A Fluid Limit for an Overloaded X Model via a Stochastic Averaging Principle

Ohad Perry

Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208,
ohad.perry@northwestern.edu, http://users.iems.northwestern.edu/~perry/

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027,
ww2040@columbia.edu, http://www.columbia.edu/~ww2040/

We prove a many-server heavy-traffic fluid limit 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 in a recent paper 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. After the control is activated, it aims to keep the two queues at a prespecified fixed ratio. For large systems that fixed ratio is achieved approximately. For the fluid limit, or FWLLN (functional weak law of large numbers), we consider a sequence of properly scaled X models in overload operating under FQR-T. Our proof of the FWLLN follows the compactness approach, i.e., we show that the sequence of scaled processes is tight and then show that all converging subsequences have the specified limit. The characterization step is complicated because the queue-difference processes, which determine the customer-server assignments, need to be considered without spatial scaling. Asymptotically, these queue-difference processes operate on a faster time scale than the fluid-scaled processes. In the limit, because of a separation of time scales, the driving processes converge to a time-dependent steady state (or local average) of a time-varying fast-time-scale process (FTSP). This averaging principle allows us to replace the driving processes with the long-run average behavior of the FTSP.

Key words: many-server queues; averaging principle; separation of time scales; state-space collapse; heavy-traffic fluid limit; overload control
MSC2000 subject classification: Primary: 60F17, 60K25; secondary: 60G70, 90B22
OR/MS subject classification: Primary: queues; secondary: limit theorems, transient results
History: Received June 1, 2011; revised April 10, 2012. Published online in *Articles in Advance* December 20, 2012.

1. Introduction. In this paper we prove that the deterministic fluid approximation for the overloaded X call-center model, suggested in Perry and Whitt [36] and analyzed in Perry and Whitt [37], arises as the many-server heavy-traffic fluid limit of a properly scaled sequence of overloaded Markovian X models under the *fixed-queue-ratio-with-thresholds* (FQR-T) control. (A list of all the acronyms appears in §F in the appendix.) The X model has two classes of customers and two service pools, one for each class, but with both pools capable of handling customers from either class. The service-time distributions depend on both the class and the pool. The FQR-T control was suggested in Perry and Whitt [35] as a way to automatically initiate sharing (i.e., sending customers from one class to the other service pool) when the system encounters an unexpected overload, while ensuring that sharing does not take place when it is not needed.

1.1. A series of papers. This paper is the fourth in a series. First, in Perry and Whitt [35] we heuristically derived a stationary fluid approximation, whose purpose was to approximate the steady state of a large many-server X system operating under FQR-T during the overload incident. More specifically, in Perry and Whitt [35] we assumed that a convex holding cost is incurred on both queues whenever the system is overloaded, and our aim was to develop a control designed to minimize that cost. (That deterministic cost approximates the long-run average cost during the overload incident in the stochastic model.) We further assumed that the system becomes overloaded because of a sudden, unexpected shift in the arrival rates, with new levels that may not be known to the system managers, and that the staffing of the service pools cannot be changed quickly enough to respond to that sudden overload.

Under the heuristic stationary fluid approximation, in Perry and Whitt [35] we proved that a queue-ratio control with thresholds (QR-T) is optimal, and showed how to calculate the optimal control parameters. We also showed that the QR-T control outperforms the optimal static control when the arrival rates are known. In general, that optimal QR-T control is a function of the arrival rates during the overload incident, which are assumed to be unknown. In the special case of a separable quadratic cost, i.e., for $C(Q_1, Q_2) = c_1 Q_1^2 + c_2 Q_2^2$, with c_1, c_2 being two constants, we proved that the FQR-T control is optimal, so that two queue ratios—one for each direction of overload—are optimal for all possible overload scenarios. More generally, we found that

a FQR-T control was approximately optimal, as illustrated by Figure 4 of Perry and Whitt [35]. Thus, in all subsequent work we have focused on the FQR-T control.

Second, in Perry and Whitt [36] we applied a heavy-traffic stochastic *averaging principle* (AP) as an engineering principle to describe the transient (time-dependent) behavior of a large overloaded X system operating under FQR-T. The suggested fluid approximation was expressed via an *ordinary differential equation* (ODE), which is driven by a stochastic process. Specifically, the expression of the fluid ODE as a function of time involves the local steady state of a stochastic process at each time point $t \geq 0$, which we named the *fast-time-scale process* (FTSP). As the name suggests, the FTSP operates on (an infinitely) faster time scale than the processes approximated by the ODE, thus converges to its local steady state instantaneously at every time $t \geq 0$. Extensive simulation experiments showed that our approximations work remarkably well, even for surprisingly small systems, having as few as 25 servers in each pool.

Third, in Perry and Whitt [37] we investigated the ODE suggested in Perry and Whitt [36] using a dynamical-system approach. The dynamical-system framework could not be applied directly, because the ODE is driven by a stochastic process, and its state space depends on the distributional characteristics of the FTSP. Nevertheless, we showed that a unique solution to the ODE exists over the positive halfline $[0, \infty)$ for each specified initial condition. The stationary fluid approximation, derived heuristically in Perry and Whitt [35], was shown to exist as the unique fixed point (or stationary point) for the fluid approximation. Moreover, we proved that the solution to the ODE converges to this stationary point, with the convergence being exponentially fast. (That supports the steady-state approximation used in Perry and Whitt [35].) In addition, a numerical algorithm to solve the ODE was developed, based on a combination of a matrix-geometric algorithm and the classical forward Euler method for solving ODEs.

1.2. Overview. In this fourth paper, we will prove that the solution to the ODE in Perry and Whitt [36, 37] for specified initial condition is indeed the many-server heavy-traffic fluid limit of the overloaded X model, which we also call a *functional weak law of large numbers* (FWLLN); see Theorem 4.1; see §3.4 for the key assumptions. In doing so, we will prove a strong version of *state-space collapse* (SSC) for the server-assignment processes; see Corollary 6.2 and Theorem 6.4. We will also prove a strong SSC result for the two-dimensional queue process in Corollary 4.1. In a subsequent paper (Perry and Whitt [38]) we prove a *functional central limit theorem* (FCLT) refinement of the FWLLN here, which describes the stochastic fluctuations about the fluid path.

We only consider the X model *during* the overload incident, once sharing has begun; that will be captured by our main assumptions, Assumptions 3.1 and 3.2. As a consequence, the model is stationary (without time-varying arrival rates), but the evolution is transient, because the system does not start in steady state. Because of customer abandonment, the stochastic models will all be stable, approaching proper steady-state distributions. As a further consequence, during the overload incident sharing will occur in only one direction, so that the overloaded X model actually behaves as an overloaded N model, but that requires proof; that follows from Corollary 6.3 and Theorem 6.4. Our FWLLN serves as an approximation for the time-dependent behavior of the model as it approaches steady state. In addition, we prove a weak law of large numbers for the stationary distributions, showing that the unique fixed point of the fluid limit is also the limit of the scaled stationary model. Proving that latter result builds on a novel limit-interchange argument, which requires the established FWLLN.

Convergence to the fluid limit will be established in roughly three steps: (i) representing the sequence of systems (§§5.1 and 6), (ii) proving that the sequence considered is $\mathscr{C}$-tight (§5.2), and (iii) uniquely characterizing the limit (Perry and Whitt [37] and §7). The first representation step in §5.1 starts out in the usual way, involving rate-1 Poisson processes, as reviewed in Pang et al. [34]. However, the SSC part in §6 requires a delicate analysis of the unscaled sequence. The second tightness step in §5.2 is routine, but the final characterization step is challenging. These last two steps are part of the standard compactness approach to proving stochastic-process limits; see Billingsley [7], Ethier and Kurtz [12], Pang et al. [34] and §11.6 in Whitt [47]. As reviewed in Ethier and Kurtz [12] and Pang et al. [34], uniquely characterizing the limit is usually the most challenging part of the proof, but it is especially so here. Characterizing the limit is difficult because the FQR-T control is driven by a queue-difference process that is not being scaled and hence does not converge to a deterministic quantity with spatial scaling. However, the driving process operates in a different time scale than the fluid-scaled processes, asymptotically achieving a (time-dependent) steady state at each instant of time, yielding the AP.

As was shown in Perry and Whitt [37], the AP and the FTSP also complicate the analysis of the limiting ODE. First, it requires that the steady state of a *continuous-time Markov chain* (CTMC), whose distribution depends on the solution to the ODE, be computed at every instant of time. (As explained in Perry and Whitt [37], this argument may seem circular at first, because the distribution of the FTSP is determined by the solution to the

ODE, while the evolution of the solution to the ODE is determined by the behavior of the FTSP. However, the separation of time scales explains why this construction is consistent.) The second complication is that the AP produces a singularity region in the state space, causing the ODE to be discontinuous in its full state space. Hence, both the convergence to the many-server heavy-traffic fluid limit and the analysis of the solution to the ODE depend heavily on the state space of the ODE, which is characterized in terms of the FTSP. For that reason, many of the results in Perry and Whitt [37] are needed for proving convergence.

1.3. Literature. Our previous papers discuss related literature; see especially §2 of Perry and Whitt [35]. Our FQR-T control extends the FQR control and other queue-and-idleness ratio controls suggested and studied in Gurvich and Whitt [14, 15, 16], but the limits there were established for a different regime under different conditions. Here we study the FQR-T control and establish limits for overloaded systems. Unlike that previous work, here the service rates may depend on both the customer class and the service pool in a very general way. In particular, our X model does not satisfy the conditions of the previous theorems even under normal loads.

There is a substantial literature on averaging principles, e.g., see Khasminskii and Yin [25] and references therein, but there is not one unified framework that can easily be applied to any model. Moreover, it is common practice to use averaging principles as direct approximations, i.e., to simply replace a fast process by its long-term average behavior. That is the classic approach for deterministic dynamical systems, e.g., see chapters 10 and 11 of Khalil [24]. We ourselves took that approach in Perry and Whitt [36]. It is significant that, unlike the classical dynamical-systems approach, the AP in our case here is a result of the fast oscillations of a *stochastic* process (the FTSP discussed above). We thus named this phenomenon a *stochastic averaging principle* in Perry and Whitt [37], although we refer to it simply as an AP.

Averaging principles are relatively rare in operations research. See p. 71 of Whitt [47] for a discussion related to the queueing literature. Two notable papers in the queueing literature are Coffman et al. [10], which considers the diffusion limit of a polling system with zero switch-over times, and Hunt and Kurtz [18], which considers large loss networks under a large family of controls. The limits via an AP in Hunt and Kurtz [18] are the basis for other papers studying loss networks. We refer to Antunes et al. [1] and Zachary and Ziedins [49], and references therein. The work in Hunt and Kurtz [18] is also closely related to our work because it considers the fluid limits of such loss systems, with the control-driving process moving on a faster time scale than the other processes considered.

For the important characterization step, we give two proofs, one in the main paper and the other in the appendix. The shorter proof in the main paper closely follows Hunt and Kurtz [18], exploiting martingales and random measures, building on Kurtz [27]. In contrast, our second approach exploits stochastic bounds, which we also use in the important preliminary step establishing state-space collapse.

There is now a substantial literature on fluid limits for queueing models, some of which is reviewed in Whitt [47]. For recent work on many-server queues, see Kang and Ramanan [21] and Kaspi and Ramanan [23]. Because of the separation of time scales here, our work is in the spirit of fluid limits for networks of many-server queues in Bassamboo et al. [4, 5], but again the specifics are quite different. Their separation of time scales justifies using a pointwise stationary approximation asymptotically, as in Massey and Whitt [31] and Whitt [46].

2. Preliminaries. In this section we specify the queueing model, which we refer to as the X model. We then specify the FQR-T control. We then provide a short summary of the many-server heavy-traffic scaling and the different regimes. We conclude with our conventions about notation.

2.1. The original queueing model. The Markovian X model has two classes of customers, initially arriving according to independent Poisson processes with rates $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$. There are two queues, one for each class, in which customers that are not routed to service immediately upon arrival wait to be served. Customers are served from each queue in order of arrival. Each class-i customer has limited patience, which is assumed to be exponentially distributed with rate θ_i, $i = 1, 2$. If a customer does not enter service before he runs out of patience, then he abandons the queue. The abandonment keep the system stable for all arrival and service rates.

There are two service pools, with pool j having m_j homogenous servers (or agents) working in parallel. This X model was introduced to study two large systems that are designed to operate independently under normal loads, but can help each other in the face of unanticipated overloads. We assume that all servers are cross trained, so that they can serve both classes. The service times depend on both the customer class i and the server type j and are exponentially distributed; the mean service time for each class-i customer by each pool-j agent is $1/\mu_{i,j}$. All service times, abandonment times, and arrival processes are assumed to be mutually independent. The FQR-T control described below assigns customers to servers.

We assume that, at some unanticipated time, the arrival rates change instantaneously, with at least one increasing. At this time the overload incident has begun. We consider the system only after the overload incident has begun, assuming that it is in effect at the initial time 0. We further assume that the staffing cannot be changed (in the time scale under consideration) to respond to this unexpected change of arrival rates. Hence, the arrival processes change from Poisson with rates $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ to Poisson processes with rates λ_1 and λ_2, where $\tilde{\lambda}_i < m_i/\mu_{i,i}$, $i = 1, 2$ (normal loading), but $\lambda_i > \mu_{i,i} m_i$ for at least one i (the unanticipated overload). (These new arrival rates may not be known by the system manager.) Without loss of generality, we assume that pool 1 (and class 1) is the overloaded (or more overloaded) pool. The fluid model (ODE) is an approximation for the system performance during the overload incident, so that we start with the new arrival rate pair (λ_1, λ_2). (The overload control makes sense much more generally; we study its performance in this specific scenario.)

The two service systems may be designed to operate independently under normal conditions (without any overload) for various reasons. In Perry and Whitt [35, 36] we considered the common case in which there is no efficiency gain from service by cross-trained agents. Specifically, in Perry and Whitt [35] we assumed the *strong inefficient sharing condition*

$$\mu_{1,1} > \mu_{1,2} \qquad \text{and} \qquad \mu_{2,2} > \mu_{2,1}. \tag{1}$$

Under condition (1), customers are served at a faster rate when served in their own service pool than when they are being served in the other-class pool. However, many results in Perry and Whitt [35] hold under the weaker *basic inefficient sharing condition*: $\mu_{1,1}\mu_{2,2} \geq \mu_{1,2}\mu_{2,1}$.

It is easy to see that some sharing can be beneficial if one system is overloaded, while the other is underloaded (has some slack), but sharing may not be desirable if both systems are overloaded. To motivate the need for sharing when both systems are overloaded, in Perry and Whitt [35] we considered a convex-cost framework. With that framework, in Perry and Whitt [35] we showed that sharing can be optimal in the fluid approximation, even if it causes the total queue length (queue 1 plus queue 2) to increase. Despite the optimality of the control in the framework of Perry and Whitt [35], in this paper we do not assume that either the strong or the weak inefficient sharing condition holds, because the FWLLN holds regardless of the service rates. We mention that there can be other operational reasons for not sharing customers between pools during normal loads (e.g., to avoid too much agent distraction) but to share during overloads (e.g., to provide some minimal service-level constraints for both classes).

Let $Q_i(t)$ be the number of customers in the class-i queue at time t, and let $Z_{i,j}(t)$ be the number of class-i customers being served in pool j at time t, $i, j = 1, 2$. Given a stationary (state-dependent) routing policy, the six-dimensional stochastic process

$$X_6(t) \equiv (Q_1(t), Q_2(t), Z_{1,1}(t), Z_{1,2}(t), Z_{2,1}(t), Z_{2,2}(t)), \quad t \geq 0, \tag{2}$$

becomes a six-dimensional CTMC. ($\equiv$ means equality by definition.) In principle, the optimal control could be found from the theory of Markov decision processes, but that approach seems prohibitively difficult. For a complete analysis, we would need to consider the unknown transient interval over which the overload occurs, and the random initial conditions, depending on the model parameters under normal loading. In summary, there is a genuine need for the simplifying approximation we develop.

2.2. The FQR-T control for the original queueing model.
The purpose of FQR-T is to prevent sharing when the system is not overloaded and to rapidly start sharing when the arrival rates shift. If sharing is elected, then we allow sharing in only one direction.

ASSUMPTION 2.1 (ONE-WAY SHARING).   *Sharing is allowed in only one direction at any one time.*

When sharing takes place, FQR-T aims to keep the two queues at a certain ratio, depending on the direction of sharing. Thus, there is one ratio, $r_{1,2}$, which is the target ratio if class 1 is being helped by pool 2, and another target ratio, $r_{2,1}$, when class 2 is being helped by pool 1. As explained in Perry and Whitt [35], appropriate ratios can be found using the steady-state fluid approximation. In particular, the specific FQR-T control is optimal in the special case of a separable quadratic cost function. More generally, fixed ratios are often approximately optimal.

We now describe the control. The FQR-T control is based on two positive thresholds, $k_{1,2}$ and $k_{2,1}$, and the two queue-ratio parameters, $r_{1,2}$ and $r_{2,1}$. We define two queue-difference stochastic processes $D_{1,2}(t) \equiv Q_1(t) - r_{1,2}Q_2(t)$ and $D_{2,1} \equiv r_{2,1}Q_2(t) - Q_1(t)$. As shown in Perry and Whitt [35], in that convex cost framework there is no incentive for sharing simultaneously in both directions, implying that these ratio parameters should

satisfy $r_{1,2} \geq r_{2,1}$; see Proposition EC.2 and (EC.11) of Perry and Whitt [35]. However, even without the cost framework, we do not want sharing to ever occur in both directions simultaneously. Hence we make the following assumption.

ASSUMPTION 2.2 (ORDERED RATIO PARAMETERS). *The ratio parameters are assumed to satisfy* $r_{1,2} \geq r_{2,1}$.

As long as $D_{1,2}(t) \leq k_{1,2}$ and $D_{2,1}(t) \leq k_{2,1}$ we consider the system to be normally loaded (i.e., not over-loaded) so that no sharing is allowed. Hence, in that case, the two classes operate independently. Once one of these inequalities is violated, the system is considered to be overloaded, and sharing is initialized. For example, if $D_{1,2}(t) > k_{1,2}$ and $Z_{2,1}(t) = 0$, then class 1 is judged to be overloaded and service-pool 2 is allowed to start helping queue 1: If a type-2 server becomes available at this time t, then it will take its next customer from the head of queue 1. When $D_{1,2}(t) - k_{1,2} \leq 0$, new sharing is not initiated. Then new sharing stops until one of the thresholds is next exceeded. However, sharing in the opposite direction with pool 1 servers helping class 2 is not allowed until both $Z_{1,2}(t) = 0$ and $D_{2,1}(t) > k_{2,1}$.

It can be of interest to consider alternative variants of the FQR-T control just defined. For example, it may be desirable to relax the one-way sharing rule imposed above. We might use additional lower thresholds for $Z_{1,2}(t)$ and $Z_{1,2}(t)$ to allow sharing to start more quickly in the opposite direction when the queue lengths indicate that is desirable. However, we do not discuss such control variants here.

With the FQR-T control just defined, the six-dimensional stochastic process $X_6 \equiv \{X_6(t): t \geq 0\}$ in (2) is a CTMC. (The control depends only on the process state.) This is a stationary model, but we are concerned with its transient behavior, because it is not starting in steady state. We aim to describe that transient behavior. The control keeps the two queues at approximately the target ratio, e.g., if queue 1 is being helped, then $Q_1(t) \approx r_{1,2}Q_2(t)$. If sharing is done in the opposite direction, then $r_{2,1}Q_2(t) \approx Q_1(t)$ for all $t \geq 0$. That is substantiated by simulation experiments, some of which are reported in Perry and Whitt [35, 36]. In this paper we will prove that the $\approx$ signs are replaced with equality signs in the fluid limit.

2.3. Many-server heavy-traffic scaling. To develop the fluid limit, we consider a sequence of X systems, $\{X_6^n: n \geq 1\}$ defined as in (2), indexed by n (denoted by superscript), using the standard many-server heavy-traffic scaling, i.e., with arrival rates and number of servers growing proportionally to n.

ASSUMPTION 2.3 (MANY-SERVER HEAVY-TRAFFIC SCALING). *As* $n \to \infty$,

$$\bar{\lambda}_i^n \equiv \frac{\lambda_i^n}{n} \to \lambda_i \quad and \quad \bar{m}_i^n \equiv \frac{m_i^n}{n} \to m_i, \qquad 0 < \lambda_i, \ m_i < \infty, \tag{3}$$

and the service and abandonment rates held fixed.

We then define the associated fluid-scaled stochastic processes

$$\bar{X}_6^n(t) \equiv n^{-1} X_6^n(t), \quad t \geq 0, \tag{4}$$

where X_6^n is defined as in (2) for each n.

For each system n there are thresholds $k_{1,2}^n$ and $k_{2,1}^n$ scaled as follows:

ASSUMPTION 2.4 (SCALED THRESHOLDS). *For* $k_{1,2}, k_{2,1} > 0$ *and a sequence of positive numbers* $\{c_n: n \geq 1\}$,

$$k_{i,j}^n / c_n \to k_{i,j}, \quad (i,j) = (1,2) \ or \ (2,1), \ where \ c_n/n \to 0 \ and \ c_n/\sqrt{n} \to \infty \ as \ n \to \infty. \tag{5}$$

The first scaling by n in Assumption 2.4 is chosen to make the thresholds asymptotically negligible in many-server heavy-traffic fluid scaling, so they have no asymptotic impact on the steady-state cost (in the cost framework of Perry and Whitt [35]). The second scaling by $\sqrt{n}$ in Assumption 2.4 is chosen to make the thresholds asymptotically infinite in many-server heavy-traffic diffusion scaling, so that asymptotically the thresholds will not be exceeded under normal loading. It is significant that the scaling shows that we should be able to simultaneously satisfy both conflicting objectives in large systems.

We consider an overload incident in which class 1 is overloaded, and more overloaded than class 2 if both are overloaded. Hence we primarily focus on the queue difference processes $D_{1,2}^n$. We redefine the queue-difference process by subtracting $k_{1,2}^n$ from Q_1^n, i.e.,

$$D_{1,2}^n(t) \equiv Q_1^n(t) - k_{1,2}^n - r_{1,2}Q_2^n(t), \quad t \geq 0. \tag{6}$$

(Similarly, we write $D_{2,1}^n(t) \equiv r_{2,1}Q_2^n(t) - k_{2,1} - Q_1^n(t)$.) We now apply FQR using the process $D_{1,2}^n$ in (6): if $D_{1,2}^n(t) > 0$, then every newly available agent (in either pool) takes his new customer from the head of the class-1 queue. If $D_{1,2}^n(t) \leq 0$, then every newly available agent takes his new customer from the head of his own queue.

Let

$$\rho_i^n \equiv \frac{\lambda_i^n}{\mu_{i,i}m_i^n}, \qquad \text{and} \qquad \rho_i \equiv \lim_{n\to\infty} \rho_i^n = \frac{\lambda_i}{\mu_{i,i}m_i}, \quad i = 1, 2. \tag{7}$$

Then ρ_i^n is the traffic intensity of class i to pool i, and ρ_i can be thought of as its fluid counterpart.

Our results depend on the system being overloaded. However, in our case, a system can be overloaded even if one of the classes is not overloaded by itself. We define the following quantities:

$$q_i^a \equiv \frac{(\lambda_i - \mu_{i,i}m_i)^+}{\theta_i} \qquad \text{and} \qquad s_i^a \equiv \left(m_i - \frac{\lambda_i}{\mu_{i,i}}\right)^+, \quad i = 1, 2, \tag{8}$$

where $(x)^+ \equiv \max\{x, 0\}$. It is easy to see that $q_i^a s_i^a = 0$, $i = 1, 2$. Note that q_i^a is the steady state of the class-i fluid-limit queue when there is no sharing, i.e., when both classes operate independently. Similarly, s_i^a is the steady state of the class-i fluid-limit idleness process. For the derivation of the quantities in (8) see Theorem 2.3 in Whitt [48], especially equation (2.19), and §5.1 in Perry and Whitt [35]. See also §6 in Perry and Whitt [37].

2.4. Conventions about notation. We use capital letters to denote random variables and stochastic processes, and lowercase letters for deterministic counterparts. For a sequence $\{Y^n: n \geq 1\}$ (of stochastic processes or random variables) we denote its fluid-scaled version by $\bar{Y}^n \equiv Y^n/n$. The fluid limit of stochastic processes $\bar{Y}^n$ will be denoted by a lowercase letter y and sometimes by $\bar{Y}$. Let $\Rightarrow$ denote convergence in distribution. Then the fluid limits will be expressed as $\bar{Y}^n \Rightarrow y$ as $n \to \infty$.

We use the usual $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{Z}_+$ notation for the real numbers, integers, and nonnegative integers, respectively. Let $\mathbb{R}_k$ denote all k-dimensional vectors with components in $\mathbb{R}$. For a subinterval I of $[0, \infty)$, let $\mathcal{D}_k(I) \equiv \mathcal{D}(I, \mathbb{R}_k)$ be the space of all right-continuous $\mathbb{R}_k$ valued functions on I with limits from the left everywhere, endowed with the familiar Skorohod J_1 topology. We let d_{J_1} denote the metric on $\mathcal{D}_k(I)$. Because we will be considering continuous limits, the topology is equivalent to uniform convergence on compact subintervals of I. If I is an arbitrary compact interval, we simply write $\mathcal{D}_k$. Let $\mathcal{C}_k$ be the subset of continuous functions in $\mathcal{D}_k$. Let e be the identity function in $\mathcal{D} \equiv \mathcal{D}_1$, i.e., $e(t) \equiv t$, $t \in I$.

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

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

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

The same notation is used for sequences, replacing x with $n \in \mathbb{Z}_+$.

For a, $b \in \mathbb{R}$, let $a \wedge b \equiv \min\{a, b\}$, $a \vee b \equiv \max\{a, b\}$, $(a)^+ \equiv a \vee 0$, and $(a)^- \equiv 0 \vee -a$. For a function $x: [0, \infty) \to \mathbb{R}$ and $0 < t < \infty$, let

$$\|x\|_t \equiv \sup_{0 \leq s \leq t} |x(s)|.$$

Let $Y \equiv \{Y(t): t \geq 0\}$ be a stochastic process, and let $f: [0, \infty) \to [0, \infty)$ be a deterministic function. We say that Y is $O_P(f(t))$, and write $Y = O_P(f)$, if $\|Y\|_t/f(t)$ is *stochastically bounded*, i.e., if

$$\lim_{a\to\infty} \limsup_{t\to\infty} P(\|Y\|_t/f(t) > a) = 0.$$

We say that Y is $o_P(f(t))$ if $\|Y\|_t/f(t)$ converges in probability (and thus, in distribution) to 0, i.e., if $\|Y\|_t/f(t) \Rightarrow 0$ as $t \to \infty$. If $f(t) \equiv 1$, then $Y = O_P(1)$ if it is stochastically bounded, and $Y = o_P(1)$ if $\|Y\|_t \Rightarrow 0$. We define $O_P(f(n))$ and $o_P(f(n))$ in a similar way, but with the domain of f being $\mathbb{Z}_+$, i.e., $f: \mathbb{Z}_+ \to [0, \infty)$. These properties extend to sequences of random variables and processes indexed by n if the property holds uniformly in n.

3. Representation of the fluid limit. In this section we represent the fluid limit as a solution to an ODE that is driven by a FSTP. In contrast to the six-dimensional scaled process $\bar{X}_6^n$ in (4), the ODE is only three-dimensional. Hence, we start by briefly discussing the dimension reduction in §3.1. Afterward, we define the FTSP in §3.2 and then present the ODE in §3.3. We have studied the FTSP and the ODE in Perry and Whitt [37], to which we refer for more details. In §3.4 we state three main assumptions.

3.1. Dimension reduction. We will making assumptions implying that we consider the system during an overload incident in which class 1 is overloaded, and more so than class 2 if it is also overloaded. We will thus be considering sharing in which only pool 2 may help class 1. We will thus have both service pools fully occupied, with service pool 1 serving only class 1. We will thus have $P(B_T^n) \to 1$ as $n \to \infty$, for all T, $0 < T < \infty$, where B_T^n is the subset of the underlying probability space defined by

$$B_T^n \equiv \{Z_{1,1}^n(t) = m_1^n,\ Z_{2,1}^n(t) = 0,\ Z_{1,2}^n(t) + Z_{2,2}^n(t) = m_2^n \text{ for } 0 \leq t \leq T\}. \tag{9}$$

On the set B_T^n the effective dimension is reduced from six to three. Carefully justifying this SSC will be the topic of §6. Thus, in addition to the process X_6^n in (2) for each n, we also consider the six-dimensional processes

$$X_6^{n,*} \equiv (Q_1^n, Q_2^n, m_1^n e, Z_{1,2}^n, 0e, m_2^n e - Z_{1,2}^n) \quad \text{in } \mathcal{D}_6 \tag{10}$$

and the associated three-dimensional processes

$$X_3^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n) \quad \text{in } \mathcal{D}_3, \tag{11}$$

obtained by truncating the process $X_6^{n,*}$, keeping only the essential first, second, and fourth coordinates. (Note that $P(X_6^{n,*} = X_6^n \text{ in } \mathcal{D}_6([0, T])) = P(B_T^n) \to 1$ as $n \to \infty$.) We obtain a further alternative representation for the associated three-dimensional fluid-scaled processes $\bar{X}_3^n$, denoted by $\bar{X}^n$ in §6; see (45).

3.2. The fast-time-scale process (FTSP). Because we consider the system during an overload incident in which class 1 is overloaded, and more so than class 2 if it is also overloaded, we will primarily consider only the one queue difference processes $D_{1,2}^n$ in (6). The FTSP can perhaps be best understood as being the limit of a family of *time-expanded queue-difference processes*, defined for each $n \geq 1$ by

$$D_e^n(\Gamma^n, s) \equiv D_{1,2}^n(t_0 + s/n), \quad s \geq 0, \tag{12}$$

where X^n is the three-dimensional process in (11) and we condition on $X^n(t_0) = \Gamma^n$ for some deterministic vector Γ^n assuming possible values of $X^n(t_0) \equiv (Q_1^n(t_0), Q_2^n(t_0), Z_{1,2}^n(t_0))$. (The time t_0 is an arbitrary initial time.) We choose Γ^n so that $\Gamma^n/n \to \gamma$ as $n \to \infty$, where $\gamma \equiv (q_1, q_2, z_{1,2})$ is an appropriate fixed state (in three dimensions) because we will have sharing in only one direction. The formal statement of the limit for D_e^n in (12) is Theorem 4.4. Because we divide s in (12) by n, we are effectively dividing the rates by n. (See (A1)–(A2) for the transition rates of $D_{1,2}^n$ itself.) We are applying a "microscope" to "expand time" and look at the behavior after the initial time more closely; that is in contrast to the usual time contraction with conventional heavy-traffic limits. See Whitt [45] for a previous limit using time expansion.

Let $r \equiv r_{1,2}$ and let $\gamma \equiv (q_1, q_2, z_{1,2})$ be a possible state in the three-dimensional state space $\mathbb{S} \equiv [0, \infty)^2 \times [0, m_2]$. Directly, we let the FTSP $\{D(\gamma, s): s \geq 0\}$ be a pure-jump Markov process with transition rates $\lambda_-^{(r)}(\gamma)$, $\lambda_-^{(1)}(\gamma)$, $\mu_-^{(r)}(\gamma)$, and $\mu_-^{(1)}(\gamma)$ for transitions of $+r$, $+1$, $-r$, and -1, respectively, when $D(\gamma, s) \leq 0$. Similarly, let the transition rates be $\lambda_+^{(r)}(\gamma)$, $\lambda_+^{(1)}(\gamma)$, $\mu_+^{(r)}(\gamma)$, and $\mu_+^{(1)}(\gamma)$ for transitions of $+r$, $+1$, $-r$, and -1, respectively, when $D(\gamma, s) > 0$.

We define the transition rates for $D(\gamma)$ as follows: First, for $D(\gamma, s) \in (-\infty, 0]$ with $\gamma \equiv (q_1, q_2, z_{1,2})$, the upward rates are

$$\lambda_-^{(1)}(\gamma) \equiv \lambda_1, \quad \text{and} \quad \lambda_-^{(r)}(\gamma) \equiv \mu_{1,2} z_{1,2} + \mu_{2,2}(m_2 - z_{1,2}) + \theta_2 q_2. \tag{13}$$

Similarly, the downward rates are

$$\mu_-^{(1)}(\gamma) \equiv \mu_{1,1} m_1 + \theta_1 q_1 \quad \text{and} \quad \mu_-^{(r)}(\gamma) \equiv \lambda_2. \tag{14}$$

Next, for $D(\gamma, s) \in (0, \infty)$, we have upward rates

$$\lambda_+^{(1)}(\gamma) \equiv \lambda_1 \quad \text{and} \quad \lambda_+^{(r)}(\gamma) \equiv \theta_2 q_2. \tag{15}$$

The downward rates are

$$\mu_+^{(1)}(\gamma) \equiv \mu_{1,1} m_1 + \mu_{1,2} z_{1,2} + \mu_{2,2}(m_2 - z_{1,2}) + \theta_1 q_1 \quad \text{and} \quad \mu_+^{(r)}(\gamma) \equiv \lambda_2. \tag{16}$$

As in §7.1 of Perry and Whitt [37], we identify important subsets of the state space $\mathbb{S} \equiv [0, \infty)^2 \times [0, m_2]$:

$$\mathbb{S}^{\mathbf{b}} \equiv \{q_1 - r q_2 = 0\}, \quad \mathbb{S}^+ \equiv \{q_1 - r q_2 > 0\}, \quad \mathbb{S}^- \equiv \{q_1 - r q_2 < 0\}. \tag{17}$$

Let $D(\gamma, \infty)$ be a random variable that has the steady-state limiting distribution of the FTSP $D(\gamma, s)$ as $s \to \infty$ and let

$$\pi_{1,2}(\gamma) \equiv P(D(\gamma, \infty) > 0). \tag{18}$$

That is, $\pi_{1,2}(\gamma)$ is the probability that the stationary FTSP associated with $\gamma \in \mathbb{S}$ is strictly positive.

It turns out that $D(\gamma, \infty)$ and $\pi_{1,2}(\gamma)$ are well defined throughout $\mathbb{S}$. In $\mathbb{S}^b$ the function $\pi_{1,2}$ can assume its full range of values, $0 \le \pi_{1,2}(\gamma) \le 1$; the boundary subset $\mathbb{S}^b$ is where the AP is taking place. For all $\gamma \in \mathbb{S}^+$, $\pi_{1,2}(\gamma) = 1$; for all $\gamma \in \mathbb{S}^-$, $\pi_{1,2}(\gamma) = 0$. For $\mathbb{S}^-$ to be a proper subspace of $\mathbb{S}$, both service pools must be constantly full. Thus, if $\gamma \in \mathbb{S}^-$, then $z_{1,1} = m_1$ and $z_{1,2} + z_{2,2} = m_2$, but q_1 and q_2 are allowed to be equal to zero.

An important role will be played by the subset $\mathbb{A}$ of $\mathbb{S}^b$ such that the FTSP is positive recurrent. The AP takes place only in the set $\mathbb{A}$. In Theorem 6.1 of Perry and Whitt [37] we showed that positive recurrence of the FTSP, and thus the set $\mathbb{A}$, depends only on the constant drift rates in the two regions:

$$\delta_+(\gamma) \equiv r(\lambda_+^{(r)}(\gamma) - \mu_+^{(r)}(\gamma)) + (\lambda_+^{(1)}(\gamma) - \mu_+^{(1)}(\gamma))$$
$$\delta_-(\gamma) \equiv r(\lambda_-^{(r)}(\gamma) - \mu_-^{(r)}(\gamma)) + (\lambda_-^{(1)}(\gamma) - \mu_-^{(1)}(\gamma)). \tag{19}$$

The FTSP $\{D(\gamma, s): s \ge 0\}$ is positive recurrent if (and only if) the state γ belongs to the set

$$\mathbb{A} \equiv \{\gamma \in \mathbb{S}: \delta_-(\gamma) > 0 > \delta_+(\gamma)\}. \tag{20}$$

Let the other two subsets of $\mathbb{S}^b$ be

$$\mathbb{A}^+ \equiv \{x \in \mathbb{S}^b \mid \delta_+(x) \ge 0\} \qquad \text{and} \qquad \mathbb{A}^- \equiv \{x \in \mathbb{S}^b \mid \delta_-(x) \le 0\}. \tag{21}$$

From Theorem 6.2 of Perry and Whitt [37], we obtain the following lemma, giving the limiting behavior of the FTSP for any state in $\mathbb{S}$.

LEMMA 3.1 (LIMITING BEHAVIOR OF THE FTSP). *For all $\gamma \in \mathbb{S}$ and $x \in \mathbb{R}$,*

$$\lim_{s \to \infty} P(D(\gamma, s) \le x) = F(\gamma, x) \equiv P(\gamma, (-\infty, x]), \tag{22}$$

where $F(\gamma, \cdot)$ is a cumulative distribution function (cdf) associated with a possibly defective probability measure $P(\gamma, \cdot)$ depending on the state γ. Moreover,

$$\begin{aligned}
&\text{for all } \gamma \in \mathbb{A} \text{ and } x \in \mathbb{R}, &0 < F(\gamma, x) < 1 &\quad \text{and} &0 < \pi_{1,2}(\gamma) < 1; \\
&\text{for all } \gamma \in \mathbb{S}^+ \cup \mathbb{A}^+ \text{ and } x \in \mathbb{R}, &F(\gamma, x) = 0 &\quad \text{and} &\pi_{1,2}(\gamma) = 1; \\
&\text{for all } \gamma \in \mathbb{S}^- \cup \mathbb{A}^- \text{ and } x \in \mathbb{R}, &F(\gamma, x) = 1 &\quad \text{and} &\pi_{1,2}(\gamma) = 0.
\end{aligned} \tag{23}$$

Later in §7.2, we obtain a proper limiting steady-state distribution for the FTSP for all γ in $\mathbb{S}$ by appending states $+\infty$ and $-\infty$ to the state space $\mathbb{R}$ of the FTSP $\{D(\gamma, s): s \ge 0\}$. Lemma 3.1 then implies that $P(\gamma, \mathbb{R}) = 1$ for $\gamma \in \mathbb{A}$, $P(\gamma, \{+\infty\}) = 1$ for $\gamma \in \mathbb{S}^+ \cup \mathbb{A}^+$ and $P(\gamma, \{-\infty\}) = 1$ for $\gamma \in \mathbb{S}^- \cup \mathbb{A}^-$.

3.3. The ordinary differential equation (ODE). We can now present the three-dimensional ODE in terms of the FTSP D. Let $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$ be the solution to the ODE at time t; let $\dot{x} \equiv (\dot{q}_1, \dot{q}_2, \dot{z}_{1,2})$, where $\dot{x}(t)$ is the derivative evaluated at time t and

$$\begin{aligned}
\dot{q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(x(t))[z_{1,2}(t)\mu_{1,2} + z_{2,2}(t)\mu_{2,2}] - \theta_1 q_1(t) \\
\dot{q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(x(t)))[z_{2,2}(t)\mu_{2,2} + z_{1,2}(t)\mu_{1,2}] - \theta_2 q_2(t) \\
\dot{z}_{1,2}(t) &\equiv \pi_{1,2}(x(t))z_{2,2}(t)\mu_{2,2} - (1 - \pi_{1,2}(x(t)))z_{1,2}(t)\mu_{1,2},
\end{aligned} \tag{24}$$

with $\pi_{1,2}(x(t)) \equiv P(D(x(t), \infty) > 0)$ for each $t \ge 0$, where $D(x(t), \infty)$ has the limiting steady-state distribution as $s \to \infty$ of the FTSP $D(\gamma, s)$ for $\gamma = x(t)$. (Recall also that $z_{2,2} = m_2 - z_{1,2}$.) Theorem 5.2 of Perry and Whitt [37] shows that the ODE has a unique solution as a continuous function mapping $[0, \infty)$ into $\mathbb{S}$ for any initial value in $\mathbb{S}$. Lemma 3.1 shows that $\pi_{1,2}(x(t))$ is well defined for any $x(t)$ in $\mathbb{S}$.

Equivalently, we have the following integral representation of the ODE in (24):

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

$$q_1(t) \equiv q_1(0) + \lambda_1 t - m_1 t - \mu_{1,2} \int_0^t \pi_{1,2}(x(s))z_{1,2}(s)\, ds$$

$$- \mu_{2,2} \int_0^t \pi_{1,2}(x(s))(m_2 - z_{1,2}(s))\, ds - \theta_1 \int_0^t q_1(s)\, ds, \tag{25}$$

$$q_2(t) \equiv q_2(0) + \lambda_2 t - \mu_{2,2} \int_0^t (1 - \pi_{1,2}(x(s)))(m_2 - z_{1,2}(s))\, ds$$

$$- \mu_{1,2} \int_0^t (1 - \pi_{1,2}(x(s)))z_{1,2}(s)\, ds - \theta_2 \int_0^t q_2(s)\, ds.$$

We will see that the integral representation in (25) is closely related to an associated integral representation of $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$; see (45); $\bar{X}^n$ is replaced by the deterministic state x and the indicators $1_{\{D_{1,2}^n(s)>0\}}$ are replaced by $\pi_{1,2}(x(s))$.

