &\equiv \int_0^t \psi^2(x(u))\sigma^2(x(u))\, du, \quad \gamma_3(t) \equiv \int_0^t \sigma^2(x(u))\, du,
\end{aligned} \tag{19}$$

where

$$\psi(x(u)) \equiv \mu_{2,2}(m_2 - z_{1,2}(u)) + \mu_{1,2}z_{1,2}(u), \quad u \geq 0, \tag{20}$$

with $\pi_{1,2}(x(u))$ *and* $\sigma^2(x(u))$ *being the quantities associated with the FTSP* $D(x(u), \cdot)$, *defined in* (7) *and* (13), *respectively, and characterized in* (14).

Since the FCLT describes a refinement of the transient behavior of the fluid limit, it should not be surprising that the limiting stochastic process $(\hat{Q}_s, \hat{Z}_{1,2})$ would be difficult to analyze. On the positive side, we can solve for $\hat{Z}_{1,2}$ in (17) without having to simultaneously solve for $\hat{Q}_s$, but we need $\hat{Z}_{1,2}$

to solve for $\hat{Q}_s$. An additional complication for $\hat{Q}_s$ is the dependence between the driving Brownian motions for the two processes $\hat{Q}_s$ and $\hat{Z}_{1,2}$; note that the time-transformed Brownian terms $\hat{L}_{i,2}$ appear in both.

The FCLT shows the impact of system variability on the stochastic limit. First, and perhaps of greatest interest, there is a Brownian contribution $\hat{L}_2 \stackrel{\mathrm{d}}{=} B_2(\gamma_2(t))$ from the FTSP appearing in the equation for $\hat{Z}_{1,2}$; note the dependence between $\hat{L}_2$ and $\hat{I}$. However, $(\hat{L}_2, \hat{I})$ is independent of all other Brownian terms. We thus see that the fluctuations about the fixed target ratio r in the queue-difference process (1) due to FQR *do* have an impact on the stochastic limit.

On the other hand, we see that the stochastic fluctuations associated with external arrivals and abandonments only affect $\hat{Q}_s$; they have no impact on $\hat{Z}_{1,2}$. The same is true for the stochastic fluctuations of service facility 1, which is always busy, without any sharing. These fluctuations are captured by the Brownian term $\hat{L}_1 \stackrel{\mathrm{d}}{=} B_1(\gamma_1(t))$. However, as noted above, in distinct contrast, the stochastic fluctuations in the service processes at service facilty 2 have a more complicated impact, because they appear in the Brownian driving processes of both equations.

4.3 Important Corollaries

The stochastic limit in the FCLT depends critically on the fluid limit x, which typically must be computed numerically, but an efficient algorithm was developed in [23], exploiting the QBD structure of the FTSP D when $r_{1,2}$ is rational. Since we are mainly interested in the steady state variance of the diffusion limits, and since the stochastic fluctuations become more significant when the fluid is nearly constant (which happens when it is close to its stationary point) it is reasonable to initialize the fluid model at this fluid stationary point in order to simplify the expressions in (17) and (19). We do this in the next corollary.

From an application point of view, the fluid limit is "more important" than the refined stochastic limit during the fluid transient period, since then the changes in the prelimit are of order $O(n)$. It follows from Theorem 2 that after some (relatively short) time, the fluid stabilizes close to its unique stationary point x_6^* in (10). After that happens, the refined stochastic limits become the significant approximation to consider.

When we consider the stochastic refinement of the stationary fluid limit x^*, the stochastic limit process becomes much more tractable: it is a bivariate Ornstein-Uhlenbeck (BOU) process centered at the origin, as in [2,30]. Consequently, the random vector $(\hat{Q}_s(t), \hat{Z}_{1,2}(t))$ has a bivariate normal distribution with zero means for all t, and the associated steady-state random vector $(\hat{Q}_s(\infty), \hat{Z}_{1,2}(\infty))$ can be very useful in applications. It is characterized by three parameters: the two variances and the covariance, which we exhibit explicitly in (24) below.

For a matrix M, let M^t denote its transpose. The following is the key result for applications. It gives explicit Gaussian approximations for the steady-state distributions of all quantities of interest.

Corollary 1 (FCLT with a stationary fluid) *If, in addition to the conditions of Theorem 4, $x(0) = x^*$ for the stationary point x^* in (10) so that x is stationary, then the time transformations in (19) simplify by having $\gamma_i(t) = \xi_i t$, $\gamma_{i,2}(t) = \xi_{i,2}t$, and $\phi_{i,2}(t) = \eta_{i,2}t$, $i = 1, 2$, where*

$$\begin{aligned}
\xi_1 &\equiv 2(\lambda_1 + \lambda_2) - \mu_{1,2}z^*_{1,2} - \mu_{2,2}(m_2 - z^*_{1,2}),\\
\xi_{1,2} &\equiv \mu_{1,2}\pi_{1,2}(x^*)z^*_{1,2}, \quad \xi_{2,2} \equiv \mu_{2,2}(1 - \pi_{1,2}(x^*)(m_2 - z^*_{1,2}),\\
\eta_{1,2} &\equiv \mu_{1,2}(1 - \pi_{1,2}(x^*))z^*_{1,2},\\
\eta_{2,2} &\equiv \mu_{2,2}\pi_{1,2}(x^*)(m_2 - z^*_{1,2}),\\
\xi_2 &\equiv \psi^2(x^*)\sigma^2(x^*) \quad \textit{and} \quad \xi_3 \equiv \sigma^2(x^*),
\end{aligned} \tag{21}$$

for $\sigma^2(x^*)$ and $\psi(x^*)$ defined in (13) and (20) with $x(u) = x^*$. Then $(\hat{Q}_s, \hat{Z}_{1,2})$ becomes a BOU process, satisfying the two-dimensional stochastic differential equation (sde)

$$d\mathcal{X} = \mathcal{M}\mathcal{X} + \mathcal{S}d\mathcal{B}, \tag{22}$$

where $\mathcal{X} \equiv (\hat{Q}_s, \hat{Z}_{1,2})^t$, $\mathcal{B} \equiv (B_1, B_2)^t$, with B_1 and B_2 being two independent standard BM's, and

$$\begin{aligned}
&\mathcal{M}_{1,1} \equiv -(p_1\theta_1 + p_2\theta_2), \quad \mathcal{M}_{1,2} \equiv (\mu_{2,2} - \mu_{1,2}), \quad \mathcal{M}_{2,1} \equiv 0,\\
&\mathcal{M}_{2,2} \equiv \frac{-\mu_{1,2}\mu_{2,2}m_2z^*_{1,2}}{\mu_{1,2}z^*_{1,2} + \mu_{2,2}(m_2 - z^*_{1,2})} < 0,\\
&\mathcal{S}^2_{1,1} \equiv \xi_1 + \xi_{1,2} + \xi_{2,2} + \eta_{1,2} + \eta_{2,2} = 2(\lambda_1 + \lambda_2),\\
&\mathcal{S}_{1,2} \equiv \mathcal{S}_{2,1} \equiv \eta_{1,2} - \eta_{2,1} = 0, \quad \mathcal{S}^2_{2,2} \equiv \xi_2 + \xi_4,\\
&\xi_4 \equiv \eta_{1,2} + \eta_{2,2} = \frac{2\mu_{1,2}\mu_{2,2}z^*_{1,2}(m_2 - z^*_{1,2})}{\mu_{1,2}z^*_{1,2} + (m_2 - z^*_{1,2})\mu_{2,2}}.
\end{aligned} \tag{23}$$

As a consequence, $(\hat{Q}_s(t), \hat{Z}_{1,2}(t))$ has a bivariate normal distribution with zero means for each t. The covariance matrix of the steady-state random vector

$(\hat{Q}_s(\infty), \hat{Z}_{1,2}(\infty))$ *has elements*

$$\begin{aligned}
\sigma^2_{\hat{Q}_s}(\infty) &\equiv Var(\hat{Q}_s) = \mathcal{Q}_1 + \mathcal{Q}_2, \\
\mathcal{Q}_1 &\equiv \frac{\mathcal{S}^2_{1,1}}{2|\mathcal{M}_{1,1}|} = \left(\frac{\lambda_1+\lambda_2}{p_1\theta_1+p_2\theta_2}\right), \\
\mathcal{Q}_2 &\equiv \frac{\mathcal{M}_{1,2}\sigma^2_{\hat{Q}_s,Z_{1,2}}(\infty)}{|\mathcal{M}_{1,1}|} = \left(\frac{(\mu_{2,2}-\mu_{1,2})\sigma^2_{\hat{Q}_s,Z_{1,2}}(\infty)}{p_1\theta_1+p_2\theta_2}\right), \\
\sigma^2_{Z_{1,2}}(\infty) &\equiv \frac{\mathcal{S}^2_{2,2}}{2|\mathcal{M}_{2,2}|} \equiv \mathcal{Z}_1 + \mathcal{Z}_2, \\
\mathcal{Z}_1 &\equiv \frac{\xi_4}{2|\mathcal{M}_{2,2}|} = 1 - \frac{z^*_{1,2}}{m_2}, \quad \mathcal{Z}_2 \equiv \frac{\xi_2}{2|\mathcal{M}_{2,2}|} = \frac{\psi^2(x^*)\sigma^2(x^*)}{2|\mathcal{M}_{2,2}|}, \\
\sigma^2_{\hat{Q}_s,Z_{1,2}}(\infty) &\equiv Cov(\hat{Q}_s, \hat{Z}_{1,2}) = \xi_5\sigma^2_{Z_{1,2}}(\infty), \quad \xi_5 \equiv \left(\frac{\mathcal{M}_{1,2}}{|\mathcal{M}_{1,1}+\mathcal{M}_{2,2}|}\right).
\end{aligned} \tag{24}$$

Proof By the definition of a stationary point, if $x(0) = x^*$ then $x(t) = x^*$ for all $t > 0$ given in (10); then $\pi_{1,2}(x^*)$ appears in (11). The expressions in (21) follow directly from the expressions in (19), by replacing the time-dependent fluid quantities by their stationary counterparts. The resulting pair of integral equations for $(\hat{Q}_s(t), \hat{Z}_{1,2}(t))$ is known to be equivalent to the BOU sde in (22). The covariance matrix of the stationary distribution, Σ, is known to satisfy the matrix equation $\mathcal{M}\Sigma + \Sigma\mathcal{M}^t = -\mathcal{V}$, where $\mathcal{V} \equiv \mathcal{S}\mathcal{S}^t$, from which (24) follows; e.g., see [2] and [16]. Algebra shows that $\xi_4/2|\mathcal{M}_{2,2}| = (1 - (z^*_{1,2}/m_2))$. □

Remark 1 (when components become null) Notice that the results in Corollary 1 simplify greatly with pool-dependent service rates, i.e., when $\mathcal{M}_{1,2} \equiv \mu_{2,2} - \mu_{1,2} = 0$. Then $\mathcal{Q}_2 = 0$ and $\xi_5 = 0$, so that $\sigma^2_{\hat{Q}_s,Z_{1,2}}(\infty) = 0$.

We now see how Theorem 4 simplifies under the condition of pool-dependent service rates (no longer assuming that $x(0) = x^*$).

Corollary 2 (FCLT with pool-dependent service rates) *If, in addition to the assumptions of Theorem 4, $\mu_{2,2} = \mu_{1,2} \equiv \nu$, then the two diffusion-limit processes $\hat{Q}_s$ and $\hat{Z}_{1,2}$ can both be represented as separate one-dimensional processes, which satisfy the following integral equations*

$$\begin{aligned}
\hat{Q}_s(t) &= \hat{Q}_s(0) - \tilde{\eta}_2 \int_0^t \hat{Q}_s(s)\, ds + B_1\left(\tilde{\gamma}_1(t)\right), \\
\hat{Z}_{1,2}(t) &= \hat{Z}_{1,2}(0) - \nu \int_0^t \hat{Z}_{1,2}(s)\, ds + B_2\left(\tilde{\gamma}_2(t)\right),
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\gamma}_1(t) &\equiv 2(\lambda_1+\lambda_2)t + \left(\frac{\tilde{\eta}_1}{\tilde{\eta}_2} - q_s(0)\right)(1 - e^{-\tilde{\eta}_2 t}) \\
\tilde{\gamma}_2(t) &\equiv \nu\left(\int_0^t \left[m_2\pi_{1,2}(x(u)) + z_{1,2}(u) - 2\pi_{1,2}(x(u))z_{1,2}(u)\right] du\right) + \gamma_2(t),
\end{aligned}$$

with $\gamma_2(t)$ defined in (19),

$$\tilde{\eta}_1 \equiv \lambda_1 + \lambda_2 - m_1\mu_{1,1} - m_2\nu, \qquad \tilde{\eta}_2 \equiv p_1\theta_1 + p_2\theta_2,$$

but B_1 and B_2 are dependent standard BM's.

Proof It is immediate from the expression for $\hat{Q}_s$ in (17) that when $\mu_{1,2} = \mu_{2,2}$ the diffusion process $\hat{Q}_s$ can be analyzed separately from $\hat{Z}_{1,2}$. Since $q_i = p_i q_s$ and $\mu_{1,2} = \mu_{2,2}$, it follows from (9) that $\dot{q}_s(t)$ satisfies the simple ordinary differential equation

$$\dot{q}_s(t) = (\lambda_1 + \lambda_2 - m_1\mu_{1,1} - m_2\mu_{2,2}) - (p_1\theta_1 + p_2\theta_2)q_s(t) \equiv \tilde{\eta}_1 - \tilde{\eta}_2 q_s(t),$$

whose solution is

$$q_s(t) = \frac{\tilde{\eta}_1}{\tilde{\eta}_2} + \left(q(0) - \frac{\tilde{\eta}_1}{\tilde{\eta}_2}\right) e^{-\tilde{\eta}_2 t}$$

for $\tilde{\eta}_1$ and $\tilde{\eta}_2$ in the statement of the lemma. Notice that $\tilde{\gamma}_1(t)$ here corresponds to $\gamma_1(t) + \gamma_{1,2}(t) + \gamma_{2,2}(t)$ in (19). Inserting $q_s(t)$ above into $\gamma_1(t)$ in (19) gives $\tilde{\gamma}_1(t)$. Notice that $\tilde{\gamma}_2(t)$ corresponds to $\phi_{1,2}(t) + \phi_{2,2}(t) + \gamma_2(t)$ in (19). Again substituting yields the conclusion.

Remark 2 (Equivalence with the single-class model.) If, in addition to the conditions of both Corollaries 1 and 2, we also have $\theta_1 = \theta_2 \equiv \theta$, then the diffusion-limit process $\hat{Q}_s$ is the same as the limit obtained for the $M/M/n+M$ model in the efficiency-driven (ED) regime, see [33]. That is, $\hat{Q}_s$ is an Ornstein-Uhlenbeck process with infinitesimal mean equal to θ and infinitesimal variance $2\lambda \equiv 2(\lambda_1 + \lambda_2)$. Thus, its steady-state distribution is normal with mean zero and variance λ/θ. However, $\hat{Z}_{1,2}$ remains somewhat complicated involving $\gamma_2(t)$ in (19).

4.4 The Case r = 1: Longer Queue First (LQF)

The most complicated feature in the FWLLN and FCLT asymptotic results in the previous two sections, inhibiting application, is the need to analyze the FTSP. Specifically, both the approximating fluid model and the stochastic refinement depend critically on the FTSP $D \equiv D(\gamma) \equiv \{D(\gamma, s) : s \geq 0\}$ at each point $\gamma \in \mathbb{A}$. In particular, both limits depend on $D(\gamma)$ through the two functions $\pi_{1,2}(\gamma)$ and $\sigma^2(\gamma)$. These two functions can be computed numerically, as indicated above. For the stationary fluid point x^*, $\pi_{1,2}(x^*)$ is given explicitly in (11).

However, there is an important special case, itself of practical value, in which the analysis simplifies greatly, which can provide insight more generally. When the target queue ratio is $r = 1$, the FTSP $D(\gamma)$ becomes an ordinary *birth-and-death* (BD) process for each $\gamma \in \mathbb{A}$. Then the quantities $\pi_{1,2}(\gamma)$ and $\sigma^2(\gamma)$ are both easily expressed. It turns out that they can be expressed in terms of the first two moments of the busy-period distributions of two $M/M/1$ queues. We consider that case now.

