Stochastic Systems
arXiv: math.PR/0000000

07/10/10

A FLUID LIMIT FOR AN OVERLOADED X MODEL VIA AN AVERAGING PRINCIPLE

By Ohad Perry[*] and Ward Whitt[†]

CWI and Columbia University

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 pre-specified fixed ratio. For large systems that fixed ratio is achieved approximately. For the fluid limit, or FWLLN, 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, remain stochastically bounded, and need to be considered without spatial scaling. Asymptotically, these queue-difference processes operate in a faster time scale than the fluid-scaled processes. In the limit, due to 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 (AP) allows us to replace the driving processes with the long-run average behavior of the FTSP.

1. Introduction. In this paper we prove that the deterministic fluid approximation for the overloaded X call-center model, suggested in [38] and analyzed in [39], arises as the *many-server heavy-traffic* (MS-HT) fluid limit of a properly scaled sequence of overloaded Markovian X models under the *fixed-queue-ratio-with-thresholds* (FQR-T) control. The X model has two classes of customers and two service pools, one for each class, but with both

[*]Centrum Wiskunde & Informatica (CWI)

[†]Department of Industrial Engineering and Operations Research, Columbia University

AMS 2000 subject classifications: Primary 60F17, 60K25secondary 60G70, 90B22

Keywords and phrases: averaging principle, separation of time scales, many-server queues, heavy-traffic limits, deterministic fluid approximations, overload control

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 [37] 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 [37] 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. More specifically, in [37] 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. We further assumed that the system becomes overloaded due to a sudden, unexpected shift in the arrival rates, 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, it was shown that FQR-T outperforms the fluid-optimal static (fixed numbers of servers) allocation, even when the new arrival rates are known.

Second, in [38] we applied a heavy-traffic *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 in (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 [39] we investigated the ODE suggested in [38] using a dynamical-system approach. The dynamical-system framework could not be applied directly, since 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 an interval $[0, \delta)$ for some $\delta > 0$, and that this interval can be extended as long as the FTSP is positive recurrent, typically all the way to $+\infty$. The stationary fluid approximation, derived heuristically in [37], was shown to exist as the unique fixed point (or stationary point) for the fluid approximation. We also provided easily-verifiable conditions for the solution of the ODE to converge to this stationary point, with the convergence being exponentially fast. 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 ODE's.

1.2. *Overview.* In this fourth paper, we will prove that the solution to the ODE in [38, 39] is indeed the MS-HT fluid limit of the overloaded X model, which we also call a *functional weak law of large numbers* (WLLN); see Theorem 6.1; and see §3 for the key assumptions. In doing so, we will also prove the AP which in turn will provide a strong version of *state-space collapse* (SSC) for the two-dimensional queue process and the server-assignment processes; for the SSC results, see Theorems 4.1, 4.2, 5.6 and 7.1. In a subsequent paper [40] 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 1 and 3 in §3. As a consequence, the model is stationary but the evolution is transient. Because of customer abandonment, the stochastic models will all be stable, approaching proper steady-state distributions. We will be proving a MS-HT limit for the system processes.

Convergence to the fluid limit will be established in roughly three steps: (i) representing the sequence of systems (§4), (ii) proving that the sequence considered is $\mathcal{C}$-tight (§8.1), and (iii) uniquely characterizing the limit ([39] and much of the rest of §3-§8).

The first representation step in §4 starts out in the usual way, involving rate-1 Poisson processes and martingales, as reviewed in [36]. However, the SSC in Theorem 4.1 requires a delicate analysis of the unscaled sequence; see §7, especially Lemma 7.4.

The second tightness step in §8.1 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 [8], [13], [36] and §11.6 in [48]. As reviewed in [13] and [36], 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 which 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 [39], 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 [39], this argument may seem circular at first, since 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 MS-HT 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 [39] are needed for proving convergence, and we summarize the essential results in §5 below.

1.3. *Literature.* Our previous papers discuss related literature; see especially §1.2 of [37]. Our FQR-T control extends the FQR control suggested and studied in [16–18], but the limits there were established for a different regime under different conditions. Here we propose FQR-T for overload 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 [26] and references therein. However, there are only a few papers in the queueing literature involving averaging principles; see p. 71 of [48] for discussion. Two notable papers are [11], which considers the diffusion limit of a polling system with zero switch-over times, and [21], which considers large loss networks under a large family of controls. Reference [21] is closely related to our work since it considers the fluid limits of such loss systems, with the control-driving process moving at a faster time scale than the other processes considered. However, the proof techniques here and in [21] are very different. In particular, the AP in [21] is proved via the martingale problem, building on [28]. In contrast, here we rely heavily on stochastic bounds, e.g., see Lemmas 7.1, 8.9 and 8.10.