It is easy to see that the right-hand side of the ODE is not a continuous function of x and, in particular, is not locally Lipschitz continuous in x. Thus, proving that the ODE possesses a unique solution is not straightforward. The proof of that statement is the main result in Perry and Whitt [37] and builds on matrix-geometric methods, as well as heavy-traffic limit theorems for the FTSP; see Theorems 5.2 and 7.1 there. The matrix-geometric representation of the FTSP also provides key tools for developing an algorithm to compute that unique solution.

3.4. Three main assumptions. We now introduce three main assumptions: Assumptions 3.1–3.3. All three assumptions are assumed to hold throughout the paper, unless explicitly stated otherwise. These assumptions are in addition to the four assumptions made in §2: Assumptions 2.1–2.4. (Here we do not require (1).) Our first new assumption is on the asymptotic behavior of the model parameters; it specifies the essential form of the overload. For the statement, recall the definitions in (3), (5), and (8), which describe the asymptotic behavior of the parameters.

ASSUMPTION 3.1 (SYSTEM OVERLOAD, WITH CLASS 1 MORE OVERLOADED). *The rates in the overload are such that the limiting rates satisfy*
(1) $\theta_1 q_1^a > \mu_{1,2} s_2^a$, *and*
(2) $q_1^a > r_{1,2} q_2^a$.

Condition (1) in Assumption 3.1 ensures that class 1 is asymptotically overloaded, even after receiving help from pool 2. To see why, first observe that, since $s_2^a \geq 0$, $q_1^a > 0$, so that $\lambda_1 > \mu_{1,1} m_1$ and $\rho_1 > 1$. Hence, class 1 is overloaded. Next observe that $\mu_{1,2} s_2^a = \mu_{1,2}(1 - \rho_2)^+$, and that $(1 - \rho_2)^+$ is the amount of (steady-state fluid) extra service capacity in pool 2, if it were to serve only class-2 customers. Thus, Condition (1) in Assumption 3.1 implies that enough class-1 customers are routed to pool 2 to ensure that pool 2 is overloaded when sharing is taking place. This conclusion will be demonstrated in §6. Condition (1) in Assumption 3.1 is slightly stronger than Condition (I) of Assumption A in Perry and Whitt [37], because here there is a strong inequality instead of a weak inequality.

Condition (2) in Assumption 3.1 ensures that class 1 is more overloaded than class 2 if class 2 is also overloaded. This condition helps ensure that there is no incentive for pool 1 to help pool 2, so that $Z_{2,1}^n$ should remain at zero.

Our second assumption is about the initial conditions. For the initial conditions, we assume that the overload, whose asymptotic character is specified by Assumption 3.1, is ongoing or is about to begin. In addition, sharing with pool 2 allowed to help class 1 has been activated by having the threshold $k_{1,2}^n$ exceeded by the queue difference process $D_{1,2}^n$ and is in process. Thus actual sharing is being controlled by the difference process $D_{1,2}^n$ in (6). Here is our specific assumption.

ASSUMPTION 3.2 (INITIAL CONDITIONS). *For each $n \geq 1$, $P(Z_{2,1}^n(0) = 0, Q_i^n(0) > a_n, i = 1, 2) = 1$,*

$$\bar{X}^n(0) \Rightarrow x(0) \in \mathbb{A} \cup \mathbb{A}^+ \cup \mathbb{S}^+ \quad as \ n \to \infty, \qquad and \qquad D_{1,2}^n(0) \Rightarrow L \quad if \ x(0) \in \mathbb{A} \cup \mathbb{A}^+,$$

where $\{a_n : n \geq 1\}$ is a sequence of real numbers satisfying $a_n / c_n \to a$, $0 < a \leq \infty$, for c_n in (5); $D_{1,2}^n$ is defined in (6); $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$; $x(0)$ is a deterministic element of $\mathbb{R}_3$; $\mathbb{A}$, $\mathbb{A}^+$, and $\mathbb{S}^+$ are the subsets of $\mathbb{S}$ in (20), (21), and (17); and L is a proper random variable, i.e., $P(|L| < \infty) = 1$.

Because we are interested in times when sharing occurs with pool 2 helping class 1, in Assumption 3.2 we assume that the scale of $Q_1^n(0)$ is at least as large as that of the threshold $k_{1,2}^n$ (so either the threshold has already been crossed, or it is about to be crossed). Note that we also assume that $D_{1,2}^n(0) \Rightarrow L$, so that it is natural to assume that $Q_2^n(0)$ has the same order as $Q_1^n(0)$; we elaborate in Remark 4.2 and Appendix E.

In Assumption 3.2 we do not allow $x(0)$ in $\mathbb{S}^-$, because such an initial condition may activate sharing in the wrong direction, with pool 1 helping class 2, causing the system to leave the state-space $\mathbb{S}$; see Remark 4.1.

As noted in §3.3, in Perry and Whitt [37] we required that the queue ratio parameter be rational in order to establish results about the FTSP and the ODE.

ASSUMPTION 3.3 (RATIONAL QUEUE RATIO PARAMETER). *The queue ratio parameters $r_{1,2}$ and $r_{2,1}$ are rational positive numbers.*

Given Assumption 3.3, without loss of generality, we let the thresholds be rational (of the form $k_{1,2}^n = m^n/k$, where $r_{1,2} = j/k$). We conjecture that Assumption 3.3 can be removed, but that condition has been used in Perry and Whitt [37] to make the pure-jump Markov FTSP a *quasi-birth-and-death* (QBD) process, which in turn was used to establish critical properties of the FTSP and the ODE. We use some of these properties in this paper as well. By Assumption 3.3, $r_{1,2} = j/k$ for positive integers j and k. The computational efficiency of the algorithm to solve the ODE developed in §11 of Perry and Whitt [37] actually depends on j and k not being too large as well, because the QBD matrices are $2m \times 2m$, where $m \equiv \max\{j, k\}$, see §6.2 of Perry and Whitt [37], and the steady state of that QBD must be calculated at each discretization step in solving the ODE. Fortunately, simulations show that the system performance is not very sensitive to small changes in $r_{1,2}$, so that having m be 5 or 10 seems adequate for practical purposes.

Relaxing Assumption 3.3 will have practical value only if an efficient algorithm for solving the ODE is developed. We remark that computing the stationary distribution of a pure-jump Markov process can in general be hard and time consuming, and that we need to compute the stationary distribution of a large number of such processes in order to solve the ODE. Hence, the ability to analyze the FTSP as a QBD has an important advantage, even if Assumption 3.3 is relaxed.

4. Main results. In this section we state the main results of the paper. In §4.1 we state the main theorem, establishing the FWLLN via the AP, proving that the (unique) solution to the ODE (24) is indeed the fluid limit of $\bar{X}_6^n$. In §4.2 we establish convergence of the stationary distributions, showing that the order of the two limits $n \to \infty$ and $t \to \infty$ can be interchanged in great generality. In §4.3 we establish asymptotic results about the queue-difference stochastic process. We conclude in §4.4 by giving a brief overview of the following proofs.

4.1. The fluid limit. We are now ready to state our main result in this paper, which is a FWLLN for scaled versions of the vector stochastic process (X_6^n, Y_8^n), where $X_6^n \equiv (Q_i^n, Z_{i,j}^n) \in \mathscr{D}_6$ as in (2) and $Y_8^n \equiv (A_i^n, S_{i,j}^n, U_i^n) \in \mathscr{D}_8$, $i, j = 1, 2$, where $A_i^n(t)$ counts the number of class-i customer arrivals, $S_{i,j}^n(t)$ counts the number of service completions of class-i customers by agents in pool j, and $U_i^n(t)$ counts the number of class-i customers to abandon from queue, all in model n during the time interval $[0, t]$. For the FWLLN, we focus on the scaled vector process

$$(\bar{X}_6^n, \bar{Y}_8^n) \equiv n^{-1}(X_6^n, Y_8^n), \tag{26}$$

as in (4). To explicitly state the AP, we also consider the functions

$$\Theta^n(t) \equiv \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} \, ds \quad \text{and} \quad \vartheta(t) \equiv \int_0^t \pi_{1,2}(x(s)) \, ds, \qquad t \geq 0, \tag{27}$$

where $\pi_{1,2}(\cdot)$ is defined in (18). In particular, $\pi_{1,2}(x(s))$ is the probability that the *stationary* FTSP $D(x(s), \cdot)$, associated with $x(s)$, is strictly positive, where $x(s)$ is the value of the fluid limit of $\bar{X}^n(s)$ at time s, $s \geq 0$.

Recall that Assumptions 2.1–2.4 and 3.1–3.3 are assumed to be in force throughout the paper.

THEOREM 4.1 (FWLLN VIA THE STOCHASTIC AVERAGING PRINCIPLE). *As $n \to \infty$,*

$$(\bar{X}_6^n, \bar{Y}_8^n, \Theta^n) \;\Rightarrow\; (x_6, y_8, \vartheta) \quad \text{in } \mathscr{D}_{15}([0, \infty)), \tag{28}$$

where (x_6, y_8, ϑ) is a deterministic element of $\mathscr{C}_{15}([0, \infty))$, $x_6 \equiv (q_i, z_{i,j})$, $y_8 \equiv (a_i, s_{i,j}, u_i)$, $i = 1, 2$; $j = 1, 2$; ϑ in (27); $z_{2,1} = s_{2,1} = m_1 - z_{1,1} = m_2 - z_{2,2} - z_{1,2} = 0e$; $x \equiv (q_1, q_2, z_{1,2})$ is the unique solution to the three-dimensional ODE in (24) mapping $[0, \infty)$ into $\mathbb{S}$. The remaining limit function y_8 is defined in terms of x_6:

$$a_i(t) \equiv \lambda_i t, \quad s_{i,j}(t) \equiv \mu_{i,j} \int_0^t z_{i,j}(s) \, ds, \quad u_i(t) \equiv \theta_i \int_0^t q_i(s) \, ds \qquad \text{for } t \geq 0, \; i = 1, 2; \; j = 1, 2. \tag{29}$$

We prove Theorem 4.1 by showing in §5.2 that the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n, \Theta^n): n \geq 1\}$ is $\mathscr{C}$-tight in $\mathscr{D}_{15}([0, \infty))$ and by showing subsequently that the limit of every convergent subsequence of $\bar{X}_6^n$ must take values in $\mathbb{S}$ and be a solution to the ODE (24), which has a unique solution by Theorem 5.2 of Perry and Whitt [37].

4.2. Limit interchange result. Under the FQR-T control operating during a single overload incident of unlimited duration, the six-dimensional stochastic process $X_6^n \equiv (Q_i^n, Z_{i,j}^n; i, j = 1, 2)$ is a positive recurrent irreducible CTMC for each n. Hence, $\bar{X}_6^n \equiv n^{-1} X_6^n$ has a unique steady-state (limiting and stationary) distribution $\bar{X}_6^n(\infty)$ for each n.

Theorem 8.2 of Perry and Whitt [37] implies that there exists a unique stationary point $x^* \equiv (q_1^*, q_2^*, z_{1,2}^*)$ in the state space $\mathbb{S}$ to the three-dimensional limiting ODE in (24), where

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

$$q_1^* = \frac{\lambda_1 - m_1\mu_{1,1} - \mu_{1,2}z_{1,2}^*}{\theta_1} \quad \text{and} \quad q_2^* = \frac{\lambda_2 - \mu_{2,2}(m_2 - z_{1,2}^*)}{\theta_2}. \tag{30}$$

Let x_6^* be the six-dimensional version of $x^* \equiv (q_1^*, q_2^*, z_{1,2}^*)$ in (30), i.e.,

$$x_6^* \equiv (q_1^*, q_2^*, m_1, z_{1,2}^*, 0, m_2 - z_{1,2}^*) \quad \text{for } x^* = (q_1^*, q_2^*, z_{1,2}^*). \tag{31}$$

Observe that, if $\lambda_2 - \mu_{2,2}m_2 > 0$, then the numerator in the expression of $z_{1,2}^*$ is equal to $\theta_1\theta_2(q_1^a - r_{1,2}q_2^a)$ and is strictly positive by Condition (2) in Assumption 3.1, so that $0 < z_{1,2}^* \leq m_2$. Moreover, by Corollary 8.2 in Perry and Whitt [37], the two conditions in Assumption 3.1 guarantee that $x^* \in \mathbb{S}^b \cup \mathbb{S}^+$, and in particular, that $x^* \in \mathbb{S}$.

We now establish a limit interchange result.

THEOREM 4.2 (INTERCHANGE OF LIMITS). *For each continuous bounded function $f: \mathbb{R}_6 \to \mathbb{R}$,*

$$\lim_{n \to \infty} \lim_{t \to \infty} E[f(\bar{X}^n(t))] = \lim_{t \to \infty} \lim_{n \to \infty} E[f(\bar{X}^n(t))] = f(x_6^*),$$

where x_6^* is defined in (31).

We will prove Theorem 4.2 by first proving the limit on the left side. For that, we can relax the assumptions. In particular, we will show that the sequence of stationary distributions converges to the unique stationary point of the ODE, without requiring Assumptions 3.2 and 3.3. Of course, Assumption 3.2 plays no role because it concerns the initial conditions.

The current proof of Theorem 8.2 of Perry and Whitt [37] used for (30) does apply to Theorem 5.2 of Perry and Whitt [37], which depends on Assumption 3.3, the technical assumption that $r_{1,2}$ and $r_{2,1}$ are rational numbers. However, we now show that Theorem 8.2 of Perry and Whitt [37] actually does not depend on Assumption 3.3.

LEMMA 4.1. *Under the conditions of Theorem 8.2 of Perry and Whitt [37], excluding Assumptions 3.2 and 3.3 here, x^* is the unique stationary point of the ODE.*

PROOF. Assume that the conditions of Theorem 8.2 of Perry and Whitt [37] are satisfied with an irrational $r \equiv r_{1,2}$. Construct a sequence of rational numbers $\{r_n: n \geq 1\}$ with $r_n \to r$ as $n \to \infty$. Then, for all n sufficiently large, the conditions of Theorem 8.2 of Perry and Whitt [37] are satisfied with r_n. Let x_n^* be the unique stationary point associated with r_n. Then, by Theorem 8.1 of Perry and Whitt [37], $x_n^* \to x^*$ as $n \to \infty$. $\square$

The existence of a stationary point of an ODE necessarily implies the existence of a (constant) solution to the ODE, but it does not require the existence of a unique solution to the ODE. Thus, the existence of a unique solution provided by Theorem 5.2 of Perry and Whitt [37], which does use Assumptions 3.2 and 3.3, is not needed. Moreover, Theorems 8.3 and 9.2 of Perry and Whitt [37] imply that x^* is globally asymptotically stable and $x(t)$ converges to x^* exponentially fast as $t \to \infty$. These too do not depend on Assumptions 3.2 and 3.3.

We now show that x_6^* is the limit of the stationary sequence $\{\bar{X}_6^n(\infty): n \geq 1\}$ *without assuming Assumptions 3.2 and 3.3. The proof of Theorem 4.3 appears in §8.1.*

THEOREM 4.3 (WLLN FOR THE STATIONARY DISTRIBUTIONS). *Under the assumptions here, excluding Assumptions 3.2 and 3.3, $\bar{X}_6^n(\infty) \Rightarrow x_6^*$ in $\mathbb{R}_6$ as $n \to \infty$, for x_6^* in (31).*

PROOF OF THEOREM 4.2. The iterated limit on the left holds by virtue of Theorem 4.3. The iterated limit on the right holds because of Theorem 4.1 together with the fact that x_6^* is a globally asymptotically stable stationary point for the fluid limit, by (31) and Theorem 8.3 of Perry and Whitt [37]. □

REMARK 4.1 (STARTING IN $\mathbb{S}^-$). It is significant that the limit interchange in Theorem 4.2 is not valid throughout $\mathbb{S}$. If Assumption 3.2 holds, except that $x(0) \in \mathbb{S}^-$, then $q_1(0) - r_{1,2}q_2(0) < 0$. Together with Assumption 2.2, that implies that, in some regions of $\mathbb{S}^-$, $d_{2,1} \equiv r_{2,1}q_2(0) - q_1(0) > 0$; that can hold in $\mathbb{S}^-$ because $r_{1,2}q_2(0)$ can be larger than $q_1(0)$. In those cases we have $P(D_{1,2}^n(0) < 0) \to 1$ and $P(D_{2,1}^n(0) > k_{2,1}^n) \to 1$ as $n \to \infty$. If we assume that $P(Z_{1,2}^n(0) = Z_{2,1}^n(0) = 0) = 1$ for all $n \geq 1$, which is consistent with Assumption 3.2, then, asymptotically, we will initially have sharing the wrong way, with pool 1 helping class 2. By the continuity, there will be an interval $[0, \delta]$ for which

$$\inf_{\{0 \leq t \leq \delta\}} \{d_{2,1}(t)\} > 0.$$

Hence, asymptotically as $n \to \infty$, there will rapidly be sharing with pool 1 helping class 2. It can be shown that there exists $\delta > 0$ and $\epsilon > 0$ such that $P(\bar{Z}_{2,1}^n(\delta) > \epsilon) \to 1$ as $n \to \infty$. This shows that the limit interchange is not valid for every initial condition in $\mathbb{S}^-$.

4.3. The limiting behavior of the queue difference process.

In this section we present important supplementary results that help "explain" the AP, which takes place in $\mathbb{A}$. The following results are not applied in the proof of Theorem 4.1, but are also not immediate corollaries of the FWLLN; their proofs are given in §8.

For each $n \geq 1$, let Γ_6^n be a random state of X_6^n that is independent of subsequent arrival, service, and abandonment processes, and let Γ^n be the random state of X_3^n associated with Γ_6^n as in (11).

THEOREM 4.4. *If* $\Gamma_6^n/n \Rightarrow \gamma_6$, *where* $\gamma_6 \equiv (q_1, q_2, m_1, z_{1,2}, 0, m_2 - z_{1,2})$ *with* $\gamma \equiv (q_1, q_2, z_{1,2}) \in \mathbb{A} \subset \mathbb{R}_3$ *for* $\mathbb{A}$ *in* (20) *and* $D_e^n(\Gamma^n, 0) \Rightarrow D(\gamma, 0)$ *in* $\mathbb{R}$ *as* $n \to \infty$, *where* D_e^n *is the time-expanded queue-difference process in* (12) *and* D *is the FTSP in* §3.2, *then*

$$\{D_e^n(\Gamma^n, s): s \geq 0\} \;\;\Rightarrow\;\; \{D(\gamma, s): s \geq 0\} \quad \text{in } \mathscr{D} \text{ as } n \to \infty; \tag{32}$$

i.e., we have convergence of the sequence of time-inhomogeneous non-Markov processes $\{D_e^n(\Gamma^n): n \geq 1\}$ *to the limiting time-homogeneous pure-jump Markov process* $D(\gamma)$.

The next results are about the queue-difference process $D_{1,2}^n$ itself (as opposed to the expanded queue difference process D_e^n). Recall the definition of stochastic boundedness in §2.4. Recall also that tightness in $\mathbb{R}$ is equivalent to stochastic boundedness in $\mathbb{R}$, but not in $\mathscr{D}$.

THEOREM 4.5 (STOCHASTIC BOUNDEDNESS OF $D_{1,2}^n$). *If* $x(t_0) \in \mathbb{A}$ *for some* $t_0 \geq 0$, *then there exists* $t_2 > t_0$ *such that* $x(t) \in \mathbb{A}$ *for all* $t \in [t_0, t_2]$ *and for all* t_1 *satisfying* $t_0 < t_1 \leq t_2$ *the following hold:*
  (i) $\{D_{1,2}^n(t): n \geq 1\}$ *is stochastically bounded in* $\mathbb{R}$ *for each* t *satisfying* $t_1 \leq t \leq t_2$.
  (ii) $\{\{D_{1,2}^n(t): t \in I\}: n \geq 1\}$ *is neither tight nor stochastically bounded in* $\mathscr{D}(I)$, $I \subseteq [t_0, t_2]$.
 (iii) *For any sequence* $\{c_n: n \geq 1\}$ *satisfying* $c_n/\log n \to \infty$ *as* $n \to \infty$, *it holds that*

$$\sup_{t_1 \leq t \leq t_2} \{D_{1,2}^n(t)/c_n\} \;\;\Rightarrow\;\; 0 \quad \text{as } n \to \infty. \tag{33}$$

If $x(t) \in \mathbb{A}$ *for all* $t \in [t_0, \infty)$ *the above statements hold for any finite* $t_2 > t_0$.

As an immediate corollary to (33) in Theorem 4.5, we have the following SSC of the queues. In particular, that claim implies SSC of the *fluid and diffusion scaled queues* when the fluid limit x is in $\mathbb{A}$.

COROLLARY 4.1 (SSC OF QUEUE PROCESS IN $\mathbb{A}$). *For the interval* $[t_1, t_2]$ *in Theorem* 4.5, $d_{J_1}(Q_1^n, r_{1,2}Q_2^n)/c_n \Rightarrow 0$ *in* $\mathscr{D}([t_1, t_2])$ *as* $n \to \infty$, *for every sequence* $\{c_n: n \geq 1\}$ *satisfying* $c_n/\log n \to \infty$ *as* $n \to \infty$. *If* $x(0) \in \mathbb{A}$ *and we consider the interval* $[0, t_2]$, *then the result is strengthened to hold on* $[t_0, t_2] \equiv [0, t_2]$.

Because the sequence of queue-difference processes is not $\mathscr{D}$ tight, by virtue of Theorem 4.5, we cannot have convergence of these processes in $\mathscr{D}$. However, we can obtain a proper limit for the tight sequence of random variables $\{D_{1,2}^n(t): n \geq 1\}$ in $\mathbb{R}$ for each fixed $t \in [t_1, t_2]$ by exploiting the AP. See Whitt [46] for a similar result.

THEOREM 4.6 (POINTWISE AP). *Consider the interval* $[t_1, t_2]$ *in Theorem* 4.5. *Then* $D_{1,2}^n(t) \Rightarrow D(x(t), \infty)$ *in* $\mathbb{R}$ *as* $n \to \infty$ *for each* t, $t_1 \leq t \leq t_2$, *where* $D(x(t), \infty)$ *has the limiting steady-state distribution of the FTSP* $D(\gamma, s)$ *for* $\gamma = x(t)$.

REMARK 4.2 (HITTING TIMES OF $\mathbb{A}$). First, the stochastic boundedness in Theorem 4.5 actually holds at time t_0 and thus in the larger interval $[t_0, t_2]$ if $t_0 = 0$ and $x(0) \in \mathbb{A}$, because of the assumed convergence of $D_{1,2}^n(0)$ in Assumption 3.2. However, we cannot get the full convergence in Theorem 4.6 at $t_0 = 0$ because the limit L in Assumption 3.2 need not be distributed the same as $D(x(0), \infty)$. Second, we may also have $t_0 > 0$ because t_0 is a hitting time of $\mathbb{A}$ from $\mathbb{S} - \mathbb{A}$. Even if $x(0) \in \mathbb{A}$, the fluid limit might leave $\mathbb{A}$ eventually, and later return to $\mathbb{A}$ at some time t_0; then $x(t_0) \in \mathbb{A}$ but $x(s) \notin \mathbb{A}$ for all $s \in (t_0 - \epsilon, t_0)$ for some $\epsilon > 0$. If t_0 is such a hitting time of $\mathbb{A}$, then we cannot obtain even a stochastic boundedness result at time t_0, but we obtain the pointwise convergence in Theorem 4.6 in the interval $(t_0, t_2]$, open on the left.

Finally, Theorem 4.6 can be applied to strengthen the conclusion of Theorem 4.4 by showing that $D_e^n(X^n(t), \cdot)$ converges to a *stationary* FTSP $D(x(t), \cdot)$, with $X^n(t) \equiv (Q_1^n(t), Q_2^n(t), Z_{1,2}^n(t))$, and $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$ is the limit of $\bar{X}^n(t)$ at the fixed time t.

COROLLARY 4.2. *Suppose that the condition of Theorem 4.5 holds. For each t such that the conclusion of Theorem 4.5 (i) holds for an interval $[t_1, t_2]$, $t_1 \le t \le t_2$,*

$$\{D_e^n(X^n(t), s): s \ge 0\} \implies \{D(x(t), s): s \ge 0\} \quad in \; \mathscr{D} \; as \; n \to \infty,$$

where the limiting FTSP $D(x(t), \cdot)$ is a stationary process, i.e., $D(x(t), s) \stackrel{d}{=} D(x(t), \infty)$ for all $s \ge 0$.

PROOF. First, for $X_6^n(t)$ as in Theorem 4.1 and $x_6(t) \equiv (q_1(t), q_2(t), m_1, z_{1,2}(t), 0, m_2 - z_{1,2}(t))$, we have $\Gamma_6^n/n \Rightarrow \gamma_6$ by Theorem 4.1, where $\Gamma_6^n \equiv X_6^n(t)$, $\gamma_6 \equiv x_6(t)$ and $\gamma \equiv x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$ is in $\mathbb{A}$ (because of our choice of t). Moreover,

$$D_e^n(X^n(t), 0) = D_{1,2}^n(t) \implies D(x(t), 0) \stackrel{d}{=} D(x(t), \infty) \quad in \; \mathbb{R} \; as \; n \to \infty,$$

where the first equality holds by the definition of D_e^n, and the limit holds by applying Theorem 4.6. Hence, the conditions in Theorem 4.4 hold, so that we have convergence in $\mathscr{D}$ of the process $D_e^n(X^n(t), \cdot)$ to the FTSP $D(x(t), \cdot)$. Since $D(x(t), 0) \stackrel{d}{=} D(x(t), \infty)$, the limiting FTSP is stationary as claimed. □

4.4. Overview of the proofs. The rest of this paper is devoted to proving Theorems 4.1–4.6. We prove Theorem 4.1 in §§5–7. Toward that end, in §5 we establish structural results for the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n): n \ge 1\}$, where $X_6^n \equiv (Q_i^n, Z_{i,j}^n) \in \mathscr{D}_6$ as in (2) and $Y_8^n \equiv (A_i^n, S_{i,j}^n, U_i^n) \in \mathscr{D}_8$, $i, j = 1, 2$ and the associated fluid-scaled process $(\bar{X}_6^n, \bar{Y}_8^n)$ in (26). In §5.1 we construct the stochastic processes (X_6^n, Y_8^n) in terms of rate-1 Poisson processes. In §5.2 we show that the sequence of stochastic processes $\{(\bar{X}_6^n, \bar{Y}_8^n): n \ge 1\}$ is C-tight in $\mathscr{D}_{14}$ and, consequently, there are convergent subsequences with smooth limits. In §6 we show that the representation established in §5.1 can be simplified under Assumptions 3.1–3.3, reducing the essential dimension from 6 to 3. The final three-dimensional representation $\bar{X}^n$ in (45) there explains the form of the ODE in (24).

Given the tightness established in §5.2, we prove the main Theorem 4.1 by characterizing the limit of all convergent subsequences in §7. Given the SSC established in §6 and given that the three-dimensional ODE in §3.3 has been shown to have a unique solution in Perry and Whitt [37], it suffices to show that the limit of any subsequence must almost surely be a solution to the ODE. For that last step, our proof in §7 follows Hunt and Kurtz [18], which draws heavily upon Kurtz [27]. It exploits our martingale representation in Theorem 6.3 and basic properties of random measures from Kurtz [27]. We also have developed an alternative proof exploiting stochastic bounds. It is given in §C in the appendix. Finally, in §8 we prove Theorems 4.3–4.6.

There is more in the appendix. In §A we present supporting technical results to prove the SSC results in §6. We start by introducing auxiliary frozen queue difference processes in §A.1. We construct useful bounding processes in §§A.2, A.3, and A.4. These are primarily for QBD processes, because we exploit a QBD representation for the FTSP; see §6 of Perry and Whitt [37] for background. We establish extreme value limits for QBD processes in §A.5. In §B we exploit the technical results in §A to prove three theorems stated in §6.

5. Preliminary results for X_6^n. In this section we establish preliminary structural results for the vector stochastic process (X_6^n, Y_8^n), where $X_6^n \equiv (Q_i^n, Z_{i,j}^n) \in \mathscr{D}_6$ as in (2) and $Y_8^n \equiv (A_i^n, S_{i,j}^n, U_i^n) \in \mathscr{D}_8$, $i, j = 1, 2$ and the associated fluid-scaled process $(\bar{X}_6^n, \bar{Y}_8^n)$ in (26). *The results in this section do not depend on Assumptions 3.1–3.3.* We do impose the many-server heavy-traffic scaling in §2.3.

In §5.1 we construct the stochastic processes (X_6^n, Y_8^n) in terms of rate-1 Poisson processes. In §5.2 we show that the sequence of stochastic processes $\{(\bar{X}_6^n, \bar{Y}_8^n): n \ge 1\}$ is C-tight in $\mathscr{D}_{14}$. In Corollary 5.1 we apply the tightness to deduce smoothness properties for the limits of convergent subsequences.

5.1. Representation of X_6^n. In this section we develop representations for the basic CTMC X_6^n with the FQR-T control. At first in this section we do not require Assumptions 3.1–3.3, so that we can have sharing in either direction, but in only one direction at any time. Let $(\lambda_1^n, \lambda_2^n)$ be the pair of fixed positive arrival rates in model n, which here are unconstrained.

Following common practice, as reviewed in §2 of Pang et al. [34], we represent the counting processes in terms of mutually independent rate-1 Poisson processes. We represent the counting processes A_i^n, $S_{i,j}^n$, and U_i^n introduced in the beginning of §4.1 as

$$A_i^n(t) \equiv N_i^a(\lambda_i^n t), \qquad S_{i,j}^n(t) \equiv N_{i,j}^s\left(\mu_{i,j} \int_0^t Z_{i,j}^n(s)\, ds\right), \qquad U_i^n(t) \equiv N_i^u\left(\theta_i \int_0^t Q_i^n(s)\, ds\right), \qquad (34)$$

for $t \geq 0$, 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.

We can then obtain a general representation of the CTMC X_6^n, which is actually valid for general arrival processes with arrivals one at a time. Let $S_1^n \equiv S_{1,1}^n + S_{2,1}^n$, $S_2^n \equiv S_{1,2}^n + S_{2,2}^n$ and $S^n \equiv S_1^n + S_2^n$. Paralleling (6), let $D_{2,1}^n(t) \equiv r_{2,1} Q_2^n(t) - k_{2,1}^n - Q_1^n(t)$.

THEOREM 5.1 (GENERAL REPRESENTATION OF X_6^n). *For each $n \geq 1$, the stochastic process X_6^n is well defined as a random element of $\mathbb{D}_6$ by (34) and*

$$Q_1^n(t) \equiv Q_1^n(0) + \int_0^t 1_{\{Z_{1,1}^n(s-) + Z_{2,1}^n(s-) = m_1^n\}}\, dA_1^n(s) - \int_0^t 1_{\{D_{1,2}^n(s-) > 0, Z_{2,1}^n(s-) = 0, Q_1^n(s-) > 0\}}\, dS^n(s)$$

$$- \int_0^t 1_{\{\{Q_1^n(s-) > 0\} \cap (\{Z_{2,1}^n(s-) > 0, D_{2,1}^n(s-) \leq 0\} \cup \{Z_{2,1}^n(s-) = 0, D_{1,2}^n(s-) \leq 0\})\}}\, dS_1^n(s) - U_1^n(t),$$

$$Z_{1,1}^n(t) \equiv Z_{1,1}^n(0) + \int_0^t 1_{\{Z_{1,1}^n(s-) + Z_{2,1}^n(s-) < m_1^n\}}\, dA_1^n(s)$$

$$- \int_0^t 1_{\{(\{Z_{1,2}^n(s-) > 0\} \cup \{D_{2,1}^n(s-) \leq 0\}) \cap \{Q_1^n(s-) = 0\}\}}\, dS_{1,1}^n(s) - \int_0^t 1_{\{D_{2,1}^n(s-) > 0, Z_{1,2}^n(s-) = 0\}}\, dS_{1,1}^n(s),$$

$$Z_{1,2}^n(t) \equiv Z_{1,2}^n(0) + \int_0^t 1_{\{D_{1,2}^n(s-) > 0, Z_{2,1}^n(s-) = 0, Q_1^n(s-) > 0\}}\, dS_{2,2}^n(s)$$

$$- \int_0^t 1_{\{\{D_{1,2}^n(s-) \leq 0\} \cup \{Z_{2,1}^n(s-) > 0\}\}}\, dS_{1,2}^n(s).$$

Symmetry yields the parallel definitions of $Q_2^n(t)$, $Z_{2,2}^n(t)$ and $Z_{2,1}^n(t)$ from $Q_1^n(t)$, $Z_{1,1}^n(t)$ and $Z_{1,2}^n(t)$ by simply switching the subscripts 1 and 2.

We remark that the representation of X_6^n in Theorem 5.1 holds even without Assumptions 3.1–3.3.

PROOF. Just as in Lemma 2.1 of Pang et al. [34], we can justify the construction by conditioning on the initial values (the first term in each display) and the counting processes. With these sample paths specified, we recursively construct the sample path of X_6^n. By applying mathematical induction over successive transition epochs of X_6^n, we show that the sample paths are right-continuous piecewise-constant functions satisfying the equations given.

To explain Q_1^n, the second term represents the increase by one at each class-1 arrival epoch when service pool 1 is fully occupied; otherwise the arrival would go directly into service pool one. The third term represents the decrease by one when any server completes service and sharing with pool 2 helping class 1 is actively taking place; that requires that the class-1 queue length be positive ($Q_1^n(s) > 0$); sharing with pool 2 helping class 1 occurs when both $D_{1,2}^n(s) > 0$ and $Z_{2,1}^n(s) = 0$. The fourth term represents the decrease by 1 when any pool-1 server completes service, provided that again the queue length is positive ($Q_1^n(s) > 0$). There are two scenarios: (i) $\{Z_{2,1}^n(s) > 0, D_{2,1}^n(s) \leq 0\}$ and (ii) $\{Z_{1,2}^n(s) = 0, D_{1,2}^n(s) \leq 0\}$. In the first, pool 1 is helping class 2, so type-1 servers take from queue 1 only when $D_{2,1}^n(s) \leq 0$. The second scenario is the relative complement within the event $\{Z_{2,1}^n(s) = 0, Q_1^n(s) > 0\}$ of the event in the third term, i.e., pool 2 is allowed to help class 1, but $D_{1,2}^n(s) \leq 0$, so that only type-1 servers take from queue 1 at time s.

To explain $Z_{1,1}^n$, the second term represents the increase by one that occurs at each class-1 arrival epoch at which service pool 1 has spare capacity ($Z_{1,1}^n(s) + Z_{2,1}^n(s) < m_1^n$). The third term represents the decrease by 1 that occurs when a server in pool 1 completes service of a class-1 customer, with pool 1 not helping class 2 ($\{Z_{1,2}^n(s) > 0\} \cup \{D_{2,1}^n(s) \leq 0\}$) when the class-1 queue is empty ($Q_1^n(s) = 0$). The fourth term represents the decrease by one that occurs when a server in pool 1 completes service of a class-1 customer, when pool 1 is helping class 2 ($\{D_{2,1}^n(s) > 0, Z_{1,2}^n(s) = 0\}$).

To explain $Z_{1,2}^n$, the second term represents the decrease by one that occurs when a server in pool 2 completes service of a class-2 customer, when class 2 is helping class 1 ($\{D_{1,2}^n(s) > 0, Z_{2,1}^n(s) = 0, Q_1^n(s) > 0\}$). The third term represents the decrease by one that occurs when a server in pool 2 completes service of a class-1 customer, when class 2 is not helping class 1 ($\{D_{1,2}^n(s) \leq 0\} \cup \{Z_{2,1}^n(s) > 0\}$).

Because the model is fully symmetric, the processes Q_2^n, $Z_{2,2}^n$, and $Z_{2,1}^n$ are the symmetric versions of Q_2^n, $Z_{2,2}^n$, and $Z_{2,1}^n$, respectively, with the indices 1 and 2 switched.   □

5.2. Tightness and smoothness of the limits.   We do part of the proof of Theorem 4.1 here by establishing tightness. For background on tightness, see Billingsley [7], Pang et al. [34], and Whitt [47]. We recall a few key facts: tightness of a sequence of k-dimensional stochastic processes in $\mathcal{D}_k$ is equivalent to tightness of all the one-dimensional component stochastic processes in $\mathcal{D}$. For a sequence of random elements of $\mathcal{D}_k$, $\mathcal{C}$-tightness implies $\mathcal{D}$-tightness and that the limits of all convergent subsequences must be in $\mathcal{C}_k$; see Theorem 15.5 of the first 1968 edition of Billingsley [7]. Alternatively, Conditions (7.6) and (7.7) of Theorem 7.3 in Billingsley [7] hold for processes in $\mathcal{D}$ if and only if conditions (13.4) and (13.5) of Theorem 13.2 of Billingsley [7] hold and the limits of all convergent subsequences are in $\mathcal{C}$; see the Corollary on p. 179 of Billingsley [7] or Theorem VI.3.26 of Jacod and Shiryaev [19].

THEOREM 5.2.   *The sequence $\{(\bar{X}_6^n, \bar{Y}_8^n): n \geq 1\}$ in* (26) *is $\mathcal{C}$-tight in $\mathcal{D}_{14}$.*

PROOF.   It suffices to verify conditions (6.3) and (6.4) of Theorem 11.6.3 of Whitt [47], namely, to show that $\bar{X}^n(0)$ is stochastically bounded (tight in $\mathbb{R}_6$) and appropriately controls the oscillations, using the modulus of continuity on $\mathcal{C}$. We obtain the stochastic boundedness at time 0 immediately from Assumption 3.2.

We now show that we can control the oscillations below. For that purpose, let $w(x, \zeta, T)$ be the modulus of continuity of the function $x \in \mathcal{D}$, i.e.,

$$w(x, \zeta, T) \equiv \sup\{|x(t_2) - x(t_1)|: 0 \leq t_1 \leq t_2 \leq T, |t_2 - t_1| \leq \zeta\}. \tag{35}$$

Using the representations in §5.1, for $t_2 > t_1 \geq 0$ we have

$$|\bar{Q}_1^n(t_2) - \bar{Q}_1^n(t_1)| \leq \frac{A_1^n(t_2) - A_1^n(t_1)}{n} + \frac{S^n(t_2) - S^n(t_1)}{n} + \frac{S_{1,1}^n(t_2) - S_{1,1}^n(t_1)}{n} + \frac{U_1^n(t_2) - U_1^n(t_1)}{n}$$

and similarly for $\bar{Q}_2^n$. Hence, for any $\zeta > 0$ and $T > 0$,

$$w(Q_1^n/n, \zeta, T) \leq w(A_1^n/n, \zeta, T) + w(S^n/n, \zeta, T) + w(S_{1,1}^n/n, \zeta, T) + w(U_1^n/n, \zeta, T).$$

Then observe that we can bound the oscillations of the service processes $S_{i,j}^n$ by the oscillations in the scaled Poisson process $N_{i,j}^s(n\cdot)$. In particular, by (34),

$$w(S_{i,j}^n/n, \zeta, T) \leq w(N_{i,j}^s(n\mu_{i,j}m_j\cdot)/n, \zeta, T) \leq w(N_{i,j}^s(n\cdot)/n, c\zeta, T) \tag{36}$$

for some constant $c > 0$. Next for the abandonment process U_i^n, we use the elementary bounds

$$Q_i^n(t) \leq Q_i^n(0) + A_i^n(t),$$

$$|U_i^n(t_2) - U_i^n(t_1)| = \left|N_i\left(\theta_i \int_{t_1}^{t_2} Q_i^n(s)\, ds\right)\right| \leq |N_i(n\theta(\bar{Q}_i^n(0) + \bar{A}_i^n(T))(t_2 - t_1))|.$$

Let $q_{bd} \equiv 2(q_i(0) + T)$, where $\bar{Q}_i^n(0) \Rightarrow q_i(0)$ by Assumption 3.2, and let B_n be the following subset of the underlying probability space:

$$B_n \equiv \{\bar{Q}_i^n(0) + \bar{A}_i^n(T) \leq q_{bd}\}.$$

Then $P(B_n) \to 1$ as $n \to \infty$ and, on the set B_n, we have

$$w(U_i^n/n, \zeta, T) \leq w(N_i^u(nq_{bd}\cdot)/n, \zeta, T) \leq w(N_i^u(n\cdot)/n, c\zeta, T) \tag{37}$$

for some constant $c > 0$.

Thus, there exists a constant $c > 0$ such that, for any $\eta > 0$, there exists n_0 and $\zeta > 0$ such that, for all $n \geq n_0$, $P(B_n) > 1 - \eta/2$ and on B_n

$$w(Q_i^n/n, \zeta, T) \leq w(N_i^a(n\cdot)/n, c\zeta, T) + 2\sum_{i=1}^{2}\sum_{j=1}^{2} w(N_{i,j}^s(n\cdot)/n, c\zeta, T) + w(N_i^u(n\cdot)/n, c\zeta, T).$$

However, by the FWLLN for the Poisson processes, we know that we can control all these moduli of continuity on the right. Thus we deduce that, for every $\epsilon > 0$ and $\eta > 0$, there exists $\zeta > 0$ and n_0 such that

$$P(w(Q_i^n/n, \zeta, T) \geq \epsilon) \leq \eta \quad \text{for all } n \geq n_0.$$

Hence, we have shown that the sequence $\{\bar{Q}_i^n\}$ is tight.