We now assume that $r = 1$, and take $\gamma \in \mathbb{A}$. In this case, the FTSP evolves as one BD process when $D(\gamma) > 0$ and evolves as another BD process when $D(\gamma) \leq 0$. We call 0 the boundary state. Let $\lambda_1(\gamma)$ denote the constant rate up (away from the boundary) and let $\mu_1(\gamma)$ denote the constant rate down (toward the boundary) of $D(\gamma)$ when $D(\gamma) > 0$. Focusing on the movement relative to the boundary, let $\lambda_2(\gamma)$ denote the constant rate *down* (away from the boundary) and let $\mu_2(\gamma)$ denote the constant rate *up* (toward the boundary) of $D(\gamma)$ when $D(\gamma) \leq 0$.

Note that we need to analyze $D(\gamma)$ only through the associated stochastic process

$$X(\gamma, t) \equiv 1_{\{D(\gamma,t)>0\}}, \quad t \geq 0,$$

which records which region $D(\gamma, t)$ is in at each time t. The stochastic process $X \equiv X(\gamma) \equiv \{X(\gamma, t) : t \geq 0\}$ is a $\{0, 1\}$-valued process associated with an alternating renewal process. Let $T_1(\gamma)$ denote a time interval between the instant of a state change from state 0 to state 1 until the next instant of a state change from state 1 back to state 0. Similarly, let $T_2(\gamma)$ denote a time interval between instant of a state change from state 1 to state 0 until the next instant of a state change from state 0 back to state 1. The successive times in the alternating renewal process are independent random variables distributed as $T_1(\gamma)$ and $T_2(\gamma)$. The process $X(\gamma)$ is a regenerative process in which the regeneration times can be the successive instant of a state change from state 0 to state 1 until the next instant of the same state change again at a later time. The intervals between successive regenerations are distributed as $T_1(\gamma) + T_2(\gamma)$.

Now observe that $T_i(\gamma)$ is distributed as a busy period in an $M/M/1$ queue with arrival rate $\lambda_i(\gamma)$ and service rate $\mu_i(\gamma)$, $i = 1, 2$. In this context, the condition $\gamma \in \mathbb{A}$ is equivalent to $\lambda_i(\gamma) < \mu_i(\gamma)$, $i = 1, 2$. Under this condition, $T_i(\gamma)$ is known to have a finite moment generating function with a positive radius of convergence, so that all moments of $T_i(\gamma)$ are finite. Let

$$m_i(\gamma) \equiv 1/\mu_i(\gamma) \quad \text{and} \quad \rho_i(\gamma) \equiv \lambda_i(\gamma)/\mu_i(\gamma), \quad i = 1, 2. \tag{25}$$

Then, from basic $M/M/1$ theory, we have

$$E[T_i(\gamma)] = \frac{m_i(\gamma)}{1 - \rho_i(\gamma)} \quad \text{and} \quad E[T_i(\gamma)^2] = \frac{2m_i(\gamma)^2}{(1 - \rho_i(\gamma))^3}. \tag{26}$$

Finally, we are interested in the cumulative process associated with $X(\gamma)$,

$$C(\gamma, t) \equiv \int_0^t X(\gamma, s)\, ds \equiv \int_0^t 1_{\{D(\gamma,s)>0\}}\, ds, \quad t \geq 0.$$

We can apply (14) to obtain the following result.

Theorem 5 (*the FTSP when $r = 1$) When $r = 1$ and $\gamma \in \mathbb{A}$, the FTSP becomes a recurrent BD process. Hence the key FTSP quantities can be expressed directly in terms of the four BD rates $\lambda_i(\gamma)$ and $\mu_i(\gamma)$ via*

$$\pi_{1,2}(\gamma) = \frac{E[T_1(\gamma)]}{E[T_1(\gamma)] + E[T_2(\gamma)]}, \quad \sigma^2(\gamma) = \frac{Var(T_1(\gamma))}{E[T_1(\gamma)] + E[T_2(\gamma)]} \tag{27}$$

for $E[T_i(\gamma)]$ *and* $E[T_i(\gamma)^2]$ *in* (26) *and* (25)*,* $i = 1, 2$*.*

In the more general QBD setting arising with $r \neq 1$, the analysis is more complicated, because the excursions of $\int_0^t 1_{\{D(\gamma,s)>0\}}$ above and below 0 depend on the entering and exit states from level 0; thus these excursions are not simply independent. Theorem 5 can be the basis for heuristic extensions to non-Markovian models in which the arrival, service and abandonment processes are non-Markovian. We may then exploit approximations for the busy period in $GI/GI/1$ queues, e.g., [1] and [25].

5 Proof of Theorem 4

First observe that the assumed convergence in $\mathbb{R}_2$ at time 0 is actually equivalent to the full convergence in $\mathbb{R}_{17}$ of the process in (16) at time 0 because of Assumption 5. Our proof has four main steps: The first step is to exploit SSC results established in [24]. In particular, we first give an asymptotically equivalent three-dimensional representation of X_6^n (without any scaling) involving rate-1 Poisson processes. Then we observe that the essential dimension is actually two (when scaling by $\sqrt{n}$) because the queue lengths are asymptotically in the fixed ratio. Thus we deduce that it is sufficient to directly prove convergence of the 2-dimensional process $(\hat{Q}_s^n, \hat{Z}_{1,2}^n)$.

The second step is to facilitate application of the continuous mapping theorem by showing that an essential mapping is continuous. The third step is to construct appropriate martingale representations, allowing application of the continuous mapping theorem. The fourth and final hardest step is to show that the driving stochastic terms in this martingale representation converge to the specified limits. This final step uses a new result of independent interest, Theorem 6, the generalization of the classical FCLT for cumulative processes in (13) to the case where the QBD parameters at time t are given by the fluid limit $x(t)$, which in general is time-varying.

5.1 Representation and SSC

Following common practice, as reviewed in §2 of [20], we represent the processes $A_i^n(t)$, $S_{i,j}^n(t)$ and $U_i^n(t)$ introduced at the beginning of §4.2 in terms of mutually independent rate-1 Poisson processes; let

$$\begin{aligned} A_i^n(t) &\equiv N_i^a(\lambda_i^n t), \\ S_{i,j}^n(t) &\equiv N_{i,j}^s\left(\mu_{i,j}\int_0^t Z_{i,j}^n(s)\,ds\right) \quad \text{and} \quad S^n \equiv \sum_{j=1}^{2}\sum_{i=1}^{2} S_{i,j}^n, \\ U_i^n(t) &\equiv N_i^u\left(\theta_i \int_0^t Q_i^n(s)\,ds\right), \quad t \geq 0, \end{aligned}$$

where N_i^a, $N_{i,j}^s$ and N_i^u for $i = 1, 2; j = 1, 2$ are eight mutually independent rate-1 Poisson processes. Theorem 5.1 of [24] gives a representation of the

CTMC in terms of these processes. Corollaries 6.1-6.3 plus Theorem 6.4 of [24] then establish state space collapse (SSC) results yielding an asymptotically equivalent three-dimensional representation of X_6^n involving these mutually independent rate-1 Poisson processes plus two others. Since we exploit that representation, we state it here. Directly, the representation of $Z_{1,2}^n$ below keeps it in the interval $[0, m_2^n]$. However, the representation directly allows the queue lengths Q_i^n to become negative. The results in [24] show that the occurrence (anywhere in a a bounded interval) is asymptotically negligible. Recall that d_{J_1} denotes the Skorohod J_1 metric.

Lemma 1 (*Representation via SSC of the service process*) *Under the assumptions in Theorem 1,* $d_{J_1}(X_6^n, X_6^{n,*}) \Rightarrow 0$ *in* $\mathcal{D}_6$ *as* $n \to \infty$*, with the three determining components of* $X_6^{n,*}$ *in* (3)*, i.e., in* X^n *in* (4)*, being represented via*

$$\begin{aligned} Z_{1,2}^n(t) &\equiv Z_{1,2}^n(0) + \int_0^t 1_{\{D_{1,2}^n(s-)>0\}}\, dS_{2,2}^n(s) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}}\, dS_{1,2}^n(s) \\ &\stackrel{\mathrm{d}}{=} Z_{1,2}^n(0) + N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(m_2^n - Z_{1,2}^n(s))\, ds\right) \\ &\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}} Z_{1,2}^n(s)\, ds\right), \end{aligned}$$

$$\begin{aligned} Q_1^n(t) &\equiv Q_1^n(0) + A_1^n(t) - \int_0^t 1_{\{D_{1,2}^n(s-)>0\}}\, dS^n(s) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}}\, dS_{1,1}^n(s) - U_1^n(t) \\ &\stackrel{\mathrm{d}}{=} Q_1^n(0) + N_1^a(\lambda_1^n t) - N_{1,1}^s(\mu_{1,1} m_1^n t) - N_{1,2}^{s,2}\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}} Z_{1,2}^n(s))\, ds\right) \\ &\quad - N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(m_2^n - Z_{1,2}^n(s))\, ds\right) - N_1^u\left(\theta_1\int_0^t Q_1^n(s)\, ds\right), \end{aligned}$$

$$\begin{aligned} Q_2^n(t) &\equiv Q_2^n(0) + A_2^n(t) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}}\, dS_{2,2}^n(s) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}}\, dS_{1,2}^n(s) - U_2^n(t) \\ &\stackrel{\mathrm{d}}{=} Q_2^n(0) + N_2^a(\lambda_2^n t) - N_{2,2}^{s,2}\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}(m_2^n - Z_{1,2}^n(s))\, ds\right) \\ &\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}} Z_{1,2}^n(s)\, ds\right) - N_2^u\left(\theta_2\int_0^t Q_2^n(s)\, ds\right). \end{aligned}$$

where $N_{1,2}^{s,2}$ *and* $N_{2,2}^{s,2}$ *are two additional rate-*1 *Poisson processes, independent of the others.*

The representation in Lemma 1 provides important simplification, but it also shows the difficulty in proving heavy traffic limit theorems; the integrals contain the indicator functions depending on $D_{1,2}^n$. We now show that the essential dimension can be reduced from three to two when we introduce scaling. The next result follows from Corollary 4.1 of [24].

Lemma 2 (*SSC to two dimensions*) *Under the conditions of Theorem* 1*, the essential dimension can be reduced from* 3 *established in Lemma* 1 *to* 2*, because* $d_{J_1}(Q_1^n, rQ_2^n)/a_n \Rightarrow 0$ *in* $\mathcal{D}([0,\delta))$ *for* δ *in Theorem* 1 *whenever* $a_n/\log n \to \infty$ *as* $n \to \infty$*. If* $x \in \mathbb{A}$ *over an interval* $[t_1, t_2)$*,* $0 < t_1 < t_2 \leq \infty$*, then the conclusion holds in* $\mathcal{D}((t_1, t_2))$*.*

Due to Assumption 6 and Lemma 2, it is sufficient to directly prove convergence of the 2-dimensional process $(\hat{Q}_s^n, \hat{Z}_{1,2}^n)$; the more general 16-dimensional limit in (16) can be obtained as a byproduct of the analysis, and in particular, $\hat{Q}_i \stackrel{\mathrm{d}}{=} p_i \hat{Q}_s$, $i = 1, 2$.

5.2 A Continuous Mapping

As in [20], our proof exploits the continuous mapping theorem. However, in our case, the stochastic processes describing the evolution of the system (the queue length and service processes) cannot be expressed directly as a continuous mapping of the primitive processes. We next establish the continuity of the mapping that we will eventually apply.

Lemma 3 (Continuity of the two-dimensional integral representation) *Consider the two-dimensional integral representation*

$$x_1(t) = b_1 + y_1(t) + \alpha_2 \int_0^t x_2(s)\, ds + \alpha_1 \int_0^t x_1(s)\, ds$$
$$x_2(t) = b_2 + y_2(t) + \int_0^t g(s) x_2(s)\, ds$$

where $g : \mathbb{R} \to \mathbb{R}$ *satisfies* $g(0) = 0$ *and is Lipschitz continuous with a Lipschitz constant* c_g*. That integral representation has a unique solution* (x_1, x_2)*, so that the integral representation constitutes a function* $f : \mathcal{D}_2 \times \mathbb{R}_2 \to \mathcal{D}_2$ *mapping* (y_1, y_2, b_1, b_2) *into* $(x_1, x_2) \equiv f(y_1, y_2, b_1, b_2)$*. In addition, the function* f *is a continuous mapping from* $\mathcal{D}_2 \times \mathbb{R}_2$ *to* $\mathcal{D}_2$*. Moreover, if* y_2 *is continuous then* x_2 *is continuous. If both* y_1 *and* y_2 *are continuous, then* x_1 *is also continuous.*

Proof By the conditions on the function g we have for all $T \geq 0$

$$\|g\|_T \leq g(0) + \|g(u) - g(0)\|_T \leq g(0) + c_g T = c_g T.$$

Note that x_2 does not depend on x_1, hence we can prove the lemma iteratively by first showing that the function $f_2 : \mathcal{D} \times \mathbb{R}$ mapping (y_2, b_2) into $x_2 \equiv f_2(y_2, b_2)$ is continuous, and then use this result to show that the function $f_1 : \mathcal{D}_2 \times \mathbb{R}$ mapping (y_1, x_2, b_1) into $x_1 \equiv f_1(y_1, x_2, b_1)$ is continuous.

To show that f_2 is continuous we use Theorem 2.11 in [27] with $h(x_2(u), u) \equiv g(u)x_2(u)$. For that purpose, choose $T > 0$ and let λ be a homeomorphism on

$[0,T]$ with strictly positive derivative $\dot{\lambda}$. Then, for every $\varphi_1, \varphi_2 \in \mathcal{D}$

$$\begin{aligned}
&\int_0^t |g(u)\varphi_1(u) - g(\lambda(u))\varphi_2(\lambda(u))|\, du \\
&\le \int_0^t |g(u)\varphi_1(u) - g(u)\varphi_2(\lambda(u))|\, du + \int_0^t |g(u)\varphi_2(\lambda(u)) - g(\lambda(u))\varphi_2(\lambda(u))|\, du \\
&\le \|g\|_T \int_0^t |\varphi_1(u) - \varphi_2(\lambda(u))|\, du + \|\varphi_2\|_T \int_0^t |g(u) - g(\lambda(u))|\, du \\
&\le \|g\|_T \int_0^t |\varphi_1(u) - \varphi_2(\lambda(u))|\, du + c_g T \|\varphi_2\|_T \|\lambda - e\|_T \\
&= c_1 \|\lambda - e\|_T + c_2 \int_0^t |\varphi_1(u) - \varphi_2(\lambda(u))|\, du.
\end{aligned}$$

where $c_1 \equiv c_g T \|\varphi_2\|_T$ and $c_2 \equiv \|g\|_T$.

For $x_1 = f_1(y_1, x_2, b_1)$ we can apply Theorem 4.1 in [20] with input $y \equiv y_1 + \alpha_2 \int_0^t x_2(u)\, du$. It follows from Theorem 2.11 in [27] that if y_2 is continuous then so is x_2. If, in addition, y_1 is continuous, then y is continuous and, by Theorem 4.1 in [20], so is x_1. □

5.3 Martingale Representations

As in Theorem 6.3 of [24], we next apply the representation in Lemmas 1 and 2 to obtain martingale representations for $\hat{Q}_s^n$ and $\hat{Z}_{1,2}^n$, but now we are interested in the FCLT instead of the FWLLN. We exploit martingale representations for the counting processes appearing in lemma 1 constructed from the rate-1 Poisson processes N_i^a, $N_{i,2}^s$, $N_{i,2}^{s,2}$ and N_i^u, $i = 1, 2$, in particular,

$$\begin{aligned}
M_{1,1}^n(t) &\equiv N_{1,1}^s(m_1^n \mu_{1,1} t) - m_1^n \mu_{1,1} t, \\
M_{1,2}^n(t) &\equiv N_{1,2}^s\left(\mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s) \le 0\}} Z_{1,2}^n(s)\, ds\right) - \mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s) \le 0\}} Z_{1,2}^n(s)\, ds, \\
M_{2,2}^n(t) &\equiv N_{2,2}^s\left(\mu_{2,2} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} (m_2^n - Z_{1,2}^n(s))\, ds\right) \\
&\quad - \mu_{2,2} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} (m_2^n - Z_{1,2}^n(s))\, ds, \\
M_{1,3}^n(t) &\equiv N_{1,2}^{s,2}\left(\mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} Z_{1,2}^n(s)\, ds\right) - \mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} Z_{1,2}^n(s)\, ds, \\
M_{2,3}^n(t) &\equiv N_{2,2}^{s,2}\left(\mu_{2,2} \int_0^t 1_{\{D_{1,2}^n(s) \le 0\}} (m_2^n - Z_{1,2}^n(s))\, ds\right) \\
&\quad - \mu_{2,2} \int_0^t 1_{\{D_{1,2}^n(s) \le 0\}} (m_2^n - Z_{1,2}^n(s))\, ds, \\
M_{a_i}^n(t) &\equiv N_i^a(\lambda_i^n t) - \lambda_i^n t, \quad i = 1, 2,
\end{aligned} \tag{28}$$

$$M_{u_i}^n(t) \equiv N_i^u \left(\theta_i \int_0^t Q_i^n(s)ds \right) - \theta_i \int_0^t Q_i^n(s)ds, \quad i = 1, 2.$$

Lemma 4 (*martingale representation for* $\hat{Q}_s^n$)

$$\hat{Q}_s^n(t) = \hat{Q}_s^n(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}_{1,2}^n(s)\, ds - (p_1\theta_1 + p_2\theta_2) \int_0^t \hat{Q}_s^n(s)\, ds + \hat{M}_s^n(t) + o_P(1) \quad as \quad n \to \infty,$$

where $\hat{M}_s^n \equiv M_s^n/\sqrt{n}$ *for the martingale*

$$M_s^n(t) \equiv \sum_{i=1}^2 M_{a_i}^n(t) - \sum_{i=1}^2 M_{u_i}^n(t) - \sum_{i=1}^2 M_{i,2}^n(t) - \sum_{i=1}^2 M_{i,3}^n(t) - M_{1,1}^n(t). \quad (29)$$

with respect to the natural filtration.