There is now a substantial literature on fluid limits for queueing models, some of which is reviewed in [48]. For recent work on many-server queues, see [25, 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 [5, 6], but again the specifics are quite different. Their separation of time scales justifies using a pointwise stationary approximation asymptotically, as in [47, 32].

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 MS-HT 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, arriving according to independent Poisson processes with rates λ_1 and λ_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 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 remains in effect.* 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 *unknown* (but fixed) 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). 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 [37, 38] we considered the common case in which there is no efficiency gain from service by cross-trained agents. Specifically, in [37] we assumed the *strong inefficient sharing condition*

$$\mu_{1,1} > \mu_{1,2} \quad \text{and} \quad \mu_{2,2} > \mu_{2,1}. \tag{2.1}$$

Under condition (2.1), 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 [37] hold under the weaker *basic inefficient sharing condition:* $\mu_{1,1}\mu_{2,2} \geq \mu_{1,2}\mu_{2,1}$.

If (2.1) holds then it is disadvantageous (from the standard quality-of-service perspective) for customers to be served in the other-class pool, since their service tends to be longer. Indeed, it is shown in [38] that there can be serious performance degradation, even in normal loading, if both pools are allowed to serve the other class. Without customer abandonment, the sharing can cause the system to become unstable, causing the queue lengths to diverge to infinity.

When there is no sharing (before the overload has occurred), the two separate systems can each be modeled as an Erlang-A $(M/M/m_i + M)$ model, having a Poisson arrival process with rate $\tilde{\lambda}_i$, m_i servers, exponential service times having mean $1/\mu_{i,i}$ and exponential times to abandon having mean $1/\theta_i$. Then standard performance analysis methods apply. We are concerned with the performance with sharing in face of the overload, including developing an effective control.

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. In order to motivate the need for sharing when both systems are overloaded, in [37] we considered a convex-cost framework. With that framework, in [37] we showed that sharing may be beneficial, even if it causes the total queue length (queue 1 plus queue 2) to increase.

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 routing policy, the six-dimensional stochastic process $X_6 \equiv \{(Q_i(t), Z_{i,j}(t) : i, j = 1, 2) : t \geq 0\}$ 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. For any given arrival rates, if sharing is desired, then we allow sharing in only one direction, so that $Z_{1,2}(t)Z_{2,1}(t) = 0$ for all $t \geq 0$. 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 [37], 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 $\tilde{D}_{1,2}(t) \equiv Q_1(t) - r_{1,2}Q_2(t)$ and $\tilde{D}_{2,1} \equiv r_{2,1}Q_2(t) - Q_1(t)$. As shown in [37], there is no incentive for sharing simultaneously in both directions. These ratio parameters should satisfy $r_{1,2} \geq r_{2,1}$; see Proposition EC.2 and (EC.11) of [37].

As long as $\tilde{D}_{1,2}(t) \leq k_{1,2}$ and $\tilde{D}_{2,1}(t) \leq k_{2,1}$ we consider the system to be normally loaded (i.e., not overloaded) 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 $\tilde{D}_{1,2}(t) > k_{1,2}$, then class 1 is judged to be overloaded and service-pool 2 is allowed to start helping queue 1. As soon as the first class-1 customer starts his service in pool 2, we drop the threshold $k_{1,2}$, but keep the other threshold $k_{2,1}$. Now, the sharing of customers is done as follows: If a type-2 server becomes available at time t, then it will take its next customer from the head of queue 1 if $\tilde{D}_{1,2}(t) > 0$. Otherwise, it will take its next customer from the head of queue 2. If at some time t after sharing has started queue 1 empties, or $\tilde{D}_{2,1}(t) = k_{2,1}$ then the threshold $k_{1,2}$ is reinstated. The control works similarly if class 2 is overloaded, but with pool-1 servers helping queue 2, and with the threshold $k_{2,1}$ dropped once it is crossed.

In addition, we impose the *one-way sharing* rule: at no time do we allow that $Z_{1,2}(t)Z_{2,1}(t) > 0$. That is, if at some time $t_0 \geq 0$ the threshold $k_{2,1}$ is crossed, we do not allow class-2 customers to be sent to service pool 1 if $Z_{1,2}(t_0) > 0$, and similarly in the other direction. The one-way sharing rule prevents sharing in both direction that may occur due to stochastic fluctuations in the finite stochastic systems.