We now turn to the sequence $\{\bar{Z}_{1,2}^n\}$. Let $A_{1,2}^n(t)$ denote the total number of class-1 arrivals up to time t, who will eventually be served by type-2 servers in system n. Let $\bar{A}_{1,2}^n \equiv A_{1,2}^n/n$ and $\bar{S}_{1,2}^n(t) \equiv S_{1,2}^n(t)/n$, for $S_{1,2}^n(t)$ in (34). Since

$$Z_{1,2}^n(t) = Z_{1,2}^n(0) + A_{1,2}^n(t) - S_{1,2}^n(t),$$

we have

$$|\bar{Z}_{1,2}^n(t_2) - \bar{Z}_{1,2}^n(t_1)| \leq \bar{A}_{1,2}^n(t_2) - \bar{A}_{1,2}^n(t_1) + \bar{S}_{1,2}^n(t_2) - \bar{S}_{1,2}^n(t_1).$$

However, for A_1^n in (34),

$$A_{1,2}^n(t_2) - A_{1,2}^n(t_1) \leq A_1^n(t_2) - A_1^n(t_1).$$

Since $\bar{A}_1^n \Rightarrow \lambda_1 e$ in $\mathcal{D}$, the sequence $\{\bar{A}_1^n\}$ is tight. Together with (36), that implies that the sequence $\{\bar{Z}_{1,2}^n\}$ is tight as well. Finally, we observe that the tightness of $\{\bar{Y}_8^n\}$ follows from (36), (37), and the convergence of $\bar{A}_i^n$. $\square$

Because the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n): n \geq 1\}$ in (26) is $\mathscr{C}$-tight by Theorem 5.2, every subsequence that a further subsequence that converges to a continuous limit. We now apply the modulus-of-continuity inequalities established in the proof of Theorem 5.2 to deduce additional smoothness properties of the limits of all converging subsequence.

COROLLARY 5.1. *If $(\bar{X}_6, \bar{Y}_8)$ is the limit of a subsequence of $\{(\bar{X}_6^n, \bar{Y}_8^n): n \geq 1\}$ in $\mathcal{D}_{14}$, then each component in $\mathcal{D}$, say $\bar{X}_i$, has bounded modulus of continuity; i.e., for each $T > 0$, there exists a constant $c > 0$ such that*

$$w(\bar{X}_i, \zeta, T) \leq c\zeta \quad w.p.1 \tag{38}$$

for all $\zeta > 0$. Hence $(\bar{X}_6, \bar{Y}_8)$ is Lipschitz continuous w.p.1, and is thus differentiable almost everywhere.

PROOF. Apply the bounds on the modulus of continuity involving Poisson processes in the proof of Theorem 5.2. For a Poisson process N, let $\hat{N}^n \equiv \sqrt{n}(\bar{N}^n - e)$, where $\bar{N}^n(t) \equiv N(nt)/n$, $t \geq 0$. By the triangle inequality, for each n, ζ, and T,

$$w(\bar{N}^n, \zeta, T) \leq \frac{w(\hat{N}^n, \zeta, T)}{\sqrt{n}} + w(e, \zeta, T) \quad \Rightarrow \quad \zeta \quad \text{as } n \to \infty.$$

Because $w(x, \zeta, T)$ is a continuous function of x for each fixed ζ and T, we can apply this bound with the inequalities in the proof of Theorem 5.2 to deduce (38). $\square$

We remark in closing this section that Theorem 5.2 and Corollary 5.1 also hold with Assumption 3.2 replaced by $\bar{X}^n(0) \Rightarrow x(0)$ as $n \to \infty$, where $x(0)$ is a deterministic element of $\mathbb{R}_3$.

6. Structural simplification. We now exploit Assumptions 3.1–3.3 to simplify the representation established in §5.1 above, reducing the essential dimension from six to three, following the plan described in §3.1. We first establish this dimension reduction over an interval $[0, \tau]$ and later, after Theorem 4.1 has been proved over the same interval $[0, \tau]$, we show that all the results here, and thus Theorem 4.1 too, can be extended to the interval $[0, \infty)$.

Let

$$\mathcal{T}_0^n \equiv \inf\{t > 0: Z_{2,1}^n(t) > 0 \text{ or } Q_1^n(t) = 0 \text{ or } Q_2^n(t) = 0\}. \tag{39}$$

By Assumption 3.2, both queues are initially strictly positive (so there is no idleness in either pool) and $Z_{2,1}^n(0) = 0$. Hence, $\mathcal{T}_0^n > 0$ for each $n \geq 1$. Theorem 5.1 with the definitions in (34) implies the following reduction from six dimensions to three over $[0, \mathcal{T}_0^n]$. Let $\stackrel{\mathrm{d}}{=}$ denote equality in distribution for processes.

COROLLARY 6.1. *On the random interval* $[0, \mathcal{T}_0^n]$, $X_6^n = X_6^{n,**}$ *w.p.*1, *where*

$$X_6^{n,**} \equiv (Q_1^n, Q_2^n, m_{1,1}^n e, Z_{1,2}^n, 0e, m_2^n e - Z_{1,2}^n),$$

with

$$Q_1^n(t) \equiv Q_1^n(0) + A_1^n(t) - \int_0^t 1_{\{D_{1,2}^n(s-)>0\}} \, dS^n(s) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}} \, dS_{1,1}^n(s) - U_1^n(t)$$

$$Q_2^n(t) \equiv Q_2^n(0) + A_2^n(t) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}} \, dS_{2,2}^n(s) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}} \, dS_{1,2}^n(s) - U_2^n(t) \tag{40}$$

$$Z_{1,2}^n(t) \equiv Z_{1,2}^n(0) + \int_0^t 1_{\{D_{1,2}^n(s-)>0\}} \, dS_{2,2}^n(s) - \int_0^t 1_{\{D_{1,2}^n(s-)\le 0\}} \, dS_{1,2}^n(s),$$

and $X_6^{n,**} \stackrel{d}{=} X_6^{n,*}$, *where* $X_6^{n,*} \equiv (Q_1^{n,*}, Q_2^{n,*}, m_{1,1}^n e, Z_{1,2}^{n,*}, 0e, m_2^n e - Z_{1,2}^{n,*})$ *with*

$$Q_1^{n,*}(t) \equiv Q_1^{n,*}(0) + N_1^a(\lambda_1^n t) - N_{1,1}^s(\mu_{1,1} m_1^n t) - N_{1,2}^{s,2}\left(\mu_{1,2} \int_0^t 1_{\{D_{1,2}^{n,*}(s)>0\}} Z_{1,2}^{n,*}(s)) \, ds\right)$$

$$- N_{2,2}^s\left(\mu_{2,2} \int_0^t 1_{\{D_{1,2}^{n,*}(s)>0\}} (m_2^n - Z_{1,2}^{n,*}(s)) \, ds\right) - N_1^u\left(\theta_1 \int_0^t Q_1^{n,*}(s) \, ds\right), \tag{41}$$

$$Q_2^{n,*}(t) \equiv Q_2^{n,*}(0) + N_2^a(\lambda_2^n t) - N_{2,2}^{s,2}\left(\mu_{2,2} \int_0^t 1_{\{D_{1,2}^{n,*}(s)\le 0\}} (m_2^n - Z_{1,2}^{n,*}(s)) \, ds\right)$$

$$- N_{1,2}^s\left(\mu_{1,2} \int_0^t 1_{\{D_{1,2}^{n,*}(s)\le 0\}} Z_{1,2}^{n,*}(s) \, ds\right) - N_2^u\left(\theta_2 \int_0^t Q_2^{n,*}(s) \, ds\right), \tag{42}$$

$$Z_{1,2}^{n,*}(t) \equiv Z_{1,2}^{n,*}(0) + N_{2,2}^s\left(\mu_{2,2} \int_0^t 1_{\{D_{1,2}^{n,*}(s)>0\}} (m_2^n - Z_{1,2}^{n,*})(s) \, ds\right)$$

$$- N_{1,2}^s\left(\mu_{1,2} \int_0^t 1_{\{D_{1,2}^{n,*}(s)\le 0\}} Z_{1,2}^{n,*}(s) \, ds\right), \tag{43}$$

where N_i^a, N_i^u, $N_{1,1}^s$, $N_{i,2}^s$, $N_{i,2}^{s,2}$ *for* $i = 1, 2$ *are mutually independent rate*-1 *Poisson Processes and* $D_{1,2}^{n,*}(t) \equiv Q_1^{n,*}(t) - k_{1,2}^n - r_{1,2} Q_2^{n,*}(t)$ *as in* (6).

Note that in two places in the three displays (41)–(43) we have introduced the new independent rate-1 Poisson processes $N_{i,2}^{s,2}$. Note that $Z_{1,2}^{n,*}(t)$ might be equal to zero for some or all t in $[0, \mathcal{T}_0^n]$.

We next prove that $\mathcal{T}_0^n$ is bounded away from zero asymptotically, i.e., that there exists a $\tau > 0$ such that $P(T_0^n \ge \tau) \to 1$ as $n \to \infty$. We do so in two parts (both proved in §B):

THEOREM 6.1 (NO SHARING IN THE OPPOSITE DIRECTION). *There exists* $\tau > 0$ *such that* $\|Z_{2,1}^n\|_\tau \Rightarrow 0$ *as* $n \to \infty$.

THEOREM 6.2 (POSITIVE QUEUE LENGTHS). *For* τ *in Theorem* 6.1,

$$P\left(\inf_{0\le t\le \tau} \min\{Q_1^n(t), Q_2^n(t)\} > 0\right) \to 1 \quad \text{as } n \to \infty.$$

As an immediate consequence of Theorems 6.1 and 6.2, we obtain the following SSC result.

COROLLARY 6.2 (SSC OF THE SERVICE PROCESS). *For* $\tau > 0$ *in Theorem* 6.1, $P(\mathcal{T}_0^n > \tau) \to 1$ *as* $n \to \infty$, *where* $\mathcal{T}_0^n$ *is defined in* (39); *i.e.*,

$$(m_1^n e - Z_{1,1}^n, Z_{2,1}^n, m_2^n e - Z_{1,2}^n - Z_{2,2}^n) \Rightarrow (0e, 0e, 0e) \quad \text{in } \mathcal{D}_3([0, \tau]) \text{ as } n \to \infty.$$

We make two important remarks about Corollary 6.2: First, the limit holds *without any scaling*. Second, here we do not yet show that a limit of $\bar{Z}_{1,2}^n$ as $n \to \infty$ exists. We only show that, when analyzing the four service processes $Z_{i,j}^n$, it is sufficient to consider $Z_{1,2}^n$.

Recall that d_{J_1} denotes the standard Skorohod J_1 metric and $X_6^{n,*}$ is the essentially three-dimensional process defined in (10). The following corollary is immediate from Corollary 6.2.

COROLLARY 6.3 (REPRESENTATION VIA SSC). *As* $n \to \infty$, $d_{J_1}(X_6^n, X_6^{n,*}) \Rightarrow 0$ *in* $\mathcal{D}_6([0, \tau])$, *for* $X_6^{n,*}$ *in* (10), τ *in Theorem* 6.1 *and* $(Q_1^{n,*}, Q_2^{n,*}, Z_{1,2}^{n,*})$ *in* (41)–(43).

We now obtain further simplification using a familiar martingale representation, again see Pang et al. [34]. *Henceforth, we work with the process $X_6^{n,*}$ defined in Corollary* 6.1, *but omit the asterisks.* Consider the representation of X_3^n in (41)–(43), and let

$$
\begin{aligned}
M_i^{n,a}(t) &\equiv N_i^a(\lambda_i^n t) - \lambda_i^n t, \\
M_i^{n,u}(t) &\equiv N_i^u\left(\theta_i \int_0^t Q_i^n(s)\,ds\right) - \theta_i \int_0^t Q_i^n(s)\,ds, \\
M_{i,2}^{n,s}(t) &\equiv N_{i,2}^s(J_{i,2}^n(t)) - J_{i,2}^n(t), \\
M_{i,2}^{n,s,2}(t) &\equiv N_{i,2}^{s,2}(J_{i,2}^n(t)) - J_{i,2}^n(t),
\end{aligned}
\tag{44}
$$

where $J_{i,2}^n(t)$ are the compensators of the point processes in (41)–(43), $i = 1, 2$, e.g.,

$$
J_{1,2}^n(t) \equiv \mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s)<0\}} Z_{1,2}^n(s)\,ds.
$$

The quantities in (44) can be shown to be martingales (with respect to an appropriate filtration); see Pang et al. [34]. The following lemma follows easily from the functional strong law of large numbers (FSLLN) for Poisson processes and the $\mathscr{C}$-tightness established in Theorem 5.2.

LEMMA 6.1 (FLUID LIMIT FOR THE MARTINGALE TERMS). *As $n \to \infty$,*

$$
n^{-1}(M_1^{n,a}, M_2^{n,a}, M_1^{n,u}, M_2^{n,u}, M_{1,2}^{n,s}, M_{2,2}^{n,s}, M_{1,2}^{n,s,2}, M_{2,2}^{n,s,2}) \;\Rightarrow\; (0e, 0e, 0e, 0e, 0e, 0e, 0e, 0e) \quad \text{in } \mathscr{D}_8([0,\tau]).
$$

PROOF. By Theorem 5.2, the sequence $\{\bar{X}_6^n: n \geq 1\}$ is tight in $\mathscr{D}$. Thus any subsequence has a convergent subsequence. By the proof of Theorem 5.2, the sequences $\{J_{i,j}^n/n\}$ are also $\mathscr{C}$-tight, so that $\{J_{i,j}^n/n\}$, $i = 1, 2$, all converge along a converging subsequence as well. Consider a converging subsequence $\{X^n\}$ and its limit $\bar{X}$, which is continuous by Theorem 5.2. Then the claim of the lemma follows for the converging subsequence from the FSLLN for Poisson processes and the continuity of the composition map at continuous limits, e.g., Theorem 13.2.1 in Whitt [47]. In this case, the limit of each fluid-scaled martingale is the zero function $0e \in \mathscr{D}$, regardless of the converging subsequence we consider, and is thus unique. Hence we have completed the proof. $\square$

Hence, instead of $\bar{X}_3^n$ (the relevant components of $\bar{X}_6^{n,*}$) in (41)–(43), we can work with $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ for

$$
\begin{aligned}
\bar{Z}_{1,2}^n(t) &\equiv \bar{Z}_{1,2}^n(0) + \mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(\bar{m}_2^n - \bar{Z}_{1,2}^n(s))\,ds - \mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\leq 0\}}\bar{Z}_{1,2}^n(s)\,ds, \\
\bar{Q}_1^n(t) &\equiv \bar{Q}_1^n(0) + \bar{\lambda}_1^n t - \bar{m}_1^n t - \mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}\bar{Z}_{1,2}^n(s)\,ds \\
&\quad - \mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(\bar{m}_2^n - \bar{Z}_{1,2}^n(s))\,ds - \theta_1\int_0^t \bar{Q}_1^n(s)\,ds, \\
\bar{Q}_2^n(t) &\equiv \bar{Q}_2^n(0) + \bar{\lambda}_2^n t - \mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)\leq 0\}}(\bar{m}_2^n - \bar{Z}_{1,2}^n(s))\,ds \\
&\quad - \mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\leq 0\}}\bar{Z}_{1,2}^n(s)\,ds - \theta_2\int_0^t \bar{Q}_2^n(s))\,ds.
\end{aligned}
\tag{45}
$$

THEOREM 6.3. *As $n \to \infty$, $d_{J_1}(\bar{X}_3^n, \bar{X}^n) \Rightarrow 0$ in $\mathscr{D}_3([0,\tau])$ as $n \to \infty$, where $\bar{X}_3^n$ is defined in* (41)–(43), *$\bar{X}^n$ is defined in* (45) *and τ is as in Theorem* 6.1.

PROOF. It suffices to show that $\bar{M}^n \Rightarrow (0e, 0e, 0e)$ in $\mathscr{D}_3([0,\tau])$ as $n \to \infty$, where

$$
\bar{M}^n \equiv \bar{X}_3^n - \bar{X}^n.
\tag{46}
$$

However, $\bar{M}^n \Rightarrow (0e, 0e, 0e)$ by virtue of Lemma 6.1 by the continuous mapping theorem with addition at continuous limits. $\square$

As a consequence of Theorem 6.3, henceforth we can focus on $\bar{X}^n$ in (45) instead of $\bar{X}_3^n$ in (41)–(43). After we prove Theorem 4.1, we can extend the interval $[0, \tau]$ over which all the previous results in this section hold to the interval $[0, \infty)$.

THEOREM 6.4 (GLOBAL SSC). *All of the results above in this section extend from the interval $[0, \tau]$ to the interval $[0, \infty)$.*

We prove Theorems 6.1 and 6.2 in §B, after establishing supporting technical results in §A. In §B.3, we prove Theorem 6.4, under the assumption that Theorem 4.1 has been proved over $[0, \tau]$, which will be done in §7, by showing that the interval over which the conclusion is valid can be extended from $[0, \tau]$ to $[0, \infty)$, once Theorem 4.1 has been proved over $[0, \tau]$. That will imply that Theorem 4.1 then holds over $[0, \infty)$ as well.

7. Completing the proof of Theorem 4.1: Characterization. The $\mathscr{C}$-tightness result in Theorem 5.2 implies that every subsequence of the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n): n \geq 1\}$ in (26) has a further converging subsequence in $\mathscr{D}_{14}([0, \infty))$, whose limit is in the function space $\mathscr{C}_{14}([0, \infty))$. To establish the convergence of the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n)\}$ we must show that every converging subsequence converges to the same limit.

Corollary 6.3 and Theorem 6.3 imply that we can simplify the framework over an initial time interval $[0, \tau]$. In particular, it suffices to focus on the sequence $\{\bar{X}^n\}$ in $\mathscr{D}_3([0, \tau])$ given in (45), where the limits of the subsequences will be in $\mathscr{C}_3([0, \tau])$, but we will later show that this restriction can be relaxed. In particular, we will show that convergence holds in $\mathscr{D}_{14}([0, \infty))$.

We achieve the desired characterization of $\bar{X}^n$ in $\mathscr{D}_3([0, \tau])$ by showing that the limit of any converging subsequence is almost surely a solution to the ODE (24) over $[0, \tau]$. The existence and uniqueness of the solution to the ODE has been established in Theorem 5.2 in Perry and Whitt [37]. As indicated in §4.4, we have developed two different proofs, the first proof exploits random measures and martingales (Hunt and Kurtz [18], Kurtz [27]); it is given here in §§7.2 and 7.3. The second exploits stochastic bounds as in the proof of Lemma B.3; it is given in Appendix C. Both proofs start from the following subsection.

7.1. Proof of Theorem 4.1: Reduction to integral terms. Let $\bar{X}$ be the limit of a converging subsequence of $\{\bar{X}^n: n \geq 1\}$ in (45) in $\mathscr{D}([0, \tau])$ for τ in Theorem 6.3. We consider $n \geq 1$ with the understanding that the limit is through a subsequence. Many of the terms in (45) converge directly to their counterparts in (25) because of the assumed many-server heavy-traffic scaling in §2.3 and the convergence $\bar{X}^n \Rightarrow \bar{X}$ through the subsequence obtained from the tightness. Indeed, the only exceptions are the integral terms involving the indicator functions. Let $\bar{I}_{z,i}^n$, $\bar{I}_{q,1,i}^n$, and $\bar{I}_{q,2,i}^n$ be the ith integral term in the respective expression for $\bar{Z}_{1,2}^n$, $\bar{Q}_1^n$, and $\bar{Q}_2^n$ in (45), $i = 1, 2$. We first observe that these sequences of integral terms are tight.

LEMMA 7.1 (TIGHTNESS OF INTEGRAL TERMS). *The six sequences of integral processes $\{\bar{I}_{z,i}^n: n \geq 1\}$, $\{\bar{I}_{q,1,i}^n: n \geq 1\}$, and $\{\bar{I}_{q,2,i}^n: n \geq 1\}$ involving the indicator functions appearing in (45) are each $\mathscr{C}$-tight in $\mathscr{D}([0, \tau])$.*

PROOF. We consider only the integral term $\bar{I}_{q,1,1}^n$, because the others are treated in the same way. First, boundedness is elementary: $0 \leq \bar{I}_{q,1,1}^n(t) \leq tm_2^n/n$. Second, the modulus in (35) is easily controlled: $w(\bar{I}_{q,1,1}^n, \zeta, T) \leq \zeta m^n/n \to \zeta m_2$. $\square$

Hence, we can consider a subsequence of our original converging subsequence in which all these integral terms converge to proper limits as well. Hence we have the following expression for $\bar{X}$, the limit of the converging subsequence:

$$\bar{Z}_{1,2}(t) = z_{1,2}(0) + \mu_{2,2}\bar{I}_{z,1}(t) - \mu_{1,2}\bar{I}_{z,2}(t)$$

$$\bar{Q}_1(t) = q_1(0) + \bar{\lambda}_1 t - \bar{m}_1 t - \mu_{1,2}\bar{I}_{q,1,1}(t) - \mu_{2,2}\bar{I}_{q,1,2}(t) - \theta_1 \int_0^t \bar{Q}_1(s)\, ds, \qquad (47)$$

$$\bar{Q}_2(t) = q_2(0) + \bar{\lambda}_2 t - \mu_{2,2}\bar{I}_{q,2,1}(t) - \mu_{1,2}\bar{I}_{q,2,2}(t) - \theta_2 \int_0^t \bar{Q}_2(s)\, ds.$$

In (47), we have exploited the assumed convergence of the initial conditions in Assumption 3.2 to replace $\bar{X}(0)$ by $x(0)$ in (47). At this point, it only remains to show that the terms $\bar{I}_{z,i}$, $\bar{I}_{q,1,i}$, and $\bar{I}_{q,2,i}$ appearing in (47) necessarily coincide almost surely with the corresponding terms in the integral representation (25) associated with the ODE in (24) over the interval $[0, \tau]$. That will uniquely characterize the limit over that initial interval $[0, \tau]$ because, by Theorem 5.2 of Perry and Whitt [37], there exists a unique solution to the ODE.

Thus, it suffices to establish the following lemma, which we do in two different ways, one in the next two sections and the other in Appendix C.

LEMMA 7.2 (REPRESENTATION OF LIMITING INTEGRAL TERMS). *For τ in Theorem 6.3, the integral terms in (47) necessarily coincide with the corresponding integral terms in (25) with $\bar{X}$ substituted for x for $0 \leq t \leq \tau$, e.g.,*

$$\bar{I}_{q,1,1}(t) = \int_0^t \pi_{1,2}(\bar{X}(s))\bar{Z}_{1,2}(s)\, ds, \quad 0 \leq t \leq \tau, \ w.p.1. \qquad (48)$$

7.2. Exploiting random measures. For the rest of our proof of Lemma 7.2, we closely follow Hunt and Kurtz [18], which applies Kurtz [27]. There are two key steps: (i) exploiting random measures and (ii) applying a martingale representation to characterize the limit in terms of the steady-state distribution of the FTSP in §3.2. This subsection is devoted to the first step; the next subsection is devoted to the second step.

We now introduce random measures in order to expose additional structure in the integral term $\bar{I}_{q,1,1}$ in (47). The random measures will be defined by setting

$$\nu^n([0, t] \times B) \equiv \int_0^t 1_{\{D_{1,2}^n(s) \in B\}} \, ds, \quad t \geq 0, \tag{49}$$

where B is a measurable subset of the set E of possible values of $D_{1,2}^n$. Given the definition of $D_{1,2}^n$ in (6), we exploit Assumption 3.3 stating that $r_{1,2}$ is rational and the assumption after it that the thresholds are rational as well, so that we can have the state space E be discrete, independent of n. (This property is not necessary for the analysis at this point, but it is a helpful simplification. The space E could even be taken to be a subset of $\mathbb{Z}$, using the construction in §6 of Perry and Whitt [37] by renaming the states in E.) Of course, we are especially interested in the case in which B is the set of positive values; then we focus on the associated random variables $\nu^n([0, t] \times (0, \infty))$, $n \geq 1$, as in (45).

As in Hunt and Kurtz [18] and Kurtz [27], it is convenient to compactify the space E. We do that here, first, by adding the states $+\infty$ and $-\infty$ to E and, second, by endowing E with the metric and associated Borel σ-field from $\mathbb{R}$ induced by the mapping $\psi: E \to [-1, 1]$ defined by $\psi(x) = x/(1 + |x|)$. That makes E a compact metric space. We then consider the space $M \equiv M(S)$ of (finite) measures μ on the product space $S \equiv [0, \delta] \times E$ for some $\delta > 0$, such that $\mu([0, t] \times E) = t$ for all $t > 0$. Moreover, we endow M with the Prohorov metric, as in (1.1) of Kurtz [27]. Because S is compact, the space M inherits the compactness; i.e., it too is a compact metric space, by virtue of Prohorov's theorem, Theorem 11.6.1 of Whitt [47].

Let $\mathcal{P} \equiv \mathcal{P}(M) \equiv \mathcal{P}(M(S))$ be the space of probability measures on $M(S)$, also made into a metric space with the Prohorov metric, so that convergence corresponds to the usual notion of weak convergence of probability measures. As a compact metric space, $M(S)$ is a complete separable metric space, so that this is a standard abstract setting for weak convergence of probability measures (Billingsley [7], Ethier and Kurtz [12], Whitt [47]). By Prohorov's theorem, this space $\mathcal{P}$ of probability measures on $M(S)$ also is a compact metric space.

Thus we have convergence of random measures $\nu^n \Rightarrow \nu$ as $n \to \infty$ if and only if $E[f(\nu^n)] \to E[f(\nu)]$ as $n \to \infty$ for all continuous bounded real-valued functions f on M. On the other hand, by the continuous mapping theorem, if we have $\nu^n \Rightarrow \nu$ as $n \to \infty$, then we also have $f(\nu^n) \Rightarrow f(\nu)$ as $n \to \infty$ for each continuous function on M. One reason that the random measure framework is convenient is that each continuous bounded real-valued function f on S corresponds to a continuous real-valued function on $M(S)$ via the integral representation

$$f(\mu) \equiv \int_S f(s) \, d\mu(s).$$

As a consequence, if $\nu^n \Rightarrow \nu$ as $n \to \infty$, then necessarily also

$$\int_S f(s) \, d\nu^n(s) \ \Rightarrow \ \int_S f(s) \, d\nu(s) \quad \text{in } \mathbb{R} \text{ as } n \to \infty$$

for all continuous bounded real-valued functions f on S.

In our context it is important to observe what are the continuous functions on E after the compactification above. The new topology requires that the functions have finite limits as $k \to +\infty$ or as $k \to -\infty$. All the functions we consider will be continuous on E because they take constant values outside bounded sets. We will use the following stronger result, which is a special case of Lemma 1.5 of Kurtz [27].

Lemma 7.3 (Extended Continuous Mapping Theorem). *If f is a continuous bounded real-valued function on S and $\{f_n: n \geq 1\}$ is a sequence of measurable real-valued functions on S such that $\|f_n - f\|_S \to 0$ as $n \to \infty$, then*

$$\int_S f_n(s) \, d\nu^n(s) \ \Rightarrow \ \int_S f(s) \, d\nu(s) \quad \text{in } \mathbb{R} \text{ as } n \to \infty.$$

Proof. First, starting from the convergence $\nu^n \Rightarrow \nu$, apply the Skorohod representation theorem to obtain versions converging w.p.1, without changing the notation. Then, by the triangle inequality,

$$\left| \int_S f_n(s) \, d\nu^n(s) - \int_S f(s) \, d\nu(s) \right| \leq \left| \int_S f_n(s) \, d\nu^n(s) - \int_S f(s) \, d\nu^n(s) \right| + \left| \int_S f(s) \, d\nu^n(s) - \int_S f(s) \, d\nu(s) \right|$$

$$\leq \|f_n - f\|_S \nu^n(S) + \left| \int_S f(s) \, d\nu^n(s) - \int_S f(s) \, d\nu(s) \right| \to 0 \quad \text{as } n \to \infty,$$

using the uniform convergence condition and the limit $\nu^n(S) \to \nu(S) < \infty$ in the first term and the continuous mapping in the second term. This shows convergence w.p.1 for the special versions and thus the claimed convergence in distribution. $\square$

The random measures we consider are random elements of $M(S)$, where S is the product space $[0, \delta] \times E$. These random measures can be properly defined by giving their definition for all product sets $[0, t] \times B$, as we have done in (49). The usual extension produces full random elements of $M(S)$.

We exploit the compactness of E introduced above to obtain compactness of $\mathcal{P}(M)$ and thus relative compactness of the sequence $\{(\bar{X}_6^n, \nu^n); n \geq 1\}$.

LEMMA 7.4 (RELATIVE COMPACTNESS). *The sequence $\{(\bar{X}_6^n, \nu^n); n \geq 1\}$ defined by (26) and (49) is relatively compact in $\mathcal{D}_6([0, \delta]) \times M(S)$ for any $\delta > 0$.*

PROOF. We already have observed that, because of the compactness imposed on E, the space $\mathcal{P}(M(S))$ is compact. By Theorem 5.2, the sequence $\{\bar{X}_{1,2}^n : n \geq 1\}$ is tight and thus relatively compact. However, relative compactness of the components implies relative compactness of the vectors. Thus the sequence $\{(\bar{X}_6^n, \nu^n); n \geq 1\}$ defined by (26) and (49) is relatively compact in $\mathcal{D}_6([0, \delta]) \times M(S)$. $\square$

Another crucial property of the random measures on the product space is that the random measures themselves admit a product representation or factorization, as indicated by Lemma 1.4 of Kurtz [27]; also see Lemma 2 of Hunt and Kurtz [18]. This result requires filtrations. For that, we observe that X_6^n is a Markov process and ν_n restricted to $[0, t] \times E$ is a function of the Markov process X_6^n over $[0, t]$. Thus, we can use the filtrations $\mathcal{F}_t^n$ generated by X_6^n for each $n \geq 1$. In our context, we have the following consequence of Lemma 1.4 of Kurtz [27].

LEMMA 7.5 (FACTORIZATION OF THE LIMITING RANDOM MEASURES). *Let $(\bar{X}, \nu)$ be the limit of a converging subsequence of $\{(\bar{X}_6^n, \nu^n); n \geq 1\}$ in $\mathcal{D}_6([0, \delta]) \times M(S)$ obtained via Lemma 7.4. Then there exists $p_s \equiv p_s(B)$, a measurable function of s for each measurable subset B in E and a probability measure on E for each s in $[0, \delta]$, such that, for all measurable subset B_1 of $[0, \delta]$ and B_2 of E,*

$$\nu(B_1 \times B_2) = \int_{B_1} p_s(B_2) \, ds. \tag{50}$$

As a consequence of the three lemmas above, we obtain the following preliminary representation.

LEMMA 7.6 (INITIAL REPRESENTATION OF LIMITING INTEGRAL TERMS). *Every subsequence of the sequence $\{(\bar{X}_6^n, \nu^n); n \geq 1\}$ defined by (26) and (49) in $\mathcal{D}_6([0, \delta]) \times M(S)$ as a further converging subsequence. Let $(\bar{X}, \nu)$ be a limit of a convergent subsequence. For any $\delta \leq \tau$ for τ in Theorem 6.3, the integral terms in (47) necessarily coincide with the corresponding integral terms in (45) with a probability $p_{1,2}(s)$ substituted for $1_{\{D_{1,2}^n(s)>0\}}$ and $\bar{X}$ substituted for x for $0 \leq t \leq \delta$, in particular,*

$$\bar{I}_{q,1,1}(t) = \int_0^t p_{1,2}(s)\bar{Z}_{1,2}(s) \, ds, \quad 0 \leq t \leq \delta, \quad w.p.1, \tag{51}$$

where $\bar{Z}_{1,2}$ is the component of $\bar{X}$ and $p_{1,2}(s) \equiv p_s((0,\infty))$ for p_s in (50), so that $p_{1,2}(s)$ is a measurable function of s with $0 \leq p_{1,2}(s) \leq 1$, $0 \leq s \leq \delta$.

PROOF. By Lemma 7.4, we are justified in focusing on a converging subsequence with limit $(\bar{X}, \nu)$, where $\bar{X} \equiv (\bar{Q}_i, \bar{Z}_{i,j})$. For (51), we focus on $(\bar{Z}_{1,2}, \nu)$. As before, apply the Skorohod representation theorem to obtain a version converging w.p.1 along the subsequence, without changing the notation. For the corresponding terms indexed by n,

$$\bar{I}_{q,1,1}^n(t) = \mu_{1,2} \int_0^t \bar{Z}_{1,2}^n(s) 1_{\{D_{1,2}^n(s)>0\}} ds = \int_0^t \bar{Z}_{1,2}^n(s) \, \nu^n(ds \times dy), \tag{52}$$

so that we can apply Lemma 7.3 to deduce that

$$\bar{I}_{q,1,1}^n(t) \to \bar{I}_{q,1,1}(t) \equiv \mu_{1,2} \int_0^t \bar{Z}_{1,2}(s) \, \nu(ds \times dy).$$

Finally, we apply Lemma 7.5 to show that the representation of ν in (52) is equivalent to (51). $\square$

It now remains to determine the term $p_{1,2}(s)$ in the integrand of the integral (51). In the next section we will show that we can write $p_{1,2}(s) = P(D(\bar{X}(s), \infty) > 0)$, thus completing the proof of Lemma 7.2.

7.3. A martingale argument to characterize the probability in the integrand. We now finish the proof of Lemma 7.2 by characterizing the probability measure p_s in Lemma 7.5.

PROOF OF LEMMA 7.2. We will prove that $p_{1,2}(s) = P(D(\bar{X}(s), \infty)) > 0$ for almost all s in the integral (51), where $D(\bar{X}(s), \infty)$ is a random variable with the steady-state distribution of the FTSP in §3.2 depending on the state $\bar{X}(s)$, which is the limit of the converging subsequence of the sequence $\{\bar{X}^n: n \geq 1\}$. This step will make (51) reduce to the desired (48).

We first comment on the exceptional sets. We establish the result w.p.1, so that there is an exceptional set, say Υ, in the underlying probability space Ω with $P(\Upsilon) = 0$, such that we claim the conclusion of Lemma 7.2 holds in $\Omega - \Upsilon$. However, for each $\omega \in \Omega - \Upsilon$, we find an exceptional set $\Psi(\omega)$ in $[0, \delta]$ where the Lebesgue measure of $\Psi(\omega)$ is 0. However, the integral in Lemma 7.2 is unchanged if we change the definition of the integrand on a set of Lebesgue measure 0. Hence, we can assume that $p_{1,2}(s) = P(D(\bar{X}(s), \infty) > 0)$ for *all* s in $[0, \delta]$ for each sample point $\omega \in \Omega - \Upsilon$. After doing that, we obtain the w.p.1 conclusion in Lemma 7.2.

We remark that it is possible to obtain a *single* exceptional set Ψ in $[0, \delta]$ such that $p_{1,2}(s) = P(D(\bar{X}(s), \infty)) > 0$ for all $s \in [0, \delta] - \Psi$ w.p.1. The construction to achieve that stronger goal is described in Example 2.3 of Kurtz [27] on p. 196. The conditions specified there hold in our context. Because that property is not required here, we do not elaborate.

Continuing with the main proof, we now aim to characterize the entire probability measure p_s on E appearing in Lemma 7.5. We will do that by showing that p_s satisfies the equation characterizing the steady-state distribution of the FTSP $D(\bar{X}(s), \cdot)$ for almost all s with respect to Lebesgue measure (consistent with the notion of an averaging principle). Because the FTSP $D(\bar{X}(s), \cdot)$, given $\bar{X}(s)$, is a CTMC with the special structure (only four transitions possible from each state, and only two different cases for these, as shown in (13)–(16)), just as for finite-state CTMCs (in elementary textbooks), it suffices to show that

$$\sum_i p_s(\{i\}) Q_{i,j}(\bar{X}(s)) = 0 \quad \text{for all } j \tag{53}$$

for almost all s in $[0, \delta]$ with respect to Lebesgue measure, where i and j are states of the FTSP and $Q_{i,j}(\bar{X}(s))$ in (53) is the (i, j)th component in the infinitesimal rate matrix (generator) of the CTMC $D(\bar{X}(s), \cdot)$.

However, we will follow Hunt and Kurtz [18] and use the framework in §§4.2 and 8.3 of Ethier and Kurtz [12]. In particular, the FTSP satisfies the assumptions in Corollary 8.3.2 on p. 379 in Ethier and Kurtz [12]. As in (9) of Hunt and Kurtz [18], this step corresponds to an application of Proposition 4.9.2 of Ethier and Kurtz [12], but the simple CTMC setting does not require all the structure there. Following the proof of Theorem 3 in Hunt and Kurtz [18] (and §2 of Kurtz [27]), we now develop a martingale representation for $f(D_{1,2}^n)$, where f is a bounded continuous real-valued function on the state space E of $D_{1,2}^n$. This construction is the standard martingale associated with functions of Markov processes, just as in Proposition 4.1.7 of Ethier and Kurtz [12]. Because, $D_{1,2}^n$ is a simple linear function of the CTMC X_6^n in (6), we can write $f(D_{1,2}^n) = g(X_6^n)$ for some continuous bounded function g. The martingale will be with respect to the filtration $\mathscr{F}_t^n$ generated by the Markov process X_6^n (as in Lemma 7.5).

Recalling that $D_{1,2}^n \equiv Q_1^n - r_{1,2}Q_2^n$, we can write $f(D_{1,2}^n(t))$ in terms of the independent rate-1 Poisson processes in (34) as follows

$$f(D_{1,2}^n(t)) \equiv f(D_{1,2}^n(0)) - \int_0^t [f(D_{1,2}^n(s-) + 1) - f(D_{1,2}^n(s-))] dN_1^a(\lambda_1^n s)$$

$$- \int_0^t [f(D_{1,2}^n(s-) - 1) - f(D_{1,2}^n(s-))](dN_1^u(\theta_1 Q_1^n(s)) + dN_{1,1}^s(\mu_{1,1}m_1^n s))$$

$$- \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} [f(D_{1,2}^n(s-) - 1) - f(D_{1,2}^n(s-))](dN_{1,2}^s(\mu_{1,2}Z_{1,2}^n(s)) + dN_{2,2}^s(\mu_{2,2}(Z_{2,2}^n(s))))$$

$$- \int_0^t 1_{\{D_{1,2}^n(s) \leq 0\}} [f(D_{1,2}^n(s-) + r) - f(D_{1,2}^n(s-))](dN_{1,2}^s(\mu_{1,2}Z_{1,2}^n(s)) + dN_{2,2}^s(\mu_{2,2}(Z_{2,2}^n(s))))$$

$$- \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} [f(D_{1,2}^n(s-) + r) - f(D_{1,2}^n(s-))] dN_s^u(\theta_2 Q_2^n(s))$$

$$- \int_0^t [f(D_{1,2}^n(s-) - r) - f(D_{1,2}^n(s-))] dN_2^a(\lambda_2^n s).$$

We next rewrite the representation for $f(D_{1,2}^n(t))$ above to achieve a martingale representation. To that end, we add and then subtract the appropriate Riemann integral from each of the integrals above, e.g.,

$$\int_0^t [f(D_{1,2}^n(s-) + 1) - f(D_{1,2}^n(s-))] dN_1^a(\lambda_1^n s)$$

$$= \int_0^t [f(D_{1,2}^n(s-) + 1) - f(D_{1,2}^n(s-))] dM_1^{n,a}(s) + \int_0^t [f(D_{1,2}^n(s-) + 1) - f(D_{1,2}^n(s-))] \lambda_1^n ds,$$

for $M_1^{n,a}$ in (44). Note that an integral of a predictable process with respect to a martingale is again a martingale. We thus achieve a modifies representation of $f(D_{1,2}^n(t))$ in terms of martingales and their associated predictable quadratic-variation processes; see, e.g., §3.5 in Pang et al. [34] for more details. Rearranging terms, so that all the martingales in the modified representation appear on the left-hand side and letting M_f^n denote the sum of those martingales, we have

$$M_f^n(t) \equiv f(D_{1,2}^n(t)) - f(D_{1,2}^n(0)) - \int_0^t \lambda_1^n [f(D_{1,2}^n(s-)+1) - f(D_{1,2}^n(s-))] \, ds$$

$$- \int_0^t (m_1^n + \theta_1 Q_1^n(s))[f(D_{1,2}^n(s-)-1) - f(D_{1,2}^n(s-))] \, ds$$

$$- \int_0^t 1_{\{D_{1,2}^n(s)>0\}} (\mu_{1,2} Z_{1,2}^n(s) + \mu_{2,2}(m_2^n - Z_{1,2}^n(s)))[f(D_{1,2}^n(s-)-1) - f(D_{1,2}^n(s-))] \, ds$$

$$- \int_0^t 1_{\{D_{1,2}^n(s)\leq 0\}} (\mu_{1,2} Z_{1,2}^n(s) + \mu_{2,2}(m_2^n - Z_{1,2}^n(s)))[f(D_{1,2}^n(s-)+r) - f(D_{1,2}^n(s-))] \, ds$$

$$- \int_0^t 1_{\{D_{1,2}^n(s)>0\}} \theta_2 Q_2^n(s)[f(D_{1,2}^n(s-)+r) - f(D_{1,2}^n(s-))] \, ds$$

$$- \int_0^t \lambda_2^n [f(D_{1,2}^n(s-)-r) - f(D_{1,2}^n(s-))] \, ds. \tag{54}$$

Note that M_f^n is a $\mathcal{F}_t^n$-martingale itself.