Proof By Theorem 1,

$$Q_s^n(t) = Q_s^n(0) + (\lambda_1^n + \lambda_2^n)t - m_1^n\mu_{1,1}t - \mu_{1,2} \int_0^t Z_{1,2}^n(s)\, ds - \mu_{2,2} \int_0^t Z_{2,2}^n(s)\, ds - \theta_1 \int_0^t Q_1^n(s)\, ds - \theta_2 \int_0^t Q_2^n(s)\, ds + M_s^n(t),$$

for $M_s^n(t)$ in (29). Observe that the indicator functions in the representation of X^n in Lemma 1 do not appear in the representation of $Q_s^n(t)$. That simplifies the analysis.

From (9) it follows that $q_s \equiv q_1 + q_2$, the fluid counterpart of Q_s^n, evolves according to the integral equation:

$$q_s(t) = q_s(0) + (\lambda_1 + \lambda_2)t - \mu_{1,1}m_1t - \mu_{1,2} \int_0^t z_{1,2}(u)\, du - \mu_{2,2} \int_0^t z_{2,2}(u)\, du - \theta_1 \int_0^t q_1(u)\, du - \theta_2 \int_0^t q_2(u)\, du,$$

so that, substituting q_1 with $p_1q_s(u)$ and $q_2(u)$ with $p_2q_s(u)$,

$$q_s(t) = q_s(0) + (\lambda_1 + \lambda_2)t - \mu_{1,1}m_1t - \mu_{2,2}m_2t + (\mu_{2,2} - \mu_{1,2}) \int_0^t z_{1,2}(u)\, du - (p_1\theta_1 + p_2\theta_2) \int_0^t q_s(u)\, du.$$

Then, by centering about nq_s and dividing by $\sqrt{n}$ as in (15), we have

$$\hat{Q}_s^n(t) = \hat{Q}_s^n(0) + \frac{[(\lambda_1^n + \lambda_2^n) - n(\lambda_1 + \lambda_2)]t}{\sqrt{n}} - \frac{\mu_{1,1}(m_1^n - nm_1)t}{\sqrt{n}} - \frac{\mu_{1,2} \int_0^t (Z_{1,2}^n(s) - nz_{1,2}(s))\, ds}{\sqrt{n}} - \frac{\mu_{2,2} \int_0^t (Z_{2,2}^n(s) - nz_{2,2}(s))\, ds}{\sqrt{n}} - \frac{\theta_1 \int_0^t (Q_1^n(s) - nq_1(s))\, ds}{\sqrt{n}} - \frac{\theta_2 \int_0^t (Q_2^n(s) - nq_2(s))\, ds}{\sqrt{n}} + \frac{M_s^n(t)}{\sqrt{n}}. \quad (30)$$

By Assumption 1, the second and third terms in the expression above converge to zero. By Corollary 6.2 and Theorem 6.4 in [24], $n^{-1/2}\|Z^n_{2,2}-(m^n_2-Z^n_{1,2})\| \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$ so that $z_{2,2} = m_2 - z_{1,2}$. Also, $(m^n_2 - nm_2)/\sqrt{n} \to 0$ as $n \to \infty$ by Assumption 1. Hence,

$$\begin{aligned}\hat{Q}^n_s = \hat{Q}^n_s(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}^n_{1,2}(s)\, ds \\ - \theta_1 \int_0^t \hat{Q}^n_1(s)\, ds - \theta_2 \int_0^t \hat{Q}^n_2(s)\, ds + \hat{M}^n_s(t) + o_P(1).\end{aligned}$$

Define

$$\begin{aligned}\hat{Q}^n_{a,s}(t) \equiv \hat{Q}^n_s(0) + (\mu_{2,2} - \mu_{1,2}) \int_0^t \hat{Z}^n_{1,2}(s)\, ds - p_1\theta_1 \int_0^t \hat{Q}^n_s(s)\, ds \\ - p_2\theta_2 \int_0^t \hat{Q}^n_s(s)\, ds + \hat{M}^n_s(t)\end{aligned}$$

By applying the SSC result in Lemma 2, we conclude that $\|\hat{Q}^n_s - \hat{Q}^n_{a,s}\|_T \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$ for any $T > 0$. That completes the proof. □

We now turn to the process $Z^n_{1,2}$.

Lemma 5 (*martingale representation for* $\hat{Z}^n_{1,2}$)

$$\begin{aligned}\hat{Z}^n_{1,2}(t) = \hat{Z}^n_{1,2}(0) - \int_0^t \left[(\mu_{2,2} - \mu_{1,2})\pi_{1,2}(x(s)) + \mu_{1,2}\right] \hat{Z}^n_{1,2}(s)\, ds \\ + \hat{L}^n + \hat{M}^n_Z + o(1),\end{aligned} \tag{31}$$

where $\hat{L}^n \equiv L^n/\sqrt{n}$, $\hat{M}^n_Z \equiv M^n_Z/\sqrt{n}$,

$$\begin{aligned}L^n(t) &\equiv \int_0^t [1_{\{D^n_{1,2}(s)>0\}} - \pi_{1,2}(x(s))]\Psi^n(s)\, ds, \\ \Psi^n(s) &\equiv \mu_{2,2}(m^n_2 - Z^n_{1,2}(s)) + \mu_{1,2}Z^n_{1,2}(s)\end{aligned} \tag{32}$$

and M^n_Z *is the martingale*

$$M^n_Z(t) \equiv M^n_{2,2}(t) - M^n_{1,2}(t). \tag{33}$$

with respect to the natural filtration, where $M^n_{2,2}$ *and* $M^n_{1,2}$ *the martingales in* (28).

Proof We start by rewriting the representation of $Z^n_{1,2}$ in Lemma 1 as

$$\begin{aligned}Z^n_{1,2}(t) = Z^n_{1,2}(0) - \mu_{1,2} \int_0^t (1 - \pi_{1,2}(x(s)))Z^n_{1,2}(s)\, ds \\ + \mu_{2,2} \int_0^t \pi_{1,2}(x(s))(m^n_2 - Z^n_{1,2}(s))\, ds + L^n + M^n_Z.\end{aligned}$$

To achieve the diffusion-scaled process, we center $Z^n_{1,2}$ about $nz_{1,2}$ and divide by $\sqrt{n}$, where, by (9), the fluid limit $z_{1,2}$ satisfies the equation

$$z_{1,2}(t) = z_{1,2}(0) + \mu_{2,2}\int_0^t \pi_{1,2}(x(s))(m_2 - z_{1,2}(s))\,ds \\ - \mu_{1,2}\int_0^t (1-\pi_{1,2}(x(s)))z_{1,2}(s)\,ds.$$

We get the representation (31) with the $o(1)$ term replacing the deterministic term $[(m_2^n - nm_2)\int_0^t \pi_{1,2}(x(s))\,ds]/\sqrt{n} \le (m_2^n - nm_2)t/\sqrt{n}$, which converges to zero by Assumption 1. □

5.4 Convergence of Stochastic Driving Terms

Given the representations in Lemmas 4 and 5, we can complete the proof of the convergence of $(\hat{Q}^n_s, \hat{Z}^n_{1,2})$ in Theorem 4 by establishing convergence of the driving terms and applying the continuous mapping theorem with the mapping in Lemma 3, i.e., with the following lemma, proved in the next section. We add an extra process, $\hat{I}^n$, also defined in (15), which is closely related to $\hat{L}^n$, but not directly needed to treat $(\hat{Q}^n_s, \hat{Z}^n_{1,2})$.

Lemma 6 (*convergence of driving terms*) *Under the assumptions of Theorem* 4*,*

$$(\hat{M}^n_s, \hat{M}^n_Z, \hat{L}^n, \hat{I}^n) \Rightarrow (\hat{M}_s, \hat{M}_Z, \hat{L}_2, \hat{I}) \quad \textit{in} \quad \mathcal{D}_4, \tag{34}$$

where

$$\hat{M}_s(t) \equiv B_1(\gamma_1(t)) - B_{1,2}(\gamma_{1,2}(t)) - B_{2,2}(\gamma_{2,2}(t)) \\ - B_{1,3}(\phi_{1,2}(t)) - B_{2,3}(\phi_{2,2}(t)),$$
$$\hat{M}_Z(t) \equiv B_{2,2}(\phi_{2,2}(t)) - B_{1,2}(\phi_{1,2}(t)),$$
$$\hat{L}_2(t) \equiv B_2(\gamma_2(t)) \quad \textit{and} \quad \hat{I}(t) \equiv B_2(\gamma_3(t)), \quad t \ge 0,$$

B_1, $B_{1,2}$, $B_{2,2}$, $B_{1,3}$, $B_{2,3}$ and B_2 are independent standard BM's as in the statement of Theorem 4 *and $\gamma_1(t)$, $\gamma_2(t)$, $\gamma_3(t)$, $\gamma_{1,2}(t)$, $\gamma_{2,2}(t)$, $\phi_{1,2}(t)$ and $\phi_{2,2}(t)$ are the increasing continuous functions defined in* (19)*.*

5.5 Overall Proof of Theorem 4

We prove convergence of $(\hat{Q}^n, \hat{Z}^n_{1,2})$ by applying the continuous mapping theorem with the continuous function in Lemma 3, exploiting the representations in Lemmas 4 and 5 and the convergence established in Lemma 6. In applying Lemma 3, we rely heavily on Theorem 7.1 in [23], which establishes that $\pi_{1,2}(\cdot)$ is locally Lipschitz continuous in $\mathbb{A}$ as a function of the fluid state x and is thus Lipschitz continuous over compact sets. Moreover, $x(\cdot)$ is itself Lipschitz continuous, as a function of the time argument s by Corollary 5.1 in [24]. It

follows that $\pi_{1,2}(x(s))$ is Lipschitz continuous as a function of the time argument s as well (using Assumption 6 implying that x lies entirely in $\mathbb{A}$). Thus the proof of Theorem 4 is complete with the exception of the proof of Lemma 6. The next four sections are devoted to that proof.

6 Proof of Lemma 6

This section is devoted to proving Lemma 6. In §6.1 we apply standard arguments to establish the convergence of the first two martingale terms $(\hat{M}_s^n, \hat{M}_Z^n)$. In preparation for treating the last two terms, in §6.2 we state two key results that we will use; they are proved in the following three sections. In §6.3 we establish convergence of the last two terms $(\hat{L}^n, \hat{I}^n)$. Finally, in §6.4 we establish joint convergence of all four terms by proving asymptotic independence of the last two terms from the first two terms.

6.1 The First Two Terms in (34)

We start by establishing convergence of the first two terms in Lemma 6, the two martingale terms.

Lemma 7 *There is joint convergence of the martingale processes*

$$(\hat{M}_s^n, \hat{M}_Z^n) \Rightarrow (\hat{M}_s, \hat{M}_Z) \quad \textit{in} \quad \mathcal{D}_2,$$

where the processes are defined in (29), (33) *and Lemma* 6.

Proof Let

$$\begin{aligned}
\hat{M}_S^n(t) &= \left(\hat{M}_{1,1}^n(t), \hat{M}_{1,2}^n(t), \hat{M}_{2,2}^n(t), \hat{M}_{1,3}^n(t), \hat{M}_{2,3}^n(t)\right) \quad \text{in} \quad \mathcal{D}_5, \\
\hat{M}_A^n(t) &= \left(\hat{M}_{a_1}(t), \hat{M}_{a_2}(t)\right) \quad \text{and} \quad \hat{M}_U^n(t) = \left(\hat{M}_{u_1}^n(t), \hat{M}_{u_2}^n(t)\right) \quad \text{in} \quad \mathcal{D}_2
\end{aligned}$$

for the martingale processes in (28). To compress the notation, for $x \in \mathcal{D}_k$ and $t \in [0,\infty)^k$, let $x(t) \equiv (x_1(t_1), x_2(t_2), \ldots, x_k(t_k))$. We start by proving that

$$\left(\hat{M}_A^n(t), \hat{M}_S^n(t), \hat{M}_U^n(t)\right) \Rightarrow \left(B_A(\lambda t), B_S(\phi(t)), B_U\left(\theta \int_0^t q(s)\, ds\right)\right) \quad \text{in} \quad \mathcal{D}_9 \tag{35}$$

as $n \to \infty$, where

$$\begin{aligned}
&\phi(t) \equiv (\phi_1(t), \phi_2(t), \phi_3(t), \phi_4(t), \phi_5(t)), \\
&\phi_1(t) \equiv m_1 \mu_{1,1}(t), \quad \phi_2(t) \equiv \phi_{1,2}(t), \quad \phi_3(t) \equiv \phi_{2,2}(t), \\
&\phi_4(t) \equiv \gamma_{1,2}(t), \quad \phi_5(t) \equiv \gamma_{2,2}(t),
\end{aligned} \tag{36}$$

for $\phi_{i,2}$ and $\gamma_{i,2}$ defined in (19). Here $B_A(t)$, $B_S(t)$ and $B_U(t)$ are vectors of independent standard Brownian motions. Using our compressed notation, we

have $\lambda t \equiv (\lambda_1 t, \lambda_2 t)$ and $\theta q(s) \equiv (\theta_1 q_1(s), \theta_2 q_2(s))$. For example, $B_A(\lambda t) = (B_{A_1}(\lambda_1 t), B_{A_2}(\lambda_2 t))$, and similarly for $B_S(\cdot)$ and $B_U(\cdot)$.

To prove (35), we apply the FCLT for Poisson processes. For the Poisson processes in Lemma 1, let

$$\tilde{M}^n_{a_i} = \frac{N^a_i(nt) - nt}{\sqrt{n}}, \quad \tilde{M}^n_{i,j} = \frac{N^s_{i,j}(nt) - nt}{\sqrt{n}} \qquad \text{and}$$
$$\tilde{M}^n_{u_i} = \frac{N^u_i(nt) - nt}{\sqrt{n}}, \qquad i = 1, 2; j = 1, 2, 3.$$

Let $\tilde{M}^n_A(t)$, $\tilde{M}^n_S(t)$ and $\tilde{M}^n_U(t)$ be the corresponding vector-valued processes. By the independence of all the unit-rate Poisson processes $N^a_i(\cdot)$, $N^s_{i,j}(\cdot)$ and $N^u_i(\cdot)$, and the FCLT for a Poisson process, the following joint convergence holds:

$$\left(\tilde{M}^n_A(t), \tilde{M}^n_S(t), \tilde{M}^n_U(t)\right) \Rightarrow \left(\tilde{B}_A(t), \tilde{B}_S(t), \tilde{B}_U(t)\right) \quad \text{in } \mathcal{D}_9 \text{ as } n \to \infty, \quad (37)$$

where $\tilde{B}_A$, $\tilde{B}_S$ and $\tilde{B}_U$ are, respectively, 2-dimensional, 5-dimensional and 2-dimensional independent Brownian motions; see Theorem 4.2 and §9.1 in [20].