It follows from essentially the same arguments as in the proof of Lemma 6.1 and the continuity of addition at continuous limits, that $\bar{M}_f^n \equiv M_f^n/n \Rightarrow 0e$ in $\mathcal{D}$ as $n \to \infty$, for M_f^n in (54). In addition, we have $n^{-1}(f(D_{1,2}^n(t)) - f(D_{1,2}^n(0))) \Rightarrow 0e$ in $\mathcal{D}$ since f is bounded. We write the remaining terms of $\bar{M}_f^n$ as

$$\bar{D}_f^n(t) \equiv \bar{M}_f^n(t) - n^{-1}(f(D_{1,2}^n(t)) - f(D_{1,2}^n(0))) = \int_{(0,t)\times E} \bar{\lambda}_1^n [f(y+1) - f(y)] \nu^n(ds \times dy)$$

$$- \int_{(0,t)\times E} (\bar{m}_1^n + \theta_1 \bar{Q}_1^n(s))[f(y-1) - f(y)] \nu^n(ds \times dy)$$

$$- \int_{(0,t)\times E} 1_{\{y>0\}} (\mu_{1,2} \bar{Z}_{1,2}^n(s) + \mu_{2,2}(\bar{m}_2^n - \bar{Z}_{1,2}^n(s)))[f(y-1) - f(y)] \nu^n(ds \times dy)$$

$$- \int_{(0,t)\times E} 1_{\{y\leq 0\}} (\mu_{1,2} \bar{Z}_{1,2}^n(s) + \mu_{2,2}(\bar{m}_2^n - \bar{Z}_{1,2}^n(s)))[f(y+r) - f(y)] \nu^n(ds \times dy)$$

$$- \int_{(0,t)\times E} 1_{\{y>0\}} \theta_2 \bar{Q}_2^n(s)[f(y+r) - f(y)] \nu^n(ds \times dy) - \int_{(0,t)\times E} \bar{\lambda}_2^n [f(y-r) - f(y)] \nu^n(ds \times dy).$$

By Lemmas 7.6 and 7.3, $\bar{D}_f^n \Rightarrow \bar{D}_f = 0e$ as $n \to \infty$ along the converging subsequence, where

$$\bar{D}_f \equiv \int_{(0,t)\times E_\Delta} \lambda_1 [f(y+1) - f(y)] \nu(ds \times dy) - \int_{(0,t)\times E} (m_1 + \theta_1 \bar{Q}_1(s))[f(y-1) - f(y)] \nu(ds \times dy)$$

$$- \int_{(0,t)\times E} 1_{\{y>0\}} (\mu_{1,2} \bar{Z}_{1,2}(s) + \mu_{2,2}(m_2 - \bar{Z}_{1,2}(s)))[f(y-1) - f(y)] \nu(ds \times dy)$$

$$- \int_{(0,t)\times E} 1_{\{y\leq 0\}} (\mu_{1,2} \bar{Z}_{1,2}(s) + \mu_{2,2}(m_2 - \bar{Z}_{1,2}(s)))[f(y+r) - f(y)] \nu(ds \times dy)$$

$$- \int_{(0,t)\times E} 1_{\{y>0\}} \theta_2 \bar{Q}_2(s)[f(y+r) - f(y)] \nu(ds \times dy) - \int_{(0,t)\times E} \lambda_2 [f(y-r) - f(y)] \nu(ds \times dy). \tag{55}$$

However, just as in Hunt and Kurtz [18], we can identify the limit $\bar{D}_f$ in (55) as the integral with respect to the random measure ν of the infinitesimal generator of the FTSP $D(\bar{X}(s), \cdot)$ applied to the test function f. In particular, for each sample point in the underlying probability space Ω supporting $(\bar{X}, \nu)$ except for a subset Υ with $P(\Upsilon) = 0$, from Lemma 7.5, we obtain

$$\int_0^t \int_E [Q(\bar{X}(s))f](y) p_s(dy) \, ds = 0 \quad \text{for all } t \geq 0, \tag{56}$$

where $[Q(\bar{X}(s))f](y)$ is the generator of the CTMC $D(\bar{X}(s), \cdot)$ applied to f as a function of y in E. As a consequence,

$$\int_E [Q(\bar{X}(s))f](y) p_s(dy) = 0 \quad \text{for almost all } s \text{ with respect to Lebesgue measure.} \tag{57}$$

It follows from Proposition 4.9.2 page 239 in Ethier and Kurtz [12], that p_s is the (unique) stationary distribution of the FTSP $D(\bar{X}(s), \cdot)$ for almost all s. (This step is equivalent to (53).) We apply Lemma 3.1 to conclude that the FTSP $D(\gamma, \cdot)$ has a unique stationary distribution on E for all $\gamma \in \mathbb{S}$. $\square$

7.4. Proof of Theorem 4.1.
We can now summarize the proof of our main result—the FWLLN via the AP.

PROOF OF THEOREM 4.1. There are two steps: (i) establishing convergence over an initial interval and (ii) expanding the interval of convergence. We consider slightly more general settings than in the statement of the theorem, by considering the vector $(\bar{X}_6^n, \bar{Y}_8^n, \nu^n)$ with the random measure ν^n replacing Θ^n. Note that $\Theta^n(t) \equiv \nu^n([0, t] \times (0, \infty))$ and $\vartheta(t) \equiv \nu([0, t] \times (0, \infty))$ for Θ^n and ϑ in (27), ν^n in (49), and ν in (50).

(i) *Establishing convergence over* $[0, \tau]$. By Theorem 5.2, the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n): n \geq 1\}$ in (26) is $\mathscr{C}$-tight in $\mathscr{D}_{14}([0, \infty))$. By Lemma 7.4, $(\bar{X}_6^n, \nu^n)$ is relative compact in $\mathscr{D}_6 \times M(S)$. By Theorem 6.3, there exists $\tau > 0$ such that the limit point of a converging subsequence of $\bar{X}_6^n$ in $\mathscr{D}_6([0, \tau])$ is also the limit point of $\bar{X}_6^{n,*}$ in (10), whose representation is specified in (41)–(43). Thus, it suffices to next characterize the limit of a converging subsequence of the sequence $(\bar{X}^n, \nu^n)$, for $\{\bar{X}^n\} \equiv \{1/n(Q_1^n, Q_2^n, Z_{1,2}^n)\}$ in (45) over an interval $[0, \tau]$. The characterization of the limit of $\bar{X}_6^n$ also characterizes the limit of ν^n because, as the proof of Lemma 7.5 demonstrates, each limit point of $(\bar{X}_6^n, \nu^n)$ is of the form $(\bar{X}_6, \nu)$, for ν in (50). In particular, for any two Borel sets B_1, B_2,

$$\nu^n(B_1, B_2) \equiv \int_{B_1} 1_{\{D_{1,2}^n(s) \in B_2\}} \Rightarrow \nu(B_1 \times B_2) \equiv \int_{B_1} p_s(B_2) \, ds \quad \text{in } \mathscr{D} \text{ as } n \to \infty,$$

where p_s is the unique stationary distribution of the FTSP $D(\bar{X}_6(s), \cdot)$ for almost all s. Hence, if $\bar{X} \equiv x_6$, and in particular, the limit of $\bar{X}_6^n$ is unique, then $p_s \equiv \pi_{1,2}(x_6(s))$ for almost all s, so that the limit ν of ν^n is unique as well.

With that characterization complete, we obtain the full convergence $(\bar{X}_6^n, \bar{Y}_8^n, \nu^n) \Rightarrow (x_6, y_8, \nu)$ in $\mathscr{D}_{15}([0, \tau])$ directly, exploiting Theorem 6.3. By Lemma 7.2, we complete the characterization step, showing that $P(\bar{X} = x) = 1$ in $\mathscr{D}_3([0, \tau])$, where x is a solution to the ODE in (24) with the initial condition $x(0)$ specified by Assumption 3.2.

(ii) *Expanding in the interval of convergence.* After establishing the convergence over an initial interval $[0, \tau]$, we can apply Theorem 6.4 (which uses Theorem 4.1 over $[0, \tau]$) to conclude that any limit point of the tight sequence $\bar{X}_6^n$ is again a limit of the tight sequence $\bar{X}_6^{n,*}$ in (10) over the entire half line $[0, \infty)$, showing that τ places no constraint on expanding the convergence interval. Moreover, by part (ii) of Theorem 5.2 in Perry and Whitt [37], any solution to the ODE, with a specified initial condition, can be extended indefinitely, and is unique. Hence that places no constraint either. Finally, the martingale argument allows us to uniquely characterize the steady-state distribution of the FTSP $D(\gamma, \cdot)$ in §3.2 even when the state γ is not in $\mathbb{A}$, provided that we have the SSC provided by Theorem 6.4. In particular, we will have either $\pi_{1,2}(\gamma) = 1$ or $\pi_{1,2}(\gamma) = 0$ if $\gamma \notin \mathbb{A}$. $\square$

8. Remaining proofs of theorems in §4.
We now provide the remaining proofs for four theorems in §4. At this point, Theorem 4.1 has been proved.

8.1. Proof of Theorem 4.3.
We will consider a sequence of stationary Markov processes $\{\{\bar{X}_6^n(t): t \geq 0\}: n \geq 1\}$, with $\bar{X}_6^n(0) \stackrel{d}{=} \bar{X}_6^n(\infty)$ for each $n \geq 1$. That initial condition makes the stochastic processes strictly stationary. We will refer to these stationary processes as *stationary versions* of the processes $\bar{X}_6^n$ and denote them by $\bar{X}_s^n$. We start by establishing tightness.

LEMMA 8.1 (TIGHTNESS OF THE SEQUENCE OF STATIONARY DISTRIBUTIONS). *The sequences* $\{\bar{X}_6^n(\infty): n \geq 1\}$ *and* $\{\{\bar{X}_s^n(t): t \geq 0\}: n \geq 1\}$ *are tight in* $\mathbb{R}_6$ *and* $\mathscr{D}_6$, *respectively.*

PROOF. First, for the tightness of $\{\bar{X}_6^n(\infty): n \geq 1\}$ in $\mathbb{R}_6$, it suffices to treat the six components separately. The tightness of $\{\bar{Z}_{i,j}^n(\infty): n \geq 1\}$ in $\mathbb{R}$ is immediate because $0 \leq \bar{Z}_{i,j}^n(\infty) \leq m_j^n/n$, where $m_j^n/n \to m_j$ as $n \to \infty$. The tightness of the queue lengths follows from Lemma A.5. In particular, since (i) $Q_i^n(t) \leq_{st} Q_{i,bd}^n(t)$ for all t and (ii) $Q_i^n(t) \Rightarrow Q_i^n(\infty)$ and $Q_{i,bd}^n(t) \Rightarrow Q_{i,bd}^n(\infty)$ as $t \to \infty$ for all $n \geq 1$, we necessarily have $Q_i^n(\infty) \leq_{st} Q_{i,bd}^n(\infty)$ for all n, because stochastic order is preserved under convergence. Since $\bar{Q}_{i,bd}^n(\infty) \Rightarrow q_{i,bd}(\infty) \equiv q_i(0) \vee (\lambda_i/\theta_i)$ as $n \to \infty$, the sequence $\{\bar{Q}_{i,bd}^n(\infty): n \geq 1\}$ is stochastically bounded, which implies that the sequence $\{\bar{Q}_i^n(\infty): n \geq 1\}$ is stochastically bounded as well. Because tightness of the marginal distributions implies tightness of vectors, the sequence of steady-state random vectors $\{\bar{X}^n(\infty): n \geq 1\}$ is tight in $\mathbb{R}_6$. Given the tightness of $\{\bar{X}_s^n(0)\}$, the proof of tightness in $\mathscr{D}_6$ is identical to the proof of Theorem 5.2. $\square$

We next establish the analogue of the structural simplification results in §6. Let $\bar{X}(\infty) \equiv (\bar{Q}_i(\infty), \bar{Z}_{i,j}(\infty))$ be a limit of the stochastically bounded sequence $\{\bar{X}_6^n(\infty)\}$. Let $\bar{X} \equiv (\bar{Q}_i, \bar{Z}_{i,j})$ be a limit of the sequence $\{\{\bar{X}_s^n(t): t \geq 0\}: n \geq 1\}$. Note that $\bar{X}$ must itself be a stationary process.

LEMMA 8.2. $P(\bar{Z}_{1,1}(\infty) = m_1, \bar{Z}_{2,1}(\infty) = 0, \bar{Z}_{2,2}(\infty) = m_2 - \bar{Z}_{1,2}(\infty)) = 1$.

PROOF.   Let $\bar{I}_j(\infty) \equiv m_j - \bar{Z}_{1,j}(\infty) - \bar{Z}_{2,j}(\infty)$, $j = 1, 2$. To prove the claim, we need to show that $P(\bar{Z}_{2,1}(\infty) > 0) = P(\bar{I}_j(\infty) > 0) = 0$. We will consider the sequence of stationary versions $\{\bar{X}_s^n\}$, and a limit $\bar{X}$ of this sequence.

(i) $P(\bar{Z}_{2,1} = 0e) = 1$. We will show that the opposite assumption leads to a contradiction. Hence, suppose that $P(\bar{Z}_{2,1}(s) > 0) > 0$ for some $s \geq 0$. The stationarity of $\bar{Z}_{2,1}$ implies that $\bar{Z}_{2,1}(0) \stackrel{d}{=} \bar{Z}_{2,1}(s)$. Hence, we can equivalently assume that $P(\bar{Z}_{2,1}(0) > 0) > 0$.

Let $B_{2,1}$ denote the set in the underlying probability space, of all sample paths of $\bar{X}$ with $\bar{Z}_{2,1}(0) > 0$. By the contradictory assumption, $P(B_{2,1}) > 0$. Following the arguments in Lemma B.1, if $\bar{Z}_{2,1}(0) > 0$, then $\bar{Z}_{1,2}(t) > 0$ for all $t \geq 0$, which implies that $\bar{Z}_{2,1}(t) > 0$ for all $t \geq 0$ in $B_{2,1}$.

Consider a sample path in $B_{2,1}$. Because of one-way sharing, $\bar{Z}_{1,2} = 0e$, so that the only departures from $\bar{Q}_1$ are due to service completions in pool 1 and abandonment. Since $\bar{Z}_{1,1} \leq m_1$ w.p.1, $\bar{Q}_1$ is stochastically bounded from below, in sample-path stochastic order, by the fluid limit of an Erlang-A model with m_1 servers. At the same time, $\bar{Q}_2$ is stochastically bounded from above by the fluid limit of an Erlang-A model with m_2 servers. Hence, there exists $\epsilon > 0$ such that, for some $s \geq 0$,

$$r\bar{Q}_2(t) - \epsilon \leq rq_2^a < q_1^a \leq \bar{Q}_1(t) + \epsilon \quad \text{for all } t \geq s, \tag{58}$$

where $rq_2^a < q_1^a$ by Assumption 3.1. Now, because $\bar{Q}_1$ and $\bar{Q}_2$ are bounded with probability 1, we can find $s_0 > 0$ such that (58) holds with this s_0 *for all possible initial condition in* $B_{2,1}$. However, this implies that $\bar{Z}_{2,1}$ is strictly decreasing for all $t \geq s_0$ and for all sample paths in $B_{2,1}$ (because no fluid can flow from queue 1 to pool 2), so that $\bar{Z}_{2,1}(t) < \bar{Z}_{2,1}(s_0)$ for all $t > s_0$ in $B_{2,1}$, contradicting the stationarity of $\bar{Z}_{2,1}$. Thus, $P(B_{2,1}) = 0$.

(ii) $P(\bar{I}_1 = \bar{I}_2 = 0e) = 1$. We follow the proof of Theorem 6.2, building on the result $P(\bar{Z}_{2,1}(\infty) = 0) = 1$ just established. Recall that $L_i^n \equiv Q_i^n + Z_{i,1}^n + Z_{i,2}^n - m_i^n$ in (B9), representing the excess number of class-i customers in the system, is stochastically bounded from below, in sample-path stochastic order, by the process $L_{b,i}^n$, with $L_{i,b}^n$ defined by imposing a reflecting upper barrier at $k_{1,2}^n$ for $i = 1, 2$ and letting $L_{i,b}^n(0) = L_i^n(0) \wedge k_{1,2}^n$. However, here we are working with stationary versions. Because, the processes (L_1^n, L_2^n) are strictly stationary, so are the reflected processes $(L_{1,b}^n, L_{2,b}^n)$. Then $U^n \geq_{st} U_b^n$, where U^n and U_b^n and the linear functions of (L_1^n, L_2^n) and $(L_{1,b}^n, L_{2,b}^n)$, respectively, defined in (B10). These processes U^n and U_b^n are also stationary processes.

However, just as in the proof of Theorem 6.2, U_b^n is a birth and death process on the integers in $(-\infty, 0]$, which is independent of $Z_{1,2}^n$, with drift δ_b^n in (B11). Since $\delta_b^n/n \to \delta_b > 0$, the birth and death U_b^n has a positive drift for all n large enough. Consequently, $-U_b^n$ has the structure of the stationary queue length process in a stable $M/M/1$ queue. Because the traffic intensity converges to a limit strictly less than one, the initial value, and the value at any time, is stochastically bounded. Moreover, we can apply essentially the same extreme value argument used in the proof of Theorem 6.2 to conclude that, for any $\tau > 0$, that $P(\|\bar{I}_i^n\|_\tau > 0) \to 0$ as $n \to \infty$. We thus conclude that $P(\|\bar{I}_i\|_\tau = 0) = 1$, from which the conclusion follows, because the interval $[0, \infty)$ can be represented as the countable union of finite intervals of finite length, and the countable sum of 0 probabilities is itself zero. □

A function $\bar{Y}: \mathbb{R}_m \to \mathbb{R}_k$, $m, k \geq 1$, is said to be locally Lipschitz continuous if for any compact set B, there exists a constant $K(B)$ such that, for any $s, t \in B$, $|\bar{Y}(s) - \bar{Y}(t)| \leq K(B)|s - t|$. A locally Lipschitz continuous function is absolutely continuous and is thus differentiable almost everywhere.

In the following we will consider two locally Lipschitz continuous functions $V: \mathbb{R}_k^+ \to \mathbb{R}^+$ and $\bar{Y}: \mathbb{R}^+ \to \mathbb{R}_k^+$, and the "Lie derivative" $\dot{V}(\bar{Y}(t)) \equiv \nabla V \cdot \bar{Y}'$, where ∇V denotes the gradient of V, and $\nabla V \cdot \bar{Y}'$ is the usual inner product of vectors. Because both functions are differentiable almost everywhere, $\dot{V}(\bar{Y}(t))$ is understood to be taken at points t for which both V and $\bar{Y}$ are differentiable.

For a vector $x \in \mathbb{R}_k$, we let $\|x\|$ denote its L_1 norm, although any other norm in $\mathbb{R}_k$ can be used in the following. For the following we draw on §8.3 of Perry and Whitt [37].

LEMMA 8.3.   *Let $V: \mathbb{R}_k^+ \to \mathbb{R}_+$ be locally Lipschitz continuous, such that $V(x) = 0$ if and only if $x = 0$. Let $\bar{Y}: \mathbb{R}^+ \to \mathbb{R}_k^+$ be Lipschitz continuous with constant N, such that $\bar{Y} \leq M$, for some $M > 0$. If $\dot{V}(\bar{Y}(t)) < 0$ for all $t \geq 0$ for which $\bar{Y}(t) \neq 0$, then $\bar{Y}(t) \to 0$ as $t \to \infty$.*

PROOF.   Let $f: \mathbb{R}_+ \to \mathbb{R}_+$ be absolutely continuous, such that $f(t) = 0$ if and only if $t = 0$. If for almost every $t \neq 0$, $f'(t) < 0$, then $f(t) \to 0$ as $t \to \infty$. We can apply that argument to the function $f(t) \equiv V(\bar{Y}(t))$ if we show that $V(\bar{Y}(t))$ is locally Lipschitz continuous. To see that is the case, note that for all $t > 0$,

$$\bar{Y}(s) \leq \bar{Y}(0) + |\bar{Y}(s) - \bar{Y}(0)| \leq M + Nt, \quad 0 \leq s \leq t.$$

Therefore, for any $0 \leq u \leq s \leq t$, $|f(s) - f(u)| \leq K(B)N(s - u)$, where $B \equiv \{x \in \mathbb{R}_k^+: \|x\| \leq M + Nt\}$.   □

PROOF OF THEOREM 4.3   We consider a converging subsequence $\{\bar{X}_s^{n'}: n' \geq 1\}$ of the sequence $\{\bar{X}_s^n: n \geq 1\}$ with limit $\bar{X}$. By Lemma 8.2, we can consider it to be three-dimensional with components $\bar{Q}_1$, $\bar{Q}_2$, and $\bar{Z}_{1,2}$. By Corollary 5.1, the limit $\bar{X}$ is Lipschitz continuous in t, so that it is differentiable almost everywhere. Specifically, $\bar{X}$ is a limit of the sequence represented by (45), with $\bar{X}^n(0) \stackrel{\mathrm{d}}{=} \bar{X}^n(\infty)$ for all $n \geq 1$. Then each component of the converging subsequence $\{\bar{X}_s^{n'}\}$ in (45) converges to its respective limit (e.g., $\bar{Q}_1^{n'}$ to $\bar{Q}_1$). For our purposes here, it is sufficient to conclude that $\bar{X}(t) \to x^*$ w.p.1 as $t \to \infty$. Hence, we will not characterize the limit of $\bar{X}(t)$ as $t \to \infty$.

For the representation of the converging subsequence $\{\bar{X}_s^{n'}\}$ in (45), let

$$\bar{W}^{n'} \equiv (C - 1)\bar{Z}_{1,2}^{n'} + C\bar{Q}_1^{n'} + \bar{Q}_2^{n'},$$

where $C \geq 1$ will be specified later. Since $\bar{X}_s^{n'} \Rightarrow \bar{X}$ by assumption, we can apply the continuous mapping theorem to conclude that $\bar{W}^n \Rightarrow \bar{W}$ in $\mathscr{D}_3$ as $n \to \infty$, where

$$\begin{aligned}
\bar{W}(t) &\equiv (C - 1)\bar{Z}_{1,2}(t) + C\bar{Q}_1(t) + \bar{Q}_2(t) \\
&= (C - 1)\bar{Z}_{1,2}(0) - (C\mu_{1,2} - \mu_{2,2})\int_0^t \bar{Z}_{1,2}(s)\,ds - \mu_{2,2}m_2 t \\
&\quad + C\bar{Q}_1(0) + C(\lambda_1 - \mu_{1,1}m_1)t - C\theta_1\int_0^t \bar{Q}_1(s)\,ds + \bar{Q}_2(0) + \lambda_2 t - \theta_2\int_0^t \bar{Q}_2(s)\,ds,
\end{aligned}$$

with derivative

$$\bar{W}'(t) = -(C\mu_{1,2} - \mu_{2,2})\bar{Z}_{1,2}(t) - C\theta_1\bar{Q}_1(t) - \theta_2\bar{Q}_2(t) + C(\lambda_1 - \mu_{1,1}m_1) + \lambda_2 - \mu_{2,2}m_2.$$

For $x \in \mathbb{R}_3$, let $V(x) \equiv Cx_1 + x_2 + (C - 1)x_3$ and note that $\dot{V}(\bar{X}) \equiv \nabla V \cdot \bar{X}' = \bar{W}'$, and that V is locally Lipschitz continuous, where ∇V denotes the gradient of V, and $\nabla V \cdot \bar{X}'$ is the usual inner product of vectors. Now let $\bar{Y}$ denote the derivative of $\bar{X}$ shifted by x^* for x^* in (30), i.e., $\bar{Y} \equiv \bar{X}' + x^*$. Hence, as in the proof of Theorem 8.3 in Perry and Whitt [37], we have

$$\dot{V}(\bar{Y}) \equiv -(C\mu_{1,2} - \mu_{2,2})\bar{Z}_{1,2}(t) - C\theta_1\bar{Q}_1(t) - \theta_2\bar{Q}_2(t).$$

If $\mu_{1,2} > \mu_{2,2}$, then we let $C = 1$, and if $\mu_{1,2} \leq \mu_{2,2}$ we let C be any number such that $C > \mu_{2,2}/\mu_{1,2} \geq 1$. With this choice of C, we see that $\dot{V}(\bar{Y}) < 0$. By Lemma 8.3, $\bar{Y}(t) \to 0$, which implies that $\bar{X}(t) \to x^*$ as $t \to \infty$ w.p.1. (Note that $\bar{Y}$ and $\bar{X}$ are Lipschitz continuous and bounded, as required.)

For $\alpha > 0$, let $\beta_V(\alpha) \equiv \{x \in \mathbb{R}_3: \|V(x) - V(x^*)\| \leq \alpha\}$. By the monotonicity of $V(\bar{X})$ established above, for each $\alpha > 0$ there exists $T(\alpha, \bar{X}(0))$, such that $\bar{X}(t) \in \beta_V(\alpha)$ for all $t \geq T(\alpha, \bar{X}(0))$. Because the queues $\bar{Q}_1$ and $\bar{Q}_2$ are bounded w.p.1, we can uniformly bound $T(\alpha, \bar{X}(0))$ (uniformly in $\bar{X}(0)$). Hence, there exists $T \equiv T(\alpha)$, such that $\bar{X}(t) \in \beta_V(\alpha)$ w.p.1 for all $t \geq T$. It follows from the stationarity of $\bar{X}$ that $\bar{X}(0) \in \beta_V(\alpha)$. Because this is true for all $\alpha > 0$, it must hold that $\bar{X}(0) = x^*$ which, by the equality in distribution $\bar{X}(\infty) \stackrel{\mathrm{d}}{=} \bar{X}(0)$, implies that $\bar{X}(\infty) = x^*$ w.p.1. We have thus shown that the limit of all converging subsequences of $\{\bar{X}^n(\infty): n \geq 1\}$ is x^* in (30), which implies the full convergence $\bar{X}^n(\infty) \Rightarrow x^*$ as $n \to \infty$. Moreover, since x^* is the limit of a stationary sequence, x^* itself must be a stationary point for each fluid limit $\bar{X}$ (i.e., if $\bar{X}(0) = x^*$, then $\bar{X}(t) = x^*$ for all $t \geq 0$), and it is globally asymptotically stable, because $\bar{X}(t) \to x^*$ as $t \to \infty$, as was shown above.   □

Note that none of the proofs in this section used the initial condition in Assumption 3.2, or the rationality of the queue ratios in Assumption 3.3.

8.2. Proof of Theorem 4.4. The claimed convergence in $\mathscr{D}$ is less complicated than it might appear, because there is no spatial scaling. Consequently, all processes are pure-jump processes, having piecewise constant sample paths, with only finitely many discontinuities in any bounded interval. Both processes have the same four possible transitions from each state: ± 1 or $\pm r$ for $r \equiv r_{1,2}$. Hence, it suffices to show convergence in distribution of the first k pairs of jump times and jump values for any $k \geq 1$. That can be done by mathematical induction on k.

We can start by applying the Skorohod representation theorem to replace the assumed convergence in distribution $\Gamma_6^n/n \Rightarrow \gamma_6$ and $D_e^n(\Gamma^n, 0) \Rightarrow D(\gamma, 0)$ as $n \to \infty$ by convergence w.p.1 for alternative random vectors having the same distribution. Hence, it suffices to assume that $\Gamma_6^n/n \to \gamma_6$ and $D_e^n(\Gamma^n, 0) \to D(\gamma, 0)$ as $n \to \infty$ and w.p.1. (We do not introduce new notation.) Thus, for each sample point, we will have $D_e^n(\Gamma^n, 0) = D(\gamma, 0)$ for all n sufficiently large. Hence, we can assume that we start with equality holding.

Moreover, we can exploit the structure of pure-jump Markov processes. The FTSP is directly such a pure-jump Markov process with transition rates given in (13)–(16). These rates were defined to be the limit of the transition rates of the queue difference processes $D_{1,2}^n$, after dividing by n. The queue difference processes $D_{1,2}^n$ in (6) are not Markov, but they are simple linear functions of the pure-jump Markov process X_6^n.

The transition rates for $D_{1,2}^n$ closely parallel (13)–(16). Because of the assumptions, we can work with the three-dimensional random state $\Gamma^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n) \in \mathbb{R}_3$ with the understanding that the remaining components of Γ_6^n are $Z_{1,1}^n = m_1^n$, $Z_{2,1}^n = 0$ and $Z_{2,2}^n = m_2^n - Z_{1,2}^n$. Let $D^n(\Gamma^n, t_0) \equiv D_{1,2}^n(t_0)$ under the condition that $X^n(t_0) = \Gamma^n$. When $D^n(\Gamma^n, t_0) \leq 0$, let the transition rates be $\lambda_-^{(r)}(n, \Gamma^n)$, $\lambda_-^{(1)}(n, \Gamma^n)$, $\mu_-^{(r)}(n, \Gamma^n)$, and $\mu_-^{(1)}(n, \Gamma^n)$ in (A1), for transitions of $+r$, $+1$, $-r$ and -1, respectively. When $D^n(\Gamma^n, t_0) > 0$, let the transition rates be $\lambda_+^{(r)}(n, \Gamma^n)$, $\lambda_+^{(1)}(n, \Gamma^n)$, $\mu_+^{(r)}(n, \Gamma^n)$, and $\mu_+^{(1)}(n, \Gamma^n)$ in (A2), for transitions of $+r$, $+1$, $-r$ and -1, respectively.

The many-server heavy-traffic scaling in (3) and the condition $\Gamma^n/n \to \gamma$ imply that the transition rates of X_6^n are of order $O(n)$ as $n \to \infty$. However, the time expansion in (12) brings those transition rates back to order $O(1)$. Indeed, from (13)–(16) and (A1)–(A2), we see that the transition rates of $D_e^n(\Gamma^n, \cdot)$, which change with every transition of the CTMC X_6^n, actually converge to the transition rates of the FTSP $D(\gamma, \cdot)$, which only depend on the region ($D > 0$ and $D \leq 0$).

The time until the first transition in the pure jump Markov process $D(\gamma, \cdot)$ is clearly exponential. Because the queue lengths can be regarded as strictly positive by Theorem 6.2, the first transition of $D_e^n(\Gamma^n, \cdot)$ coincides with the first transition time of the underlying CTMC X_6^n, which also is exponential. Because the transition rates converge, the time until the first transition of $D_e^n(\Gamma^n, \cdot)$ converges in distribution to the exponential time until the first transition of the FTSP $D(\gamma, \cdot)$. Moreover, in both processes the jump takes one of four values ± 1 or $\pm r$. The probabilities of these values converges as well. Hence the random first pair of jump time and jump value converges to the corresponding pair of the FTSP. The same reasoning applies to successive pairs of jump times and jump values, applying mathematical induction. That completes the proof.

8.3. Auxiliary results for FTSPs. Before proving Theorems 4.5 and 4.6, we prove some auxiliary lemmas that we employ in the proofs. We use stochastic bounds by frozen queue-difference stochastic processes as in §A.1. The following lemma is proved much like Lemma B.3, exploiting the bounds in §A.2.

LEMMA 8.4 (BOUNDING FROZEN PROCESSES). *Suppose that $x(t) \in \mathbb{A}$, $t_1 \leq t \leq t_2$. For any $t \in [t_1, t_2]$ and $\epsilon > 0$, there exist positive constants δ, η, state vectors x_m, $x_M \in \mathbb{A}$ and random state vectors X_m^n, X_M^n, $n \geq 1$, such that $\|x_m - x(t)\| < \epsilon$, $\|x_M - x(t)\| < \epsilon$, $n^{-1} X_m^n \Rightarrow x_m$, $n^{-1} X_M^n \Rightarrow x_M$ as $n \to \infty$,*

$$D_f^n(X_m^n, \cdot) \leq_r D_f^n(X^n(t), \cdot) \leq_r D_f^n(X_M^n, \cdot) \qquad and$$

$$D_f^n(X_m^n, \cdot) \leq_r D_{1,2}^n(t) \leq_r D_f^n(X_M^n, \cdot) \quad in \ \mathscr{D}([t_1 \vee (t-\delta), (t+\delta) \wedge t_2]) \ for all \ n \geq 1, \tag{59}$$

and $P(B^n(\delta, \eta)) \to 1$ as $n \to \infty$, where

$$B^n(\delta, \eta) \equiv \{\delta_+^n(X_M^n) < -\eta, \ \delta_-^n(X_M^n) > \eta, \ \delta_+^n(X_m^n) < -\eta, \ \delta_-^n(X_m^n) > \eta\}. \tag{60}$$

As a consequence, the bounding frozen processes $D_f^n(X_M^n, \cdot)$ and $D_f^n(X_m^n, \cdot)$ in (59), and thus also the interior frozen processes $D_f^n(X^n(t), \cdot)$, satisfy (20) on $B_n(\delta, \eta)$ and are thus positive recurrent there, $n \geq 1$.

PROOF. We use the convergence $\bar{X}^n \Rightarrow x$ provided by Theorem 4.1. Consider one $t \in [t_1, t_2]$. By the linearity of the drift functions δ_+ and δ_- in (19), $\delta_+^n(X^n(t))/n \Rightarrow \delta_+(x(t))$ and $\delta_-^n(X^n(t))/n \Rightarrow \delta_-(x(t))$ for $x(t) \in \mathbb{A}$, so that (20) holds. Hence there exists $\eta > 0$ such that

$$\lim_{n \to \infty} P(\delta_+^n(X^n(t)) < -\eta \ and \ \delta_-^n(X^n(t)) > \eta) = 1,$$

i.e., (A4) holds for $\Gamma^n = X^n(t)$ with probability converging to 1 as $n \to \infty$.

We now bound the drifts in (A3). We do that by bounding the change in the components of $X^n(t)$ in a short interval around time t. To do that, we use the stochastic-order bounds in §A.2, constructed over the interval $[t_1 \vee (t - \delta), (t + \delta) \wedge t_2]$, with the construction beginning at the left endpoint $t_1 \vee (t - \delta)$, just as in the construction after time 0 in §A.2. Let $I_{(t_1, \delta)} \equiv [t_1 \vee (t - \delta), (t + \delta) \wedge t_2]$ and let $\|\cdot\|_\delta$ denote the norm over the interval $I_{(t_1, \delta)}$. To construct X_M^n, let

$$X_{M^+}^n \equiv (Q_{1,M}^n, Q_{2,M}^n, Z_{M^+}^n) \qquad \text{and} \qquad X_{M^-}^n \equiv (Q_{1,M}^n, Q_{2,M}^n, Z_{M^-}^n),$$

where

$$Q_{1,M}^n \equiv \inf_{t \in I_{(t_1, \delta)}} Q_{1,a}^n(t) \vee 0, \qquad Q_{2,M}^n \equiv \|Q_{2,b}^n\|_\delta,$$

$$Z_{M^+}^n \equiv \inf_{t \in I_{(t_1, \delta)}} Z_+^n(t), \qquad Z_{M^-}^n \equiv \|Z_-^n\|_\delta,$$

with $Z_+^n(t) \equiv Z_b^n(t)$ and $Z_-^n(t) \equiv Z_a^n(t)$ if $\mu_{2,2} \geq \mu_{1,2}$, and $Z_+^n(t) \equiv Z_a^n(t)$ and $Z_-^n(t) \equiv Z_b^n(t)$ otherwise. We work with the final value $X_{M^+}^n \equiv X_{M^+}^n((t + \delta) \wedge t_2)$, and similarly for $X_{M^-}^n$. Let $\{D_f^n(X_M^n, s) : s \geq 0\}$ have the rates determined by $X_{M^-}^n$ when $D_f^n(X_M^n, s) \leq 0$, and the rates determined by $X_{M^+}^n$ when $D_f^n(X_M^n, s) > 0$.

We do a similar construction for X_m^n. Let

$$X_{m^+}^n \equiv (Q_{1,m}^n, Q_{2,m}^n, Z_{m^+}^n) \qquad \text{and} \qquad X_{m^-}^n \equiv (Q_{1,m}^n, Q_{2,m}^n, Z_{m^-}^n),$$

where

$$Q_{1,m}^n \equiv \|Q_{1,b}^n\|_\delta, \qquad Q_{2,m}^n \equiv \inf_{t \in I_{(t_1, \delta)}} Q_{2,a}^n(t) \vee 0,$$

$$Z_{m^+}^n \equiv \|Z_+^n\|_\delta, \qquad Z_{m^-}^n \equiv \inf_{t \in I_{(t_1, \delta)}} Z_-^n(t),$$

with $Z_+^n(t) \equiv Z_a^n(t)$ and $Z_-^n(t) \equiv Z_b^n(t)$ if $\mu_{2,2} \geq \mu_{1,2}$, and $Z_+^n(t) \equiv Z_b^n(t)$ and $Z_-^n(t) \equiv Z_a^n(t)$ otherwise (the reverse of what is done in X_M^n). Let $\{D_f^n(X_m^n, s) : s \geq 0\}$ have the rates from $X_{m^-}^n$ when $D_f^n(X_m^n, s) \leq 0$, and the rates from $X_{m^+}^n$ when $D_f^n(X_m^n, s) > 0$. By this construction, we achieve the ordering in (59). We cover the rates of $D_{1,2}^n(t)$ too because we can make the identification: the rates of $D_{1,2}^n(t)$ given $X^n(t)$ coincide with the rates of $D_f^n(X^n(t), \cdot)$.

It remains to find a δ such that both the processes $\{D_f^n(X_m^n, s) : s \geq 0\}$ and $\{D_f^n(X_M^n, s) : s \geq 0\}$ are asymptotically positive recurrent. To do so, we use Lemma A.3, which concludes that the bounding processes as functions of δ have fluid limits. By Lemma A.3, we can conclude that $\bar{X}_{m^+}^n \equiv n^{-1} X_{m^+}^n \Rightarrow x_{m^+}$, $\bar{X}_{m^-}^n \equiv n^{-1} X_{m^-}^n \Rightarrow x_{m^-}$, $\bar{X}_{M^+}^n \equiv n^{-1} X_{M^+}^n \Rightarrow x_M^+$ and $\bar{X}_{M^-}^n \equiv n^{-1} X_{M^-}^n \Rightarrow x_M^-$ in $\mathcal{D}_3$, where x_{m^+}, x_{m^-}, x_M^+ and x_M^- are all continuous with $x_m^+(t_1 \vee (t - \delta)) = x_m^-(t_1 \vee (t - \delta)) = x_M^+(t_1 \vee (t - \delta)) = x_M^-(t_1 \vee (t - \delta)) = x(t_1 \vee (t - \delta)) \in \mathbb{A}$. Hence, we can find δ' such that $x_m(\delta) \in \mathbb{A}$ and $x_M(\delta) \in \mathbb{A}$ for all $\delta \in [0, \delta']$. Hence, we can choose δ such that the constant vectors $x_m \equiv x_m(\delta)$ and $x_M \equiv x_M(\delta)$ both are arbitrarily close to $x(t_1)$.

Finally, we use the linearity of the drift function to deduce the positive recurrence of the processes depending upon n. As $n \to \infty$,

$$\delta_-^n(X_{m^-}^n)/n \Rightarrow \delta_-(x_m^-), \qquad \delta_+^n(X_{m^+}^n)/n \Rightarrow \delta_+(x_m^+),$$

$$\delta_-^n(X_{M^-}^n)/n \Rightarrow \delta_-(x_M^-), \qquad \text{and} \qquad \delta_+^n(X_{M^+}^n)/n \Rightarrow \delta_+(x_M^+) \quad \text{in } \mathbb{R}. \quad \square$$

We immediately obtain the following corollary to Lemma 8.4, exploiting Corollary A.1.

COROLLARY 8.1. *Let $\zeta \equiv (j \vee k) - 1$ using the QBD representation based on $r_{1,2} = j/k$ in §6 of Perry and Whitt [37]. If, in addition to the conditions of Lemma 8.4,*

$$D_f^n(X_m^n, 0) - \zeta \leq_{st} D_f^n(X^n(0 \vee t - \delta), 0) \leq_{st} D_f^n(X_M^n, 0) + \zeta \qquad \text{and}$$

$$D_f^n(X_m^n, 0) - \zeta \leq_{st} D_{1,2}^n(0 \vee t - \delta) \leq_{st} D_f^n(X_M^n, 0) + \zeta \quad \text{in } \mathbb{R}, \quad n \geq 1, \tag{61}$$

then, in addition to the conclusions of Lemma 8.4,

$$D_f^n(X_m^n, \cdot) - \zeta \leq_{st} D_f^n(X^n(t), \cdot) \leq_{st} D_f^n(X_M^n, \cdot) + \zeta,$$

$$D_f^n(X_m^n, t) - \zeta \leq_{st} D_{1,2}^n(t) \leq_{st} D_f^n(X_M^n, t) + \zeta \quad \text{in } \mathcal{D}([t_1 \vee (t - \delta), (t + \delta) \wedge t_2]). \tag{62}$$

The following lemmas correspond to the QBD structure of the FTSP. These results hold for QBD processes much more generally, but we state them in terms of the FTSP.

LEMMA 8.5 (CONTINUITY OF THE STATIONARY DISTRIBUTION OF THE FTSP). *The FTSP stationary random variable $D(\gamma, \infty)$ is continuous in the metric* (63) *as a function of the state γ in $\mathbb{A}$ and thus $D(x(t), \infty)$ is continuous in the metric* (63) *as a function of the time argument t when $x(t) \in \mathbb{A}$.*

PROOF. This result follows from Theorem 7.1 in Perry and Whitt [37], in particular from the stronger differentiability of the QBD R matrix in an open neighborhood of each $\gamma \in \mathbb{A}$, building on Theorem 2.3 in He [17]. The second statement follows from the first, together with the continuity of $x(t)$ as a function of t. □

REMARK 8.1 (CONTINUITY ON $\mathbb{S}$). Lemma 3.1 shows that the stationary distribution of the FTSP is well defined on all of $\mathbb{S}$ provided that we extend the set of possible values of the FTSP to the space $E \equiv \mathbb{Z} \cup \{+\infty\} \cup \{-\infty\}$, as in §7.2. Following Theorem 7.1 of Perry and Whitt [37], we can show that Lemma 8.5 holds, not only on $\mathbb{A}$, but also on $\mathbb{S}$, but that extension is not needed here.

Let $\tau(t) \equiv \inf\{u > 0: D(x(t), u) = s\}$, where s is a state in the state space of the QBD D. The next lemma establishes the existence of a finite moment generating function (mgf) for $\tau(t)$ for a positive recurrent FTSP.

LEMMA 8.6 (FINITE MGF FOR RETURN TIMES). *For $x(t) \in \mathbb{A}$, let $\tau \equiv \tau(t)$ be the return time of the positive recurrent QBD $D(x(t), \cdot)$ to a specified state s. Then there exists $\theta^* > 0$ such that $\phi_\tau(\theta) \equiv E[e^{\theta\tau}] < \infty$ for all $\theta < \theta^*$.*

PROOF. As for any irreducible positive recurrent CTMC, a positive recurrent QBD is regenerative, with successive visits to any state constituting an embedded renewal process. As usual for QBDs (see Latouche and Ramaswami [28]), we can choose to analyze the system directly in continuous time or in discrete time by applying uniformization, where we generate all potential transitions from a single Poisson process with a rate exceeding the total transition rate out of any state. In continuous time we focus on the interval between successive visits to the regenerative state; in discrete time we focus on the number of Poisson transitions between successive visits to the regenerative state.

Let N be the number of Poisson transitions (with specified Poisson rate). The number of transitions, N, has the generating function (gf) $\psi_N(z) \equiv E[z^N]$, for which there exists a radius of convergence z^* with $0 < z^* < 1$ such that $\psi_N(z) < \infty$ for $z < z^*$ and $\psi_N(z) = \infty$ for $z > z^*$.

The mgf $\phi_\tau(\theta)$ and gf $\psi_N(z)$ can be expressed directly in terms of the finite QBD defining matrices. It is easier to do so if we choose a regenerative state, say s^*, in the boundary region (corresponding to the matrix B in (6.5)–(6.6) of Perry and Whitt [37]). To illustrate, we discuss the gf. With s^* in the boundary level, in addition to the transitions within the boundary level and up to the next level from the boundary, we only need consider the number of transitions, plus starting and ending states, from any level above the boundary down one level. Because of the QBD structure, these key downward first passage times are the same for each level above the boundary, and are given by the probabilities $G_{i,j}[k]$ and the associated matrix generating function $G(z)$ on p. 148 of Latouche and Ramaswami [28]. Given $G(z)$, it is not difficult to write an expression for the generating function $\psi_{N^n}(z)$, just as in the familiar birth-and-death process case; e.g., see §4.3 of Latouche and Ramaswami [28]. □

The next lemmas establish results regarding distances between processes in a discrete state apace. In particular, we consider the state-space E of the processes $D_{1,2}^n$ and the FTSP D, which is a countable lattice. With the QBD representation (achieved by renaming the states in E) the state-space E is a subset of $\mathbb{Z}$. To measure distances between probability distributions on $\mathbb{Z}$, corresponding to convergence in distribution, we use the Lévy metric, defined for any two cdf's F_1 and F_2 by

$$\mathcal{L}(F_1, F_2) \equiv \inf\{\epsilon > 0: F_1(x - \epsilon) - \epsilon \le F_2(x) \le F_1(x + \epsilon) + \epsilon \text{ for all } x\}. \tag{63}$$

For random variables X_1 and X_2, $\mathcal{L}(X_1, X_2)$ denotes the Lévy distance between their probability distributions. (The Lévy metric is defined for probability distributions on $\mathbb{R}$, but we will use it for processes defined on the discrete space E.)

LEMMA 8.7 (UNIFORM BOUNDS ON THE RATE OF CONVERGENCE TO STATIONARITY). *For every $t_0 \in [t_1, t_2]$ there exist $\zeta > 0$ and constants $\beta_0 < \infty$ and $\rho_0 > 0$, such that*

$$\mathcal{L}(D(x(t), s), D(x(t), \infty)) \le \beta_0 e^{-\rho_0 s} \quad \text{in } \mathbb{R} \tag{64}$$

holds for all $t \in [t_1 \vee (t_0 - \zeta), t_0 + \zeta] \subset [t_1, t_2]$.

For our proof of Lemma 8.7, we also introduce the concept of *uniformly* (u, α)-*small* sets on an interval, which generalizes the concept of (u, α)-*small* sets; for background see, e.g., Kontoyiannis and Meyn [26], Meyn and Tweedie [32], and Roberts and Rosenthal [40]. For each (fixed) $t \geq 0$ let $P_{i,j}^t(u)$ denote the transition probabilities of $\{D(x(t), u): u \geq 0\}$:

$$P_{i,j}^t(u) \equiv P(D(x(t), u) = j \mid D(x(t), 0) = i), \quad i, j \in \mathbb{Z}.$$

DEFINITION 8.1. A set $C \subseteq \mathbb{Z}$ is (u, α)-*small*, for a time $u > 0$ and for some $\alpha > 0$, if there exists a probability measure $\varphi^t(\cdot)$ on $\mathbb{Z}$ satisfying the *minorization condition*

$$P_{i,j}^t(u) \geq \alpha \varphi^t(j), \quad i \in C, \ j \in \mathbb{Z}.$$

We say that a set C is *uniformly* (u, α)-*small* in the interval I, if C is (u, α)-small for each process in the family $\{D(x(t), \cdot): t \in I\}$, i.e., if C is (u, α)-small for each $D(x(t), \cdot)$ with the same u and α, $t \in I$.

Note that the probability measure φ^t in the minorization condition is allowed to depend on t in the definition of uniformly (u, α)-small.

PROOF OF LEMMA 8.7. As above, let $\tau(t)$ denote the return time of the process $D(x(t), \cdot)$ to the regeneration state s^*, which, for concreteness, we take to be state 0, i.e., $s^* = 0$. Consider the infinitesimal generator matrix $Q(t)$ of the process $D(x(t), \cdot)$. In a countable state space, every compact set is *small* (see, e.g., definition on p. 11 in Kontoyiannis and Meyn [26]) and, in particular, $\{0\}$ is a small set. (We show in (66) that $\{0\}$ is in fact uniformly (u, α)-small.) Moreover, by Theorem 2.5 in Kontoyiannis and Meyn [26], the existence of a finite mgf for the hitting time of the small set $\{0\}$ is equivalent to the existence of a Lyapunov function $V: \mathbb{Z} \to [1, \infty)$, which satisfies the exponential drift condition on the generator, Condition (**V4**) in Kontoyiannis and Meyn [26]:

$$Q(t)V \leq -c_t V + d_t \mathbf{1}_{\{0\}}, \tag{65}$$

where c_t and d_t are strictly positive constants.

Consider the time $t_0 \in [t_1, t_2]$. Then (65) holds at t_0 with constants c_{t_0} and d_{t_0}. Since $c_{t_0} > 0$ we can decrease it such that (65) holds with strict inequality and the new c_{t_0} is still strictly positive. We increase d_{t_0} appropriately, such that $\sum_j q_{0,j}(t)V(j) < -c_{t_0}V(0) + d_{t_0}\mathbf{1}_{\{0\}}$. The continuity of $Q(t)$ on $[t_1, t_2]$ as a function of t, which follows immediately from the continuity of the rates (13)–(16) as functions of the continuous function $x(t)$, implies that there exist $\zeta > 0$ and two positive constants c_0 and d_0, such that (65) holds for all $t \in [t_0 - \zeta, t_0 + \zeta]$ with the same constants c_0 and d_0. However, this is still not sufficient to conclude that the bounds in (64) are the same for all $t \in [t_0 - \zeta, t_0 + \zeta]$; see Theorem 1.1 in Baxendale [6] (for discrete-time Markov chains).

Recall that for each fixed t, $P_{i,j}^t(s)$ denotes the transition probabilities of the CTMC $\{D(x(t), s): s \geq 0\}$. We can establish uniform bounds on the convergence rates to stationarity by showing that $\{0\}$ is *uniformly* (u, α)-*small* in an interval $[t_0 - \zeta, t_0 + \zeta]$ for the family $\{D(x(t), \cdot): t \in [t_0 - \zeta, t_0 + \zeta]\}$, as in Definition 8.1. In particular, we need to show that

$$P_{0,j}^t(u) \geq \alpha \varphi^t(j), \quad j \in \mathbb{Z} \tag{66}$$

holds for all $t \in [t_0 - \zeta, t_0 + \zeta]$ with the same α (but φ^t is allowed to change with t). This step, together with the uniform bounds c_0 and d_0 in (65) established above, will be shown to be sufficient to conclude the proof.

Hence, it is left to show that (66) holds for all $t \in [t_0 - \zeta, t_0 + \zeta]$ with the same $\alpha > 0$. This step is easy because $\{0\}$ is a singleton in a countable state space. Specifically, for each t we consider, we can fix any $u > 0$ and define $\varphi^t(j) \equiv P_{0,j}^t(u)$. With this definition of φ^t we can take any $\alpha \leq 1$ in (66). As in the discrete-time case in Baxendale [6] (the *strong aperiodicity condition* (A3) in Baxendale [6] is irrelevant in continuous time), the bounds on the convergence rates in (64) depend explicitly on α in the minorization condition (66), the bounds in the drift condition (65), and the Lyapunov function V in (65). This can be justified by uniformization, but can also be justified directly for continuous-time processes, e.g., from the expressions in Theorem 3 and Corollary 4 in Roberts and Rosenthal [40].

The uniform bounds on the rate of convergence to steady state established above by applying Roberts and Rosenthal [40] are directly expressed in the total-variation metric. If the total variation metric can be made arbitrarily small, then so can the Levy metric. Hence we have completed the proof.

REMARK 8.2 (BOUNDS ON COUPLING TIMES). The bounds on the rate of convergence to stationarity of Markov processes satisfying (65) and (66) in Roberts and Rosenthal [40] are obtained via the coupling inequality. In particular, we have provided explicit bounds for the time it takes a positive recurrent FTSP $D(\gamma, \cdot)$ (with $\gamma \in \mathbb{A}$), initialized at some finite time, to couple with its stationary version.

The final auxiliary lemma relates the Lévy distance $\mathscr{L}$ in (63) between $D_e^n(X^n(u), s_0)$ in (12) and the FTSP $D(x(u), s_0)$ at a finite time s_0.

LEMMA 8.8. *Suppose that $x(t) \in \mathbb{A}$ for $t_1 \leq t \leq t_2$. Then and for any fixed $s_0 > 0$,*

$$\mathscr{L}(D_e^n(X^n(u), s_0), D(x(u), s_0)) \to 0 \quad \text{as } n \to \infty, \quad \text{uniformly in } u \in [t_1, t_2]. \tag{67}$$

PROOF. If follows from the proof of Theorem 4.4 that, for the given $\epsilon > 0$, there exists $\delta(t_1)$, $n_0(t_1)$, and $s_0(t_1)$ such that

$$\mathscr{L}(D_e^n(X^n(v), s_0(t_1)), D(x(v), s_0(t_1))) < \epsilon \quad \text{for all } t_1 \leq v \leq (t_1 + \delta(t_1)) \wedge t_2 \text{ and } n \geq n_0(t_1)$$

exploiting the convergence $\bar{X}^n \Rightarrow x$ and $D^n(\Gamma^n, 0) = D_{1,2}^n(t_1) \Rightarrow D(x(t_1), 0) = D(\gamma, 0)$ for $\Gamma^n \equiv X^n(t_1)$ and $\gamma \equiv x(t_1)$. We can apply this reasoning in an open interval about each $u \in [t_1, t_2]$. In particular, for the given $\epsilon > 0$ and $u \in [t_1, t_2]$, there exists $\delta(u)$, $n_0(u)$, and $s_0(u)$ such that

$$\mathscr{L}(D_e^n(X^n(v), s_0(u)), D(x(v), s_0(u))) < \epsilon \quad \text{for all } (u - \delta(u)) \vee t_1 \leq v \leq (u + \delta(u)) \wedge t_2 \text{ and } n \geq n_0(u).$$

However, because the interval $[t_1, t_2]$ is compact and the family of intervals $(t_1 \vee (u - \delta(u)), (u + \delta(u)) \wedge t_2)$, taken to be closed on the left at t_1 and closed on the right at t_2, is an open cover, there is a finite subcover. Hence, all time points in $[t_1, t_2]$ are contained in only finitely many of these intervals. Hence, we can achieve the claimed uniformity. Moreover, because the conclusion does not depend on the subsequence used at the initial time t_1, the overall proof is complete.

8.4. Proofs of Theorems 4.5 and 4.6. We apply the results in §8.3 to prove the theorems.

PROOF OF THEOREM 4.5. First, if $x(t_0) \in \mathbb{A}$ then there exists t_2 such that $x(t) \in \mathbb{A}$ over $[t_0, t_2]$ because $\mathbb{A}$ is an open set and x is continuous. It is possible that the fluid limit never leaves $\mathbb{A}$ after time t_0, in which case $x \in \mathbb{A}$ over $[t_0, \infty)$. In the latter case we can consider any $t_2 > t_0$.

(i) We begin by showing that there exists $\delta_0 \equiv \delta(t_0) > 0$, such that $D_{1,2}^n(t_1)$ is tight in $\mathbb{R}$ for all t_1 satisfying $t_0 < t_1 < t_0 + \delta_0$. Henceforth, t_1 denotes such a time point. We need to show that, for any $\epsilon > 0$, there exists a constant K such that $P(|D_{1,2}^n(t_1)| > K) < \epsilon$ for all $n \geq 1$. Overall we prove the claim in three steps: in the first step we bound $D_{1,2}^n$ over $[t_0, t_0 + \delta_0]$, in sample-path stochastic order, with positive recurrent QBDs but with random initial conditions. In the second step, we show that these bounding QBDs are tight for each t_1 as above, by showing that they couple with stationary versions rapidly enough. In the third step we show that we can extend the conclusion from the subinterval $[t_1, t_0 + \delta_0]$ to the entire interval $[t_1, t_2]$.

Step 1. We apply Lemma 8.4 to find a $\delta_0 > 0$ and construct random states X_m^n and X_M^n with the properties stated there, such that rate order holds as in (59) in $\mathscr{D}([t_0, t_0 + \delta_0])$, after we let $D_f^n(X_m^n, t_0) \equiv D_f^n(X_M^n, t_0) \equiv D_{1,2}^n(t_0)$. Next, for all sufficiently large n, we bound the upper bounding process above and the lower bounding process below in rate order, each by a FTSP with fixed states $\tilde{x}_m$ and $\tilde{x}_M$ but ordered so that strict rate order in Lemma A.4 holds. Because x_m and x_M are in $\mathbb{A}$ and $\mathbb{A}$ is open, these new states $\tilde{x}_m$ and $\tilde{x}_M$ can be chosen to be sufficiently near the initial states x_m and x_M that they too are in $\mathbb{A}$. We let these FTSPs be given the same initial conditions depending on n, consistent with above. For that purpose, we put the initial condition in the notation; i.e., we let $D(\gamma, b, \cdot) \equiv D(\gamma, \cdot)$ given that $D(\gamma, 0) = b$. We combine the rate order just established with Corollary 8.1 to obtain the sample-path stochastic order

$$D(\tilde{x}_m, D_{1,2}^n(0), nt) - \zeta \leq_{st} D_{1,2}^n(t) \leq_{st} D(\tilde{x}_M, D_{1,2}^n(0), nt) + \zeta \tag{68}$$

for all $n \geq n_0$ for some n_0.

So far, we have succeeded in bounding $D_{1,2}^n(t)$ above (and below) stochastically by a positive recurrent FTSP with random initial conditions, in particular starting at $D_{1,2}^n(0)$. It suffices to show that each bounding FTSP with these initial conditions in stochastically bounded at any time $t_1 \in (t_0, t_0 + \delta_0]$. We show that next.

Step 2. We now show that the two bounding processes in (68) are indeed stochastically bounded at such t_1. We do that by showing that they couple with the stationary versions of these FTSPs with probability converging to 1 as $n \to \infty$. Since $x(t_0) \in \mathbb{A}$, we can apply Theorem 4.1 to deduce that $\bar{X}^n(t_0) \Rightarrow x(0)$ as $n \to \infty$, which implies that $D_{1,2}^n(0) = o(n)$ as $n \to \infty$. If we consider the FTSP without the scaling by n in (68), it suffices to show, for t_1 as above, the bounding processes recover from the $o(n)$ initial condition within $o(n)$ time, which will be implied by showing that the coupling takes place in time $o(n)$.

We show that by applying Lemma 8.7 and Remark 8.2. They imply that the time until the FTSP in (68) couples with its stationary version is bounded by constants depending on c_{t_0} and d_{t_0} in (65) and α in (66), as well as the Lyapunov function V in (65). Now, $D(\tilde{x}_M, t_0) = D_{1,2}^n(t_0)$ by construction, and $x(t_0) \in \mathbb{A}$, so that $D(\tilde{x}_M, t_0) = o(n)$. Hence, $V(D(\tilde{x}_M, t)) = o(n)$ for all $t \geq t_0$, for V in (65). In particular, we see that the time until $D(\tilde{x}_M, t_0)$ couples with it stationary version, when initialized at $D_{1,2}^n(t_0)$, is $o(n)$. Together with (68), that implies the claim. That is, we have shown that $D_{1,2}^n(t_1)$ is tight in $\mathbb{R}$ for all t_1, $t_0 < t_1 < t_0 + \delta_0$.

Step 3. To extend the result to the interval $[t_1, t_2]$ we observe that we can repeat the reasoning above for any starting point $t \in [t_1, t_2]$ (since $x(t) \in \mathbb{A}$ for all $t \in [t_1, t_2]$), achieving an uncountably infinite cover for $[t_1, t_2]$ of intervals of the form $[t, t + \delta(t)]$. Because the interval $[t_1, t_2]$ is compact, the uncountably infinite cover of these intervals, made open at the left unless the left endpoint is t_1 and made open on the right unless the right endpoint is t_2, has a finite subcover. As a consequence, the entire interval $[t_1, t_2]$ is covered by only finitely many of these constructions, and we can work with the finite collection of closures of these intervals.

In particular, since the sequence $\{D_{1,2}^n(t_1): n \geq 1\}$ is stochastically bounded, we can apply the construction over an interval of the form $[t_1, t']$ for $t_1 < t' \leq t_2$. As a consequence we obtain stochastic boundedness for each t in $[t_1, t']$. If $t' < t_2$, then we continue. We then can choose a second interval $[t'', t''']$ such that $t_1 < t'' \leq t' < t'''$. We thus can carry out the construction over $[t'', t''']$. Since $t'' \leq t'$, we already know that the sequence of random variables $\{D_{1,2}^n(t''); n \geq 1\}$ is stochastically bounded from the first step. Thus, in finitely many steps, we will deduce the first conclusion in (i). If $x(0) \in \mathbb{A}$ and we take $t_0 = 0$, then $D_{1,2}^n(t_0)$ is tight in $\mathbb{R}$ because $D_{1,2}^n(0) \Rightarrow L$ by Assumption 3.2, so the result holds on $[t_0, t_2] \equiv [0, t_2]$.

(ii) To prove the statement in (ii) we observe that Theorem 4.4 implies that the oscillations are asymptotically too rapid for the sequence $\{D_{1,2}^n: n \geq 1\}$ to be tight in $\mathcal{D}$ over any finite interval. Moreover, there is a finite interval over which a single frozen-difference process serves as a lower bound, by Lemma 8.4. Because of the scaling by n, the maximum in the lower bound QBD over this interval will be unbounded above (actually of order $O(\log n)$ by reasoning as in Lemma A.6). Thus, the sequence of stochastic processes $\{D_{1,2}^n: n \geq 1\}$ is not even stochastically bounded over any finite subinterval of $[t_0, t_2]$.

(iii) We apply Lemma A.6 to establish (33). We can bound the supremum of $D_{1,2}^n(t)$ over the interval $[t_1, t_2]$ by the supremum of the finitely many frozen queue-difference processes. Because the rates are of order n, Lemma A.6 implies (33). Once again, the result can be extended to hold on $[t_0, t_2] \equiv [0, t_2]$ if $x(0) \in \mathbb{A}$ by the assumed convergence $D_{1,2}^n(0) \Rightarrow L$ in Assumption 3.2. $\quad\square$

PROOF OF THEOREM 4.6. For any given t with $t_1 < t \leq t_2$ and $\epsilon > 0$, we will show that we can choose n_0 such that $\mathcal{L}(D_{1,2}^n(t), D(x(t), \infty)) < \epsilon$ for all $n \geq n_0$, where $\mathcal{L}$ is the Lévy metric in (63). Since $\{D_{1,2}^n(t_1): n \geq 1\}$ is stochastically bounded, it is tight. Hence, we start with a converging subsequence, without introducing subsequence notation. The result will not depend on the particular converging subsequence we choose.

Hence, we start with $D_{1,2}^n(t_1) \Rightarrow L$, where L is a proper (almost surely finite) random variable. We will then let $D(x(t_1), 0) = L$ to obtain $D_{1,2}^n(t_1) \Rightarrow D(x(t_1), 0)$, as required to apply Theorem 4.4 at t_1. If $t > t_1$, then for all sufficiently large n, $t - s/n \geq t_1$, in which case we can write $D_{1,2}^n(t) = D_{1,2}^n(t - s/n + s/n) = D_e^n(X^n(t - (s/n)), s)$, where D_e^n is the expanded queue-difference process defined in (12), with $t - s/n \geq t_1$ so that $x(t - s/n) \in \mathbb{A}$. Hence we can write

$$
\begin{aligned}
\mathcal{L}(D_{1,2}^n(t), D(x(t), \infty)) &\leq \mathcal{L}(D_e^n(X^n(t - s/n), s), D(x(t - s/n), s)) \\
&\quad + \mathcal{L}(D(x(t - s/n), s), D(x(t - s/n), \infty)) \\
&\quad + \mathcal{L}(D(x(t - s/n), \infty), D(x(t), \infty)).
\end{aligned} \tag{69}
$$

For t and any $\epsilon > 0$ given, we first choose δ so that $t - \delta \geq t_1$ and x remains in $\mathbb{A}$ throughout $[0, t + \delta]$, which is always possible because $x(t) \in \mathbb{A}$, $t > t_1$, x is continuous and $\mathbb{A}$ is an open subset of $\mathbb{S}$. (We use the condition that $t > t_1$ here.) In addition, we choose δ sufficiently small that the third term in (69) is bounded as

$$
\mathcal{L}(D(x(u), \infty)), D(x(t), \infty)) < \epsilon/3 \quad \text{for all } u, \ t - \delta \leq u \leq t + \delta,
$$

which is possible by virtue of Lemma 8.5. Given t, ϵ, and δ, we choose s_0 sufficiently large that the second term in (69) is bounded as

$$
\mathcal{L}(D(x(u), s_0), D(x(u), \infty)) < \epsilon/3 \quad \text{for all } u, \ t - \delta \leq u \leq t + \delta,
$$

which is possible by Lemma 8.7. Finally, we choose n_0 sufficiently large that $s_0/n < \delta$ for all $n \geq n_0$ and the first term in (69) is bounded as

$$
\mathcal{L}(D_e^n(X^n(u, s_0)), D(x(u), s_0)) < \epsilon/3 \quad \text{for all } u, \ t - \delta \leq u \leq t + \delta,
$$

which is possible by Lemma 8.8. That choice of δ, s_0, and n_0 makes each term in (69) less than or equal to $\epsilon/3$ for all $n \geq n_0$, thus completing the proof.

9. Conclusion and further research. In this paper we proved a FWLLN (Theorem 4.1) for an overloaded X model operating under the FQR-T control (§2.2), with many-server heavy-traffic scaling (§2.3). Theorem 4.1 shows that the fluid-scaled version of the six-dimensional Markov chain X_6^n, whose sample-path representation appears in Theorem 5.1, converges to a deterministic limit, characterized by the unique solution to the three-dimensional ODE (24), which in turn is driven by the FTSP in §3.2. We also proved a WLLN for the stationary distributions (Theorem 4.3) that justifies a limit interchange in great generality (Theorem 4.2).

Finally, in §4.3 we presented results regarding the queue difference process and SSC for the queues when the fluid limit is in $\mathbb{A}$. in particular, Theorem 4.5 proved statements regarding the tightness of $D_{1,2}^n(t)$ in $\mathbb{R}$, and nontightness in $\mathcal{D}$. Corollary 4.1 shows that SSC for the queues hold under any scaling larger than $O(\log n)$, depending only on whether the fluid limit x is in $\mathbb{A}$. Theorem 4.6 establishes a pointwise AP result, which is not an immediate corollary of the AP in the FWLLN.

Proof of the FWLLN. We proved the FWLLN in three steps, (i) showing that the sequence of processes $\{\bar{X}_6^n \colon n \geq 1\}$ is tight and every limit is continuous (Theorem 5.2); (ii) simplifying the representation in Theorem 5.1 to the essentially three-dimensional representation in (45) (Corollary 6.4); and (iii) characterizing the limit via the averaging principle (AP) (§7). Characterizing the fluid limit of that three-dimensional process was challenging because the sequence of queue-difference processes $\{\{D_{1,2}^n(t) \colon t \geq 0\} \colon n \geq 1\}$ in (6) does not converge to any limiting process as $n \to \infty$. Instead, we have the AP, which can be better understood through Theorems 4.4–4.6.

Because of the AP, the indicator functions $1_{\{D_{1,2}^n(s)>0\}}$ and $1_{\{D_{1,2}^n(s)\leq 0\}}$ in the representation (45) are replaced in the limit with appropriate steady-state quantities related to the FTSP $D(x(t), \cdot)$, e.g., $1_{\{D_{1,2}^n(s)>0\}}$ is replaced in the limit with

$$\pi_{1,2}(x(s)) \equiv P(D(x(s), \infty) > 0) = \lim_{t \to \infty} \frac{1}{t} \int_0^t 1_{\{D(x(s), u)>0\}} \, du.$$

Our proof of the AP in §7 is based on the framework established by Kurtz [27]. In the appendix we present a different proof of the AP that is weaker, because it only characterizes the FWLLN in the set $\mathbb{A}$. However, it has the advantage of being intuitive, with an explicit demonstration of the separation of time scales, which takes place in $\mathbb{A}$. Moreover, this proof continues the bounding logic of §§A and B, and thus follows naturally from previous results established in the paper. Both proofs have merits, and can be useful in other models where separation of time scales occur in the limit.

Results for QBD processes. In the process of proving the SSC in Theorem 6.4 and Theorems 4.5 and 4.6, we established some general results for QBDs, which are interesting in their own right. First, we established an extreme-value result in Lemma A.6. Then, in §8.3 we established the continuity of stationary distributions (Lemma 8.5), finite mgf for return times (Lemma 8.6), and uniform ergodicity for a family of ergodic QBD processes (Lemma 8.7). To the best of our knowledge, those results are not stated in the existing literature.

The control. The FQR-T control is appealing, not only because it is optimal during the overload incident in the fluid limit Perry and Whitt [35], but also because the FQR-T control produces significant simplification of the limit, because it produces a strong form of SSC. Specifically, by Theorem 6.4, the four-dimensional service process is asymptotically one dimensional, with $Z_{1,2}^n$ alone characterizing limits under any scaling. By Theorem 4.5, the two-dimensional queue process is asymptotically one dimensional under any scaling larger than $O(\log n)$ (in particular, under fluid and diffusion scaling). This latter result holds unless class 1 is so overloaded that there is not enough service capacity in both pools to keep the desired ratio between the two queues (in which case the fluid limit will not be in $\mathbb{A}$). Hence, the six-dimensional Markov chain describing the system during overloads in the prelimit is replaced by a simplified lower dimensional deterministic function, which is easier to analyze. In addition, these SSC results are crucial to proving the FCLT for the system (Perry and Whitt [38]).

Further research. As Lemma B.1 and the proof of Theorem 6.4 reveal, one-way sharing, which we suggested in Perry and Whitt [35] as a means to prevent unwanted simultaneous sharing of customers in finite systems (where the thresholds themselves may not be sufficient to prevent two-way sharing) has shortcomings; see Remark B.1. For example, in the limiting fluid model, if after some time the overload switches over, so that queue 2 should start receiving help from pool 1, then the one-way sharing rule will not allow class-2 customers to be sent to pool 1 because, as was shown in the proof of Theorem 6.4, $z_{1,2}(t) > 0$ for all $t > 0$. As a consequence, in large systems, a significant amount of time must pass before sharing in the opposite direction is allowed. This problem with one-way sharing can be remedied by dropping that rule completely, or by introducing lower thresholds on the service process; again see Remark B.1. We have begun studying such alternative controls.

Appendix. This Appendix has six sections. In §A we establish important technical results to be used to prove the results in §6. In §B we apply the results in §A to prove the results in §6. In §C we present an alternative proof of Lemma 7.2 and thus an alternative proof of the FWLLN in Theorem 4.1. In §D we show how to present the process D_*^n in step 1 of the proof of Lemma B.3 as a QBD for each n. In §E we explain why Assumption 3.2 about the initial conditions is reasonable. Finally, in §F we list all the acronyms used in this paper.

Appendix A. Supporting technical results. In this section we establish supporting technical results that we will apply to prove the results in §6. In §A.1 we introduce a frozen queue difference process, which "freezes" the state of the slow process ($\bar{X}^n$ in (45)), so that the fast process (the queue difference process $D_{1,2}^n$ in (6)) can be considered separately. Like the FTSP in §3.2, the frozen process is a pure-jump Markov process. To carry out the proofs, we exploit stochastic bounds. Hence, we discuss them next in §§A.2, A.3, and A.4. Because the FTSP and the frozen processes can be represented as QBD processes, as indicated in §6 of Perry and Whitt [37], we next establish extreme value limits for QBDs in §A.5. We establish continuity results for QBDs, used in the alternative proof of characterization, in §C.3.

A.1. Auxiliary frozen processes. Our proof of Theorem 6.1 will exploit stochastic bounds for the queue difference process $D_{2,1}^n$. Our alternative proof of Theorem 4.1 will exploit similar stochastic bounds for the queue difference process $D_{1,2}^n$. These processes, $D_{2,1}^n$ and $D_{1,2}^n$, are non-Markov processes whose rates at each time t are determined by the state of the system at time t, i.e., by $X_6^n(t)$, and are thus hard to analyze directly. To circumvent this difficulty, we consider related auxiliary processes, with *constant* rates determined by a fixed initial state $\Gamma^n \equiv X_6^n(0)$. The construction is essentially the same for the two processes $D_{2,1}^n$ and $D_{1,2}^n$. Because we are primarily concerned with $D_{1,2}^n$, we carry out the following in that context, with the understanding that there is a parallel construction for the process $D_{2,1}^n$.

When we work with $D_{1,2}^n$, we exploit the reduced representation involving X^n in (45). Let $D_f^n(\Gamma^n) \equiv \{D_f^n(\Gamma^n, t): t \geq 0\}$ denote this new process with fixed state $\Gamma^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n)$. Conditional on Γ^n, $D_f^n(\Gamma^n)$ is a QBD with the same fundamental structure as the FTSP defined in §3.2. We use the subscript f because we refer to this constant-rate pure-jump Markov process as the *frozen queue-difference process*, or alternatively, as the *frozen process*, thinking of the constant transition rates being achieved because the state has been frozen at the state Γ^n.

As in §3.2, the rates of the frozen process are determined by $\Gamma^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n)$ and by its position. Also, as in §3.2, let $r \equiv r_{1,2}$. The frozen process $D_f^n(\Gamma^n)$ has jumps of size $+r, +1, -r, -1$ with respective rates $\lambda_+^{(r)}(n, \Gamma^n), \lambda_+^{(1)}(n, \Gamma^n), \mu_-^{(r)}(n, \Gamma^n), \mu_-^{(1)}(n, \Gamma^n)$ when $D_f^n \leq 0$, and jumps of size $+r, +1, -r, -1$ with respective rates $\lambda_+^{(r)}(n, \Gamma^n), \lambda_+^{(1)}(n, \Gamma^n), \mu_+^{(r)}(n, \Gamma^n), \mu_+^{(1)}(n, \Gamma^n)$ when $D_f^n > 0$.

Analogously to (13)–(16), in the nonpositive state space these rates are equal to

$$\lambda_-^{(1)}(n, \Gamma^n) \equiv \lambda_1^n \qquad \text{and} \qquad \lambda_-^{(r)}(n, \Gamma^n) \equiv \mu_{1,2}Z_{1,2}^n + \mu_{2,2}(m_2^n - Z_{1,2}^n) + \theta_2 Q_2^n,$$
$$\mu_-^{(1)}(n, \Gamma^n) \equiv \mu_{1,1}m_1^n + \theta_1 Q_1^n \qquad \text{and} \qquad \mu_-^{(r)}(n, \Gamma^n) \equiv \lambda_2^n, \tag{A1}$$

and in the positive state space, these rates are equal to

$$\lambda_+^{(1)}(n, \Gamma^n) \equiv \lambda_1^n \qquad \text{and} \qquad \lambda_+^{(r)}(n, \Gamma^n) \equiv \theta_2 Q_2^n,$$
$$\mu_+^{(1)}(n, \Gamma^n) \equiv \mu_{1,1}m_1^n + \mu_{1,2}Z_{1,2}^n + \mu_{2,2}(m_2^n - Z_{1,2}^n) + \theta_1 Q_1^n \qquad \text{and} \qquad \mu_+^{(r)}(n, \Gamma^n) \equiv \lambda_2^n. \tag{A2}$$

Using these transition rates, we can define the *drift rates* for $D^n(\Gamma^n)$, paralleling (19). Let these drift rates in the regions $(0, \infty)$ and $(-\infty, 0]$ be denoted by $\delta_+^n(\Gamma^n)$ and $\delta_-^n(\Gamma^n)$, respectively, then

$$\delta_+^n(\Gamma^n) \equiv [\lambda_1^n - \mu_{1,1}m_1^n + (\mu_{2,2} - \mu_{1,2})Z_{1,2}^n(t) - \mu_{2,2}m_2^n(t) - \theta_1 Q_1^n(t)] - r[\lambda_2^n - \theta_2 Q_2^n(t)],$$
$$\delta_-^n(\Gamma^n) \equiv [\lambda_1^n - \mu_{1,1}m_1^n - \theta_1 Q_1^n(t)] - r[\lambda_2^n + (\mu_{2,2} - \mu_{1,2})Z_{1,2}^n(t) - \mu_{2,2}m_2^n(t) - \theta_2 Q_2^n(t)]. \tag{A3}$$

Just as for the FTSP, $D_f^n(\Gamma^n)$ is, conditional on Γ^n, a positive recurrent QBD if an only if

$$\delta_+^n(\Gamma^n) < 0 < \delta_-^n(\Gamma^n). \tag{A4}$$

The constant-rate pure-jump Markov process $D_f^n(\Gamma^n)$ will frequently appear with Γ^n being a state of some process, such as $X^n(t)$, We then write $D_f^n(X^n(t)) \equiv \{D_f^n(X^n(t), s): s \geq 0\}$, where it is understood that $D_f^n(X^n(t)) \stackrel{d}{=} D_f^n(\Gamma^n)$ under the condition that $\Gamma^n \stackrel{d}{=} X^n(t)$. It is important that this frozen difference process $D_f^n(\Gamma^n)$ can be directly identified with a version of the FTSP defined in §3.2, because both are pure-jump Markov processes with the same structure. Indeed, the frozen-difference process can be defined as a version of the FTSP with special state and basic model parameters λ_i^n and m_j^n, and transformed time. To express the relationship, we first introduce appropriate notation. Let $D(\lambda_i, m_j, \gamma, s)$ denote the FTSP, as defined in §3.2 with transition rates in (13)–(16) as a function of the arrival rates λ_i, $i = 1, 2$, the staffing levels m_j, $j = 1, 2$, and the state γ, where s is the time parameter as before. (We now will allow the parameters λ_i and m_j to vary as well as the state.) With that new notation, we see that the frozen process is equal in distribution to the corresponding FTSP with new parameters, in particular,

$$\{D_f^n(\lambda_i^n, m_j^n, \Gamma^n, s): s \geq 0\} \stackrel{d}{=} \{D(\lambda_i^n/n, m_j^n/n, \Gamma_n/n, ns): s \geq 0\}, \tag{A5}$$

with the understanding that the initial differences coincide, i.e.,

$$D(\lambda_i^n/n, m_j^n/n, \Gamma_n/n, 0) \equiv D_f^n(\lambda_i^n, m_j^n, \Gamma^n, 0).$$

This can be checked by verifying that the constant transition rates are indeed identical for the two processes, referring to (13)–(16) and (A1)–(A2). Since $\lambda_i^n/n \to \lambda_i$, $i = 1, 2$ and $m_j^n/n \to m_j$, $j = 1, 2$, by virtue of the many-server heavy-traffic scaling in (3), we will have the transition rates of $D(\lambda_i^n/n, m_j^n/n, \Gamma_n/n, \cdot)$ converge to those of $D(\gamma) \equiv D(\lambda_i, m_j, \gamma, \cdot)$ whenever $\Gamma_n/n \to \gamma$.

A.2. Bounding processes. We will use bounding processes in our proof of Theorem 6.1 and later results. We construct the bounding processes so that they have the given initial conditions at time 0 and satisfy FWLLNs with easily determined continuous fluid limits, coinciding at time 0. We thus control the initial behavior.

We first construct w.p.1 lower and upper bounds for $Z_{1,2}^n$. Recall that $N_{i,j}^s$, $i, j = 1, 2$, are the independent rate-1 Poisson processes used in (34). Let

$$Z_a^n(t) = Z_{1,2}^n(0) - N_{1,2}^s \left(\mu_{1,2} \int_0^t Z_a^n(s)\, ds \right),$$

$$Z_b^n(t) = Z_{1,2}^n(0) + N_{2,2}^s \left(\mu_{2,2} \int_0^t (m_2^n - Z_b^n(s))\, ds \right). \tag{A6}$$

LEMMA A.1. *For all $n \geq 1$ and $t \geq 0$, $Z_a^n(t) \leq Z_{1,2}^n(t) \leq Z_b^n(t)$ w.p.1.*

PROOF. The bounding processes Z_a^n and Z_b^n are both initialized as $Z_{1,2}^n(0)$ at time 0. They are defined in terms of the same rate-1 Poisson processes as $Z_{1,2}^n$, so that the three processes can be compared for each sample path. The lower-bound process Z_a^n is the pure death process obtained by routing no new class-1 customers to pool 2 and letting all initial ones depart after receiving service. Hence, Z_a^n is decreasing monotonically to zero. The upper-bound process Z_b^n is a pure birth process obtained by having pool 2 not serve any of its initial class 1 customers and by assigning every server in pool 2 that completes service of a class 2 customer to a new class 1 customer, assuming that such customers are always available. Hence, Z_b^n is increasing monotonically to m_2^n. The given process $Z_{1,2}^n$ necessarily falls in between, where $Z_{1,2}^n$ is defined in (43) with the asterisk omitted. That is so because, whenever $Z_{1,2}^n$ is equal to Z_b^n, every jump up in $Z_{1,2}^n$ is also a jump up in Z_b^n, but not vice versa; whenever $Z_{1,2}^n$ is equal to Z_a^n every jump down in $Z_{1,2}^n$ is also a jump down in Z_a^n, but not vice versa. This is because jumps are generated by the same Poisson processes, but for $Z_{1,2}^n$ jumps occur only if the indicator functions in the respective Poisson processes are equal to one. $\square$

We next construct w.p.1 lower and upper bounds for Q_i^n. The upper bound processes will have the specified arrivals but no departures, whereas the lower bound process will have no arrivals but maximum possible departures. Both processes will start at the initial values.

For $i, j = 1, 2$, let N_i^a, $N_{i,j}^s$, and N_i^u be the previously specified independent rate-1 Poisson processes used in (34), and let

$$Q_{i,a}^n(t) = Q_i^n(0) - \sum_{k=1}^2 \sum_{j=1}^2 N_{k,j}^s(\mu_{k,j} m_j^n t) - N_i^u(\theta_i Q_i^n(0)) t,$$

$$Q_{i,b}^n(t) = Q_i^n(0) + N_i^a(\lambda_i^n t), \quad t \geq 0. \tag{A7}$$

LEMMA A.2. *For all $n \geq 1$ and $t \geq 1$,*

$$(Q_{1,a}^n(t), Q_{2,a}^n(t)) \leq (Q_1^n(t), Q_2^n(t)) \leq (Q_{1,b}^n(t), Q_{2,b}^n(t)) \quad w.p.1.$$

PROOF. Just as in Lemma A.1, we get a w.p.1 comparison because we construct both systems using the same rate-1 Poisson processes. The upper bound is immediate, because the two systems being compared have the same initial value and the same arrivals, but the upper bound system has no service completions or abandonments. For the lower bound, we separately consider the fate of new arrivals, customers initially in service and customers initially in queue. However, we allow the identity of departing customers to shift, which does not affect the result. In the lower bound system new arrivals never enter, so they necessarily leave sooner. Allowing for identity shift, customers initially in queue abandon as rapidly as possible in the lower bound system, because the rate is fixed at the initial rate $\theta_i Q_i^n(0)$. Of course, some customers from queue may enter service. But all customers in service leave at least as quickly in the lower bound system because all servers are working continuously. In the lower bound system we act as if all servers in each pool are simultaneously serving customers of both classes. Hence, we do not need to pay attention to the service assignment rule. The identity of customers may change in this comparison, but the order will hold for the numbers. Hence the proof is complete.