We now introduce random time changes. Let

$$\begin{aligned}
\Phi^n_{A,i}(t) &\equiv n^{-1}\lambda^n_i t, \qquad \Phi^n_{S,1}(t) \equiv n^{-1}\mu_{1,1} m^n_1 t, \\
\Phi^n_{S,2}(t) &\equiv n^{-1}\mu_{1,2} \int_0^t 1_{\{D^n_{1,2}(s) \le 0\}} Z^n_{1,2}(s)\, ds, \\
\Phi^n_{S,3}(t) &\equiv n^{-1}\mu_{2,2} \int_0^t 1_{\{D^n_{1,2}(s) > 0\}} (m^n_2 - Z^n_{1,2}(s))\, ds, \\
\Phi^n_{S,4}(t) &\equiv n^{-1}\mu_{1,2} \int_0^t 1_{\{D^n_{1,2}(s) > 0\}} Z^n_{1,2}(s)\, ds, \\
\Phi^n_{S,5}(t) &\equiv n^{-1}\mu_{2,2} \int_0^t 1_{\{D^n_{1,2}(s) \le 0\}} (m^n_2 - Z^n_{1,2}(s))\, ds, \\
\Phi^n_{U,i}(t) &\equiv n^{-1}\theta_i \int_0^t Q^n_i(s)\, ds, \quad i, j = 1, 2. \qquad (38)
\end{aligned}$$

By Assumption 1 on the arrival rates, $\Phi^n_{A_i} \to \lambda_i e$ in $\mathcal{D}$, $i = 1, 2$. From the initial conditions in the statement of Theorem 4, the fluid limit and the continuity of the integral mapping, it follows that $\Phi^n_{S,i} \Rightarrow \phi_i$, $1 \le i \le 5$, and $\Phi^n_{U,i}(t) \Rightarrow \theta_i \int_0^t q_i(s)\, ds$ in $\mathcal{D}$ as $n \to \infty$.

Let $\Phi^n_A(t)$, $\Phi^n_S(t)$ and $\Phi^n_U(t)$ be the corresponding vector-valued processes. By Theorem 11.4.5 of [32], these limits hold jointly, yielding

$$(\Phi^n_A(t), \Phi^n_S(t), \Phi^n_U(t)) \Rightarrow \left(\lambda t,\, \phi(t),\, \theta \int_0^t q(s)\, ds\right) \quad \text{in} \quad \mathcal{D}_9 \qquad (39)$$

as $n \to \infty$. By Theorem 11.4.5 of [32], the limits in (37) and (39) also hold jointly. By definition,

$$\left(\hat{M}^n_A(t), \hat{M}^n_S(t), \hat{M}^n_U(t)\right) = \left(\tilde{M}^n_A\left(\Phi^n_A(t)\right), \tilde{M}^n_S\left(\Phi^n_S(t)\right), \tilde{M}^n_U\left(\Phi^n_U(t)\right)\right).$$

Thus, the convergence in (35) follows from the continuity of the composition mapping at continuous limits, Theorem 13.2.1 in [32]. Finally, the conclusion of the lemma itself then follows from the definition of $\hat{M}_s^n$ and $\hat{M}_Z^n$ in (29) and (33), and the continuity of addition under continuous limits, e.g., Corollary 12.7.1 in [32]. □

6.2 Key Supporting Results for the Last Two Terms

In §4.1 we indicated that the stochastic limit will depend on the FCLT for cumulative processes associated with the FTSP, as stated in (13). As indicated in §6 of [23], the FTSP with fixed state γ is a QBD; its parameters (transition rates) are given explicitly in (13)-(16) of [24]. Since the FTSP $D(\gamma, \cdot)$ is a QBD for each state γ, it is a relatively simple regenerative stochastic process for each state γ, assuming that γ makes the QBD positive recurrent. However, in our application, the fluid state is *not* fixed at γ, but is instead given by the fluid limit $x(t)$, which is a function of time t. That means that the parameters of the FTSP are actually time-varying. By Assumption 6, the FTSP with fluid state $x(t)$ is a positive recurrent QBD for all states $x(t)$ considered. Moreover, by Lemma C.5 of [24], the infinitesimal generator and the asymptotic variance of the QBD are continuous functions of the underlying state $x(t)$. Since the essential matrix structure (e.g., the dimension of the matrices) of the QBD's depends only on the rational ratio parameter $r_{1,2}$, and thus does not change, the QBD is characterized by only finitely many parameters. As a consequence, we can establish a variant of the FCLT in (13), allowing the FTSP to have a time-varying state.

In our remaining proof of Lemma 6, in particular for Lemma 8 below, we will want to generalize the state of the QBD. The parameters of the QBD depend on the fluid state $\gamma \equiv (q_1, q_2, z_{1,2})$, but also on the rest of the QBD parameters, in particular, also upon $\zeta \equiv (\lambda_i, m_j; i, j = 1, 2)$. In order to establish Lemma 8 below, we will want to allow the parameters λ_i and m_j to vary, because they vary with n in the many-server heavy-traffic scaling in Assumption 1. The QBD also depends on the other model parameters θ_i and $\mu_{i,j}$, but they are fixed, so we do not include them. Thus, we will consider the more general "full" *parameter state function* $\eta \equiv (\zeta, \gamma)$ for $\eta \equiv \eta(t)$ and $\gamma \equiv \gamma(t)$ above, which we understand to be an element of the functions space $\mathcal{D}$. We obtain a conventional stationary QBD model for each full parameter state $\eta(t)$.

Now we will establish a FCLT for

$$\hat{C}^n(t;\eta) \equiv n^{-1/2} \int_0^{nt} \left(1_{\{D(\eta(s/n),s)>0\}} - \pi_{1,2}(\eta(s/n))\right) ds, \quad t \geq 0, \qquad (40)$$

where the state function η is an element of $\mathcal{D}$ and $D(\eta(0), 0)$ is some fixed finite initial value. Note that in the special case of a constant parameter state function, with $\eta(t) = \gamma$, $0 \leq t \leq T$, this new process reduces to the previous one in §4.1; i.e.,

$$\hat{C}^n(t;\eta) = \hat{C}^n_{QBD}(t;\gamma), \quad 0 \leq t \leq T.$$

for $\hat{C}^n_{QBD}(t;\gamma)$ in (13).

However, more generally, the process $\hat{C}^n(t;\eta)$ in (40) is more complicated, so that the new FCLT is by no means immediate. The non-constant function η makes the new process $\{D(\eta(s/n),s) : s \geq 0\}$ appearing in the integrand of (40) neither a QBD nor a regenerative process. Nevertheless, we establish the following generalization of the FCLT in (13). The proof is given in §7.

Theorem 6 (*FCLT for FTSP with time-varying parameter state*) *Consider the FTSP D as a function of its parameter state function η specified above, where η is a function in $\mathcal{D}$. Suppose that the QBD $D(\eta(t),\cdot)$ is positive recurrent for all $\eta(t)$, $0 \leq t \leq T$. Then*

$$\hat{C}^n(\cdot;\eta) \Rightarrow \hat{C}(\cdot;\eta) \quad in \quad \mathcal{D}([0,T]) \quad as \quad n \to \infty,$$

where $\hat{C}^n(\cdot;\eta)$ is given in (40) *and*

$$\hat{C}(t;\eta) \equiv B\left(\int_0^t \sigma^2(\eta(u))\,du\right), \quad t \geq 0,$$

with B being a standard BM and, for each u, $\sigma^2(\eta(u))$ is the asymptotic variance of the cumulative process with constant full parameter state $\eta(u)$, as in (13)-(14).

For Lemma 8 below, we will also want to extend the FCLT in Theorem 6 to full parameter state functions that are suitably near a given deterministic one. For that purpose, we use the following elementary corollary to Theorem 6 and its proof. (Also see §6 of [23] and §C.3 of[24]. We use the Prohorov metric $d_{\mathcal{P},T}(Y_1,Y_2)$ characterizing convergence in distribution in $\mathcal{D}([0,T])$; see p. 77 of [32]. We say that a parameter-state function η is positive recurrent if the associated FTSP $D(\eta,\cdot)$ is positive recurrent.

Corollary 3 (*continuity of the FCLT for the FTSP with time-varying parameter state*) *Consider the FTSP D as a function of its parameter state function η specified above, where the parameter state function η is a positive-recurrent element of $\mathcal{D}$. For all $\epsilon > 0$ and $T > 0$, there exists $\delta > 0$ such that, if η' is a parametric state function satisfying $\|\eta - \eta'\|_T < \delta$, then η' is positive recurrent for all t in $[0,T]$ and $d_{\mathcal{P},T}(\hat{C}(\cdot;\eta),\hat{C}(\cdot,\eta')) < \epsilon$ where $\hat{C}(\cdot;\eta)$ is the limit process associated with $D(\eta,\cdot)$ in Theorem 6.*

Proof We exploit the criterion for recurrence in terms of the drift rates given in (8). The drift rates $\delta_+(\eta)$ and $\delta_-(\eta)$ for constant η in the regions $\{s : D(\eta,s) > 0\}$ and $\{s : D(\eta,s) \leq 0\}$, respectively, are linear functions of the components of the vector η. We can thus express the drifts as the inner products $\delta_\pm(\eta) = a_\pm \cdot \eta$, where a_+ and a_- are constant vectors. Hence, if $|\eta - \eta'| \leq \epsilon$, then $|\delta_\pm(\eta) - \delta_\pm(\eta')| \leq \epsilon(|a_\pm| \cdot 1)$, where here 1 is a vector of $1's$ of the appropriate dimension. This property for constant parameter states extends immediately to more general state functions in $\mathcal{D}$ using the uniform norm; i.e., if $\|\eta - \eta'\|_T \leq \epsilon$, then $\|\delta_\pm(\eta) - \delta_\pm(\eta')\|_T \leq \epsilon(|a_\pm| \cdot 1)$. Thus, for any positive recurrent state function η, there exists $\epsilon > 0$ such that $\delta_+(\eta') < 0$ and $\delta_-(\eta') > 0$ if $\|\eta - \eta\|_T < \epsilon$, implying that η' is also positive recurrent. □

In Lemma 8 below, we will apply Corollary 3 to random state functions $\tilde{\eta}_n$ which converge weakly to η as $n \to \infty$, i.e., for which $\tilde{\eta}_n \Rightarrow \eta$ in $\mathcal{D}$ as $n \to \infty$. To do so, we need to connect the queue-difference processes $D^n_{1,2}$ appearing in $\hat{I}^n$ in (15) to the FTSP. We do that via the associated *frozen processes*, introduced in §A.1 of [24]. The frozen process $\{D^n_f(X^n(t), s) : s \geq 0\}$ corresponds to the queue-difference process $D^n_{1,2}$ starting at time t, conditioned on the state $X^n(t)$ at time t under the assumption that the transition rates are fixed ("frozen") at the rates associated with the initial state $X^n(t)$. A key property, for applying Theorem 6 and Corollary 3 above, is that the frozen process can be represented as the FTSP with modified parameters. To express the connection, we write the frozen process and the FTSP as functions of the parameters $(\lambda_i, m_j, \gamma, s)$. As in equation (74) of [24], we have the representation

$$\{D^n_f(\lambda^n_i, m^n_j, X^n(t), s) : s \geq 0\} \stackrel{\mathrm{d}}{=} \{D(\lambda^n_i/n, m^n_j/n, X^n(t)/n, ns) : s \geq 0\}, \tag{41}$$

where D^n_f on the left of (41) is the frozen process described above, and D on the right of (41) is the FTSP.

Like the queue-difference process, the frozen process has $O(n)$ transition rates, whereas the FTSP has $O(1)$ transition rates, because of the time scaling in (5). Thus the time variable s on the right in (41) is scaled by n.

However, we need to construct a process that is made up of different frozen processes over different subintervals. Thus, for each $n \geq 1$, we will construct a process that is a different frozen process over each successive interval of length $1/n$, but identical to the queue-difference process at each interval endpoint. In particular, we will construct the overall frozen process by setting

$$\tilde{D}^n_f(t) \equiv D^n_f(X^n((k-1)/n), t-(k-1)/n), \quad \frac{k-1}{n} \leq t < \frac{k}{n}, \tag{42}$$

$0 \leq t \leq T$, where D^n_f is the frozen process defined above. That is, we use a different frozen state and thus frozen process for each interval $[(k-1)/n, k/n)$ in $[0, T]$. As a consequence, the frozen process state for the process $\tilde{D}^n_f$ as a function of t is thus

$$X^n_f(t) \equiv X^n(\lfloor nt \rfloor/n), \quad 0 \leq t \leq T. \tag{43}$$

As a consequence of (41)-(43) above, we can simply write

$$\{\tilde{D}^n_f(t) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{D(\tilde{\eta}_n(t), nt) : t \geq 0\}, \tag{44}$$

where $\tilde{\eta}_n$ is a random full parameter state function with the special parameter function given in (41) above, with the frozen state at time t given by (43). Corollary 3 is relevant because, by virtue of Assumption 1 and Theorem 1, for each $T > 0$, we have

$$\|\tilde{\eta}_n - \eta\|_T \Rightarrow 0 \quad \text{as} \quad n \to \infty,$$

where η has fixed components λ_i, m_j and $x(t)$, $t \geq 0$.

Hence, the FCLT for fixed positive recurrent state function η, which holds by Theorem 6, also holds with η replaced by $\tilde{\eta}_n$ by virtue of Corollary 3. However, it remains to show that the newly constructed frozen processes approximate the queue-difference processes suitably well. For that, we will use a special coupling construction, similar to the coupling constructions used in [24]. The following result is proved in §10.

Lemma 8 *For each n, we can construct the new frozen processes $\tilde{D}_f^n$ defined by* (41)-(44) *on the same underlying probability space with the queue-difference processes $D_{1,2}^n$ so that, $\Delta^n \Rightarrow 0$ in $\mathcal{D}$ as $n \to \infty$, where*

$$\Delta^n(t) \equiv \sqrt{n} \int_0^t \left(1_{\{D_{1,2}^n(s)>0\}} - 1_{\{\tilde{D}_f^n(s)>0\}} \right) ds, \quad t \geq 0. \tag{45}$$

6.3 The Last Two Terms in (34)

We now establish joint convergence of the last two terms in Lemma 6.

Lemma 9 *There is joint convergence of the last two terms in Lemma* 6*, i.e.,*

$$(\hat{L}^n, \hat{I}^n) \Rightarrow (\hat{L}_2, \hat{I}) \quad \textit{in} \quad \mathcal{D}_2,$$

where the converging processes $\hat{L}^n$ and $\hat{I}^n$ are defined, respectively, in (32) *and* (15)*, while the vector limit process is $(\hat{L}_2(t), \hat{I}(t)) \equiv (B_2(\gamma_2(t), B_2(\gamma_3(t))$ for B_2 a standard Brownian motion and $(\gamma_2(t), \gamma_3(t))$ in* (19)*, as in* (18)*.*

Proof We start by considering just $\hat{I}^n$. We make a change of variables in (15) to get

$$\hat{I}^n(t) \equiv \frac{1}{\sqrt{n}} \int_0^{nt} [1_{\{D_{1,2}^n(s/n)>0\}} - \pi_{1,2}(x(s/n))] \, ds, \quad 0 \leq t \leq T. \tag{46}$$

From either the original representation of $\hat{I}^n$ in (15) or the equivalent alternative expression in (46), the main line of the proof should be evident: We show that the time-scaled queue-difference process $D_{1,2}^n(s/n)$ in (46) is asymptotically equivalent to the scaled FTSP $D(x(s/n), s)$, making the expression in (46) be essentially of the form of $\hat{C}^n$ in (40). If we could just directly make that substitution, then the desired limit $\hat{I}^n \Rightarrow \hat{I}$ would be an immediate consequence of Theorem 6. However, the desired substitution is only valid asymptotically. We actually achieve the desired approximation by the FTSP indirectly by approximating the queue-difference process and applying Lemma 8 and Corollary 3 in addition to Theorem 6. In particular, we can write

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_0^{nt} [1_{\{D_{1,2}^n(s/n)>0\}} - \pi_{1,2}(x(s/n))] \, ds &= \sqrt{n} \int_0^t [1_{\{D_{1,2}^n(s)>0\}} - \pi_{1,2}(x(s))] \, ds \\ &= \sqrt{n} \int_0^t \left(1_{\{D_{1,2}^n(s)>0\}} - 1_{\{\tilde{D}_f^n(s)>0\}} \right) ds \end{aligned}$$

$$+\frac{1}{\sqrt{n}}\int_0^{nt} 1_{\{\tilde{D}_f^n(s/n)>0\}} - \pi_{1,2}(x(s/n))]\,ds.$$

We then apply Lemma 8 to the first component in the RHS of the equality, and Corollary 3 to the second component, using (44).