Unlike the processes in Corollary 6.1, we can easily establish stochastic-process limits for the associated fluid-scaled bounding processes, and these continuous limits coincide at time 0. Let $X_{a,b}^n \equiv (Q_{1,a}^n, Q_{2,a}^n, Z_a^n, Q_{1,b}^n, Q_{2,b}^n, Z_b^n)$.

LEMMA A.3. *As $n \to \infty$, $\bar{X}_{a,b}^n \Rightarrow x_{a,b}$ in $\mathscr{D}_6$, where $x_{a,b} \equiv (q_{1,a}, q_{2,a}, z_a, q_{1,b}, q_{2,b}, z_b)$ is an element of $\mathscr{C}_6$ with*

$$z_a(t) = z_{1,2}(0) - \mu_{1,2} \int_0^t z_a(s)\, ds,$$
$$z_b(t) = z_{1,2}(0) + \mu_{2,2} \int_0^t z_b(s)\, ds, \tag{A8}$$

$$q_{i,a}(t) = q_i(0) - \sum_{k=1}^2 \sum_{j=1}^2 \mu_{k,j} m_j t + \theta_i q_i(0) t, \tag{A9}$$
$$q_{i,b}(t) = q_i(0) + \lambda_i t, \quad t \geq 0.$$

PROOF. The stated convergence is a relatively simple application of the continuous mapping theorem. In particular, we first exploit the continuity of the integral representation, Theorem 4.1 in Pang et al. [34], to establish the convergence $(\bar{Z}_a^n, \bar{Z}_b^n) \Rightarrow (z_a, z_b)$. The queue length bounds are simple linear functions. $\square$

We will apply the bounding results above in §B.1, starting with the proof of Lemma 6.1.

A.3. Rate order for FTSPs. Given that we can represent frozen processes as FTSPs with appropriate state parameters γ, as shown in (A5), it is important to be able to compare FTSPs with different state parameters. We establish such a comparison result here for the FTSP in §3.2 using rate order. We say that one pure-jump Markov process Y_1 is less than or equal to another Y_2 in rate order, denoted by $Y_1 \leq_r Y_2$, if all the upward transition rates (with same origin and destination states) are larger in Y_2 and all the downward transition rates (with same origin and destination states) are larger for Y_1. The following lemma is an immediate consequence of the definition of rates in (13)–(16).

LEMMA A.4 (RATE ORDER FOR FTSPs). *Consider the FTSP in §3.2 for candidate states $\gamma^{(i)} \equiv (q_1^{(i)}, q_2^{(i)}, z_{1,2}^{(i)})$, $i = 1, 2$. (a) If $\mu_{1,2} \geq \mu_{2,2}$ and*

$$(-q_1^{(1)}, q_2^{(1)}, z_{1,2}^{(1)}) \leq (-q_1^{(2)}, q_2^{(2)}, z_{1,2}^{(2)}) \quad in\ \mathbb{R}^3,$$

then

$$D(\gamma^{(1)}, \cdot) \leq_r D(\gamma^{(2)}, \cdot).$$

(b) *If $\mu_{1,2} \leq \mu_{2,2}$ and*

$$(-q_1^{(1)}, q_2^{(1)}, -z_{1,2}^{(1)}) \leq (-q_1^{(2)}, q_2^{(2)}, -z_{1,2}^{(2)}) \quad in\ \mathbb{R}^3,$$

then

$$D(\gamma^{(1)}, \cdot) \leq_r D(\gamma^{(2)}, \cdot).$$

We will apply the rate order to get a sample path stochastic order, involving coupling; see Kamae et al. [20], Whitt [44], Ch. 4 of Lindvall [30], and §2.6 of Müller and Stoyan [33]. We briefly discuss those bounds for a sequence of stochastic processes $\{Y^n: n \geq 1\}$. We will bound the process Y^n, for each $n \geq 1$, by a process Y_b^n; i.e., for each n, we will establish conditions under which it is possible to construct stochastic processes $\tilde{Y}_b^n$ and $\tilde{Y}^n$ on a common probability space, with $\tilde{Y}_b^n$ having the same distribution as Y_b^n, $\tilde{Y}^n$ having the same distribution as Y^n, and every sample path of $\tilde{Y}_b^n$ lies below (or above) the corresponding sample path of $\tilde{Y}^n$. We will then write $Y_b^n \leq_{st} (\geq_{st}) Y^n$. However, we will not introduce this "tilde" notation; instead, we will use the original notation Y^n and Y_b^n. As a first step, we will directly give both processes, Y^n and Y_b^n identical arrival processes, the Poisson arrival processes specified for Y^n. We will then show that the remaining construction is possible by increasing (decreasing) the departure rates so that, whenever $Y^n = Y_b^n$, any departure in Y^n also leads to a departure in Y_b^n. That is justified by having the conditional departure rates, given the full histories of the systems up to time t, be ordered.

When $r_{1,2} = 1$, rate order directly implies the stronger sample path stochastic order, but not more generally, because the upper (lower) process can jump down below (up above) the lower (upper) process when the lower process is at state 0 or below, while the upper process is just above state 0. Nevertheless, we can obtain the following stochastic order bound, involving a finite gap. For the following we use the rational form $r_{1,2} = j/k$, $j, k \in \mathbb{Z}_+$ and the associated integer-valued QBD, as in §6 of Perry and Whitt [37]. (Recall Assumption 3.3.) There is no gap when $r_{1,2} = 1$ because then $j = k = 1$ and the jump Markov process and associated QBD process both are equivalent to a simple birth-and-death process.

COROLLARY A.1 (STOCHASTIC BOUNDS FROM RATE ORDER FOR FTSPs). *Consider the FTSP in §3.2 with QBD representation based on $r_{1,2} = j/k$ as in §6 of Perry and Whitt [37], for candidate states $\gamma^{(i)} \equiv (q_1^{(i)}, q_2^{(i)}, z_{1,2}^{(i)})$, $i = 1, 2$.*
   (a) *If $\mu_{1,2} \geq \mu_{2,2}$ and*

$$(-q_1^{(1)}, q_2^{(1)}, z_{1,2}^{(1)}) \leq (-q_1^{(2)}, q_2^{(2)}, z_{1,2}^{(2)}) \quad in \; \mathbb{R}^3,$$

then

$$D(\gamma^{(1)}, \cdot) \leq_{st} D(\gamma^{(2)}, \cdot) + (j \vee k) - 1.$$

   (b) *If $\mu_{1,2} \leq \mu_{2,2}$ and*

$$(-q_1^{(1)}, q_2^{(1)}, -z_{1,2}^{(1)}) \leq (-q_1^{(2)}, q_2^{(2)}, -z_{1,2}^{(2)}) \quad in \; \mathbb{R}^3,$$

then

$$D(\gamma^{(1)}, \cdot) \leq_{st} D(\gamma^{(2)}, \cdot) + (j \vee k) - 1.$$

PROOF. We can do the standard sample path construction: Provided that the processes are on the same side of state 0 in the QBD representation, we can make all the processes jump up by the same amount whenever the lower one jumps up, and make all the processes jump down by the same amount whenever the upper one jumps down. However, there is a difficulty when the processes are near the state 0 in the QBD representation (which involves the matrix B for the QBD). When the upper process is above zero and the lower process is at or below zero, the lower process can jump over the upper process by at most $(j \vee k) - 1$, and the upper process can jump below the lower process by this same amount. But the total discrepancy cannot exceed $(j \vee k) - 1$, because of the rate order. Whenever the desired order is switched, no further discrepancies can be introduced. □

The complexity of the proof in Appendix C is primarily because of the fact that we allow general rational ratio parameters. If $r_{1,2} = r_{2,1} = 1$, the proof can be much shorter, directly exploiting the sample path stochastic order in Corollary A.1 (where there is no gap).

REMARK A.1 (RATE ORDER COMPARISONS FOR QUEUE DIFFERENCE PROCESSES). In the rest of this paper, in particular in the proof of Lemma B.3 in §B.1, Theorem 4.1 in §C, and Theorem 4.5 in §8.4, we will combine the results in this subsection and earlier subsections to establish rate order and sample path stochastic order comparisons between queue difference processes and associated frozen difference processes. A typical initial rate order statement will be of the form

$$D_{1,2}^n \leq_r D_f^n(\Gamma^n) \quad in \; \mathcal{D}([0, \delta]) \tag{A10}$$

for some $\delta > 0$, which we now explain. First, $D_{1,2}^n$ is a function of the Markov process X_6^n, which has state-dependent rates. Thus the "transition rates" of $D_{1,2}^n$ are understood to be functions of time t and $X_6^n(t)$, the state of the Markov process X_6^n at time t, which includes the value of $D_{1,2}^n(t)$. However, the right side of (A10) is interpreted quite differently. We regard $D_f^n(\Gamma^n)$, conditional on the random state vector Γ^n, as a homogenous

pure-jump Markov process constructed independently of X_6^n, with new rate-1 Poisson processes, as in §5.1. However, we deliberately construct the random fixed state vector Γ^n as a function of X_6^n in order to facilitate comparison of rates. Thus, $D_f^n(\Gamma^n)$ and X_6^n, and thus $D_{1,2}^n$, are dependent, but they are conditionally independent given the random state vector Γ^n. Thus, the conclusion in (A10) means that the transition rates at each time t for each value of $X_6^n(t)$ are ordered. If the two processes are in the same state at some time t, then the two processes can make transitions to the same states, and each upward transition rate of $D_f^n(\Gamma^n)$ is greater than or equal to the corresponding upward transition rate of $D_{1,2}^n$, whereas each downward transition rate of $D_f^n(\Gamma^n)$ is less than or equal to the corresponding downward transition rate of $D_{1,2}^n$. That rate ordering then allows the sample path stochastic order comparisons, as in Corollary A.1.

A.4. A sample path stochastic order bound. For the proof of Theorem 4.3 in §8, we also need an upper bound process, unlike $Q_{i,b}^n(t)$ in (A7), that does not explode as $t \to \infty$. Hence, we now establish an elementary sample path stochastic order bound on the queue lengths that is stronger than the w.p.1 upper bound in Lemma A.2. Each of the two upper bound stochastic processes has the structure of the queue length in an $M/M/\infty$ model, with a service rate equal to the abandonment rate here, for which asymptotic results have been established (Pang et al. [34]).

LEMMA A.5 (SAMPLE PATH STOCHASTIC ORDER FOR THE QUEUE LENGTHS). *For $i = 1, 2$ and $n \geq 1$, let*

$$Q_{i,bd}^n(t) = Q_i^n(0) + N_i^a(\lambda_i^n t) - N_i^u\left(\theta_i \int_0^t Q_{i,bd}^n(s)\,ds\right), \quad t \geq 0.$$

Then

$$Q_{i,bd}^n(t) \;\Rightarrow\; Q_{i,bd}^n(\infty) \quad as \; t \to \infty,$$

where $Q_{i,bd}^n(\infty)$ has a Poisson distribution with mean λ_i^n/θ_i and

$$\bar{Q}_{i,bd}^n \equiv n^{-1} Q_{i,bd}^n \;\Rightarrow\; q_{i,bd} \quad in \; \mathcal{D}([0,\infty)) \; as \; n \to \infty,$$

where $q_{i,bd}$ evolves deterministically according to the ODE $\dot{q}_{i,bd}(t) = \lambda_i - \theta_i q_{i,bd}(t)$, starting at $q_{i,bd}(0) \equiv q_i(0)$ for $q_i(0)$ part of $x(0)$ in Assumption 3.2. Thus,

$$q_{i,bd}(t) \leq q_{i,bd}(\infty) \equiv q_i(0) \vee (\lambda_i/\theta_i).$$

Moreover,

$$(Q_1^n, Q_2^n) \leq_{st} (Q_{1,bd}^n, Q_{2,bd}^n) \quad in \; \mathcal{D}_2([0,\infty)).$$

PROOF. We apply Assumption 3.2 to get the intial queue lengths to converge. Just as for $Q_{i,b}^n$ in Lemma A.2, the upper bound system here provides no service completion at all. However, unlike the upper bound $Q_{i,b}^n$ in Lemma A.2, here abandonment from queue is allowed. Here we have sample path stochastic order because we can construct the two systems together, keeping the upper bound system greater than or equal to $Q_i^n(t)$ for all t. Whenever the constructed processes are equal, they can have the same abandonments, because the abandonment rate in both systems will be identical. □

A.5. Extreme-value limits for QBD processes. To prove Theorem 6.1 in §B.1, we exploit extreme-value limits for QBD processes. Because we are unaware of any established extreme-value limits for QBD processes, we establish the following result here. Recall that a QBD has states (i, j), where i is the level and j is the phase. If we only consider the level, we get the level process; it is an elementary function of a QBD.

LEMMA A.6 (EXTREME VALUE FOR QBD). *If $\mathcal{L}$ is the level process of a positive recurrent (homogeneous) QBD process (with a finite number of phases), then there exists $c > 0$ such that*

$$\lim_{t \to \infty} P(\|\mathcal{L}\|_t / \log t > c) = 0.$$

PROOF. Our proof is based on regenerative structure. The intervals between successive visits to the state $(0, j)$ constitute an embedded renewal process for the QBD. Because the QBD is positive recurrent, these cycles have finite mean. Given the regenerative structure, our proof is based on the observation that, if the process $\mathcal{L}$ were continuous real valued with an exponential tail, instead of integer valued with a geometric tail, then we could establish the conventional convergence in law of $\|\mathcal{L}\|_t - c \log t$ to the Gumbel distribution, which implies

our conclusion. Hence, we bound the process $\mathscr{L}$ above w.p.1 by another process $\mathscr{L}_b$ that is continuous real valued with an exponential tail and which inherits the regenerative structure of $\mathscr{L}$.

We first construct the bounding process $\mathscr{L}_b$ and then afterward explain the rest of the reasoning. To start, choose a phase determining a specific regenerative structure for the level process $\mathscr{L}$. Let S_i be the epoch cycle i ends, $i \geq -1$, with $S_{-1} \equiv 0$, and let $L(n)$ be the set of states in level n. For each cycle i, we generate an independent exponential random variable X_i and take the maximum between $\mathscr{L}(t)$ and X_i for all $S_{i-1} \leq t < S_i$ such that $\mathscr{L}(t) \notin L(0)$; i.e., letting $\{X_i : i \geq 0\}$ be an i.i.d. sequence of exponential random variables independent of $\mathscr{L}$ and letting $C(t)$ be the cycle in progress at time t, $\mathscr{L}_b(t) \equiv \mathscr{L}(t) \vee X_{C(t)} 1_{\{\mathscr{L}(t) \notin L(0)\}}$. Clearly, $\mathscr{L}_b$ inherits the regenerative structure of $\mathscr{L}$ and satisfies $\mathscr{L} \leq \mathscr{L}_b$ almost surely. Moreover, by the assumed independence, for each $x > 0$ and $t \geq 0$,

$$P(\mathscr{L}_b(t) > x) = P(\mathscr{L}(t) > x) + P(X > x) - P(\mathscr{L}(t) > x)P(X > x),$$

where X is an exponential random variable distributed as X_i that is independent of $\mathscr{L}(t)$. We now consider the stationary version of $\mathscr{L}$, which makes $\mathscr{L}_b$ stationary as well. We let the desired constant c be the mean of the exponential random variables X_i. If we make c sufficiently large, then we clearly have $P(\mathscr{L}_b(t) > x) \sim e^{-x/c}$ as $x \to \infty$, because the first and third terms become asymptotically negligible as $x \to \infty$. (We choose c to make $\mathscr{L}(t)$ asymptotically negligible compared to X.)

It now remains to establish the conventional extreme-value limit for the bounding process $\mathscr{L}_b$. For that, we exploit the exponential tail of the stationary distribution, just established, and regenerative structure. There are two approaches to extreme-value limits for regenerative processes, which are intimately related, as shown by Rootzén [41]. One is based on stationary processes, and the other is based on the cycle maxima, i.e., the maximum values achieved in successive regenerative cycles. First, if we consider the stationary version, then we can apply classical extreme-value limits for stationary processes as in Leadbetter et al. [29]. The regenerative structure implies that the mixing condition in Leadbetter et al. [29] is satisfied; see Section 4 of Rootzén [41].

However, the classical theory in Leadbetter et al. [29] and the analysis in Rootzén [41] applies to sequences of random variables as opposed to continuous-time processes. In general, the established results for stationary sequences in Leadbetter et al. [29] do not extend to stationary continuous-time processes. That is demonstrated by extreme-value limits for positive recurrent diffusion processes in Borkovec and Klüppelberg [9] and Davis [11]. Proposition 3.1, Corollary 3.2, and Theorem 3.7 of Borkovec and Klüppelberg [9] show that, in general, the extreme-value limit is not determined by the stationary distribution of the process.

However, continuous time presents no difficulty in our setting, because the QBD is constant between successive transitions, and the transitions occur in an asymptotically regular way. It suffices to look at the embedded discrete-time process at transition epochs. That is a standard discrete-time Markov chain associated with the continuous-time Markov chain represented as a QBD. Let $N(t)$ denote the number of transitions over the interval $[0, t]$. Then $\mathscr{L}_b(t) = \mathscr{L}_d(N(t))$, where $\mathscr{L}_d(n)$ is the embedded discrete-time process associated with $\mathscr{L}_b$. Since $N(t)/t \to c' > 0$ w.p.1 as $t \to \infty$ for some constant $c' > 0$, the results directly established for the discrete-time process D_d are inherited with minor modification by $\mathscr{L}_b$. Indeed, the maximum over random indices already arises when relating extremes for regenerative sequences to extremes of i.i.d. sequences; see p. 372 and Theorem 3.1 of Rootzén [41]. In fact, there is a substantial literature on extremes with a random index, e.g., see Proposition 4.20 and (4.53) of Resnick [39] and also Silvestrov and Teugels [42]. Hence, for the QBD we can initially work in discrete time, to be consistent with Leadbetter et al. [29] and Rootzén [41]. After doing so, we obtain extreme-value limits in both discrete and continuous time, which are essentially equivalent.

So far, we have established an extreme-value limit for the stationary version of $\mathscr{L}_b$, but our process $\mathscr{L}_b$ is actually not a stationary process. So it is natural to apply the second approach based on cycle maxima, which is given in Rootzén [41], Asmussen [2], and Section VI.4 of Asmussen [3]. We would get the same extreme-value limit for the given version of $\mathscr{L}_b$ as the stationary version if the cycle maximum has an exponential tail. Moreover, this reasoning would apply directly to continuous time as well as discrete time. However, Rootzén [41] has connected the two approaches (see p. 380 of Rootzén [41]), showing that all the versions of the regenerative process have the same extreme-value limit. Hence, the given version of the process $\mathscr{L}_b$ has the same extreme-value limit as the stationary version, already discussed. Moreover, as a consequence, the cycle maximum has an exponential tail if and only if the stationary distribution has an exponential tail. Hence, we do not need to consider the cycle maximum directly. □

Appendix B. Proofs of three theorems from §6. In this section we prove the three theorems in §6.

B.1. Proof of Theorem 6.1. Theorem 6.1 is an immediate consequence of the following three lemmas: Lemmas B.1, B.2, and B.3.

LEMMA B.1. *If $z_{1,2}(0) > 0$, then, for all $T > 0$, $P(\inf_{0 \le t \le T} \bar{Z}_{1,2}^n(t) > 0) \to 1$ as $n \to \infty$. As a consequence, $Z_{2,1}^n \Rightarrow 0$ as $n \to \infty$.*

PROOF. By Assumption 3.2, $\bar{Z}_{1,2}^n(0) \Rightarrow z_{1,2}(0)$. From Lemma A.1 we know that $Z_{1,2}^n \ge Z_a^n$ in $\mathscr{D}$ for all $n \ge 1$ w.p.1. From Lemma A.3, we know that $\bar{Z}_a^n \Rightarrow z_a$ in $\mathscr{D}$ as $n \to \infty$, for z_a in (A8). However, the integral equation for z_a in (A8) is equivalent to the ODE $\dot{z}_a(t) = -\mu_{1,2} z_a(t)$ with initial value $z_a(0) = z_{1,2}(0)$. Since $z_{1,2}(0) > 0$ by assumption, it follows that $z_a(t) \ge z_a(0) e^{-\mu_{1,2} t} > 0$ for all $t \ge 0$. Thus $P(\inf_{0 \le s \le t} Z_a^n(s) > 0) \to 1$ as $n \to \infty$. Lemma A.1 implies that the same is true for $Z_{1,2}^n$, which proves the first claim of the lemma. The second claim that $Z_{2,1}^n \Rightarrow 0$ as $n \to \infty$ follows from the first together with the one-way sharing rule. $\square$

REMARK B.1 (IMPLICATIONS FOR ONE-WAY SHARING RULE). The conclusion of Lemma B.1 reveals a disadvantage of the one-way sharing rule for very large systems. The lemma concludes that, for large n, if for some $\epsilon > 0$ and $t_0 \ge 0$ $Z_{1,2}^n(t_0) > \epsilon n$, then $Z_{1,2}^n(t)$ is very likely not to reach zero for a long time, thus preventing sharing in the opposite direction, even if that would prove beneficial to do so at a later time, e.g., because there is a new overload incident in the opposite direction.

In practice, we thus may want to relax the one-way sharing rule. One way of relaxing the one-way sharing rule is by dropping it entirely, and relying only on the thresholds $k_{1,2}^n$ and $k_{2,1}^n$ to prevent sharing in both directions simultaneously (at least until the arrival rates change again). Another modification is to introduce lower thresholds on the service processes, denoted by $s_{i,j}^n$, $i \ne j$, such that pool 2 is allowed to start helping class 1 at time t if $D_{2,1}^n > k_{2,1}^n$ and $Z_{1,2}^n(t) < s_{1,2}^n$, and similarly in the other direction. We do not analyze either of these modified controls in this paper.

Given Lemma B.1, it remains to consider only the case $z_{1,2}(0) = 0$. Hence, we assume that $z_{1,2}(0) = 0$ for the rest of this section. Here is the outline of the proof: We first prove (Lemmas B.2 and B.3) that $Z_{2,1}^n$ is asymptotically null over an interval $[0, \tau]$, for some $\tau > 0$. We then prove that $\bar{Z}_{1,2}^n(t)$ must become strictly positive before time τ in fluid scale. By Assumption 2.2, the optimal ratios for FQR-T satisfy $r_{1,2} \ge r_{2,1}$. In Lemma B.2 we consider the cases (i) $x(0) \in \mathbb{A} \cup \mathbb{A}^+$ with $r_{1,2} > r_{2,1}$ and $q_1(0) > 0$ and (ii) $x(0) \in \mathbb{S}^+$; in Lemma B.3 we consider the remaining cases, i.e., $x(0) \in \mathbb{A} \cup \mathbb{A}^+$ with $r_{1,2} = r_{2,1}$ or $q_1(0) = 0$. Unlike the definition of $D_{1,2}^n$ in (6), let $D_{2,1}^n$ be defined by

$$D_{2,1}^n(t) \equiv r_{2,1} Q_2^n(t) - Q_1^n(t), \quad t \ge 0. \tag{B1}$$

LEMMA B.2. *Assume that $z_{1,2}(0) = 0$. If either one of the following two conditions hold (i) $x(0) \in \mathbb{A} \cup \mathbb{A}^+$, $r_{1,2} > r_{2,1}$ and $q_1(0) > 0$, or (ii) $x(0) \in \mathbb{S}^+$, then there exists τ, $0 < \tau \le \infty$, such that*

$$\lim_{n \to \infty} P\left(\sup_{t \in [0, \tau]} D_{2,1}^n(t) \le 0\right) = 1$$

for $D_{2,1}^n$ in (B1), so that $\|Z_{2,1}^n\|_\tau \Rightarrow 0$ as $n \to \infty$.

PROOF. We first show that the appropriate conditions hold in fluid scale at the origin. We start by assuming that $x(0) \in \mathbb{A} \cup \mathbb{A}^+$, which implies that $d_{1,2}(0) \equiv q_1(0) - r_{1,2} q_2(0) = 0$. Since $q_1(0) > 0$ by assumption, $q_2(0) > 0$ too. Since $r_{1,2} > r_{2,1}$ by assumption, $d_{2,1}(0) = r_{2,1} q_2(0) - q_1(0) < r_{1,2} q_2(0) - q_1(0) = 0$, so that we also have $d_{2,1}(0) < 0$. If (ii) holds, so that $x(0) \in \mathbb{S}^+$, then $d_{2,1}(0) < 0$ by definition of $\mathbb{S}^+$.

Given Assumption 3.2, we also have $\bar{X}^n(0) \Rightarrow x(0)$ in $\mathbb{A} \cup \mathbb{A}^+ \cup \mathbb{S}^+$. Hence, the fluid-scaled queueing processes converge to these initial values. In particular, we necessarily have $D_{2,1}^n(0)/n \Rightarrow d_{2,1}(0) < 0$ as $n \to \infty$. Hence, there exists $c > 0$ such that $P(D_{2,1}^n(0) < -cn) \to 1$ as $n \to \infty$. Our goal now is to show that there exists $\tau > 0$ such that $P(\sup_{0 \le t \le \tau} D_{2,1}^n(t) > 0) \to 0$. That will imply the desired conclusion.

It only remains to show that the change in these quantities has to be continuous in fluid scale. For the purpose of bounding $D_{2,1}^n = r_{2,1} Q_2^n - Q_1^n$ above, it suffices to bound Q_2^n above and Q_1^n below, as we have done in Corollary A.1. Hence $D_{2,1}^n \le_{st} D_{u,2,1}^n \equiv r_{2,1} Q_{2,b}^n - Q_{1,a}^n$. By Lemma A.3, $\bar{D}_{u,2,1}^n \Rightarrow d_{2,1}^u \equiv r_{2,1} q_{2,b} - q_{1,a}$, where the limit function $d_{2,1}^u$ evolves continuously, starting with $d_{2,1}^u(0) = d_{2,1}(0) < 0$. Hence there is a time τ' such that $d_{2,1}^u(t) < 0$ for all $0 \le t \le \tau'$. Asymptotically, by the FWLLN, the same will be true for the fluid scaled queue difference process $\bar{D}_{u,2,1}^n$. Hence, we deduce that $P(\sup_{0 \le t \le \tau}\{D_{u,2,1}^n(t)\} > 0) \to 0$ as $n \to \infty$ for any τ with $0 < \tau < \tau'$. Since $D_{2,1}^n \le_{st} D_{u,2,1}^n$, the same is true for $D_{2,1}^n$. That completes the proof. $\square$

The proof of Lemma B.2 relies on a fluid argument because, under its assumptions and Assumption 3.2, $D_{2,1}^n(0)/n$ converges to a strictly negative number as $n \to \infty$. In particular, the difference $D_{2,1}^n(0)$ without centering by $k_{2,1}^n$ is order $O_P(n)$ away from the threshold $k_{2,1}^n$. That fluid reasoning fails when $r_{2,1} = r_{1,2} \equiv r$ or when $q_1(0) = 0$ because in either of these cases, $q_1(0) - r_{1,2}q_2(0) = q_1(0) - r_{2,1}q_2(0) = 0$. (By Assumption 3.2, $q_2(0) = 0$ if $q_1(0) = 0$.) In these cases we will rely on the threshold $k_{2,1}^n$ to prevent class-2 customers to be sent to pool 1, and construct a finer sample-path stochastic-order bound for the stochastic system. Below we remove the centering by the threshold in $D_{2,1}^n$.

LEMMA B.3. *Assume that $z_{1,2}(0) = 0$ and that $q_1(0) - r_{2,1}q_2(0) = 0$ (necessarily, $x(0) \notin \mathbb{S}^+$, i.e., $x(0) \in \mathbb{A} \cup \mathbb{A}^+$). In this case, there exists τ, $0 < \tau \leq \infty$, such that*

$$\lim_{n \to \infty} P\left(\sup_{t \in [0,\tau]} D_{2,1}^n(t) \leq k_{2,1}^n\right) = 1,$$

for $D_{2,1}^n$ in (B1). Hence, $\|Z_{2,1}^n\|_\tau \Rightarrow 0$ as $n \to \infty$.

PROOF. We start by showing that $P(D_{2,1}^n(0) \leq 0) \to 1$ as $n \to \infty$. That follows because, by Assumption 2.2 and the definitions (6) and (B1),

$$D_{2,1}^n(0) \equiv r_{2,1}Q_2^n(0) - Q_1^n(0) \leq r_{1,2}Q_2^n(0) - Q_1^n(0) = -D_{1,2}^n(0) - k_{1,2}^n.$$

Assumptions 2.4 and 3.2 then imply that $D_{1,2}^n(0) \Rightarrow L$ and $k_{1,2}^n \to \infty$ as $n \to \infty$, which together imply the initial conclusion. Going forward, it suffices to assume that we initialize by $D_{2,1}^n(0) = 0$.

The rest of our proof follows three steps: In the first step, paralleling Lemma A.4 and Corollary A.1, we construct a QBD that bounds $D_{2,1}^n$ (without centering by $k_{2,1}^n$) in rate order, which enables us to obtain a stochastic order bound for $D_{2,1}^n$; see Remark A.1. The bound is constructed over an interval $[0, \tau]$. In the second step, we show how to choose τ small enough so that the QBD bound is asymptotically positive recurrent. In the third step, we "translate" the QBD bound to a time-accelerated QBD, in the spirit of (A5), and employ the extreme-value result for the time-accelerated QBD in Lemma A.6 to conclude the proof.

Step 1. We construct a stochastic-order bound for $D_{2,1}^n$, building on rate order. For $\tau > 0$, let

$$X_*^n(t) \equiv (Q_{1,b}^n(t), Q_{2,a}^n(t), Z_b^n(t)) \qquad \text{and} \qquad \Gamma_\tau^n \equiv X_\tau^n \equiv (Q_{1,\tau}^n, Q_{2,\tau}^n, Z_\tau^n), \tag{B2}$$

where

$$Q_{1,\tau}^n \equiv \|Q_{1,b}^n\|_\tau, \qquad Q_{2,\tau}^n \equiv \inf_{0 \leq t \leq \tau} Q_{2,a}^n(t) \vee 0, \qquad \text{and} \qquad Z_\tau^n \equiv \|Z_b^n\|_\tau$$

using the processes defined in (A6) and (A7). By Lemmas A.1 and A.2, for all $n \geq 1$,

$$(Q_1^n(t), -Q_2^n(t), Z_{1,2}^n(t)) \leq (Q_{1,\tau}^n, -Q_{2,\tau}^n, Z_\tau^n) \text{ in } \mathbb{R}_3 \text{ for all } t \in [0, \tau] \quad \text{w.p.1.} \tag{B3}$$

As in §A.1, let $D_f^n(\Gamma_\tau^n)$ be the frozen difference process associated with $D_{2,1}^n$ and Γ_τ^n in (B2). (Recall that we are considering $D_{2,1}^n$ here and not $D_{1,2}^n$.) However, to obtain positive results, we want to consider the process $D_f^n(\Gamma_\tau^n)$ only for nonnegative values. We obtain such a process by working with the associated reflected process, denoted by

$$D_f^{n,*} \equiv D_f^{n,*}(\Gamma) \equiv \{D_f^{n,*}(\Gamma, t): t \geq 0\}, \quad \Gamma \in \mathbb{R}_3$$

obtained by imposing a reflecting lower barrier at 0, where Γ specifies the fixed rates of $D_f^{n,*}(\Gamma)$. We omit Γ from the notation for statements that hold for all $\Gamma \in \mathbb{R}_3$. The reflected process $D_f^{n,*}$ is always nonnegative and has the same state space as the nonnegative part of the state space of $D_{2,1}^n$. Within the QBD framework used in §6 of Perry and Whitt [37], we obtain the reflected process by omitting all transitions down below level 0; the specific QBD construction is given in Appendix D.

It follows from (B3) and Theorem A.4 that we have rate order. By the analog of Lemma A.1, there exists a constant K such that we have sample path stochastic order in $\mathcal{D}([0, \tau])$, i.e.,

$$D_{2,1}^n \leq_{st} D_f^{n,*}(\Gamma_\tau^n) + K \text{ in } \mathcal{D}([0, \tau]) \quad \text{for all } n \geq 1, \tag{B4}$$

for Γ_τ^n in (B2).

Step 2. We now show that we can choose $\tau > 0$ so that there exist sets E^n in the underlying probability space such that $D_f^{n,*}(\Gamma_\tau^n)$ is positive recurrent in E^n and $P(E^n) \to 1$ as $n \to \infty$. Paralleling (A4) (for the frozen process associated with $D_{1,2}^n$), the set E^n here is

$$E^n \equiv \{\delta_*(\Gamma_\tau^n) < 0\}, \tag{B5}$$

where δ_* is the drift for $D_f^{n,*}$. To find $\tau > 0$ such that $P(E^n) \to 0$ for E^n in (B5), we analyze the asymptotic behavior of $\Gamma_\tau^n \equiv X_\tau^n$ in (B2).

First, by Lemma A.3, $\bar{X}_*^n \Rightarrow x_* \equiv (q_{1,b}, q_{2,a}, z_b)$ in $\mathcal{D}_3$ as $n \to \infty$, where the components of x_* are given in (A9) and $x_*(0) = x(0)$ by construction. Then, by the continuous mapping theorem for the supremum function, e.g., Theorem 12.11.7 in Whitt [47],

$$\bar{X}_\tau^n \equiv X_\tau^n/n \Rightarrow x_\tau \equiv (q_{1,\tau}, q_{2,\tau}, z_\tau) \quad \text{in } \mathbb{R}_3, \text{ as } n \to \infty, \tag{B6}$$

where

$$q_{1,\tau} \equiv \|q_{1,b}\|_\tau, \qquad q_{2,\tau} \equiv \inf_{0 \le t \le \tau} q_{2,a}(t) \vee 0 \qquad \text{and} \qquad z_\tau \equiv \|z_b\|_\tau.$$

Since $x(0) \in \mathbb{A} \cup \mathbb{A}^+$ by the assumption of the lemma, $\delta_-(x(0)) > 0$, where δ_- is the drift for the FTSP in (19) associated with $D_{1,2}^n$.

For $\gamma \equiv \gamma(t) \equiv (q_{1,b}(t), q_{2,a}(t) \vee 0, z_b(t))$ let $\hat{\lambda}_r(\gamma) \equiv \lambda_2$, $\hat{\mu}_r(\gamma) \equiv \mu_{2,2}(m_2 - z_b(t)) + \theta_2(q_{2,a}(t) \vee 0)$, $\hat{\lambda}_1(\gamma) \equiv \mu_{1,1}m_1 + \mu_{1,2}z_b(t) + \theta_1 q_{1,b}(t)$ and $\hat{\mu}_1(\gamma) \equiv \lambda_1$. Let $D^*(\gamma)$ be the reflected FTSP corresponding to $D_f^{n,*}(\Gamma_\tau^n)$. The process $D^*(\gamma) \equiv \{D^*(\gamma, t): t \ge 0\}$ has upward jumps of size $r_{2,1}$ with rate $\hat{\lambda}_r(\gamma)$ and downward jumps of size r with rate $\hat{\mu}_r(\gamma)$. It has upward jumps of size 1 with rate $\hat{\lambda}_1(\gamma)$, and downward jumps of size 1 with rate $\hat{\mu}_1(\gamma)$. By Theorem 7.2.3 in Latouche and Ramaswami [28], $D^*(\gamma)$ is positive recurrent if and only if $\delta_*(\gamma) < 0$, where

$$\delta_*(\gamma) \equiv r_{2,1}(\hat{\lambda}_r(\gamma) - \hat{\mu}_r(\gamma)) + (\hat{\lambda}_1(\gamma) - \hat{\mu}_1(\gamma)), \quad \gamma \in \mathbb{R}_3. \tag{B7}$$

Replacing γ with $\gamma_\tau \equiv x_\tau$, we have that $D^*(\gamma_\tau)$ is positive recurrent if and only if $\delta_*(\gamma_\tau) < 0$. Hence, it suffices to show that $\delta_*(x_\tau) < 0$ for some $\tau > 0$. We do that next.

We consider the two possible cases of the condition imposed in the lemma: (i) $r_{2,1} = r_{1,2}$ and (ii) $q_1(0) = q_2(0) = 0$ (with $z_b(0) = z_{1,2}(0) = 0$ in both cases). First, in case (i), $\delta_*(x_*(0)) = -\delta_-(x(0))$ for $\delta_-(\gamma)$ in (19), $\delta_*(\gamma)$ in (B7) and $\gamma = x(0)$. Because we have already observed that $\delta_-(x(0)) > 0$, we necessarily have $\delta_*(x_*(0)) < 0$.

In case (ii) with $r_{1,2} > r_{2,1}$ and $q_1(0) = q_2(0) = 0$ (and again $z_{1,2}(0) = 0$), $\delta_*(x_*(0)) = r_{2,1}(\lambda_1 - \mu_{1,1}m_1) + (\lambda_2 - \mu_{2,2}m_2)$, so that $\delta_*(x_*(0)) < 0$ if and only if $(\lambda_1 - \mu_{1,1}m_1) + r_{2,1}(\lambda_2 - \mu_{2,2}m_2) < 0$. However, this inequality must hold because, by Assumption 3.2, $\delta_-(x(0)) > 0$, so that $\delta_-(x(0)) = (\lambda_1 - \mu_{1,1}m_1) - r_{1,2}(\lambda_2 - \mu_{2,2}m_2) > 0$. Since $r_{2,1} < r_{1,2}$, it follows that here too $\delta_*(x_*(0)) < 0$.

Finally, the continuity of x_* and $\delta_*(x_*)$ imply that we can find $\tau > 0$ and $\eta > 0$ such that $\sup_{s \in [0,\tau]} \delta_*(x_*(s)) < -\eta < 0$. In particular, for that choice of τ, $\delta_*(x_\tau) < -\eta$. Hence, $P(E^n) = P(\delta_*(\Gamma_\tau^n) < 0) \to 1$, because $\delta_*(\bar{X}_*^n) \Rightarrow \delta_*(x_*)$ in $\mathcal{D}$ and $\delta_*(\Gamma_\tau^n) \Rightarrow \delta_*(\gamma_\tau)$ in $\mathbb{R}$ as $n \to \infty$ by the continuous mapping theorem.

Step 3. Finally, we apply the extreme-value result in Lemma A.6 to the stochastic upper bound $K + D_f^{n,*}(\Gamma_\tau^n)$ in (B4) for Γ_τ^n in (B2). For that purpose, observe that, paralleling (A5), $\{D_f^{n,*}(\Gamma_\tau^n, s): s \ge 0\} \stackrel{d}{=} \{D^*(\Gamma_\tau^n/n, ns): s \ge 0\}$ for $D_f^{n,*}$ and D^* defined above, with $\Gamma_\tau^n/n \Rightarrow x_\tau$ in (B6). For the rest of the proof, we apply the Skorohod representation theorem to replace the convergence in distribution by convergence $\Gamma_\tau^n/n \to x_\tau$ w.p.1, without changing the notation.

By the arguments above

$$P(\|D_{2,1}^n\|_\tau/\log n > c) \le P((K + \|D_f^{n,*}(\Gamma_\tau^n)\|_\tau)/\log n > c) = P((K + \|D^*(\Gamma_\tau^n/n)\|_{n\tau})/\log n > c). \tag{B8}$$

To apply Lemma A.6 to the final term in (B8), we want to replace $\Gamma_\tau^n/n \equiv \bar{X}_\tau^n$ by a vector independent of n, say x_ϵ, such that $D^*(x_\epsilon)$ is positive recurrent and, for some n_0, $\bar{X}_\tau^n \le x_\epsilon$ for all $n \ge n_0$. Then we can apply Lemma A.6 to get, for $n \ge n_0$,

$$P((K + \|D^*(\Gamma_\tau^n/n)\|_{n\tau})/\log n > c) \le P((2K + \|D^*(x_\epsilon)\|_{n\tau})/\log n > c) \to 0 \quad \text{as } n \to \infty.$$

That implies the claim because of the way the thresholds are scaled in Assumption 2.4.

We conclude by showing how to construct the vector x_ϵ such that $D^*(x_\epsilon)$ is positive recurrent and, for some n_0, $\Gamma_\tau^n/n = \bar{X}_\tau^n \le x_\epsilon$ for all $n \ge n_0$. If $q_{2,\tau} > 0$, then choose ϵ such that $0 < \epsilon < q_{2,\tau}$ and let $x_\epsilon \equiv (q_{1,\epsilon}, q_{2,\epsilon}, z_\epsilon) \equiv (q_{1,\tau} + \epsilon, q_{2,\tau} - \epsilon, z_b + \epsilon)$. Otherwise, let $x_\epsilon \equiv (q_{1,\tau} + \epsilon, 0, z_b + \epsilon)$. Clearly $x_\epsilon \le x_\tau$ for all $\epsilon > 0$, so that $\{D^*(x_\tau, t): 0 \le t \le \tau\} \le_{st} K + \{D^*(x_\epsilon, t): 0 \le t \le \tau\}$ and, by the choice of τ, we can find $\epsilon > 0$ small enough so that $\delta^*(x_\epsilon) < 0$, so that $D^*(x_\epsilon)$ is positive recurrent. $\square$

B.2. Proof Theorem 6.2. Define the processes

$$L_1^n \equiv Q_1^n + Z_{1,1}^n + Z_{1,2}^n - m_1^n \qquad \text{and} \qquad L_2^n \equiv Q_2^n + Z_{2,1}^n + Z_{2,2}^n - m_2^n, \tag{B9}$$

representing the excess number in system for each class. Note that $(L_i^n)^+ = Q_i^n$, $i = 1, 2$. For all n sufficiently large, we will bound the two-dimensional process (L_1^n, L_2^n) below in sample-path stochastic order by another two-dimensional process $(L_{1,b}^n, L_{2,b}^n)$, $n \geq 1$.