Having established the limit for $\hat{I}^n$, we turn to $\hat{L}^n$. From (32), we know that L^n differs from I^n by having the extra term Ψ^n in the integrand. However, by the FWLLN, Theorem 1, $\Psi^n \Rightarrow \psi$ as $n \to \infty$, where $\psi(t) \equiv \mu_{2,2}(m_2 - z_{1,2}(t)) + \mu_{1,2}z_{1,2}(t)$. Hence, we first write

$$\hat{L}_1^n(t) \equiv \sqrt{n}\int_0^t [1_{\{D_{1,2}^n(s)>0\}} - \pi_{1,2}(x(s))]\psi(s)\,ds, \quad t \geq 0.$$

Since ψ is continuous, we can approximate it uniformly closely by a piecewise constant function, with all discontinuities occurring at multiples of a small positive ϵ. Hence, by approximation, we can assume without loss of generality that

$$\hat{L}_1^n(t) \equiv \sum_{j=1}^{\lfloor t/\epsilon \rfloor + 1} \psi_j \hat{I}_j^n(t),$$

where ψ_j is a constant for each j and $\hat{I}_j^n(t)$ has the structure of $\hat{I}^n$ over the subinterval $[(j-1)\epsilon, j\epsilon)$ and is 0 outside that interval. Hence, we have convergence of $\hat{L}_1^n$, jointly with $\hat{I}^n$, by essentially the same argument as for $\hat{I}^n$. Finally, we can write

$$\|\hat{L}^n - \hat{L}_1^n\|_T \leq \|\bar{\Psi}^n - \psi\|_T \|\hat{I}^n\|_T \Rightarrow 0 \quad \text{as} \quad n \to \infty,$$

because $\|\bar{\Psi}^n - \psi\|_T \Rightarrow 0$ and $\hat{I}^n \Rightarrow \hat{I}$ as $n \to \infty$, so that $\|\hat{I}^n\|_T \Rightarrow \|\hat{I}\|_T$, where $\|\hat{I}\|_T$ is finite by the continuous mapping theorem with the function $\|\cdot\|_T$. Hence, we obtain the claimed joint limit for $(\hat{L}^n, \hat{I}^n)$. □

6.4 Joint Convergence in Lemma 6

To complete the proof of Lemma 6, it remains to show that the two limits established in Lemmas 7 and 9 actually hold jointly. The two separate limits directly imply the associated tightness, which in turn implies the tightness for the sequence of four-dimensional processes. Thus, to prove convergence it suffices to show that the limits of all converging subsequences coincide. We uniquely characterize the joint limit by showing that the limit for every convergent subsequence of the sequence $\{(\hat{M}_s^n, \hat{M}_Z^n, \hat{L}^n, \hat{I}^n)\}$ must be of the form $(\hat{M}_s, \hat{M}_Z, \hat{L}, \hat{I})$, where $(\hat{M}_s, \hat{M}_Z)$ is independent of $(\hat{L}, \hat{I})$, having distributions as determined above. Thus it suffices to establish the following lemma.

Lemma 10 (*independent limits*) *The limits* $(\hat{M}_s, \hat{M}_Z)$ *and* $(\hat{L}, \hat{I})$ *for every convergent subsequence of the sequence* $\{(\hat{M}_s^n, \hat{M}_Z^n, \hat{L}^n, \hat{I}^n)\}$ *are independent.*

In order to prove Lemma 10, we use the following lemma.

Lemma 11 (*basis for independent limits*) *If $\hat{D}^n \Rightarrow 0e$, $\hat{I}^n \Rightarrow \hat{I}$ and $\hat{V}^n \Rightarrow \hat{V}$ as $n \to \infty$ for random vectors $(\hat{I}^n, \hat{D}^n, \hat{V}^n)$ in $\mathcal{D}_3$, where $\hat{D}^n = f(\hat{V}^n)$ for some function f and $P(\hat{I}^n \in B|\hat{V}^n = v) = P(\hat{I}^n \in B|\hat{D}^n = f(v))$ for all Borel sets B almost surely with respect to $dP(\hat{V}^n = v)$, then $\hat{I}$ is independent of $\hat{V}$.*

Proof Let g_i be a continuous bounded real-valued function on $\mathcal{D}$ for $i = 1, 2, 3$. By the assumptions above,

$$\begin{aligned} E[g_1(\hat{I}^n)g_2(\hat{D}^n)g_3(\hat{V}^n)] &= E[E[g_1(\hat{I}^n)g_2(\hat{D}^n)|\hat{V}^n]g_3(\hat{V}^n)] \\ &= E[E[g_1(\hat{I}^n)g_2(\hat{D}^n)|\hat{D}^n]g_3(\hat{V}^n)] \\ &= E[E[g_1(\hat{I}^n)|\hat{D}^n]g_2(\hat{D}^n)g_3(\hat{V}^n)]. \end{aligned} \tag{47}$$

Since $\hat{I}^n \Rightarrow \hat{I}$ and $\hat{D}^n \Rightarrow 0e$, we also have $(\hat{I}^n, \hat{D}^n) \Rightarrow (\hat{I}, 0e)$ by Theorem 11.4.5 of [32]. Thus, $E[g_1(\hat{I}^n)g_2(\hat{D}^n)] \Rightarrow E[g_1(\hat{I})g_2(\hat{D})] = g_2(0e)E[g_1(\hat{I})]$ as $n \to \infty$, so that $\hat{I}^n$ is asymptotically independent of $\hat{D}^n$ and

$$E[g_1(\hat{I}^n)|\hat{D}^n]g_2(\hat{D}^n) \Rightarrow E[g_1(\hat{I})]g_2(0e) \in \mathbb{R} \quad \text{as} \quad n \to \infty.$$

By Theorem 11.4.5 of [32] once again,

$$(E[g_1(\hat{I}^n)|\hat{D}^n]g_2(\hat{D}^n), \hat{V}^n) \Rightarrow (E[g_1(\hat{I})]g_2(0e), \hat{V}) \quad \text{in} \quad \mathbb{R} \times \mathcal{D} \quad \text{as} \quad n \to \infty,$$

so that, applying the continuous mapping theorem with the function $h : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ defined by $h(x, y) \equiv xg_3(y)$, we obtain

$$E[g_1(\hat{I}^n)|\hat{D}^n]g_2(\hat{D}^n)g_3(\hat{V}^n) \Rightarrow E[g_1(\hat{I})]g_2(0e)g_3(\hat{V}) \quad \text{in} \quad \mathbb{R}. \tag{48}$$

Since the random variables in (48) are bounded, we can apply the bounded convergence theorem, combined with (47) and (48), to get

$$E[g_1(\hat{I}^n)g_2(\hat{D}^n)g_3(\hat{V}^n)] \to E[g_1(\hat{I})]g_2(0e)E[g_3(\hat{V})] \quad \text{as} \quad n \to \infty.$$

From the special case $g_2 \equiv 1e$, $E[g_1(\hat{I}^n)g_3(\hat{V}^n)] \to E[g_1(\hat{I})]E[g_3(\hat{V})]$ as $n \to \infty$. Since the product g_1g_3 is a continuous bounded real-valued function, we also have $E[g_1(\hat{I}^n)g_3(\hat{V}^n)] \to E[g_1(\hat{I})g_3(\hat{V})]$ as $n \to \infty$. Hence, $E[g_1(\hat{I})g_3(\hat{V})] = E[g_1(\hat{I})]E[g_3(\hat{V})]$ for all continuous bounded real-valued functions g_1 and g_3, so that $\hat{I}$ is independent of $\hat{V}$. □

Proof of Lemma 10 We show that the conditions of Lemma 11 are satisfied in our case. For that, we rely strongly on the SSC result in Corollary 4.1 of [24]. We first observe that, for each n, the stochastic process $\{D^n_{1,2}(t) : t \geq 0\}$, and thus also the stochastic processes $\{1_{\{D^n_{1,2}(t)>0\}} : t \geq 0\}$ and $\hat{I}^n$ in (15) are directly functions of $\hat{D}^n$ in (15). Thus, for each n, the conditional distribution of $\hat{I}^n$ in $\mathcal{D}$ conditional on $\hat{V}^n \equiv \left(\hat{A}^n_i, \hat{U}^n_i, \hat{S}^n_{i,j}, \hat{D}^n, \hat{Q}^n_i, \hat{Q}^n_s, \hat{Z}^n_{i,j}\right)$ in (15) coincides with the conditional distribution of $\hat{I}^n$ in $\mathcal{D}$ conditional on $\hat{D}^n$. Moreover, $\hat{D}^n$ is scaled in the same way as the other processes in $\hat{V}^n$ in (15). However, Theorem 4.5 (iii) and its corollary 4.1, both from [24], imply that $\hat{D}^n \Rightarrow 0e$. Thus all the conditions of Lemma 11 are satisfied, and the statement of the Lemma follows. □

7 Proof of Theorem 6

First, if the parameter state function η is a constant function, with $\eta(t) = \gamma$, $0 \leq t \leq T$, then $\hat{C}^n(t;\eta) = \hat{C}^n_{QBD}(t,\gamma)$ for $\hat{C}^n(t;\eta)$ in (40) and $\hat{C}^n_{QBD}(t,\gamma)$ in (13), as noted in §6.2. Moreover, if the QBD $D(\gamma,\cdot)$ is positive recurrent, then the conclusion in Theorem 6 reduces to the standard FCLT for a cumulative process in (13). To consider more general time-varying parameter state functions $\eta \equiv \{\eta(u) : 0 \leq u \leq T\}$, we require that η be positive recurrent where, as before, we say that a state function η is positive recurrent if the associated FTSP $D(\eta(t),\cdot)$ is positive recurrent for all t, $0 \leq t \leq T$.

Next, we observe that the conclusion in Theorem 6 is also valid for all positive-recurrent piecewise constant parameter state functions, where we include the condition that there be only finitely many discontinuities in each bounded interval. Let $\mathcal{D}_{pc}$ be the subspace of $\mathcal{D}$ containing all such piecewise constant functions. To see that the conclusion holds for each positive recurrent $\eta \in \mathcal{D}_{pc}$, note that, because of the time scaling, each subinterval $[a,b)$ of length $O(1)$ for the state function η corresponds to an interval of length $O(n)$ for the stochastic process $\{D(\eta(s/n),s) : s \geq 0\}$, which has transition rates of $O(1)$. Moreover, the convergence on each successive interval implies that the initial distributions converge on the next interval. Hence, the initial conditions on each subinterval do not alter the limit. Thus, the separate subintervals can be treated separately, as if we were considering the first case of a constant parameter state function.

Intuitively, it should be evident that the result extends to positive recurrent state functions η in $\mathcal{D}$ because each such function is the uniform limit over bounded intervals of piecewise-constant state functions; see p. 393 of [32]. However, a complete proof for this seemingly minor extension seems quite complicated. The remaining proof will be based on a series of lemmas, which are proved in the next section.

First, we exploit Corollary 3 showing that the subset of positive recurrent state functions in $\mathcal{D}$ is an open subset. With Corollary 3, we then exploit the continuity of QBD's established in Lemma C.5 of [24] to complete the proof. We complete the proof in several steps, requiring further lemmas. In doing so, we will exploit frozen processes to simplify the argument. As before, we use a coupling construction to show that they serve as suitable asymptotic approximations.

Here we consider a modification of the process $\hat{C}^n$ in (40), having a parameter state that is frozen over each successive cycle, where as before a cycle is the period between successive visits to a fixed state. As remarked before, in the case of a constant parameter state function η, with $\eta(t) = \gamma$, $0 \leq t \leq T$, these are the regeneration cycles associated with the regenerative process $D(\gamma,\cdot)$, as in §4.1, but here we have a more general case. For each n, let $\hat{C}^n_f$ denote this modification of $\hat{C}^n$, having a parameter state that is frozen over each successive cycle. We use a coupling construction to show that it suffices to consider $\hat{C}^n_f$ in order to establish the desired convergence of $\hat{C}^n$ in (40).

Lemma 12 (*frozen cumulative processes*) *The processes* $\hat{C}_f^n$ *and* $\hat{C}^n$ *can be constructed on the same underlying space so that* $d_{J_1}(\hat{C}_f^n, \hat{C}^n) \Rightarrow 0$.

Now we want to establish the convergence $\hat{C}_f^n \Rightarrow \hat{C}$ as $n \to \infty$. To do so, we apply modified versions of the reasoning used to prove the FCLT in (13), as given in [9]. In particular, as in (1.1)-(1.4) of [9], we observe that $\hat{C}_f^n$ is asymptotically equivalent to a random sum, ignoring remainder terms, and we then establish convergence for the sequence of random sums. To set the stage, let the i^{th} full cycle in system n end at time T_i^n. (Recall that the cycle begins upon transition into the designated state, while the next cycle begins upon first returning to that state after first leaving the state, which is well defined because the processes are pure-jump processes.)

As in §4.1, the key random variables associated with these cycles are the *cycle lengths*

$$\tau_i^n \equiv T_i^n - T_{i-1}^n, \quad i \geq 1, \tag{49}$$

and the integrals of the centered process over the cycle, which we call the *cycle variables*,

$$Y_i^n \equiv \int_{T_{i-1}^n}^{T_i^n} \left(1_{\{D(\gamma_i,s)>0\}} - \pi_{1,2}(\gamma_i)\right) \, ds, \quad i \geq 0, \tag{50}$$

where $\gamma_i \equiv \eta(T_{i-1}^n)$, with T_{i-1}^n being the random time at which the i^{th} full cycle begins and $T_0^n = 0$, so that Y_0^n is the cycle variable for the first partial cycle. We do not need to make additional assumptions for the analog of the variables $W_i(f)$ in (1.2) of [9] because

$$W_i^n \equiv \int_{T_{i-1}^n}^{T_i^n} \left|1_{\{D(\gamma_i,s)>0\}} - \pi_{1,2}(\gamma_i)\right| \, ds \leq \tau_i^n. \tag{51}$$

With this construction, we can write

$$\hat{C}_f^n(T_i^n; \eta) = \hat{C}^n(T_i^n; \tilde{\eta}_f^n), \quad i \geq 0,$$

for

$$\tilde{\eta}_f^n(t) = \gamma_i, \quad T_{i-1}^n \leq t < T_i^n, \quad t <\geq 0.$$

Unlike for a regenerative process, as in [9], here the random *cycle vectors* (τ_i^n, Y_i^n) are in general neither independent nor identically distributed. However, the sequence of cycle variables $\{(\tau_j^n, Y_j^n) : j \geq i\}$ is conditionally independent of the entire system history up to time T_{i-1}^n, which we denote by $\mathcal{F}_{i-1}^n$, given only T_{i-1}^n, for each $i \geq 0$ and $n \geq 1$. Of course, in general these conditional distributions vary with i because the parameter state function η is not constant, but they change little if η changes little, by the QBD continuity.

Let $N^n(t)$ count the number of full cycles up to time t. As in (1.4) of [9], we can write

$$\hat{C}_f^n(t) = \hat{R}^n(t) + \hat{R}_1^n(t) + \hat{R}_2^n(t), \quad t \geq 0,$$

where $\hat{R}^n(t)$ is the random sum

$$\hat{R}^n(t) \equiv n^{-1/2} \sum_{i=1}^{N^n(t)} Y_i^n, \quad t \geq 0,$$

while $\hat{R}_1^n(t)$ and $\hat{R}_2^n(t)$ are remainder terms involving the initial and final partial cycle, if any, also scaled by dividing by $\sqrt{n}$.

Just as in the standard regenerative setting, we are able to show that $\hat{C}_f^n$ is asymptotically equivalent to $\hat{R}^n$, so that it suffices to work with $\hat{R}^n$.

Lemma 13 (*reduction to random sums*) *As $n \to \infty$, $\hat{R}_1^n \Rightarrow 0e$ and $\hat{R}_2^n \Rightarrow 0e$, so that $d_{J_1}(\hat{R}^n, \hat{C}_f^n) \Rightarrow 0$.*

It now suffices to show that $\hat{R}^n(\cdot;\eta) \Rightarrow \hat{C}(\cdot;\eta)$ as $n \to \infty$ for each positive recurrent η in $\mathcal{D}$. By virtue of Corollary 3, given such an η, we can find a sequence of piecewise-constant state functions $\{\eta_{pc}^m : m \geq 1\}$ where $\|\eta_{pc}^m - \eta\|_T \to 0$ as $m \to \infty$ with η_{pc}^m being positive recurrent for all sufficiently large m. For those m, we have the desired convergence $\hat{C}^n(\cdot;\eta_{pc}^m) \Rightarrow \hat{C}(\cdot;\eta_{pc}^m)$ as $n \to \infty$, as observed in the beginning of the proof. Thus, by Lemma 12 and 13 above, we also have $\hat{R}^n(\cdot;\eta_{pc}^m) \Rightarrow \hat{C}(\cdot;\eta_{pc}^m)$ as $n \to \infty$ for these m as well. We now want to show that the established convergence also holds when η_{pc}^m is replaced by η. For that purpose, we need to establish convergence as $n \to \infty$ and $m \to \infty$ jointly. In order to justify that joint convergence, we establish the following result.

Lemma 14 (*tightness and bounds for the random sums*) *Consider a parameter state function η in $\mathcal{D}$ and a piecewise-constant parameter state function η_{pc}, where both η and η_{pc} are positive recurrent. Let $T > 0$ and $\delta > 0$ be such that $\|\eta - \eta_{pc}\|_T < \delta$. Then the sequence $\{\hat{R}^n(\cdot,\eta)\}$ is C-tight in $\mathcal{D}([0,T^*])$ for some constant $T^* > 0$ and there exist functions $\sigma_l(\eta_{pc}(\cdot),\delta)$ and $\sigma_u(\eta_{pc}(\cdot),\delta)$ such that the limit, say $\hat{R}(\cdot,\eta)$, of any convergent subsequence of $\{\hat{R}^n(\cdot,\eta)\}$ can be represented as*

$$\hat{R}(t,\eta) = B(\bar{W}(t),\eta), \quad 0 \leq t \leq T^*, \tag{52}$$

where B is standard BM and $\bar{W}$ can be bounded above and below by

$$\int_{t_1}^{t_2} \sigma_l^2(\eta_{pc}(s),\delta)\, ds \leq \bar{W}(t_2,\eta) - \bar{W}(t_1,\eta) \leq \int_{t_1}^{t_2} \sigma_u^2(\eta_{pc}(s),\delta)\, ds \tag{53}$$

for all t_1 and t_2 with $0 \leq t_1 < t_2 \leq T^*$, where $0 \leq \sigma_l^2(\eta_{pc}(s),\delta) \leq \sigma_u^2(\eta_{pc}(s),\delta) < \infty$ for all s, $0 \leq s \leq T$, and having the form in (14) determined by the state $\eta_{pc}(s)$. Moreover, for any $\epsilon > 0$ and $T^* > 0$, there exist $\delta > 0$ and $T > 0$ as above, such that

$$\|\sigma_u^2(\eta_{pc}(\cdot),\delta) - \sigma_l^2(\eta_{pc}(\cdot),\delta)\|_{T^*} < \epsilon. \tag{54}$$

Lemma 14 is based on associated lemmas for partial sums from triangular arrays of the cycle lengths and cycle variables τ_i^n and Y_i^n, exploiting martingale structure; these results are stated in §8 and proved in §9. Given these lemmas, we now can complete the proof of Theorem 6. First, we have observed that $\hat{C}^n(\cdot, \eta_{pc}) \Rightarrow \hat{C}(\cdot, \eta_{pc})$ in $\mathcal{D}$ for any positive-recurrent piecewise-constant parameter state function η_{pc}. By Lemmas 12 and 13, $\hat{R}^n(\cdot, \eta_{pc}) \Rightarrow \hat{C}(\cdot, \eta_{pc})$ in $\mathcal{D}$ as well. We can then apply Lemma 14 to deduce that the sequence of random sums $\{\hat{R}^n(\cdot, \eta)\}$ is tight. Hence, each subsequence has a convergent subsequence. Let $\hat{R}(\cdot, \eta)$ be the limit of such a convergent subsequence. Next we construct a sequence $\{\eta_{pc}^m\}$ of positive-recurrent piecewise-constant state functions with $\|\eta_{pc}^m - \eta\|_T \to 0$ as $m \to \infty$. As shown above, for each of them, we have $\{\hat{R}^n(\cdot, \eta_{pc}^m)\} \Rightarrow \hat{C}(\cdot, \eta_{pc}^m)$ as $n \to \infty$. However, again by Lemma 14, we have $\hat{R}(\cdot, \eta)$ bounded above and below by the limits $\hat{C}(\cdot, \eta_{pc}^m)$ which converge to $\{\hat{C}(\cdot, \eta)\}$ as $m \to \infty$. Hence, we must have $\hat{R}(\cdot, \eta) = \{\hat{C}(\cdot, \eta)\}$. Hence all convergent subsequences must have the same limit, which implies that we must have the full convergence, $\hat{R}^n(\cdot, \eta) \Rightarrow \hat{C}(\cdot, \eta)$ in $\mathcal{D}$ as $n \to \infty$. By Lemmas 12 and 13, we must also have $\hat{C}^n(\cdot, \eta) \Rightarrow \hat{C}(\cdot, \eta)$ in $\mathcal{D}$. Hence, Theorem 6 is proved. □

8 Proof of Lemma 14: Using the Martingale FCLT

We have indicated that Lemma 14 is based on associated lemmas for partial sums from triangular arrays of the cycle lengths and cycle variables τ_i^n and Y_i^n, exploiting martingale structure; in particular, we apply the martingale FCLT for triangular arrays. We can treat these two components of $\hat{R}^n(\cdot, \eta)$ separately because, just as in the familiar setting of renewal reward processes discussed in §§7.4 and 13.2 of [32], the FCLT for $\hat{R}^n(\cdot, \eta)$ depends on a FCLT for partial sums of Y_i^n and a FWLLN for $N^n(t)$ separately. By the inverse relation discussed in §§7.3 and 13.6 of [32], a FWLLN for $N^n(t)$ is equivalent to a corresponding FWLLN for the partial sums of τ_i^n. Since we can reduce the case of piecewise-constant η_{pc} to the case of constant η_c by focusing on the subintervals separately, we now relate the given η to a constant η_c.

Consider the cycle variables Y_i^n in (50) associated with a parameter state function η. Let $\mathcal{F}_k^n$ be the σ-field generated by $X_6^n(t) : 0 \le t \le T_k^n$, $k \ge -1$. Let

$$M_Y^n(k) \equiv \sum_{i=1}^{k} Y_i^n, k \ge 1, \quad \text{and} \quad \hat{M}_Y^n(t) \equiv n^{-1/2} M_Y^n(\lfloor nt \rfloor), t \ge 0. \tag{55}$$

For $i \ge 0$, let

$$\begin{aligned} \sigma_{n,i}^2 &\equiv E[(Y_i^n)^2 | \mathcal{F}_{i-1}^n], \\ \bar{V}^n(t) &\equiv n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \sigma_{n,i}^2 \quad \text{and} \quad \mathcal{V}^n(t) \equiv \sup\{s : \bar{V}^n(s) \le t\}, \quad t \ge 0. \end{aligned} \tag{56}$$

We will be strongly exploiting the QBD continuity to obtain regularity in the variables Y_i^n.

Lemma 15 (*sums of cycle variables*) *Consider a parameter state function η in $\mathcal{D}$ and an associated constant parameter state function η_c, where $\|\eta - \eta_c\|_T < \delta$ for some $T > 0$ and $\delta > 0$, and both η and η_c are positive recurrent. Consider the cycle variables Y_i^n in* (50) *and the associated variables in* (55) *and* (56)*, all associated with η. Then there exist constants $\sigma_l^2(\eta_c, \delta)$, $\sigma_u^2(\eta_c, \delta)$ and $\delta' > 0$ such that, for all i and n,*

$$\sigma_l^2(\eta_c, \delta) \leq \sigma_{n,i}^2 \leq \sigma_u^2(\eta_c, \delta) \quad \textit{and} \quad \sigma_u^2(\eta_c, \delta) - \sigma_l^2(\eta_c, \delta) < \delta', \tag{57}$$

for $\sigma_{n,i}^2$ in (56)*, associated with η, so that*

$$\sigma_l^2(\eta_c, \delta)(t_2 - t_1) \leq \bar{V}^n(t_2) - \bar{V}^n(t_1) \leq \sigma_u^2(\eta_c, \delta)(t_2 - t_1) \tag{58}$$

for all $n \geq 1$ and $0 \leq t_1 < t_2 \leq T$, for $\bar{V}^n$ in (56)*. As a consequence,*

$$\frac{(t_2 - t_1)}{\sigma_u^2(\eta_c, \delta)} \leq \mathcal{V}^n(t_2) - \mathcal{V}^n(t_1) \leq \frac{(t_2 - t_1)}{\sigma_l^2(\eta_c, \delta)} \tag{59}$$

for all $n \geq 1$ and $0 \leq t_1 < t_2 \leq T'$ for $T' \equiv T/\sigma_u^2(\eta, \delta)$. Hence, the sequences $\{\bar{V}^n\}$ and $\{\mathcal{V}^n\}$ associated with η, defined in (56)*, are C-tight in $\mathcal{D}([0, T])$ and $\mathcal{D}([0, T'])$, respectively. Moreover, the limits of convergent subsequences, say $\bar{V}$ and $\mathcal{V}$ must satisfy corresponding inequalities, i.e.,*

$$\begin{aligned} \sigma_l^2(\eta_c, \delta)(t_2 - t_1) \leq \bar{V}(t_2) - \bar{V}(t_1) \leq \sigma_u^2(\eta_c, \delta)(t_2 - t_1) \quad \textit{and} \\ \frac{(t_2 - t_1)}{\sigma_u^2(\eta_c, \delta)} \leq \mathcal{V}(t_2) - \mathcal{V}(t_1) \leq \frac{(t_2 - t_1)}{\sigma_l^2(\eta_c, \delta)} \end{aligned} \tag{60}$$

for the same ranges of t_1 and t_2 above, so that $\bar{V}$ and $\mathcal{V}$ are both continuous and strictly increasing. In addition,

$$\hat{M}_Y^n \circ \mathcal{V}^n \Rightarrow B \quad \textit{in} \quad \mathcal{D}([0, T']) \tag{61}$$

for $\hat{M}_Y^n$ in (55)*, where B is standard Brownian motion. Thus, the sequence $\{\hat{M}_Y^n\}$ is C-tight in $\mathcal{D}([0, T])$ with the limit of any convergent subsequence, say $\hat{M}_Y$, being of the form*

$$\hat{M}_Y(t) = B(\bar{V}(t)), \quad 0 \leq t \leq T, \tag{62}$$

where $\bar{V}$ is bounded above and below over all subintervals as in (60)*. If we are free to choose the bounding constant δ above, then for any $\epsilon > 0$, we can find $\delta > 0$ so that $\delta' < \epsilon$ for δ' in* (57)*.*

We now state the corresponding result for the partial sums of the cycle lengths.

Lemma 16 (*sums of cycle lengths*) *Consider a parameter state function η in $\mathcal{D}$ and a constant state function η_c, where both η and η_c are positive recurrent. Let $T > 0$ and $\delta > 0$ be such that $\|\eta - \eta_c\|_T < \delta$. Consider the cycle lengths τ_i^n in (49) associated with η. Let $U_k^n \equiv \tau_1^n + \cdots + \tau_k^n$, $k \geq 1$, and $\bar{U}^n(t) \equiv n^{-1} U_{\lfloor nt \rfloor}^n$, $t \geq 0$. Let $M_{U,i}^n \equiv E[\tau_i^n | \mathcal{F}_{i-1}^n]$, $\bar{M}_U^n(t) \equiv n^{-1}(M_{U,1}^n + \cdots + M_{U,\lfloor nt \rfloor}^n)$. Then the sequence $\{\bar{U}^n\}$ is C-tight in $\mathcal{D}([0,T''])$ for an appropriate time $T'' > 0$, and if $\bar{U}$ is the limit of a convergent subsequence, then necessarily it is bounded above and below with probability 1 by linear functions, i.e.,*

$$P(m_l(\eta_c, \delta)t \leq \bar{U}(t) \leq m_u(\eta_c, \delta)t, \quad 0 \leq t \leq T'') = 1.$$

where $m_l(\eta_c, \delta)$ and $m_u(\eta_c, \delta)$ are constants depending on δ such that $0 < m_l(\eta_c, \delta) \leq m_u(\eta_c, \delta) < \infty$. If we are free to choose the time $T > 0$ and the bounding constant δ above, then for any $\epsilon > 0$ and T'', $0 < T'' < \infty$, we can find $\delta > 0$ so that the conclusions above hold with $m_u(\eta_c, \delta) - m_l(\eta_c, \delta) < \epsilon$.

As a consequence of the inverse relation between the partial sums and the associated counting processes, as in Chapter 13 of [32], we obtain the following corollary for the counting processes associated with the partial sums. Let $\bar{N}^n(t) \equiv n^{-1} N^n(nt)$, $t \geq 0$. In the next section we combine Corollary 4 below with Lemma 15 to prove Lemma 14.

Corollary 4 (*counting process for cycle lengths*) *Under the assumptions of Lemma 16, the sequence of scaled counting processes $\{\bar{N}^n\}$ is C-tight in $\mathcal{D}([0,T'''])$ for any time $T''' < T''/m_l(\delta)$, where T'' is as in Lemma 16. If $\bar{N}$ is the limit of a convergent subsequence of $\{\bar{N}^n\}$, then necessarily it is bounded above and below with probability 1 by linear functions, i.e.,*

$$P(t/m_u(\eta_c, \delta) \leq \bar{N}(t) \leq t/m_l(\eta_c, \delta), \quad 0 \leq t \leq T''') = 1. \tag{63}$$

where $m_l(\eta_c, \delta)$ and $m_u(\eta_c, \delta)$ are the constants depending on δ from Lemma 16 above. If we are free to choose the time $T > 0$ and the bounding constant δ above, then for any $\epsilon > 0$ and T''', we can find $\delta > 0$ so that the conclusions above hold with $m_u(\eta_c, \delta) - m_l(\eta_c, \delta) < \epsilon$.

9 Remaining Proofs of Lemmas in §§7 and 8

In this section we prove five lemmas in the previous two sections, which were used in the proof of Theorem 6. We prove them in the order needed for the proof. We prove the one remaining lemma, Lemma 12 justifying the approximation by the frozen process $\hat{C}_f^n$, afterwards in §10.

Proof of Lemma 15 The key observation is that the sequence of random vectors $\{(\tau_j^n, Y_j^n) : j \geq i\}$ associated with the general parametric state function η is conditionally independent of the entire system history up to time T_{i-1}^n for

each i, which we have denoted by $\mathcal{F}^n_{i-1}$, given only T^n_{i-1}. As a consequence, paralleling the regenerative case in [9] and (14),

$$E\left[\int_{T^n_{i-1}}^{T^n_i} \left(1_{\{D(\eta_i,s)>0\}}\right)\, ds|\mathcal{F}^n_{i-1}\right] = \pi_{1,2}(\eta_i)E[\tau^n_i|\mathcal{F}^n_{i-1}]$$

for $i \geq 1$, where $\eta_i \equiv \eta(T^n_{i-1})$, so that $E[Y^n_i|\mathcal{F}^n_{i-1}] = 0$ for each i. Hence, the stochastic process $\{M^n_Y(k) : k \geq 1\}$ is a square integrable martingale with respect to the filtration $\{\mathcal{F}^n_k : k \geq 1\}$.

Moreover, by the QBD continuity, the variances $\sigma^2_{n,i} \equiv E[(Y^n_i)^2|\mathcal{F}^n_{i-1}]$ in (56) cannot differ too much from the corresponding variance for the constant parameter function η_c. For a fixed $t \geq 0$, let $\sigma^2_Y(\eta(t))$ be $\sigma^2_{n,i}$ under the condition that $T^n_{i-1} = t$, so that $\eta_i \equiv \eta(T^n_{i-1}) = \eta(t)$. Since $\|\eta - \eta_c\|_T < \delta$, we can apply the QBD continuity to obtain the relations in (57), where

$$\begin{aligned}
\sigma^2_l(\eta_c,\delta) &\equiv \min\{\sigma^2_Y(\eta(t)) : \eta \in A(\eta_c,\delta)\} \quad \text{and} \\
\sigma^2_u(\eta_c,\delta) &\equiv \max\{\sigma^2_Y(\eta(t)) : \eta \in A(\eta_c,\delta)\} \quad \text{with} \\
A(\eta_c,\delta) &\equiv \{\eta : \|\eta - \eta_c\|_T \leq \delta\}.
\end{aligned} \tag{64}$$

These in turn imply that the inequalities in (58) and (59) hold for $\bar{V}^n$ and $\mathcal{V}^n$ for all n, implying the tightness of the sequences $\{\bar{V}^n\}$ and $\{\mathcal{V}^n\}$ and the inequalities stated in (60) for the limits of all convergent subsequences. However, we cannot conclude that in general either $\bar{V}^n$ or $\mathcal{V}^n$ converges.