We construct the lower-bound process $(L_{1,b}^n, L_{2,b}^n)$ by having $L_{i,b}^n(0) = L_i^n(0) \wedge k_{1,2}^n$, $i = 1, 2$, and by increasing the departure rates in both processes L_1^n and L_2^n, making it so that each goes down at least as fast, regardless of the state of the pair. First, we place reflecting upper barriers on the two queues. This is tantamount to making the death rate infinite in these states and all higher states. We place the reflecting upper barrier on L_i^n at $k_{1,2}^n$, where $k_{1,2}^n \geq 0$. This necessarily produces a lower bound for all n. By the initial conditions assumed for the queue lengths in Assumption 3.2, we have $P((L_{1,b}^n(0), L_{2,b}^n(0)) = (k_{1,2}^n, k_{1,2}^n)) \to 1$ as $n \to \infty$.

With the upper barrier at $k_{1,2}^n$, the departure rate of $L_1^n(t)$ at time $t > 0$ is bounded above by $\mu_{1,1} m_1^n + \mu_{1,2} Z_{1,2}^n(t) + \theta_1 k_{1,2}^n$, based on assuming that pool 1 is fully busy serving class 1, that $L_1^n(t)$ is at its upper barrier, and that $Z_{1,2}^n(t)$ agents from pool 2 are currently busy serving class 1 in the original system. With the upper barrier at $k_{1,2}^n$, the departure rate of $L_2^n(t)$ at time t is bounded above by $\mu_{2,2}(m_2^n - Z_{1,2}^n(t)) + \theta_2 k_{1,2}^n$, based on assuming that pool 2 is fully busy with $Z_{1,2}^n(t)$ agents from pool 2 currently busy serving class 1, and that $L_2^n(t)$ is at its upper barrier $k_{1,2}^n$. Thus, we give $L_{1,b}^n$ and $L_{2,b}^n$ these bounding rates at all times.

As constructed, the evolution of $(L_{1,b}^n, L_{2,b}^n)$ depends on the process $Z_{1,2}^n$ associated with the original system, which poses a problem for further analysis. However, we can avoid this difficulty by looking at a special linear combination of the processes. Specifically, let

$$U^n \equiv \mu_{2,2}(L_1^n - k_{1,2}^n) + \mu_{1,2}(L_2^n - k_{1,2}^n) \qquad \text{and} \qquad U_b^n \equiv \mu_{2,2}(L_{1,b}^n - k_{1,2}^n) + \mu_{1,2}(L_{2,b}^n - k_{1,2}^n). \tag{B10}$$

By the established sample-path stochastic order $(L_{1,b}^n, L_{2,b}^n) \leq_{st} (L_1^n, L_2^n)$, the initial conditions specified above and the monotonicity of the linear map in (B10), we get the associated sample-path stochastic order $U_b^n \leq_{st} U^n$. The lower-bound stochastic process U_b^n has constant birth rate $\lambda_b^n = \mu_{2,2}\lambda_1^n + \mu_{1,2}\lambda_2^n$ and constant death rate

$$\mu_b^n \equiv \mu_{2,2}(\mu_{1,1} m_1^n + \mu_{1,2} Z_{1,2}^n(t) + \theta_1 k_{1,2}^n) + \mu_{1,2}(\mu_{2,2} m_2^n - \mu_{2,2} Z_{1,2}^n(t) + \theta_2 k_{1,2}^n)$$

$$= \mu_{2,2}(\mu_{1,1} m_1^n + \theta_1 k_{1,2}^n) + \mu_{1,2}(\mu_{2,2} m_2^n + \theta_2 k_{1,2}^n).$$

In particular, unlike the pair of processes $(L_{1,b}^n, L_{2,b}^n)$, the process U_b^n is independent of the process $Z_{1,2}^n$. Consequently, U_b^n is a birth and death process on the set of all integers in $(-\infty, 0]$. Since $P((L_{1,b}^n(0), L_{2,b}^n(0)) = (k_{1,2}^n, k_{1,2}^n)) \to 1$ as $n \to \infty$, $P(U_b^n(0) = 0) \to 1$ as $n \to \infty$.

The drift in U_b^n is

$$\delta_b^n \equiv \lambda_b^n - \mu_b^n = \mu_{2,2}(\lambda_1^n - m_1^n \mu_{1,1} - \theta_1 k_{1,2}^n) + \mu_{1,2}(\lambda_2^n - m_2^n \mu_{2,2} - \theta_2 k_{1,2}^n). \tag{B11}$$

Hence, after scaling, we get $\delta_b^n/n \to \delta$ (recall that $k_{1,2}^n$ is $o(n)$), where

$$\delta_b \equiv \mu_{2,2}(\lambda_1 - m_1 \mu_{1,1}) + \mu_{1,2}(\lambda_2 - m_2 \mu_{2,2}) > 0, \tag{B12}$$

with the inequality following from Assumption 3.1.

Now we observe that $-U_b^n$ is equivalent to the number in system in a stable $M/M/1$ queueing model with traffic intensity $\rho_*^n \to \rho_* < 1$, starting out empty, asymptotically. Let Q_* be the number-in-system process in an $M/M/1$ system having arrival rate equal to $\lambda_* \equiv \mu_{2,2}\mu_{1,1} m_1 + \mu_{1,2}\mu_{2,2} m_2$, service rate $\mu_* \equiv \mu_{2,2}\lambda_1 + \mu_{1,2}\lambda_2$, and traffic intensity $\rho_* \equiv \lambda_*/\mu_* < 1$. Observe that the scaling in U_b^n is tantamount to accelerating time by a factor of order $O(n)$ in Q_*. That is, $\{-U_b^n(t): t \geq 0\}$ can be represented as $\{Q_*(c_n t): t \geq 0\}$, where $c_n/n \to 1$ as $n \to \infty$. We can now apply the extreme-value result in Lemma A.6 for the $M/M/1$ queue above (since the $M/M/1$ birth and death process is trivially a QBD process) to conclude that $\|Q_*\|_t = O_P(\log t)$. This implies that $-U_b^n/\log n$ is stochastically bounded.

From the way that the reflecting upper barriers were constructed, we know at the outset that $L_{1,b}^n(t) \leq k_{1,2}^n$ and $L_{2,b}^n(t) \leq k_{1,2}^n$. Hence, we must have both $L_{1,b}^n - k_{1,2}^n$ and $L_{2,b}^n - k_{1,2}^n$ nonpositive. Combining this observation with the result that $(-U_b^n)/\log n$ is stochastically bounded, we deduce that both $(k_{1,2}^n - L_{1,b}^n)/\log n$ and $(k_{1,2}^n - L_{2,b}^n)/\log n$ are stochastically bounded, i.e., the fluctuations of $L_{i,b}^n$ below $k_{1,2}^n$ are $O_P(\log n)$. The result follows because $-L_i^n \leq_{st} -L_{i,b}^n$, $i = 1, 2$, and from the choice of $k_{1,2}^n$, which satisfies $k_{1,2}^n/\log n \to \infty$ as $n \to \infty$ by Assumption 2.4.

B.3. Proof of Theorem 6.4. We now show that the interval $[0, \tau]$ over which the conclusions in §6 are valid can be extended from $[0, \tau]$ to $[0, \infty)$ after Theorem 4.1 has been proved over the interval $[0, \tau]$.

For this purpose, we use the processes L_1^n and L_2^n defined in (B9). By Lemma B.1, we only need to consider the case $z_{1,2}(0) = 0$. By Lemmas B.2 and B.3, there exists $\tau > 0$ such that

$$\lim_{n \to \infty} P(\|D_{2,1}^n\|_\tau < k_{2,1}^n) = 1.$$

Hence, the claim of the theorem will follow from Lemma B.1 if we show that for some t_0 satisfying $0 < t_0 \leq \tau$ it holds that $z_{1,2}(t_0) > 0$, where $z_{1,2}$ is the (deterministic) fluid limit of $\bar{Z}_{1,2}^n$ as $n \to \infty$, which exists by Theorem 4.1. We will actually show a somewhat stronger result, namely, that for any $0 < \epsilon \leq \delta$, where δ is chosen from Lemma C.1 and thus Theorem 4.1, there exists $t_0 < \epsilon$ such that $z_{1,2}(t_0) > 0$. We prove that by assuming the contradictory statement: for some $0 < \epsilon \leq \delta$ and for all $t \in [0, \epsilon]$, $z_{1,2}(t) = 0$.

By our contradictory assumption above, $\bar{Z}_{1,2}^n = o_P(1)$, i.e., $\|Z_{1,2}^n/n\|_\epsilon \Rightarrow 0$. Recall also that $Z_{2,1}^n = o_P(1)$ over $[0, \epsilon]$ (since $\epsilon \leq \tau$, and τ is chosen according to Lemmas B.2 and B.3). Hence, by our contradictory assumption and by our choice of ϵ, there is negligible sharing of customers over the interval $[0, \epsilon]$. We can thus represent L_i^n in (B9), $i = 1, 2$ by

$$L_i^n(t) = L_i^n(0) + N_i^a(\lambda_i^n t) - N_{i,i}^s\left(\mu_{i,i}\int_0^t (L_i^n(s) \wedge 0)\,ds\right) - N_i^u\left(\theta_i\int_0^t (L_i^n(s) \vee 0)\,ds\right) + o_P(n), \qquad \text{(B13)}$$

for $i = 1, 2$ and $0 \leq t \leq \delta$ as $n \to \infty$, where N_i^a, $N_{i,i}^s$, and N_i^u are independent rate-1 Poisson processes. The $o_P(n)$ terms are replacing the (random-time changed) Poisson processes related to $Z_{1,2}^n$ and $Z_{2,1}^n$, which can be disregarded when we consider the fluid limits of (B13). The negligible sharing translates into the $o_P(n)$ term in (B13) by virtue of the continuous mapping theorem and Gronwall's inequality, as in §4.1 of Pang et al. [34].

Letting $\bar{L}_i^n \equiv L_i^n/n$, $i = 1, 2$, and applying the continuous mapping theorem for the integral representation function in (B13), Theorem 4.1 in Pang et al. [34], (see also Theorem 7.1 and its proof in Pang et al. [34]), we have $(\bar{L}_1^n, \bar{L}_2^n) \Rightarrow (\bar{L}_1, \bar{L}_2)$ in $\mathscr{D}([0, \epsilon])$ as $n \to \infty$, where, for $i = 1, 2$,

$$\bar{L}_i(t) = \bar{L}_i(0) + (\lambda_i - \mu_{i,i} m_i)t - \int_0^t [\mu_{i,i}(\bar{L}_i(s) \wedge 0) + \theta_i(\tilde{L}_i(s) \vee 0)]\,ds, \quad 0 \leq t \leq \epsilon,$$

so that

$$\bar{L}_i'(t) \equiv \frac{d}{dt}\bar{L}_i(t) = (\lambda_i - \mu_{i,i} m_i) - \mu_{i,i}(\bar{L}_i(t) \wedge 0) - \theta_i(\tilde{L}_i(t) \vee 0), \quad 0 \leq t \leq \epsilon.$$

By Assumption 3.2, both pools are full at time 0, so that $L_i(0) \geq 0$. Moreover, for $i = 1, 2$, $\bar{L}_i^e \equiv (\lambda_i - \mu_{i,i})/\theta_i$ is an equilibrium point of the ODE $\bar{L}_i'$, in the sense that, if $\bar{L}_i(t_0) = \bar{L}_i^e$, then $\bar{L}_i(t) = \bar{L}_i^e$ for all $t \geq t_0$. (That is, $\bar{L}_i^e$ is a fixed point of the solution to the ODE.) It also follows from the derivative of $\bar{L}_i$ that $\bar{L}_i$ is strictly increasing if $\bar{L}_i(0) < \bar{L}_i^e$, and strictly decreasing if $\bar{L}_i(0) > \bar{L}_i^e$, $i = 1, 2$.

Recall that $\rho_1 > 1$, so that $\lambda_1 - \mu_{1,1} m_1 > 0$. Together with the initial condition, $L_1(0) \geq 0$, we see that, in that case, $\bar{L}_1(t) \geq 0$ for all $t \geq 0$. First assume that $\rho_2 \geq 1$. Then, by similar arguments, $\bar{L}_2(t) \geq 0$ for all $t \geq 0$. In that case, we can replace $\bar{L}_i$ with $q_i(t) = (\bar{L}_i(t))^+$, $i = 1, 2$, where q_i is the fluid limit of $\bar{Q}_i^n$ over $[0, \epsilon]$. We can then write, for $t \in [0, \epsilon]$,

$$q_1(t) = q_1(0) - (\lambda_1 - \mu_{1,1} m_1)t - \theta_1\int_0^t q_1(s)\,ds,$$

$$q_2(t) = q_2(0) - (\lambda_2 - \mu_{2,2} m_2)t - \theta_2\int_0^t q_2(s)\,ds,$$

so that, for $t \in [0, \epsilon]$,

$$\begin{aligned}
d_{1,2}(t) &= q_1^a + (q_1(0) - q_1^a)e^{-\theta_1 t} - r(q_2^a + (q_2(0) - q_2^a)e^{-\theta_2 t}) \\
&= (q_1^a - r q_2^a) + (q_1(0) - q_1^a)e^{-\theta_1 t} - r(q_2(0) - q_2^a)e^{-\theta_2 t}. \qquad \text{(B14)}
\end{aligned}$$

First assume that $x(0) \in \mathbb{A} \cup \mathbb{A}^+$, so that $d_{1,2}(0) = 0$. From (B14),

$$d_{1,2}'(t) \equiv \frac{d}{dt}d_{1,2}(t) = -\theta_1(q_1(0) - q_1^a)e^{-\theta_1 t} + r\theta_2(q_2(0) - q_2^a)e^{-\theta_2 t}.$$

Hence, $d_{1,2}'(0) = \lambda_1 - \mu_{1,1} m_1 - \theta_1 q_1(0) - r(\lambda_2 - \mu_{2,2}) + r\theta_2 q_2(0)$. If follows from (19) and the assumption $z_{1,2}(0) = 0$, that $d_{1,2}'(0) = \delta_-(x(0))$. By assumption, $x(0) \in \mathbb{A} \cup \mathbb{A}^+$, so that $d_{1,2}'(0) = \delta_-(x(0)) > 0$ (that follows

from the definition of $\mathbb{A}$ and $\mathbb{A}^+$ in (20) and (21) and the fact that $\delta_- > \delta_+$). Hence, $d_{1,2}$ is strictly increasing at 0. Now, since $d_{1,2}(0) = 0$, we can find $t_1 \in (0, \epsilon]$, such that $d_{1,2}(t) > 0$ for all $0 < t < t_1$. This implies that $P(\inf_{0 < t \leq t_1} D_{1,2}^n(t) > 0) \to 1$ as $n \to \infty$. The same is true if $x(0) \in \mathbb{S}^+$.

Since $\|Z_{1,2}^n/n\|_\epsilon \Rightarrow 0$, as a consequence of our contradictory assumption, it follows from the representation of $Z_{1,2}^n$ in Theorem 5.1 that

$$\bar{Z}_{1,2}^n(t) = n^{-1} N_{2,2}^s(\mu_{2,2} m_2^n t) + o_P(1) \quad \text{as } n \to \infty. \tag{B15}$$

However, by the FSLLN for Poisson processes, the fluid limit $z_{1,2}$ of $\bar{Z}^n$ in (B15) satisfies $z_{1,2}(t) = \mu_{2,2} m_2 t > 0$ for every $0 < t \leq t_1$. We thus get a contradiction to our assumption that $z_{1,2}(t) = 0$ for all $t \in [0, \epsilon]$.

For the case $\rho_2 < 1$ the argument above still goes through, but we need to distinguish between two cases: $\bar{L}_2 = 0$ and $\bar{L}_2 > 0$. In both cases $\bar{L}_2$ is strictly decreasing. In the first case, this implies that $\bar{L}_2$ is negative for every $t > 0$. It follows immediately that $q_1(t) - rq_2(t) > 0$ for every $t > 0$. If $\bar{L}_2(0) > 0$, then necessarily $\bar{L}_1(0) > 0$, and we can replace $\bar{L}_i$ with q_i, $i = 1, 2$, on an initial interval (before $\bar{L}_2$ becomes negative). We then use the arguments used in the case $\rho_2 \geq 1$ above. $\square$

In the proof of Theorem 6.4 we have shown that, if $z_{1,2}(0) = 0$, then $z_{1,2}(t) > 0$ for all $t > 0$. We prove in Appendix E that the two queues must also become strictly positive right after time 0, if they are not strictly positive at time 0.

Appendix C. Alternative proof of characterization using stochastic order bounds.

In this appendix we present an alternative proof of Lemma 7.2 and thus an alternative proof of the FWLLN in Theorem 4.1. At the beginning of §7 we observed that it suffices to characterize the limiting integral terms in (47); i.e., it suffices to prove Lemma 7.2. In §7 we accomplished that goal by using the martingale argument of Hunt and Kurtz [18] and Kurtz [27]. Here we show that same goal can be achieved with stochastic bounds, exploiting Lemma 8.4 and similar reasoning. However, we prove less than the full Theorem 4.1 here. Our proof here is under the special case of Assumption 3.2 for which $x(0) \in \mathbb{A}$. For this special case we carry out the characterization over the full interval $[0, \infty)$ if $x(t)$ remains within $\mathbb{A}$ for all $t \geq 0$. Otherwise, we complete the characterization proof over $[0, T_A]$, where

$$T_A \equiv \inf \{t > 0 \colon x(t) \notin \mathbb{A}\}. \tag{C1}$$

Since x is continuous and $\mathbb{A}$ is an open subset of $\mathbb{S}$, we know that $T_A > 0$. A first step is to do the characterization over an interval $[0, \delta]$ for some $\delta > 0$. We start in §C.1 by indicating how the interval of convergence can be extended given that the first step has been carried out. Next in §C.2 we prove Theorem 4.1 subject to Lemma C.2, establishing convergence of integral terms over the interval $[0, \delta]$. For that purpose, we state Lemma C.4 establishing a sample path stochastic order bound that we will use to prove Lemma C.2 in §C.3 we establish continuity results for QBD processes. We then prove Lemmas C.2 and C.4, respectively, in §C.4 and §C.5.

C.1. Extending the interval of convergence.

Unlike the first proof, with this second proof we only establish convergence in $\mathcal{D}_{14}([0, T_A])$ if $T_A < \infty$, for T_A in (C1). We now show how we achieve this extension.

As in the first proof, after establishing the convergence over an initial interval $[0, \delta]$ with $\delta \leq \tau$, we apply Theorem 6.4 to conclude that any limit point of the tight sequence $\bar{X}_6^n$ is again a limit of the tight sequence $\bar{X}_6^{n,*}$ in (10) over the entire half line $[0, \infty)$, showing that τ places no constraint on expanding the convergence interval. Moreover, by part (ii) of Theorem 5.2 in Perry and Whitt [37], any solution to the ODE, with a specified initial condition, can be extended indefinitely, and is unique. Hence that places no constraint either.

However, for this second method of proof, we do critically use that fact that $x(t) \in \mathbb{A}$, $0 \leq t \leq \delta$, in order to prove the characterization. (This is proved in Lemma C.1.) However, we can extend the interval of convergence further. Given that we have shown that $\bar{X}(t) = x(t) \in \mathbb{A}$ for $\mathbb{A}$ in (20) over a time interval $[0, \delta]$, and thus established the desired convergence $\bar{X}^n \Rightarrow x$ over that time interval $[0, \delta]$, we can always extend the time interval to a larger interval $[0, \delta']$ for some δ' with $\delta' > \delta$. To do so, we repeat the previous argument treating time δ as the new time origin. That directly yields $\bar{X}(t) = x(t) \in \mathbb{A}$, and thus convergence $\bar{X}^n \Rightarrow x$, over the time interval $[\delta, \delta']$. However, we can combine that with the previous result to obtain $\bar{X}(t) = x(t) \in \mathbb{A}$, and thus convergence $\bar{X}^n \Rightarrow x$, over the longer time interval $[0, \delta']$. Let ν be the supremum of all δ for which the expansion of convergence to $[0, \delta]$ is valid. We must have $\bar{X}(t) = x(t) \in \mathbb{A}$, and thus convergence $\bar{X}^n \Rightarrow x$, over the interval $[0, \nu)$, open on the right. The interval is $[0, \infty)$ if $\nu = \infty$. Suppose that $\nu < \infty$. In that case, we can next apply continuity to extend the interval of convergence to the closed interval $[0, \nu]$. Since $\bar{X}(t) = x(t)$, $0 \leq t < \nu$, x is continuous and $\bar{X}$ is almost surely continuous, we necessarily have $\bar{X}(\nu) = x(\nu)$ w.p.1 as well. We claim that $\nu \geq T_A \equiv \inf\{t \geq 0 \colon x(t) \notin \mathbb{A}\}$. If not, we can do a new construction yielding $\bar{X}(t) = x(t)$, first for $\nu \leq t < \nu'$ and then for $0 \leq t \leq \nu'$, $\nu' > \nu$, contradicting the definition of ν. Hence, we have extended the domain of convergence to $[0, T_A]$ if $T_A < \infty$ and to $[0, \infty)$ otherwise, as claimed.

C.2. Reduction to convergence of integral terms. Because each of these integrals in (47) can be treated in essentially the same way, we henceforth focus only on the single term $\bar{I}_{q,1,1}(t)$ and establish (48).

Recall that $\bar{X}$ is the limit of the converging subsequence of $\{\bar{X}^n : n \geq 1\}$ in $\mathcal{D}_3([0,\tau])$ for τ in Theorem 6.3. An important first step is to identify an initial interval $[0,\delta]$, $0 < \delta < \tau$, over which $\bar{X} \in \mathbb{A}$. We apply the proof of Lemma 8.4 to prove the following lemma.

LEMMA C.1 (STATE IN $\mathbb{A}$ OVER $[0,\delta]$). *There exists δ with $0 < \delta \leq \tau$ for τ in Theorem 6.3 such that*

$$P(\bar{X}^n(t) \in \mathbb{A}, 0 \leq t \leq \delta) \to 1 \quad as \ n \to \infty,$$

for $\mathbb{A}$ in (20), so that $P(\bar{X}(t) \in \mathbb{A}\ 0 \leq t \leq \delta) = 1$.

PROOF. Recall that we have assumed that Assumption 3.2 holds with $x(0) \in \mathbb{A}$. We can apply the first step of the proof of Lemma 8.4 to obtain the stochastic bound over an initial interval $[0,\delta]$ for some $\delta > 0$. Since the FTSP $D(\gamma, \cdot)$ is positive recurrent if and only if $\gamma \in \mathbb{A}$ for $\mathbb{A}$ in (20). By Lemma 8.4, $D_f(X_m^n, \cdot)$ and $D_f(X_M^n, \cdot)$ are positive recurrent on $B^n(\delta, \eta)$ in (60). Thus, by Corollary 8.1, $D_f(X^n(t), \cdot)$ is positive recurrent on $B^n(\delta, \eta)$ as well for $0 \leq t \leq \delta$. Hence, the claim holds. $\square$

The next important step is the following lemma, proved in §C.4, after establishing preliminary bounding lemmas.

LEMMA C.2 (CONVERGENCE OF THE INTEGRAL TERMS). *There exists δ with $0 < \delta \leq \tau$ for τ in Theorem 6.3, such that for any $\epsilon > 0$ and t with $0 \leq t < \delta$, there exists $\sigma \equiv \sigma(\epsilon, \delta, t)$ with $0 < \sigma < \delta - t$ and n_0 such that*

$$P\left(\left| \frac{1}{\sigma} \int_t^{t+\sigma} 1_{\{D_{1,2}^n(s) > 0\}} \bar{Z}_{1,2}^n(s)\ ds - \pi_{1,2}(\bar{X}(t))\bar{Z}_{1,2}(t) \right| > \epsilon \right) < \epsilon \quad for \ all \ n \geq n_0. \tag{C2}$$

To apply Lemma C.2 to prove Lemma 7.2, we exploit the absolute continuity of $\bar{I}_{q,1,1}$, established now.

LEMMA C.3 (ABSOLUTE CONTINUITY OF $\bar{I}_{q,1,1}$). *The limiting integral term $\bar{I}_{q,1,1}$ almost surely satisfies*

$$0 \leq \bar{I}_{q,1,1}(t+u) - \bar{I}_{q,1,1}(t) \leq m_2 u \quad for \ all \ 0 \leq t < t + u < \tau, \tag{C3}$$

and so $\bar{I}_{q,1,1}$ is the cumulative distribution function corresponding to a finite measure, having a density h depending on $\bar{X}$. As a consequence, for all $\epsilon > 0$, there exists $u_0 > 0$ such that

$$\left| \frac{\bar{I}_{q,1,1}(t+u) - \bar{I}_{q,1,1}(t)}{u} - h(t) \right| < \epsilon \tag{C4}$$

for all $u < u_0$.

PROOF. Since

$$\bar{I}_{q,1,1}^n(t+u) - \bar{I}_{q,1,1}^n(t) = \int_t^{t+u} 1_{\{D_{1,2}^n(s) > 0\}} \bar{Z}_{1,2}^n(s)\ ds \leq um_2^n/n,$$

$I_{q,1,1}^n(t)$ is a nondecreasing function with $0 \leq I_{q,1,1}^n(t+u) - I_{q,1,1}^n(t) \leq m_2^n$ for all $0 \leq t < t + u \leq \tau$. Hence, $I_{q,1,1}^n(t)$ is a cumulative distribution function associated with a finite measure. The convergence obtained along the subsequence based on tightness then yields

$$0 \leq \bar{I}_{q,1,1}(t+u) - \bar{I}_{q,1,1}(t) \leq m_2 u \quad for \ all \ 0 \leq t < t + u \leq \tau.$$

Hence, $\bar{I}_{q,1,1}$ has a density with respect to Lebesgue measure, as claimed. $\square$

PROOF OF LEMMA 7.2. Given Lemma C.2, for any $\epsilon > 0$, we can find σ and n_0 such that (C2) is valid for all $n \geq n_0$. Hence, we can let $n \to \infty$ and conclude that, for any $\epsilon > 0$, we can find σ such that

$$P\left(\left| \frac{1}{\sigma}(\bar{I}_{q,1,1}(t+\sigma) - \bar{I}_{q,1,1}(t)) - \pi_{1,2}(\bar{X}(t))\bar{Z}_{1,2}(t) \right| > \epsilon \right) < \epsilon. \tag{C5}$$

However, given that (C4) and (C5) both hold, we conclude that we must almost surely have $h(t) = \pi_{1,2}(\bar{X}(t))\bar{Z}_{1,2}(t)$, which completes the proof. $\square$

We now apply bounds to prove Lemma C.2. The comparisons in Lemma 8.4 and Corollary 8.1 are important, but they are not directly adequate for our purpose. The sample-path stochastic order bound in Corollary 8.1 enables us to prove Lemma C.2, and thus Theorem 4.1, for the special case of $r_{1,2} = 1$, because then $\zeta = 0$, where ζ is the gap in Corollary 8.1, but not more generally when the gap is positive. However, we now show that an actual gap will only be present rarely, if we choose the interval length small enough and n big enough. We use the construction in the proof of Lemma 8.4 exploiting the fact that we have rate order, where the bounding rates can be made arbitrarily close to each other by choosing the interval length suitably small. We again construct a sample path stochastic order, but for the time averages of the time spent above zero. We prove the following lemma in §C.5. We use it to prove Lemma C.2 in §C.4.

LEMMA C.4.   *Suppose that conditions of Lemma 8.4 hold. Then, for any $\epsilon > 0$, there exists n_0 and ξ with $0 < \xi \le \delta$ for δ in Lemma 8.4, such that, in addition to the conclusions of Lemma 8.4 and Corollary 8.1, the states $x_m, x_M \in \mathbb{A}$ and random vectors X_m^n, X_M^n, and associated frozen processes $D_f^n(X_m^n)$ and $D_f^n(X_M^n)$ can be chosen so that on $B_n(\xi, \eta)$ (defined as in (60) with ξ instead of δ and necessarily $P(B_n(\xi, \eta)) \to 1$ as $n \to \infty$), first*

$$D_f^n(X_m^n, 0) = D_f^n(X_M^n, 0) = D_{1,2}^n(0) \tag{C6}$$

and, second,

$$\frac{1}{t} \int_0^t 1_{\{D_f^n(X_m^n, s) > 0\}} \, ds - \epsilon \le_{st} \frac{1}{t} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} \, ds \le_{st} \frac{1}{t} \int_0^t 1_{\{D_f^n(X_M^n, s) > 0\}} \, ds + \epsilon \tag{C7}$$

for $n \ge n_0$ and $0 \le t \le \xi$; i.e., there is sample path stochastic order in $\mathbb{D}([0, \xi])$ for $n \ge n_0$.

C.3. Continuity of the FTSP QBD.   In the current proof of Lemma 7.2 and thus Theorem 4.1, we will ultimately reduce everything down to the behavior of the FTSP D. First, we intend to analyze the inhomogeneous queue-difference processes $D_{1,2}^n(\Gamma^n)$ in (6) in terms of associated frozen processes $D_f^n(\Gamma^n)$ introduced in §A.1, obtained by freezing the transition rates at the transition rates in the initial state Γ^n. In (A5), we showed that the frozen-difference processes can be represented directly in terms of the FTSP, by transforming the model parameters (λ_i, m_j) and the fixed initial state γ and scaling time. We will appropriately bound the queue-difference processes $D^n(\Gamma^n)$ above and below by associated frozen-queue difference processes, and then transform them into versions of the FTSP D. For the rest of the proof of Theorem 4.1 in §7, we will exploit a continuity property possessed by this family of pure-jump Markov processes, which exploits their representation as QBD processes using the construction in §6 of Perry and Whitt [37]. We will be applying this continuity property to the FTSP D.

To set the stage, we review basic properties of the QBD process. We refer to §6 of Perry and Whitt [37] for important details. From the transition rates defined in (13)–(16), we see that there are only eight different transition rates overall. The generator Q (in (65) of Perry and Whitt [37]) is based on the four basic $2m \times 2m$ matrices B, A_0, A_1, and A_2, involving the eight transition rates (as shown in (66) of Perry and Whitt [37]). By Theorem 6.4.1 and Lemma 6.4.3 of Latouche and Ramaswami [28], when the QBD is positive recurrent, the FTSP steady-state probability vector has the matrix-geometric form $\alpha_n = \alpha_0 R^n$, where α_n and α_0 are $1 \times 2m$ probability vectors and R is the $2m \times 2m$ rate matrix, which is the minimal nonnegative solutions to the quadratic matrix equation $A_0 + RA_1 + R^2 A_2 = 0$, and can be found efficiently by existing algorithms, as in Latouche and Ramaswami [28]; see Perry and Whitt [37] for applications in our settings. If the drift condition (20) holds, then the spectral radius of R is strictly less than one and the QBD is positive recurrent (Corollary 6.2.4 of Latouche and Ramaswami [28]). As a consequence, we have $\sum_{n=0}^{\infty} R^n = (I - R)^{-1}$. Also, by Lemma 6.3.1 of Latouche and Ramaswami [28], the boundary probability vector α_0 is the unique solution to the system $\alpha_0(B + RA_2) = 0$ and $\alpha \mathbf{1} = \alpha_0(I - R)^{-1} \mathbf{1} = 1$. See §6.4 of Perry and Whitt [37] for explicit expressions for $\pi_{1,2}(\gamma)$ for $\gamma \in \mathbb{A}$.

As in Lemma 8.6, we also use the return time to a fixed state, τ, and its mgf $\phi_\tau(\theta)$ with a critical value $\theta^* > 0$ such that $\phi_\tau(\theta) < \infty$ for $\theta < \theta^*$ and $\phi_\tau(\theta) = \infty$ for $\theta > \theta^*$. We will be interested in the *cumulative process*

$$C(t) \equiv \int_0^t (f(D(s)) - E[f(D(\infty))]) \, ds \quad t \ge 0, \tag{C8}$$

for the special function $f(x) \equiv 1_{\{x \ge 0\}}$. Cumulative processes associated with regenerative processes obey CLTs and FCLTs, depending upon assumptions about the basic cycle random variables τ and $\int_0^\tau f(D(s)) \, ds$, where we assume for this definition that $D(0) = s^*$; see §VI.3 of Asmussen [3] and Glynn and Whitt [13]. From Bolthausen [8], we have the following CLT with a Berry-Esseen bound on the rate of convergence (stated in continuous time, unlike Bolthausen [8]): For any bounded measurable function g, there exists t_0 such that

$$|E[g(C(t))/\sqrt{t}] - E[g(N(0, \sigma^2))]| \le \frac{K}{\sqrt{t}} \quad \text{for all } t > t_0, \tag{C9}$$

where

$$\sigma^2 \equiv E\left[\left(\int_0^\tau f(D(s)) - E[f(D(\infty))]\, ds\right)^2\right], \tag{C10}$$

again assuming for this definition that $D(0) = s^*$. The constant K depends on the function g (as well as the function f in (C8)) and the third absolute moments of the basic cycle variables defined above, plus the first moments of the corresponding cycle variables in the initial cycle if the process does not start in the chosen regenerative state.

There is significant simplification in our case, because the function f in (C8) is an indicator function. Hence, we have the simple domination

$$\int_0^\tau |f(D(s)|\, ds = \int_0^\tau f(D(s))\, ds \le \tau \quad \text{w.p.1}. \tag{C11}$$

As a consequence, boundedness of absolute moments of both cycle variables reduces to the moments of the return times themselves, which are controlled by the mgf.

We will exploit the following continuity result for QBDs, which parallels previous continuity results for Markov processes, e.g., Karr [22] and Whitt [43].

LEMMA C.5 (CONTINUITY OF QBDs). *Consider a sequence of irreducible, positive recurrent QBDs having the structure of the fundamental QBD associated with the FTSP in §3.2 and §6 in Perry and Whitt [37] with generator matrices $\{Q_n: n \ge 1\}$ of the same form. If $Q_n \to Q$ as $n \to \infty$ (which is determined by the convergence of the eight parameters), where the positive-recurrence drift condition (20) holds for Q, then there exists n_0 such that the positive-recurrence drift condition (20) holds for Q_n for $n \ge n_0$. For $n \ge n_0$, the quantities $(R, \alpha_0, \alpha, \phi_\tau, \theta^*, \psi_N, z^*, \sigma^2, K)$ indexed by n are well defined for Q_n, where σ^2 and K are given in (C9) and (C10), and converge as $n \to \infty$ to the corresponding quantities associated with the QBD with generator matrix Q.*

PROOF. First, continuity of R, α_0, and α follows from the stronger differentiability in an open neighborhood of any $\gamma \in \mathbb{A}$, which was shown to hold in the proof of Theorem 5.1 in Perry and Whitt [37], building on Theorem 2.3 in He [17]. The continuity of σ^2 follows from the explicit representation in (C10) (which corresponds to the solution of Poisson's equation). We use the QBD structure to show that the basic cycle variables τ and $\int_0^\tau f(D(s))\, ds$ are continuous function of Q, in the sense of convergence in distributions (or convergence of mgfs and gfs) and then for convergence of all desired moments, exploiting (C11) and the mgf of τ to get the required uniform integrability. Finally, we get the continuity of K from Bolthausen [8] and the continuity of the third absolute moments of the basic cycle variables, again exploiting the uniform integrability. We will have convergence of the characteristic functions used in Bolthausen [8]. However, we do not get an explicit expression for the constants K. $\square$

We use the continuity of the steady-state distribution α in §C.4; specifically in (C19). In addition, we use the following corollary to Lemma C.5 in (C18) in §C.4. We use the notation in (A5).

COROLLARY C.1. *If $(\bar\lambda_i^n, \bar m_j^n, \bar\gamma_n) \to (\lambda_i, m_j, \gamma)$ as $n \to \infty$ for our FTSP QBDs, where (20) holds for (λ_i, m_j, γ), then for all $\epsilon > 0$ there exist t_0 and n_0 such that*

$$P\left(\left|\frac{1}{t}\int_0^t 1_{\{D(\lambda_i^n, m_j^n, \gamma_n, s) > 0\}}\, ds - P(D(\lambda_i, m_j, \gamma, \infty) > 0)\right| > \epsilon\right) < \epsilon$$

for all $t \ge t_0$ and $n \ge n_0$.

PROOF. First apply Lemma C.5 for the steady-state probability vector α, to find n_0 such that $|P(D(\lambda_i^n, m_j^n, \gamma_n, \infty) > 0) - P(D(\lambda_i, m_j, \gamma, \infty) > 0)| < \epsilon/2$ for all $n \ge n_0$. By the triangle inequality, henceforth it suffices to work with $P(D(\lambda_i^n, m_j^n, \gamma_n, \infty) > 0)$ in place of $P(D(\lambda_i, m_j, \gamma, \infty) > 0)$ in the statement to be proved. By (C9), for any M, there exists t_0 such that for all $t \ge t_0$,

$$P\left(\left|\frac{1}{t}\int_0^t 1_{\{D(\lambda_i^n, m_j^n, \gamma_n, s) > 0\}}\, ds - P(D(\lambda_i^n, m_j^n, \gamma_n, \infty) > 0)\right| > \frac{M}{\sqrt{t}}\right)$$

$$< P(|N(0, \sigma^2(\lambda_i^n, m_j^n, \gamma_n))| > M) + \frac{K(\lambda_i^n, m_j^n, \gamma_n)}{\sqrt{t}}. \tag{C12}$$

We get (C12) from (C9) by letting $f(x) = 1_{(0,\infty)}(x)$ in (C8) and letting $g(x) = 1_{\{|x| > M\}}(x)$ in (C9).

Next, choose M so that $P(|N(0, \sigma^2(\lambda_i, m_j, \gamma))| > M) < \epsilon/2$. Then, invoking Lemma C.5, increase n_0 and t_0 if necessary so that $|\sigma^2(\lambda_i^n, m_j^n, \gamma_n) - \sigma^2(\lambda_i, m_j, \gamma))|$ and $|K(\lambda_i^n, m_j^n, \gamma_n) - K(\lambda_i, m_j, \gamma)|$ are sufficiently small so that the right side of (C12) is less than $\epsilon/2$ for all $n \geq n_0$ and $t \geq t_0$. If necessary, increase t_0 and n_0 so that $M/\sqrt{t_0} < \epsilon/2$. With those choices, the objective is achieved. $\square$

C.4. Proof of Lemma C.2. In this subsection we give the first of two long proofs of previous lemmas. We now prove Lemma C.2, which completes the alternate proof of Theorem 4.1 in the case $r_{1,2} = 1$, because then Corollary 8.1 holds with gap $\zeta = 0$. In that case, Corollary 8.1 directly implies that Lemma C.4 holds with $\epsilon = 0$. Lemma C.4, which is proved in the following subsection, is required to treat more general $r_{1,2}$.

First, let $\delta > 0$, $\epsilon > 0$ and t with $0 < t < \delta$ be given, where the δ is chosen as in Lemma 8.4 (with $\delta \leq \tau$ for τ in Theorem 6.3). Here we will be introducing a new interval $[t, t + \sigma]$, where $0 < \sigma \equiv \sigma(t) \equiv \sigma(t, \epsilon, \delta) < \delta - t$, so that $[t, t + \sigma] \subset [0, \delta]$. Moreover, we will make $\sigma < \xi$ for ξ existing via Lemma C.4. Lemmas 8.4 and C.4 hold on the interval $[t, t + \xi]$, where $\xi \equiv \xi(t)$ satisfies $0 < \xi < \delta - t$. We will be choosing σ with $0 < \sigma < \xi$.

Before we started with regularity conditions at time 0 provided by Assumption 3.2. We now will exploit tightness to get corresponding regularity conditions at time t here. In particular, before we started with the convergence $\bar{X}^n(0) \Rightarrow x(0)$ in $\mathbb{R}^3$, where $x(0) \in \mathbb{A}$ based on Assumption 3.2. Now, instead, we base the convergence $\bar{X}^n(t) \Rightarrow \bar{X}(t)$ at time t on the convergence established along the converging subsequence (without introducing new subsequence notation). We apply Lemma C.1 to deduce that $P(\bar{X}(t) \in \mathbb{A}) = 1$. Because the frozen processes to be constructed are Markov processes, we can construct the processes after time t, given only the value of $X^n(t)$, independently of what happens on $[0, t]$. Before, we also started with $D_{1,2}^n(0) \Rightarrow L$, where L is a finite random variable. Instead, here we rely on the stochastic boundedness (and thus tightness) of $\{D_{1,2}^n(t): n \geq 1\}$ in $\mathbb{R}$ provided by Theorem 4.5. As a consequence, the sequence $\{(X^n(t), D_{1,2}^n(t)): n \geq 1\}$ is tight in $\mathbb{R}_2$. Thus there exists a convergent subsequence of the latest subsequence we are considering. Hence, without introducing subsequence notation, we have $(\bar{X}^n(t), D_{1,2}^n(t)) \Rightarrow (\bar{X}(t), L(t))$ in $\mathbb{R}_2$ as $n \to \infty$, where $(\bar{X}(t), L(t))$ is a finite random vector with $P(\bar{X}(t) \in \mathbb{A}) = 1$.