Nevertheless, we can apply an appropriate martingale FCLT to deduce that the limit in (61) holds; e.g., see Theorems 2.1 and 2.2 of [5], Theorem 5 of [26] and p. 98 of [14]. The QBD continuity and the bounds in (57) imply that the technical regularity conditions are satisfied in this case. Hence, for any $\delta > 0$, we can apply the martingale FCLT to get the convergence in (61).

Given that $\mathcal{V}$ is a strictly increasing continuous function with bounded slope, as in (60), we can deduce from the tightness of $\{\hat{M}^n_Y \circ \mathcal{V}^n\}$, which follows from the convergence in (61), that the sequence $\{\hat{M}^n_Y\}$ itself must be tight. That is most easily done by letting $\mathcal{V}^n$ be a continuous function constructed by linear interpolation under which we still have the convergence in (61). Then, $\bar{V}^n$ itself is a continuous strictly increasing function with modulus bounds in (60). Hence, we can deduce that the sequence $\{\hat{M}^n_Y\}$ must be tight.

The sequence $\{(\hat{M}^n_Y, \mathcal{V}^n, \bar{V}^n)\}$ is tight because the component sequences are all tight. Starting from the joint convergence $(\hat{M}^n_Y, \mathcal{V}^n, \bar{V}^n) \Rightarrow (\hat{M}_Y, \mathcal{V}, \bar{V})$ in $\mathcal{D}_3$ for any convergent subsequence, we can deduce from (61) that $\hat{M}_Y = B \circ \bar{V}$, as claimed in (62). The final ϵ bound follows from the QBD continuity in Lemma C.5 of [24]. □

Proof of Lemma 16 The proof is similar to the proof of Lemma 15 above, but now we need a FWLLN instead of a FCLT. However, it is convenient to apply the FCLT in order to deduce the FWLLN. Indeed, by the same reasoning used to prove Lemma 15 above, we can obtain a martingale FCLT for the

sums of the centered variables $\tau_i^n - E[\tau_i^n|\mathcal{F}_{i-1}^n]$, paralleling (61). Here we use the conditional variances and their sums, defined by

$$\sigma_{n,i}^2 \equiv E[(\tau_i^n - E[\tau_i^n|\mathcal{F}_{i-1}^n])^2|\mathcal{F}_{i-1}^n], \quad \bar{V}^n(t) \equiv n^{-1}\sum_{i=1}^{\lfloor nt \rfloor}\sigma_{n,i}^2.$$

instead of (56). We then obtain analogs of (57), (58) and (60).

Given that FCLT, we scale further, essentially dividing by $\sqrt{n}$, to get the associated FWLLN for the centered variables. As a consequence, we obtain the FWLLN $\bar{U}^n - \bar{M}_U^n \Rightarrow 0e$ in $\mathcal{D}([0,T''])$ as $n \to \infty$, for an appropriate finite time T'', not necessarily equal to T or T' in the previous proof above. Then, in direct analogy with (64), we apply the QBD continuity to obtain $m_l(\eta_c,\delta) \le M_{U,i}^n \le m_u(\eta_c,\delta)$ for all i and n. Hence, $m_l(\eta_c,\delta)t \le \bar{M}_U^n(t) \le m_u(\eta_c,\delta)t$ for all n and t, $0 \le t \le T$. We can then combine these bounds with the FWLLN to obtain the conclusions stated in the lemma. By the QBD continuity, $m_u(\eta_c,\delta) - m_l(\eta_c,\delta) \to 0$ as $\delta \downarrow 0$. □

Proof of Lemma 14 First, Lemmas 15 and 16 and Corollary 4 can be extended directly to piecewise-constant state functions as well as constant state functions. Thus, for η in $\mathcal{D}$, they imply that the sequences $\{\hat{M}_Y^n\}$ and $\{\bar{N}^n\}$ are each C-tight in $\mathcal{D}$. Consequently, the associated sequence of vector processes $\{(\hat{M}_Y^n, \bar{N}^n)\}$ is C-tight in $\mathcal{D}_2$. Hence, every subsequence has a further convergent subsequence. Moreover, by Lemma 15 and Corollary 4, any limit, say $(\hat{M}_Y, \bar{N})$, can be represented as $(B \circ \bar{V}, \bar{N})$, where $\bar{V}$ and $\bar{N}$ are bounded as in (60) and (63) over each subinterval where the piecewise-constant parametric state function is constant. Hence, overall they can be bounded above and below by

$$(\bar{V}_{Y,l}, \bar{N}_l) \le (\bar{V}_Y, \bar{N}) \le (\bar{V}_{Y,u}, \bar{N}_u),$$

where $(\bar{V}_{Y,l}(0), \bar{N}_l(0)) = (\bar{V}_{Y,u}(0), \bar{N}_u(0)) = (0,0)$ and

$$(\bar{V}_{Y,l}(t), \bar{N}_l(t)) \equiv (\bar{V}_{Y,l}(t_{i-1}) + \sigma_{l,i}^2(t - t_{i-1}), \bar{N}_l(t_{i-1}) + (1/m_{u,i})(t - t_{i-1})),$$
$$(\bar{V}_{Y,u}(t), \bar{N}_u(t)) \equiv (\bar{V}_{Y,u}(t_{i-1}) + \sigma_{u,i}^2(t - t_{i-1}), \bar{N}_u(t_{i-1}) + (1/m_{l,i})(t - t_{i-1})),$$

for $t_{i-1} \le t < t_i$, where $0 \equiv t_0 < t_1 < \ldots < t_k \equiv T$, so that t_i are the endpoints of a piecewise constant state function η_{pc}, with $\sigma_{l,i}^2$ and $1/m_{u,i}$ being the lower bounds and $\sigma_{u,i}^2$ and $1/m_{l,i}$ being the upper bounds on the i^{th} subinterval, depending on η_{pc} and δ. Hence, we can apply the continuous mapping theorem to obtain the corresponding convergence for the random sum for all convergent subsequences, with the limit of all convergent subsequences represented as claimed in (52) with $\bar{W}$ there bounded as in (53). The bounding variance functions are given explicitly by $\sigma_u^2(\eta_{pc}(s),\delta) = \sigma_{u,i}^2/m_{l,i}$ and $\sigma_l^2(\eta_{pc}(s),\delta) = \sigma_{l,i}^2/m_{u,i}$ for $t_{i-1} \le s < t_i$. Thus, by having $\|\eta - \eta_{pc}\|_T < \delta$ and choosing δ sufficiently small, we can obtain the desired variance inequality (54). □

Proof of Lemma 13 The reasoning follows the regenerative case as in [9]. First, the remainder term $\hat{R}_1^n(t)$ is relatively easy to treat since it involves the initial cycle and is thus independent of t. Since $D(\eta(0), 0)$ has been specified as some fixed state after (40), the initial partial cycle until hitting time of the designated state is clearly $O(1)$ and becomes asymptotically negligible when we divide by $\sqrt{n}$.

As in [9], to treat the second remainder term, we exploit the representation

$$\begin{aligned} |\hat{R}_2^n(t)| &\le n^{-1/2} W^n_{N^n(t)+1} \le n^{-1/2} \tau^n_{N^n(t)+1} \\ &\le n^{-1/2} \max\{\tau_i^n : 1 \le i \le N^n(t)+1\}, \quad t \ge 0, \end{aligned} \tag{65}$$

for W_i^n and τ_i^n defined in (51) and (49). However, the last term in (65) is asymptotically negligible because of the FCLT for the cycle lengths used in the proof of Lemma 16 above. The last term is the maximum discontinuity in the prelimit process indexed by n. Since the limit is continuous, that term is asymptotically negligible. □

10 Proof of Lemmas 8 and 12: Coupling Constructions

In this section we prove the two lemmas justifying approximation by frozen processes, using coupling constructions.

Proof of Lemma 8 By the construction in (42), we have forced the new frozen processes $\tilde{D}_f^n$ to coincide with the queue-difference processes $D_{1,2}^n$ for all time points t of the form k/n. To complete the proof, we employ a special coupling construction to construct these two processes on the same underlying probability space to make the processes have the same transitions within each interval $[(k-1)/n, k/n)$ with high probability. As usual [19] [29], this coupling construction produces an artificial joint distribution, but leaves the distributions of each of the two processes individually unchanged.

We start by focusing on a single interval $[(k-1)/n, k/n)$. It suffices to focus on any one interval, because we will show that the construction is uniform over the n intervals. Since the transition rates in system n are of order $O(n)$ and the interval is of length $1/n$, it is convenient to start by rescaling time as in the fluid limit in Theorem 1. By doing a change of variables, we have

$$\begin{aligned} &\sqrt{n} \int_{(k-1)/n}^{k/n} \left(1_{\{D_{1,2}^n(s)>0\}} - 1_{\{\tilde{D}_f^n(s)>0\}}\right) ds, \\ &\quad = \frac{1}{\sqrt{n}} \int_0^1 \left(1_{\{D_{1,2}^n((k-1)/n+s/n)>0\}} - 1_{\{\tilde{D}_f^n((k-1)/n+s/n)>0\}}\right) ds. \end{aligned}$$

Then recall that both processes inside the integral converge appropriately to the FTSP. To expose the connection, let k go to infinity with n so that $k/n \to t$ as $n \to \infty$. First, by Theorem 1, $\bar{X}^n((k-1)/n) \Rightarrow x_6(t)$. Then, by Theorem 4.4 of [24],

$$D_{1,2}^n((k-1)/n + s/n) \equiv D_e^n(X^n((k-1)/n), s) \Rightarrow D(x_6(t), s).$$

Second, by (41),

$$\begin{aligned}&\{\tilde{D}_f^n((k-1)/n+s/n):0\le s\le 1\}\\&\quad\stackrel{\mathrm{d}}{=}\{D(\lambda_i^n/n,m_j^n/n,X^n((k-1)/n),s):0\le s\le 1\}\\&\quad\Rightarrow\{D(x_6(t),s):0\le s\le 1\}.\end{aligned}$$

The main point for the coupling is that, after the change of time scale, both processes have transition rates of order $O(1)$ that differ by $O(1/n)$. Moreover, the processes are identical w.p.1 at the left end point of the interval $[0,1]$.

However, we need to apply the argument above to all n intervals, where $n\to\infty$. It is thus important that the conclusions are valid uniformly over the n subintervals. Those conclusions are justified because the fluid limit in Theorem 1 implies that $\bar{X}_6^n\Rightarrow x_6$ uniformly over each finite interval. Moreover, the limit x_6 is a continuous function over a bounded interval with values in in a compact subset of $\mathbb{A}$. Finally, the limiting transition rates are a continuous function of the state.

Let $\nu^n(T)$ be the number of k for which the $nk\le T$ and the sample paths of $\tilde{D}_f^n$ and $D_{1,2}^n$ fail to be identical over the interval $[(k-1)/n,k/n)$. As a consequence of the asymptotically equivalent transition rates after changing the time scale above, we show below that $\nu^n(T)=O(1)$ as $n\to\infty$. Thus, to complete the proof, we use the elementary bound $\|\Delta^n\|_T\le\nu^n(T+\epsilon)/\sqrt{n}$ for all $n\ge 1/\epsilon$, where $T>0$ and $\epsilon>0$ are arbitrary constants.

We now discuss the coupling in more detail. Since the transitions in the queue-difference process $D_{1,2}^n$ are generated from state changes in the CTMC X_6^n, we do the special construction from the perspective of the CTMC X_6^n. We use the device of uniformization to generate the transitions of the CTMC; i.e., we construct the transitions by thinning a Poisson process. Without loss of generality, we use different independent Poisson processes to generate potential transitions for each kind of transition, each interval $[(k-1)/n,k/n)$ and each n. Since the transition rate of the CTMC is not uniformly bounded, there is a possibility that this direct construction will be invalid, but by choosing these Poisson process rates sufficiently high, we can make the likelihood of a violation asymptotically negligible. In the actual construction, we can change the Poisson process when the constructed process hits a state from which a further transition could lead to a violation. The detailed construction does not matter because we declare a difference occurring throughout the entire subinterval if the Poisson rate needs to be adjusted, thus contributing the maximum possible to the bound above. Since the integrand in (45) is bounded by 1, the total impact upon (45) by such rate violations can clearly be made asymptotically negligible.

The coupling is achieved by using the same Poisson processes to generate the transitions in both $D_{1,2}^n$ and $\tilde{D}_f^n$ over each subinterval $[(k-1)/n,k/n)$. These are done with respect to the states of $X_6^n(t)$ and $X_6^n((k-1)/n)$. For $D_{1,2}^n$, the transitions rates of the various transitions (arrivals, abandonments from each queue and service completions of each class from each pool) are determined by the actual state $X_6^n(t)$, which changes throughout the interval

$[(k-1)/n, k/n)$. For, $\tilde{D}_f^n$, we do the same construction, but we leave the state fixed at its initial value $X_6^n((k-1)/n)$ throughout the interval $[(k-1)/n, k/n)$, so that the transition rates do not change. However, we match the transitions in the two systems as much as possible. We make the transitions differ only to the extent that the state of $X_6^n(t)$ differs from $X_6^n((k-1)/n)$.

As stated above, we use different independent Poisson processes for each kind of transition. We have one Poisson process generate potential arrivals for each n. Since the arrival rates are unaffected by the state, the Poisson process for generating potential arrivals of class i can have rate λ_i^n, so that every potential arrival corresponds to an actual arrival in both systems. Thus no difference is caused by any arrival. That arrival in turn affects the constructed processes $D_{1,2}^n$ and $\tilde{D}_f^n$ in the obvious way: an arrival of class 1 increases them by 1, while an arrival of class 2 decreases them by r.

For service completions of class 1 by pool 2, we let the Poisson process generating potential transitions have rate $\mu_{1,2}m_2^n$. The actual transition rate at time t for $X_6^n(t)$ is $\mu_{1,2}Z_{1,2}^n(t)$, so that the Poisson rate is an upper bound on the actual transition rate for all states. If the Poisson process with rate $\mu_{1,2}m_2^n$ has a transition at time t, where $(k-1)/n \le t < k/n$, then we let both systems have an actual service completion of class 1 by pool 2 at time t with probability $(Z_{1,2}^n(t) \wedge Z_{1,2}^n((k-1)/n)/m_2^n$; we let only the system associated with $D_{1,2}^n$ have an actual service completion of class 1 by pool 2 at time t with probability $[Z_{1,2}^n(t) - (Z_{1,2}^n(t) \wedge Z_{1,2}^n((k-1)/n)]/m_2^n$; we let only the system associated with D_f^n have an actual service completion of class 1 by pool 2 at time t with probability $[Z_{1,2}^n((k-1)/n) - (Z_{1,2}^n(t) \wedge Z_{1,2}^n((k-1)/n)]/m_2^n$; and we let neither system have an actual service completion of class 1 by pool 2 with probability $[m_2^n - (Z_{1,2}^n(t) \vee Z_{1,2}^n((k-1))/n)]/m_2^n$. Thus, a difference in the sample path is caused by this transition with probability $[(Z_{1,2}^n(t) \vee Z_{1,2}^n((k-1)/n) - (Z_{1,2}^n(t) \wedge Z_{1,2}^n((k-1)/n)]/m_2^n$, which clearly is of order $O(1/n)$.

We do similar constructions with independent Poisson processes for each of the other transitions. The abandonments are where the transition rate is unbounded, because the queue lengths $Q_i^n(t)$ are unbounded above. However, the maximum queue length over the interval is bounded above by the initial queue length plus the number of arrivals over the interval, so that the probability of violation is easily controlled by the Poisson arrival process for that class. Hence, for the Poisson process generating potential abandonments from the class-i queue over the interval $[(k-1)/n, k/n)$, we can give it rate $(Q_i^n((k-1)/n) + cn^3)\theta_i$ for $c > \lambda_i$. (The exponent 3 is chosen to make careful calculations unnecessary.) This is sufficient, because the initial number in queue i is $Q_i^n((k-1)/n)$ and new class-i arrivals occur at rate λ_i^n, which is $O(n)$. The higher power of n ensures that a violation of the rate-order uniformization condition is asymptotically negligible as $n \to \infty$. If the Poisson process generates a potential abandonment at time t, then it is a real abandonment for at least one system with probability $a_n/c_n = O(1/n^2)$, a real abandonment for both systems with probability $b_n/c_n = O(1/n^2)$ and a real abandonment for only one of the two systems with probability $(a_n - b_n)/c_n = O(1/n^3)$, where $a_n \equiv Q_i^n((k-1)/n) \vee Q_i^n(t)$, $b_n \equiv Q_i^n((k-1)/n) \wedge Q_i^n(t)$ and

$c_n \equiv Q_i^n((k-1)/n)+cn^3$. The main point is that $(a_n - b_n) = O(1)$ because the two queues differ by arrivals at rate $O(n)$ over the interval of length $1/n$. Hence, the probability that a **real transition** at t (not counting transitions from a state to itself, which are generated by the common Poisson process) produces an abandonment for only one of the two systems is $(a_n - b_n)/a_n = O(1/n)$. At the same time, the probability that the uniformization condition is violated during the entire interval is $o(1/n)$, so that it is asymptotically negligible in the relevant scale.

We now assess the impact of this construction. Both processes have transition rates of order $O(n)$ because the relevant processes Q_i^n and $Z_{i,j}^n$ in X_6^n are $O(n)$. Thus, the processes $D_{1,2}^n$ and $\tilde{D}_f^n$ have $O(1)$ transitions over each interval of length $1/n$. Hence, the state of $X^n(t)$ will only change an amount of order $O(1)$ within each interval $[(k-1)/n, k/n)$. Consequently, the probability of any one transition being different is $O(1/n)$, and the probability that there is any difference over the interval $[(k-1)/n, k/n)$ is also of $O(1/n)$. Hence, $\nu^n(T)$ – the total number of intervals having any difference over the interval $[0, T]$ – will be of order $O(1)$, as claimed at the beginning of the proof.

Elaborating on the last step, observe that conditional upon $\bar{X}_6^n$, which converges to x_6, we can regard $\nu^n(T)$ as the sum of at most $\lfloor nT \rfloor + 1$ independent Bernoulli random variable, assuming the value 1 with probability $p_{n,i}$ and 0 otherwise, where $p_l/n \leq p_{n,i} \leq p_u/n$ for all $i = 1, \ldots, \lfloor nT \rfloor + 1$, provided that n is suitably large, where p_l/n and p_u/n are the minimum and maximum "success probabilities" among those Bernoulli random variables. The bounds hold because $t \mapsto x_6(t)$ is a continuous function that is considered over a compact interval. Hence, all the transition rates described above, producing the probabilities $p_{n,i}$ over each interval i, also have continuous limits which can be bounded uniformly for all n large enough. Using the upper bound, we can bound $\nu^n(T)$ above stochastically by $\nu_u^n(T)$, defined as the partial sum of i.i.d. Bernoulli random variables. taking the value 1 with probability p_u/n. By the LLN for partial sums from triangular arrays $\nu_u^n(T) \Rightarrow p_u T$ as $n \to \infty$, which implies that $\nu^n(T)$ is indeed properly $O(1)$ as $n \to \infty$. Hence the proof is complete. □

Proof of Lemma 12 The reasoning here is similar to the proof of Lemma 8. As before, we can use a coupling construction to make the two processes have identical sample paths over the vast majority of the cycles. We exploit the oscillation property for functions in $\mathcal{D}([0, T])$, Corollary 12.2 of [32], concluding that, for any $\epsilon > 0$, there are k time points t_i with $0 \equiv t_0 < t_1 < \cdots < t_{k-1} < t_k \equiv T$ such that $|\eta(s_1) - \eta(s_2)| < \epsilon$ for all $s_1, s_2 \in [t_{i-1}, t_i)$ for all i. Hence, with the time scaling by $1/n$ in (40), we see that, except for at most k cycles in $[0, T]$ containing the k boundary points t_i, the oscillation of η over the cycle is at most ϵ/n. Hence, the coupling can be performed as in the proof of Lemma 8, making the probability that the sample paths differ over any one cycle among all except the k be of order $O(1/n)$. Since there are $O(n)$ cycles in $[0, T]$, as substantiated by Corollary 4, there are order $O(1)$ among the $O(n)$ cycles that

have any difference in the sample paths. Hence, with the spatial scaling by $\sqrt{n}$, we clearly have $d_{J_1}(\hat{C}_f^n, \hat{C}^n) \Rightarrow 0$ as $n \to \infty$ as claimed. □

11 Comparisons with Simulation

To both support the validity of the theorems and their applicability to the intended engineering problems, we now compare the approximations stemming from the FWLLN and the FCLT to the results of simulation experiments. Specifically, we will compare the Gaussian approximations for the steady-state queue lengths with simulation estimates of these quantities, obtained by simulating the actual queueing model over a large time interval. The approximate mean values come directly from the stationary point of the fluid limit, x^* in Theorem 2; the approximate variances come from Corollary 1, specifically, from (24).

Our simulation examples will have parameters related to a *base case*. First, scale is described by the parameter n, which is the scaling parameter in our limit theorems. The abandonment and service rate parameters, which describe the behavior of individual customers and servers, are independent of n: $\theta_1 = \theta_2 = 0.2$, $\mu_{1,1} = \mu_{2,2} = 1.0$ and $\mu_{1,2} = \mu_{2,1} = 0.8$. The service rates are chosen so that it is less efficient to serve a customer from a different class.

The parameters that scale as the service system grows depend on n; they are chosen to be directly proportional to n: $m_i^{(n)} \equiv nm_i$, $\lambda_i^{(n)} \equiv n\lambda_i$ and $k_{1,2}^{(n)} \equiv nk_{1,2}$. We take $k_{1,2}^n$ to be order $O(n)$ so it is easy to compare different system sizes. Our base case then has $m_1 \equiv m_2 \equiv 1$, $\lambda_1 = 1.3$, $\lambda_2 = 0.9$ and $k_{1,2} = 0.1$. The arrival rates are chosen to put class 1 in a focused overload, while class 2 is initially normally loaded or slightly underloaded, but becomes overloaded too after the sharing. (These model parameters satisfy case 1 of Assumption 3.1 of [24].) We use the FQR-T control with ratio parameter $r = 1.0$, which allows us to apply the simple asymptotic formulas from §4.4.

From (10) and (11), we see that the stationary fluid solution for this base case yields $z_{1,2}^* = 0.2111$, $q_1^* = 0.6556$, $q_2^* = 0.5556$ and $\pi_{1,2}^* \equiv \pi_{1,2}(x^*) = 0.1763$. Without any sharing, the fluid approximation for queue 1 would be 1.5000. Hence the sharing reduces the first fluid queue from 1.5000 to 0.6556, at the expense of causing the second class to have a fluid queue of 0.5556.

We now turn to the variances, for which we need to analyze the FTSP more carefully. The FTSP has BD parameters: $\lambda_1(x^*) = 1.411$, $\mu_1(x^*) = 2.989$, $\lambda_2(x^*) = 2.031$ and $\mu_2(x^*) = 2.369$. The associated $M/M/1$ traffic intensities are $\rho_1(x^*) = 0.472$ and $\rho_2(x^*) = 0.8574$. The associated mean busy periods are $E[T_1(x^*)] = 0.6338$ and $E[T_2(x^*)] = 2.9603$. Hence, the alternative formula for $\pi_{1,2}(x^*)$ in (27) agrees with the value 0.1763 given above (providing a check on our calculations).

Turning to the FCLT, from (20), we see that $\psi(x^*) = 0.6200$, so that $\psi^2(x^*) = 0.3844$. For $\sigma^2(x^*)$, from (26), we see that $E[T_1(x^*)^2] = 1.5218$, so that $Var(T_1(x^*)) = 1.1201$, and $\sigma^2(x^*) = 1.1201/3.5941 = 0.3116$. Then

$\xi_2 \equiv \psi^2(x^*)\sigma^2(x^*) = 0.1198$. Since $|\mathcal{M}_{2,2}| = 0.176$, $\mathcal{Z}_2 = 0.3403$. Hence, $\sigma^2_{Z_{1,2}}(\infty) = 1 - 0.2111 + 0.3403 = 1.1292$.

As a consequence, $\sigma^2_{Q_s,Z_{1,2}}(\infty) = (1.1292)(0.5319) = 0.6006$. Since $\mu_{2,2} - \mu_{1,2} = p_1\theta_1 + p_2\theta_2 = 0.2$, $\mathcal{Q}_2 = \sigma^2_{Q_s,Z_{1,2}}(\infty) = 0.6006$. Since $\mathcal{Q}_1 = 11.0$, we have $\sigma^2_{Q_s}(\infty) = 11.6006$, so that the associated standard deviation is 3.41. (Without $\mathcal{Q}_2$, we would approximate the standard deviation by $\sqrt{11} = 3.32$, so $\mathcal{Q}_2$ contributes only 3% to the standard deviation approximation in this case.)

By the SSC, the diffusion approximations for Q_1 and Q_2 are linearly related to Q_s; in particular, $\sigma^2_{Q_i}(\infty) = (p_i)^2\sigma^2_{Q_s}(\infty)$, so that $\sigma^2_{Q_i}(\infty) = 11.6006/4 = 2.900$ and the associated standard deviation is 1.70.

We now turn to the simulations. We simulate the actual queueing system obtained by scaling up the appropriate parameters by n. We consider three cases: $n = 25$, $n = 100$, and $n = 400$. (Since $k^n_{1,2}$ must be an integer, we let $k^n_{1,2} = 3$ when $n = 25$.)

In all our simulation experiments, we used 5 independent runs, each with $300,000$ arrivals. We report averages together with the half widths of the 95% confidence intervals, based on a t statistic with four degrees of freedom. Simulation results for the base case above are presented in Table 1 below.

The first four rows of Table 1 show mean values. We display both the steady-state mean values and the associated scaled values (i.e., divided by n). The unscaled values helps us evaluate the performance of the actual system, while the scaled values show the convergence in the FWLLN. Table 1 clearly shows that the accuracy improves as n gets larger, but even for relatively small systems, the fluid approximation gives reasonable results.

Rows $5 - 10$ of Table 1 show the standard-deviations of the total queue length $Q_s = Q_1 + Q_2$ as well as the two queues. As before, we treat both the actual values and the scaled values, but now we are scaling in diffusion scale (dividing by $\sqrt{n}$ after subtracting the order-$O(n)$ mean), as in (15), so that we will be substantiating the FCLT, specifically Corollary 1 and the variance formulas in (24). To save space, we omit the confidence intervals for the scaled standard deviations; these can be computed from the confidence intervals of the actual queues by dividing the half widths by $\sqrt{n}$.

Overall, we conclude that Table 1 shows that the approximations are remarkably accurate.

Acknowledgments

This research is part of the first author's doctoral dissertation in the IEOR Department at Columbia University. Additional work was done subsequently, including while the first author had a postdoctoral fellowship at CWI in Amsterdam. This research was partly supported by NSF grants DMI-0457095, CMMI 0948190 and CMMI 1066372.

	n=25		n=100		n=400	
perf. meas.	Approx.	Sim.	Approx.	Sim.	Approx.	Sim.
$E[Q_1]$	16.6	15.7 ±0.3	65.6	63.6 ±1.9	262.2	258.3 ±5.0
$E[Q_1/n]$	0.656	0.629 ±0.013	0.656	0.636 ±0.019	0.656	0.646 ±0.013
$E[Q_2]$	13.6	15.9 ±0.4	55.6	58.6 ±1.8	222.2	223.9 ±5.0
$E[Q_2/n]$	0.556	0.636 ±0.016	0.556	0.586 ±0.018	0.556	0.560 ±0.013
$std(Q_s)$	17.1	16.0 ±0.3	34.1	33.7 ±1.4	68.2	67.6 ±2.9
$std(\hat{Q}_s)$	3.41	3.21	3.41	3.37	3.41	3.38
$std(Q_1)$	8.5	8.8 ±0.1	17.0	17.2 ±0.7	34.0	33.9 ±1.4
$std(\hat{Q}_1)$	1.70	1.75	1.70	1.72	1.70	1.70
$std(Q_2)$	8.5	8.6 ±0.1	17.0	17.1 ±0.7	34.0	33.9 ±1.5
$std(\hat{Q}_2)$	1.70	1.73	1.70	1.71	1.70	1.69

Table 1 A comparison of approximations to simulation results for the means and standard deviations of the steady-state queue lengths as a function of the scale parameter n in the base case with $\lambda_1^n = 1.3n$, $\lambda_2^n = 0.9n$, $k_{1,2}^n = 0.1n$, $r = 1$ and other parameters defined above.

References

1. Abate, J., Whitt, W.: Limits and approximations for the busy-period distribution in single-server queues. *Prob. Engr. Inf. Sci.* **9** 581–602 (1995)
2. Arnold, L. : *Stochastic Differential Equations: Theory and Applications*, Wiley, New York (1974)
3. Asmussen, S.: *Applied probability and Queues*, second ed., Wiley, New York (2003)
4. Coffman, E. G., Puhalskii, A. A., Reiman, M. I.: Polling systems with zero switchover times: a heavy-traffic averaging principle. *Annals of Applied Probability* **5**, 681–719 (1995)
5. Durrett, R., Resnick, S. I.: Functional limit theorems for dependent random variables. *Annals of Probability* **6** (5), 829-846 (1978)
6. Fricker, C., Robert, P., Tibi, D.: A degenerate central limit theorem for single resource loss systems *Annals of Applied Probability* **13** (2), 561–575 (2003)
7. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review and research prospects. *Manuf. Serv. Oper. Mgmnt.* **5**, 79–141 (2003)
8. Garnet, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. *Manuf. Serv. Oper. Mgmnt.* **4** (3), 208–227 (2002)
9. Glynn, P. W., Whitt, W.: Limit theorems for cumulative processes. *Stochastic Processes and Their Applications* **47**, 299–314 (1993)
10. Gurvich, I., Whitt, W.: Scheduling flexible servers with convex delay costs in many-server service systems. *Manuf. Serv. Oper. Mgmnt.* **11**, 237–253 (2009a)
11. Gurvich, I., Whitt, W.: Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* **34**, 363–396 (2009b)
12. Gurvich, I., Whitt, W.: Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res. 58*, 316–328 (2010)
13. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** (3), 567–588 (1981)
14. Hall, P., Heyde, C. C.: *Martingale Limit Theory and its Applications.* Academic Press, New York (1980)

15. Hunt, P.J., Kurtz, T. G.: Large loss networks. *Stochastic Processes and their Applications* **53**, 363–378 (1994)
16. Karlin, S., Taylor H. M.: *A Second Course in Stochastic Processes.* Academic Press, New York (1981)
17. Karr, A.F.: Weak Convergence of a Sequence of Markov Chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **33**, 41–48 (1975)
18. Latouche, G., Ramaswami, V.: *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM and ASA, Philadelphia (1999)
19. Lindvall, T.: *Lectures on the Coupling Method*, Wiley, New York (1992)
20. Pang, G., Talreja, R., Whitt, W.: Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys.* **4**, 193–267 (2007)
21. Perry, O., Whitt, W.: Responding to unexpected overloads in large-scale service systems. *Management Sci.*, **55** (8), 1353–1367 (2009)
22. Perry, O., Whitt, W.: A fluid approximation for service systems responding to unexpected overloads. *Operations Res.*, vol. 59 (5), 1159-1170 (2011a)
23. Perry, O., Whitt, W.: An ODE for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems*, vol. 1 (1), 17–66 (2011b)
24. Perry, O., Whitt, W.: A fluid limit for an overloaded X call center model via an averaging principle. *Math. Oper. Res.*, forthcoming (2012) Available at: http://www.columbia.edu/∼ww2040/allpapers.html
25. Salminen, P., Norros, I.: On busy periods of the unbounded Brownian storage. *Queueing Systems* **39**, 317–333 (2001)
26. Rootzen, H.: On the functional central limit theorem for Martingales. *Zeit. Wahrscheinlichkeitsth. werv. Gebiete* **38**, 199-210 (1977)
27. Talreja, R., Whitt, W.: Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Prob.* **19** (6), 2137–2175 (2009)
28. Whitt, W.: Continuity of generalized semi-Markov processes. *Math. Oper. Res.* **5** (4), 494-501 (1980)
29. Whitt, W.: Comparing counting processes and queues. *Adv. Appl. Prob.* **13** (1), 207–220 (1981)
30. Whitt, W.: On the heavy-traffic limit theorem for GI/G/infinity queues. *Advances in Applied Probability* **14** (1), 171–190 (1982)
31. Whitt, W.: Asymptotic formulas for Markov processes with applications to simulation. *Oper. Res.* **40** (2), 279–291 (1992)
32. Whitt, W.: *Stochastic-Process Limits*, New York: Springer (2002)
33. Whitt, W.: Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50** (10), 1449–1461 (2004)