We use the same construction used previously in the proofs of Lemmas 8.4 and C.4, letting σ decrease to achieve new requirements in addition to the conclusions deduced before. We now regard t as the time origin for the frozen difference processes. Given $D_{1,2}^n(t)$, let the two associated frozen difference processes be $\{D_f^n(X_M^n, s): s \geq t\}$ and $\{D_f^n(X_m^{n,\xi}, s): s \geq 0\}$. We directly let

$$D_f^n(X_M^n, t) = D_f^n(X_m^n, t) = D_f^n(X^n(u), t) = D_{1,2}^n(t), \quad u \geq t, \tag{C13}$$

so that we can invoke property (C6) at time t in our application of Lemma C.4 here. As before, the initial random states X_M^n and X_m^n and their fluid-scaled limits are chosen to achieve the goals before and here.

We now successively decrease upper bounds on σ and increase lower bounds on n until we achieve (C2) in Lemma C.2. First, we can apply Lemma 8.4 to find an n_1 such that $P(B^n(\delta, \epsilon)) > 1 - \epsilon/6$ for $n \geq n_1$. We next apply Lemma C.4 to conclude that there exists σ_1 such that the following variants of the integral inequalities in (C7) hold with probability at least $1 - \epsilon/6$ as well:

$$\frac{1}{\sigma} \int_t^{t+\sigma} 1_{\{D_f^n(X_m^n, s) > 0\}} \, ds - \frac{\epsilon}{6m_2} \leq \frac{1}{\sigma} \int_t^{t+\sigma} 1_{\{D_{1,2}^n(s) > 0\}} \, ds \leq \frac{1}{\sigma} \int_t^{t+\sigma} 1_{\{D_f^n(X_M^n, s) > 0\}} \, ds + \frac{\epsilon}{6m_2} \tag{C14}$$

for all $\sigma \leq \sigma_1$. (We divide by m_2 because we will be multiplying by $z_{1,2}(t)$.)

We now present results only for X_M^n, with the understanding that corresponding results hold for X_m^n. We represent the bounding frozen queue-difference processes directly in terms of the FTSP, using the relation (A5), with the notation introduced there:

$$\{D_f^n(\lambda_i^n, m_j^n, X_M^n, t + s): s \geq 0\} \stackrel{\mathrm{d}}{=} \{D(\lambda_i^n/n, m_j^n/n, X_M^n/n, t + sn): s \geq 0\}. \tag{C15}$$

Upon making a change of variables, the bounding integrals in (C14) become

$$\frac{1}{\sigma} \int_t^{t+\sigma} 1_{\{D_f^n(\lambda_i^n, m_j^n, X_M^n, s) > 0\}} \, ds \stackrel{\mathrm{d}}{=} \frac{1}{n\sigma} \int_t^{t+n\sigma} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\sigma}/n, s) > 0\}} \, ds. \tag{C16}$$

For each integer k, we can apply Lemma C.5 to obtain the iterated limits

$$\lim_{n \to \infty} \lim_{s \to \infty} P(D(\lambda_i^n/n, m_j^n/n, X_M^n/n, s) = k) = \lim_{s \to \infty} \lim_{n \to \infty} P(D(\lambda_i^n/n, m_j^n/n, X_M^n/n, s) = k), \tag{C17}$$

where the limit is $P(D(x_M, \infty) = k) \equiv P(D(\lambda_i, m_j, x_M, \infty) = k)$. In particular, the limit for the model parameters first implies convergence of the generators. The convergence of the generators then implies convergence of the processes. Finally Lemma C.5 implies convergence of the associated steady-state distributions.

By Corollary C.1, we also have the associated double limit for the averages over intervals of length $O(n)$ as $n \to \infty$. As $n \to \infty$,

$$\frac{1}{n\sigma} \int_t^{t+n\sigma} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_M^n/n, s) > 0\}} \, ds \;\Rightarrow\; P(D(\lambda_i, m_j, x_M, \infty) > 0) \equiv \pi_{1,2}(x_M). \tag{C18}$$

Invoking Lemma C.5, choose σ_2 less than or equal to the previous value σ_1 such that

$$|\pi_{1,2}(x_m) - \pi_{1,2}(\bar{X}(t))| \leq \frac{\epsilon}{6m_2} \quad \text{for all } \sigma \leq \sigma_2. \tag{C19}$$

For that σ_2, applying (C18), choose $n_2 \geq n_1$ such that for all $n \geq n_2$,

$$P\left(\left|\frac{1}{n\sigma_2} \int_t^{t+n\sigma_2} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_M^n/n, s) > 0\}} \, ds - \pi_{1,2}(x_M)\right| > \frac{\epsilon}{6m_2}\right) < \frac{\epsilon}{6}. \tag{C20}$$

We now use the convergence of $\bar{X}^n$ along the subsequence over $[0, t]$ together with the tightness of the sequence of processes $\{\bar{X}^n : n \geq 1\}$ to control $\bar{Z}_{1,2}^n$ in an interval after time t. In particular, there exists $\sigma_3 \leq \sigma_2$ and $n_3 \geq n_2$ such that

$$P\left(\sup_{u : t \leq u \leq t + \sigma_3} \{|\bar{X}^n(u) - \bar{X}(t)|\} > \epsilon/6\right) < \epsilon/6 \quad \text{for all } n \geq n_3. \tag{C21}$$

For the current proof, we will use the consequence

$$P\left(\sup_{u : t \leq u \leq t + \sigma_3} \{|\bar{Z}_{1,2}^n(u) - \bar{Z}_{1,2}(t)|\} > \epsilon/6\right) < \epsilon/6 \quad \text{for all } n \geq n_3. \tag{C22}$$

We let the final value of σ be σ_3. We now show the consequences of the selections above. We will directly consider only the upper bound; the reasoning for the lower bound is essentially the same. Without loss of generality, we take $\epsilon \leq 1 \wedge m_2$. From above, we have the following relations (explained afterward) holding with probability at least $1 - \epsilon$ (counting $\epsilon/6$ once each to achieve $\|x_M - \bar{X}(t)\| \leq \epsilon/6$, $\|x_m - \bar{X}(t)\| \leq \epsilon/6$, (C14), (C22) and twice for (C20)):

$$
\begin{aligned}
\text{(a)} \quad & \int_t^{t+\sigma} 1_{\{D_{1,2}^n(s) > 0\}} \bar{Z}_{1,2}^n(s) \, ds \leq \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \int_t^{t+\sigma} 1_{\{D_{1,2}^n(s) > 0\}} \, ds \\
\text{(b)} \quad & \leq \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \left(\int_t^{t+\sigma} 1_{\{D_f^n(\lambda_i^n, m_j^n, X_M^n, s) > 0\}} \, ds + \frac{\epsilon\sigma}{6m_2}\right) \\
\text{(c)} \quad & \stackrel{\text{d}}{=} \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \left(\int_0^{\sigma} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_M^n/n, t+sn) > 0\}} \, ds + \frac{\epsilon\sigma}{6m_2}\right) \\
\text{(d)} \quad & \stackrel{\text{d}}{=} \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \sigma \left(\frac{1}{n\sigma} \int_0^{n\sigma} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_M^n/n, t+s) > 0\}} \, ds + \frac{\epsilon}{6m_2}\right) \\
\text{(e)} \quad & \leq \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \sigma \left(\pi_{1,2}(x_M) + \frac{2\epsilon}{6m_2}\right) \\
\text{(f)} \quad & \leq \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \sigma \left(\pi_{1,2}(\bar{X}(t)) + \frac{3\epsilon}{6m_2}\right) \\
\text{(g)} \quad & \leq \bar{Z}_{1,2}(t) \pi_{1,2}(\bar{X}(t)) \sigma + \frac{\pi_{1,2}(\bar{X}(t))}{6} \epsilon\sigma + \frac{1}{2}\epsilon\sigma + \frac{\sigma\epsilon^2}{12m_2} \\
\text{(h)} \quad & \leq \bar{Z}_{1,2}(t) \pi_{1,2}(\bar{X}(t)) \sigma + \frac{3}{4}\epsilon\sigma \\
& \leq (\bar{Z}_{1,2}(t) \pi_{1,2}(\bar{X}(t)) + \epsilon)\sigma \quad \text{for all } n \geq n_0 \equiv n_3.
\end{aligned}
\tag{C23}
$$

We now explain the steps in (C23): First, for (a) we replace $\bar{Z}_{1,2}^n(s)$ by $\bar{Z}_{1,2}(t)$ for $t \leq s \leq t + \xi$ by applying (C22). For (b), we apply Lemma C.4. For (c), we use the alternative representation in terms of the FTSP in (C15). For (d), we use the change of variables in (C16). For (e), we use (C20), exploiting the convergence in (C18). For (f), we use (C19). Step (g) is simple algebra, exploiting $\bar{Z}_{1,2}(t) \leq m_2$. Step (h) is more algebra, exploiting $\pi_{1,2}(\bar{X}(t)) \leq 1$, and $\epsilon \leq 1 \wedge m_2$. That completes the proof of the lemma. $\quad\square$

C.5. Proof of Lemma C.4. We will directly show how to carry out the construction of the sample path order for stochastic processes having the same distributions as the terms in (C7). We will show that the interval $[0, \xi]$ can be divided into a large number $(O(n))$ of alternating subintervals, each of length $O(1/n)$, such that full sample path order holds on one subinterval, and then the processes are unrelated on the next subinterval. We will then show, for an appropriate choice of ξ, that these intervals can be chosen so that the first intervals where order holds are much longer than the second intervals where the processes are unrelated. Hence we will deduce for this construction that the inequalities in (C7) holds w.p.1 for all n sufficiently large. We will specify the steps in the construction and explain why they achieve their goal, while minimizing the introduction of new notation. We highlight twelve key steps in the argument.

An integer state space. We exploit Assumption 3.3 to have $r_{1,2}$ be rational, and then use the integer state space associated with the QBD representation in §6 of Perry and Whitt [37]. However, we do not directly exploit the QBD structure; instead we directly work with the CTMCs with transitions $\pm j$ and $\pm k$, where j and k are positive integers. We will reduce the analysis to consideration of homogeneous CTMCs.

Exploiting Lemma 8.4, but reducing the length of the interval. We construct the states $x_m, x_M \in \mathbb{A}$, and random vectors X_m^n, X_M^n, and associated frozen processes $D_f^n(X_m^n)$ and $D_f^n(X_M^n)$ exactly as in Lemma 8.4, but we adjust the specific values to make them closer together at time 0 as needed by choosing ξ to be suitably smaller than the δ needed in Lemma 8.4. It is significant that here ξ can be arbitrarily small; we only require that $\xi > 0$.

Initializing at time 0. By Assumption 3.2, we have $\bar{X}^n \Rightarrow x(0) \in \mathbb{A} \subset \mathbb{R}_3$ and $D_{1,2}^n(0) \Rightarrow L \in \mathbb{R}$. Since $x(0)$ is deterministic, we immediately have the joint convergence $(\bar{X}^n, D_{1,2}^n(0)) \Rightarrow (x(0), L) \in \mathbb{R}_4$; apply Theorem 11.4.5 of Whitt [47]. We then apply the Skorohod representation theorem to replace the convergence in distribution with convergence w.p.1., without changing the notation. Hence, we can start with $(\bar{X}^n, D_{1,2}^n(0)) \to (x(0), L) \in \mathbb{R}_4$ w.p.1. With that framework, we condition on $D_{1,2}^n(0)$. We then initialize the processes $D_f^n(X_m^n, 0)$ and $D_f^n(X_M^n, 0)$ to satisfy $D_f^n(X_m^n, 0) = D_f^n(X_M^n, 0) = D_{1,2}^n(0)$, as in (C6).

Coupling over an initial random interval. Given identical initial values, we can apply the rate order in Lemma 8.4 to construct versions of the processes that will be ordered w.p.1 over an initial interval of random positive length, just as in Corollary 8.1. Let ν_n be the first time that the sample path order is violated. The rate order implies that $\nu_n > 0$ w.p.1 for each $n \geq 1$. The coupling is performed over the interval $[0, \nu_n]$. At random time ν_n, we must have all three processes in the boundary set of states $\{j: -\beta + 1 \leq j \leq \beta\}$, where $\beta \equiv j \vee k$. That is so, because violation of order only need occur when, just prior to the order violation, we have $-\beta < D_f^n(X_m^n) \leq 0 < 1 \leq D_f^n(X_M^n) < \beta$, where the rates are no longer ordered properly, because the processes are in different regions. At time ν_n, either the upper process jumps down below the lower process or the lower process jumps up over the upper process.

A new construction using independent versions when order is first violated. At time ν_n, the order is first violated and would remain violated over an interval thereafter. However, at this time ν_n, we alter the construction. At this random time ν_n we temporarily abandon the coupling based on rate order. Instead, going forward from time ν_n, we construct independent versions of the three processes being considered. More precisely, the three processes are conditionally independent, given their initial values at time ν_n. The idea is to let them evolve in this independent manner until the three processes reach a state in which the desired sample path ordering does again hold. We do this in a simple controlled manner. We wait until, simultaneously, the upper bound process $D_f^n(X_M^n, \nu_n + t)$ exceeds a suitably high threshold, the lower bound process $D_f^n(X_m^n, \nu_n + t)$ falls below a suitably low threshold, and the interior processes $D_{1,2}^n(\nu_n + t)$ is in a middle region. At such a random (stopping) time, the three processes will necessarily be ordered in the desired way. After that time, we can use the coupling again.

Avoiding working directly with $D_{1,2}^n$. However, since $D_{1,2}^n(\nu_n + t)$ is an inhomogeneous CTMC, and thus difficult to work with, we avoid working with it directly. Instead, we simultaneously construct new upper and lower bound frozen processes, $\tilde{D}_f^n(X_M^n)$ and $\tilde{D}_f^n(X_m^n)$ starting at time ν_n, coupled with $\{D_{1,2}^n(\nu_n + t): t \geq 0\}$, initialized by stipulating that $\tilde{D}_f^n(X_M^n, \nu_n) = \tilde{D}_f^n(X_m^n, \nu_n) = D_{1,2}^n(\nu_n)$. These new processes are coupled with $\{D_{1,2}^n(\nu_n + t): t \geq 0\}$, but conditionally independent of the independent versions of the other two processes $\{D_f^n(X_M^n, \nu_n + t): t \geq 0\}$ and $\{D_f^n(X_m^n, \nu_n + t): t \geq 0\}$. We again can apply Lemma 8.4 to obtain rate order, but again we cannot have full sample path order, because of the gap. Nevertheless, we can now apply Corollary 8.1 to conclude that we have the sample path stochastic order

$$\tilde{D}_f^n(X_m^n, \nu_n + t) - \beta \leq_{st} D_{1,2}^n(\nu_n + t) \leq_{st} \tilde{D}_f^n(X_M^n, \nu_n + t) + \beta \quad \text{in } D([\nu_n, \delta]).$$

We thus can do a sample path construction to achieve

$$\tilde{D}_f^n(X_m^n, \nu_n + t) - \beta \leq D_{1,2}^n(\nu_n + t) \leq \tilde{D}_f^n(X_M^n, \nu_n + t) + \beta \quad \text{for } 0 \leq t \leq \delta - \nu_n \quad w.p.1. \tag{C24}$$

Stopping times for desired order to be achieved. We now define stopping times for the order to be achieved. Let

$$\phi_n \equiv \inf \{t > 0: D_f^n(X_M^n, \nu_n + t) > 2\beta, \; D_f^n(X_m^n, \nu_n + t) < -2\beta, \; -\beta < \tilde{D}_f^n(X_M^n, \nu_n + t) < \beta\}, \qquad (\text{C25})$$

with the three processes $D_f^n(X_M^n)$, $D_f^n(X_m^n)$, and $\tilde{D}_f^n(X_M^n)$ in (C25) being mutually conditionally independent, given that all three are initialized at time ν_n with the final values there obtained from the evolution of the coupled processes up until time ν_n. (We have conditional independence given the vector of initial values at time ν_n.) We can combine (C25) with (C24) to conclude that we have the appropriate order at time $\nu_n + \phi_n$, in particular,

$$D_f^n(X_m^n, \nu_n + \phi_n) \le D_{1,2}^n(\nu_n + \phi_n) \le D_f^n(X_M^n, \nu_n + \phi_n) \quad \text{w.p.1.}$$

We not only obtain the desired order at time $\nu_n + \phi_n$, but the stopping time ϕ_n depends on the three processes $\{D_f^n(X_M^n, \nu_n + t): t \ge 0\}$, $\{D_f^n(X_m^n, \nu_n + t): t \ge 0\}$, and $\{\tilde{D}_f^n(X_M^n, \nu_n + t): t \ge 0\}$, which are independent homogeneous CTMCs conditional on their random state vectors X_M^n and X_m^n, respectively. Thus the stopping time is with respect to the three-dimensional vector-valued CTMC (conditional on the random state vectors X_M^n and X_m^n).

Alternating cycles using coupled and independent versions. Going forward after the time $\nu_n + \phi_n$ for ϕ_n in (C25), we again construct coupled processes just as we did at time 0. However, now we initialize by having the specified states coincide at time $\nu_n + \phi_n$; i.e., $D_f^n(X_M^n, \nu_n + \phi_n) = D_f^n(X_m^n, \nu_n + \phi_n) = D_{1,2}^n(\nu_n + \phi_n)$. Hence, we can repeat the coupling done over the initial interval $[0, \nu_n]$. In particular, we can apply the ordering of the initial states and the rate ordering to construct new coupled versions after time $\nu_n + \phi_n$. We can thus repeat the construction already done over the time interval $[0, \nu_n]$ after time $\nu_n + \phi_n$. The coupled construction beginning at time $\nu_n + \phi_n$ will end at the first subsequent time that the order is violated. At this new random time (paralleling ν_n), we must have all three processes in the boundary set of states $\{j: -\beta + 1 \le j \le \beta\}$. In this way, we produce a sequence of alternating intervals, where we perform coupling on one interval and then create independent versions on the next interval.

Applying a FWLLN after scaling time by n. Because the transition rates of $D_{1,2}^n$ and all the other processes are $O(n)$, there necessarily will be order $O(n)$ of these cycles in any interval $[0, \xi]$. However, we scale time by n, replacing t by nt, just as in (A5), (C15), and (C16) to represent all processes as FTSPs with random parameters depending on n, converging to finite deterministic limits. Once we scale time by n in that way, the scaled transition rates converge to finite limits as $n \to \infty$, but the relevant time interval becomes $[0, n\xi]$ instead of $[0, \xi]$, as in (C16). We can thus apply a FWLLN to complete the proof.

Regenerative structure as a basis for the FWLLN. As a formal basis for the FWLLN, we can apply regenerative structure associated with successive epochs of order violation starting from coupling, but a regenerative cycle must contain more than two successive intervals; we do not have an alternating renewal process. Such order-violation epochs necessarily occur in the boundary region $\{j: -\beta + 1 \le j \le \beta\}$, where $\beta \equiv j \vee k$. Hence, there are at most $(2\beta)^3$ vectors of values for the three processes. Moreover, there necessarily will be one of these vectors visited infinitely often. Successive visits to that particular state vector after coupling thus serve as regeneration points for the entire process. That is, there is an embedded delayed renewal process. We next ensure that the times between successive regeneration times have finite mean values, with the correct asymptotic properties as $n \to \infty$. That justifies the FWLLN. In particular, we can apply a FWLLN for cumulative processes, as in Glynn and Whitt [13]. For the particular QBD processes being considered, there is continuity of the asymptotic parameters, as indicated in Lemma C.5.

Finite mean time between regenerations of order $O(1/n)$. From basic CTMC theory, it follows that the first passage times ϕ_n in (C25) have finite mean values that are of order $O(1/n)$. In particular, $E[\phi_n] < c_{n,1} < \infty$, where $nc_{n,1} \to c_1 > 0$. It is more difficult to treat the mean time over which the order is valid during a single coupling, i.e., the initial time ν_n and the subsequent random times that which the order is valid during the coupling. However, we can truncate the variables at finite constant times in order to ensure that the FWLLN reasoning can be applied; e.g., we replace ν_n by $\tilde{\nu}_n \equiv \nu_n \wedge c_{n,2} < \infty$, where $nc_{n,2} \to c_2 > 0$. We can later choose c_2 large enough so that the inequalities in (C7) hold for the specified ϵ. Finally, the total number of these cycles until the designated initial state vector appears again is a random variable. Because the successive vector of initial state vectors visited on successive cycles is a finite-state discrete-time Markov chain, the random number of cycles within a regeneration interval is a random variable with finite mean, say $c_{n,3}$, with $c_{n,3} \to c_3 > 0$ as $n \to \infty$. As a consequence, the mean time between successive regenerations is $c_{n,4} < \infty$, where $nc_{n,4} \to c_4 > 0$ as $n \to \infty$.

The proportion of time that order holds. Finally, it is important that we can control the proportion of time that order holds in any interval $[0, t]$, $0 < t \le \xi$, as stated in (C7), for an appropriate choice of ξ. As a basis

for that control, we exploit the fact that the distribution of ϕ_n, the length of the first interval during which order violation is allowed, and the random lengths of all subsequent intervals on which the desired sample path order is violated, depend on the states of the three processes at time ν_n, i.e., upon $D_f^n(X_m^n, \nu_n)$, $D_f^n(X_M^n, \nu_n)$, and $D_{1,2}^n(\nu_n)$. However, because the violation necessarily occurs in the boundary region $\{j: -\beta + 1 \leq j \leq \beta\}$, where $\beta \equiv j \vee k$, there are at most 2β possible values for each process, and thus at most $(2\beta)^3$ vectors of values for the four processes. Because this number is finite, the violation interval lengths (like ϕ_n) can be stochastically bounded above, independent of the specific values at the violation point. Moreover, this can be done essentially independently of ξ. On the other hand, the length of the intervals on which the coupling construction remains valid necessarily increases as we decrease ξ and appropriately increase c_2 and $c_{n,2}$ in the paragraph above. Thus, for any $\epsilon > 0$, we can achieve the claimed ordering in (C7) for all $n \geq n_0$ by an appropriate choice of ξ, c_2, and n_0. $\square$

Appendix D. The bounding QBD in Lemma B.3. We now describe how to present the process D_*^n in Step 1 of the proof of Lemma B.3 as a QBD for each n. We assume that $r_{2,1} = j/k$ for $j, k \in \mathbb{Z}_+$, and let $m \equiv j \vee k$. We define the process $\tilde{D}_*^n \equiv kQ_{2,*}^n - jQ_{1,*}^n$. Thus, $\tilde{D}_*^n$ is D_*^n with an altered state space. In particular, each process is positive recurrent if and only if the other one is.

We divide the state space $\mathbb{Z}_+ \equiv \{0, 1, 2, \ldots\}$ into level of size m: Denoting level i by $L(i)$, we have

$$L(0) = (0, 1, \ldots, m-1), \qquad L(1) = (m, m+1, \ldots, 2m-1) \quad \text{etc.}$$

The states in $L(0)$ are called the boundary states. Then the generator matrix $Q^{(n)}$ of the process D_*^n has the QBD form

$$Q^{(n)} \equiv \begin{pmatrix} B^{(n)} & A_0^{(n)} & 0 & 0 & \cdots \\ A_2^{(n)} & A_1^{(n)} & A_0^{(n)} & 0 & \cdots \\ 0 & A_2^{(n)} & A_1^{(n)} & A_0^{(n)} & \cdots \\ 0 & 0 & A_2^{(n)} & A_1^{(n)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}.$$

(All matrices are functions of X_*^n. However, to simplify notation, we drop the argument X_*^n, and similarly in the example below.)

For example, if $j = 2$ and $k = 3$, then

$$B^{(n)} = \begin{pmatrix} -\sigma^n + \hat{\mu}_\Sigma^n & 0 & \hat{\lambda}_2^n \\ \hat{\mu}_\Sigma^n & -\sigma^n & 0 \\ \hat{\mu}_\Sigma^n & 0 & -\sigma^n \end{pmatrix}, \qquad A_0^{(n)} = \begin{pmatrix} \hat{\lambda}_3^n & 0 & 0 \\ \hat{\lambda}_2^n & \hat{\lambda}_3^n & 0 \\ 0 & \hat{\lambda}_2^n & \hat{\lambda}_3^n \end{pmatrix},$$

$$A_1^{(n)} = \begin{pmatrix} -\sigma^n & 0 & \hat{\lambda}_2^n \\ 0 & -\sigma^n & 0 \\ \hat{\mu}_2^n & 0 & -\sigma^n \end{pmatrix}, \qquad A_2^{(n)} = \begin{pmatrix} \hat{\mu}_3^n & \hat{\mu}_2^n & 0 \\ 0 & \hat{\mu}_3^n & \hat{\mu}_2^n \\ 0 & 0 & \hat{\mu}_2^n \end{pmatrix},$$

where $\hat{\mu}_\Sigma^n \equiv \hat{\mu}_3^n + \hat{\mu}_2^n$ and $\sigma^n \equiv \hat{\mu}_\Sigma^n + \hat{\lambda}_2^n + \hat{\lambda}_3^n$.

Let $A^{(n)} \equiv A_0^{(n)} + A_1^{(n)} + A_2^{(n)}$. Then $A^{(n)}$ is an irreducible CTMC infinitesimal generator matrix. It is easy to see that its unique stationary probability vector, $\nu^{(n)}$, is the uniform probability vector, attaching probability $1/m$ to each of the m states. Then by Theorem 7.2.3 in Latouche and Ramaswami [28], the QBD is positive recurrent if and only if $\nu A_0^{(n)} \mathbf{1} < \nu A_2^{(n)} \mathbf{1}$, where $\mathbf{1}$ is the vector of all 1s. This translates to the stability condition given in the proof of Lemma B.3.

Appendix E. Positivity of fluid limit. In this appendix we explain why Assumption 3.2 about the initial conditions is reasonable. First, the assumed convergence $D_{1,2}^n(0) \Rightarrow L$ if $x(0) \in \mathbb{A}$ is natural, since that convergence holds whenever $x(t) \in \mathbb{A}$, provided that t is not a hitting time of the set $\mathbb{A}$ from $\mathbb{S} - \mathbb{A}$, by virtue of Theorem 4.6 and Remark 4.2. We can take time 0 to be a time shortly after a hitting time of $\mathbb{A}$.

We now elaborate on another part of Assumption 3.2, namely, the assumed initial values for each n. Observe that, given the convergence $\bar{X}^n \Rightarrow x$ established in Theorem 4.1, we necessarily have $Z_{1,2}^n(t) = O_P(n)$ (and thus $Z_{2,1}^n(t) = 0$) and $Q_i^n(t) = O_P(n)$, $i = 1, 2$, if $z_{1,2}(t) > 0$ and $q_i(t) > 0$, $i = 1, 2$, where $x \equiv (q_1, q_2, z_{1,2})$ is the fluid limit. However, in general we may not have such strict positivity in the components of the fluid limit at time 0. (Recall that Assumption 3.2 allows $Q_i^n(0) = o(n)$, $i = 1, 2$, and $Z_{1,2}^n(0)$ is allowed to be 0.) Nevertheless, we now show that we do necessarily have strict positivity of the components of $x(t)$ for all $t > 0$ small enough, even if that property does not hold at time 0.

For a vector $y \in \mathbb{R}_3$, we write $y > 0$ if the (strict) inequality holds componentwise.

PROPOSITION E.1. *The fluid limit x satisfies $x(t) > 0$ for all $t > 0$ sufficiently small.*

PROOF. The statement follows immediately for the case $x(0) > 0$ because of the continuity of x. We thus assume that at least one component of $x(0)$ is not strictly positive. If $z_{1,2}(0) = 0$, then it was shown in the proof of Theorem 6.4 that $z_{1,2}(t) > 0$ for all $t > 0$. Now note that, by the assumption on the initial condition $x(0)$ in Assumption 3.2, $q_1(0) \geq r q_2(0)$, so that if $q_1(0) = 0$, then we must have $q_2(0) = 0$ as well. We prove the result for the three possible regions of $x(0)$ in Assumption 3.2, namely, $\mathbb{S}^+$, $\mathbb{A}^+$, and $\mathbb{A}$, separately.

(i) $x(0) \in \mathbb{S}^+$. In this case, $q_1(0) > 0$ and $\pi_{1,2}(x(0)) = 1$, so we need to consider the case $q_2(0) = 0$. Plugging these values of $\pi_{1,2}(x(0))$ and $q_2(0)$ in the equation for $\dot{q}_2$ in (24), we see that $\dot{q}_2(0) = \lambda_2 > 0$, so that q_2 is strictly increasing at time 0, and the result follows.

(ii) $x(0) \in \mathbb{A}^+$. It is shown in Lemma 7.2 in Perry and Whitt [37] that if $x(0) \in \mathbb{A}^+$, then for all $t > 0$ sufficiently small, $x(t) \in \mathbb{S} - \mathbb{S}^- - \mathbb{A}^-$, i.e., $q_1(t) \geq r q_2(t)$ for all $t \in (0, \epsilon_1]$, for some $\epsilon_1 > 0$. As in the previous case, $\pi_{1,2}(x(0)) = 1$ and, assuming $q_2(0) = 0$, we have that q_2 is strictly increasing at time 0, so that $q_2(t) > 0$ for all t satisfying $0 < t \leq \epsilon_2$, for some $\epsilon_2 > 0$. Then, with $\varepsilon \equiv \epsilon_1 \wedge \epsilon_2$, $q_1(t) > 0$ for all $t \in (0, \varepsilon]$.

(iii) $x(0) \in \mathbb{A}$. By Theorem 5.2 in Perry and Whitt [37], the fluid limit x is right differentiable at zero and is differentiable on an open interval $(0, \varepsilon)$ for some $\varepsilon > 0$. For $i = 1, 2$, if $q_i(t) = 0$, then $\dot{q}_i(t)$ cannot be negative by Theorem 5.1 in Perry and Whitt [37]. Hence, because we are considering the case $q_i(0) = 0$, we have two possibilities: either $\dot{q}_i(0) = 0$ or $\dot{q}_i(0) > 0$.

Our proof builds on the fact that $\dot{x}(t)$ is itself a continuous function of t. That is so because the right-hand side of each component of $\dot{x}(t)$ in (24) includes the system's parameters, the Lipschitz-continuous fluid limit $x(t)$, and the function $\pi_{1,2}(x(t))$. However, $\pi_{1,2}(x)$ is locally Lipschitz continuous as a function of x by Theorem 7.1 in Perry and Whitt [37], and in particular, $\pi_{1,2}(x(t))$ is a continuous function of $x(t)$, which is itself a continuous function of t. Hence, $\pi_{1,2}(x(t))$ is continuous in t, and so is $\dot{q}_i(t)$, $i = 1, 2$. As a consequence, if $\dot{q}_i(0) > 0$, then $\dot{q}_i(t) > 0$ for all t in some neighborhood of 0, so that q_i is strictly increasing on a positive interval, $i = 1, 2$.

Hence it remains to consider the case in which $\dot{q}_i(0) = 0$ for $i = 1$ or $i = 2$. Assuming that to be the case, we will show that $\dot{q}_j(0) = 0$ for $j \neq i$, $j = 1, 2$. Indeed, if $\dot{q}_j(0) > 0$, then q_j must be increasing at a neighborhood of time 0, so that $x(t) \notin \mathbb{A}$ for all $t > 0$ small enough, contradicting the fact that $\mathbb{A}$ is an open set. Hence, the proof of the proposition for the case $x(0) \in \mathbb{A}$ will follow if we assume that $q_i(0) = \dot{q}_i(0) = 0$ for both $i = 1$ and $i = 2$, and show that we reach a contradiction. We consider two subcases, depending on whether $\mu_{1,2} > \mu_{2,2}$ or $\mu_{1,2} \leq \mu_{2,2}$.

If $\mu_{1,2} > \mu_{2,2}$, then define the function $V(x(t)) = q_1(t) + q_2(t)$. It is shown in the proof of Theorem 5.1 in Perry and Whitt [37] that V is strictly increasing whenever $x(t) \in \mathbb{A}$, so that at least one queue is increasing at time 0. Once again, this means that the second queue must also be strictly increasing, for otherwise x will leave $\mathbb{A}$ immediately after time 0, so the statement is proved.

If $\mu_{1,2} \leq \mu_{2,2}$, then define the function $V(x(t)) = (1 + \epsilon)q_1(t) + q_2(t) + \epsilon z_{1,2}(t)$, for arbitrary $\epsilon > 0$. We again use the proof of Theorem 5.1 in Perry and Whitt [37] to conclude that V is strictly increasing at time 0. However, in that case, even though $\dot{q}_i(0) = 0$, $i = 1, 2$, V can be increasing because $\dot{z}_{1,2}(0) > 0$. Assume that is the case, and consider the ODE (24). (Recall that $q_i(0) = \dot{q}_i(0) = 0$, $i = 1$ and 2.)

It follows from the assumption $\dot{z}_{1,2}(0) > 0$ and the equation for $\dot{z}_{1,2}$ in (24), that

$$\pi_{1,2}(x(0))(\mu_{1,2}z_{1,2}(0) + \mu_{2,2}z_{2,2}(0)) > \mu_{1,2}z_{1,2}. \tag{E1}$$

Then, by the assumption $\dot{q}_1(0) = q_1(0) = 0$, we have

$$0 = \dot{q}_1(0) = \lambda_1 - \mu_{1,1}m_1 - \pi_{1,2}(x(0))(\mu_{1,2}z_{1,2}(0) + \mu_{2,2}z_{2,2}(0)) < \lambda_1 - \mu_{1,1}m_1 - \mu_{1,2}z_{1,2}(0),$$

where the inequality follows from (E1). Hence, $\mu_{1,2}z_{1,2}(0) > \lambda_1 - \mu_{1,1}m_1$, which further implies that

$$z_{1,2}(0) > s_2^a, \tag{E2}$$

for s_2^a in (8). From the equation for $\dot{q}_2(0)$ in (24) (recalling that $q_2(0) = 0$ by assumption), we get

$$\dot{q}_2(0) = \lambda_2 - \mu_{1,2}z_{1,2}(0) - \mu_{2,2}z_{2,2}(0) + \pi_{1,2}(x(0))(\mu_{1,2}z_{1,2}(0) + \mu_{2,2}z_{2,2}(0)) > \lambda_2 - \mu_{2,2}z_{2,2}(0),$$

where again, the inequality follows from (E1). Now, since $z_{2,2}(0) = m_2 - z_{1,2}(0)$, we have

$$\dot{q}_2(0) > \lambda_2 - \mu_{2,2}m_2 + \mu_{2,2}z_{1,2}(0) > \lambda_2 - \mu_{2,2}m_2 + \mu_{2,2}s_2^a \geq 0,$$

where the second inequality follows from (E2), and the third from Assumption 3.1. This contradicts the assumption that $\dot{q}_2(0) = \dot{q}_1(0) = 0$. Hence the proof is complete. $\square$

Appendix F. List of acronyms. In this appendix we list all the acronyms used in the paper and refer to the sections where they are introduced and discussed.

AP—averaging principle (§1 and Perry and Whitt [36, 37])
CTMC—continuous time Markov chain (§1)
FCLT—functional central limit theorem (§1)
FQR-T—fixed queue ratio *control* with thresholds (§§1 and 2.2)
FSLLN—functional strong law of large numbers (§6)
FTSP—fast-time-scale process (§§1 and 3.2)
FWLLN—functional weak law of large numbers (§§1 and 4.1)
ODE—ordinary differential equation (§§1 and 3.3)
QBD—quasi-birth-and-death *process* (§3.3 and Perry and Whitt [37])
QR-T—*not fixed* queue ratio *control* with thresholds (§1 and Perry and Whitt [35])
SSC—state-space collapse (§§1 and 6)

References

[1] Antunes N, Fricker C, Robert P, Tibi D (2006) Analysis of loss networks with routing. *Ann. Appl. Prob.* 16(4):2007–2026.
[2] Asmussen S (1998) Extreme value theory for queues via cycle maxima. *Extremes* 1(2):137–168.
[3] Asmussen S (2003) *Applied Probability and Queues*, 2nd ed. (Springer, New York).
[4] Bassamboo A, Harrison JM, Zeevi A (2005) Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems* 51(3–4):249–285.
[5] Bassamboo A, Harrison JM, Zeevi A (2006) Dynamic and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* 54(3):419–435.
[6] Baxendale P (2005) Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Prob.* 15(1):700–738.
[7] Billingsley P (1999) *Convergence of Probability Measures*, 2nd ed. (John Wiley & Sons, New York).
[8] Bolthausen E (1980) The Berry-Esseen theorem for functionals of discrete Markov chains. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 54(1):59–73.
[9] Borkovec M, Klüppelberg C (1998) Extremal behavior of diffusion models in finance. *Extremes* 1(1):47–80.
[10] Coffman EG, Puhalskii AA, Reiman MI (1995) Polling systems with zero switchover times: A heavy-traffic averaging principle. *Ann. Appl. Probab.* 5(3):681–719.
[11] Davis RA (1982) Maximum and minimum of one-dimensional diffusions. *Stochastic Processes and Their Appl.* 13(1):1–9.
[12] Ethier SN, Kurtz TG (1985) *Markov Processes: Characterization and Convergence* (John Wiley & Sons, New York).
[13] Glynn PW, Whitt W (1993) Limit theorems for cumulative processes. *Stoch. Proc. Appl.* 47(2):299–314.
[14] Gurvich I, Whitt W (2009) Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* 34(2):363–396.
[15] Gurvich I, Whitt W (2009) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* 11(2):237–253.
[16] Gurvich I, Whitt W (2010) Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* 58(2):316–328.
[17] He Q (1993) Differentiability of the matrices R and G in the matrix analytic method. *Stochastic Models* 11(1):123–132.
[18] Hunt PJ, Kurtz TG (1994) Large loss networks. *Stochastic Processes and Their Appl.* 53(2):363–378.
[19] Jacod J, Shiryaev AN (1987) *Limit Theorems for Stoschastic Processes* (Springer, Berlin).
[20] Kamae T, Krengel U, O'Brien GL (1977) Stochastic inequalities on partially ordered spaces. *Ann. Prob.* 5(6):899–912.
[21] Kang W, Ramanan K (2010) Fluid limits of many-server queues with reneging. *Ann. Appl. Prob.* 20:2204–2260.
[22] Karr AF (1975) Weak convergence of a sequence of Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 33(1):41–48.
[23] Kaspi H, Ramanan K (2011) Law of large numbers limits for many-server queues. *Ann. Appl. Prob.* 21(1):33–114.
[24] Khalil HK (2002) *Nonlinear Systems* (Prentice Hall, Englewood Cliffs, NJ).
[25] Khasminskii RZ, Yin G (2004) On averaging principles: An asymptotic expansion approach. *SIAM J. Math. Anal.* 35(6):1534–1560.
[26] Kontoyiannis I, Meyn SP (2003) Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Prob.* 13(1):304–362.
[27] Kurtz TG (1992) Averaging for martingale problems and stochastic approximations, *Appl. Stochastic Anal., Proc. US-French Workshop*, Lecture Notes in Control Inform. Sci., Vol. 177 (Springer, Berlin), 186–209.
[28] Latouche G, Ramaswami V (1999) *Introduction to Matrix Analytic Methods in Stochastic Modeling* (Siam and ASA, Philadelphia).

[29] Leadbetter MR, Lindgren G, Rootzén H (1983) *Extremes and Related Properties of Random Sequences and Processes* (Springer-Verlag, New York).

[30] Lindvall T (1992) *Lectures on the Coupling Method* (John Wiley & Sons, New York).

[31] Massey WA, Whitt W (1998) Uniform acceleration expansions for Markov chains with time-varying rates. *Ann. Appl. Prob.* 8(4):1130–1155.

[32] Meyn SP, Tweedie RL (1993) *Markov Chains and Stochastic Stability* (Springer, New York).

[33] Müller A, Stoyan D (2002) *Comparison Methods for Stochastic Models with Risks* (John Wiley & Sons, New York).

[34] Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.

[35] Perry O, Whitt W (2009) Responding to unexpected overloads in large-scale service systems. *Management Sci.* 55(8):1353–1367.

[36] Perry O, Whitt W (2011) A fluid approximation for service systems responding to unexpected overloads. *Oper. Res.* 59(5):1159–1170.

[37] Perry O, Whitt W (2011) An ODE for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems* 1(1):17–66.

[38] Perry O, Whitt W (2012) Diffusion approximation for an overload X model via a stochastic averaging principle. Working paper, Columbia University. Available at: http:// www.columbia.edu/~ww2040/allpapers.html.

[39] Resnick SI (1987) *Extreme Values, Regular Variation and Point Processes* (Springer-Verlag, New York).

[40] Roberts GO, Rosenthal JS (1996) Quantative bounds for convergence rates of continuous time Markov processes. *Electronic J. Appl. Probab.* 1(9):1–21.

[41] Rootzén H (1988) Maxima and exceedances of stationary Markov chains. *Adv. Appl. Probab.* 20(2):371–390.

[42] Silvestrov DS, Teugels JL (1998) Limit theorems for extremes with random sample size. *Adv. Appl. Prob.* 30(3):777–806.

[43] Whitt W (1980) Continuity of generalized semi-Markov processes. *Math. Oper. Res.* 5(4):494–501.

[44] Whitt W (1981) Comparing counting processes and queues. *Adv. Appl. Prob.* 13(1):207–220.

[45] Whitt W (1984) Departures from a queue with many busy servers. *Math. Oper. Res.* 9(4):534–544.

[46] Whitt W (1991) The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Sci.* 37(3):307–314.

[47] Whitt W (2002) *Stochastic-Process Limits* (Springer, New York).

[48] Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* 50(10):1449–1461.

[49] Zachary S, Ziedins J (2002) A refinement of the Hunt-Kurtz theory of large loss networks, with an application to virtual partitioning. *Ann. Appl. Prob.* 12(1):1–22.