It can be of interest to consider alternative variants of the FQR-T control just defined. In very large systems, the thresholds can be chosen to be large enough compared to the stochastic fluctuations, so that they are very rarely crossed under normal loads. At the same time, the thresholds can be chosen to be small enough compared to the queue size when the system becomes overloaded so that they do not affect the cost during the overload; see §2.3 and the scaling in (2.5). In such circumstances one can choose to rely on the thresholds alone to prevent unwanted two-way sharing, without applying the one-way sharing rule. We might also elect not to drop the threshold after it is crossed.

During the overload, after the sharing has begun in one specified direction and remains in effect, the six-dimensional stochastic process

$$X_6(t) \equiv (Q_i(t), Z_{i,j}(t); i,j = 1,2), \quad t \geq 0 \tag{2.2}$$

is a CTMC. 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 [37, 38]. 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 (MS-HT) Scaling.* We develop the fluid limit for the system after sharing has begun, which we assume is during an overload incident. To develop the fluid limit, we consider a sequence of X systems, $\{X_6^n : n \geq 1\}$ defined as in (2.2), indexed by n (denoted by superscript), with arrival rates and number of servers growing proportionally to n, i.e.,

$$\bar{\lambda}_i^n \equiv \frac{\lambda_i^n}{n} \to \lambda_i \quad \text{and} \quad \bar{m}_i^n \equiv \frac{m_i^n}{n} \to m_i \quad \text{as} \quad n \to \infty, \tag{2.3}$$

and the service and abandonment rates held fixed. We then define the associated fluid-scaled stochastic processes

$$\bar{Q}_i^n(t) \equiv \frac{Q_i^n(t)}{n} \quad \text{and} \quad \bar{Z}_{i,j}^n(t) \equiv \frac{Z_{i,j}^n(t)}{n}, \quad i,j = 1,2, \quad t \geq 0,$$

$$\bar{X}_6^n(t) \equiv (\bar{Q}_i^n(t), \bar{Z}_{i,j}^n(t) : i,j = 1,2), \quad t \geq 0. \tag{2.4}$$

In this framework, with additional regularity conditions, we will prove that $\bar{X}_6^n \Rightarrow x_6$ in an appropriate framework (see §2.5), where x_6 is a deterministic continuous function. We call this a FWLLN. We do not state this FWLLN until §6, because the limit x_6 is quite complicated.

We now return to the description of our systems. For each system n, there are thresholds $k_{1,2}^n$ and $k_{2,1}^n$, scaled as suggested in [37, 38]:

$$\frac{k_{i,j}^n}{n} \to 0 \quad \text{and} \quad \frac{k_{i,j}^n}{\sqrt{n}} \to \infty \quad \text{as} \quad n \to \infty, \quad i,j = 1,2. \tag{2.5}$$

The first scaling by n is chosen to make the thresholds asymptotically negligible in MS-HT fluid scaling, so they have no asymptotic impact on the steady-state cost. The second scaling by $\sqrt{n}$ is chosen to make the thresholds asymptotically infinite in MS-HT diffusion scaling, so that asymptotically the thresholds will not be exceeded under normal loading. It is significant that MS-HT scaling shows that we should be able to simultaneously satisfy both conflicting objectives in large systems.

Primarily motivated by [37], we will also consider additional variants of the model. Specifically, We intorduce shifting thresholds $\kappa_{i,j}^n$, satisfying

$$\frac{\kappa_{i,j}^n}{n} \to \kappa_{i,j} \geq 0 \quad \text{as } n \to \infty, \quad i,j = 1,2. \tag{2.6}$$

These shifting thresholds can be of order n, i.e., $\kappa_{i,j} > 0$, if a version of FQR-T, the *shifted FQR-T* control, is employed. Shifted FQR-T is designed to keep the relation between the queues at $Q_1 \approx r_{1,2}Q_2 + \kappa_{1,2}$, or $Q_1 \approx r_{2,1}Q_2 + \kappa_{2,1}$, which is the optimal relation in the stationary fluid model for the important class of separable quadratic cost functions; see EC.4 in [37].

We use the original thresholds $k_{1,2}^n$ and $k_{2,1}^n$ to activate sharing. If threshold $k_{1,2}^n$ is passed to activate sharing, then instead of simply dropping it, we replace it with the new shifting threshold $\kappa_{1,2}^n$ (and similarly in the other direction). When the shifting thresholds are of order $O(n)$, they implement shifted FQR-T, as discussed above. These shifting constants can also stand for the original thresholds $k_{i,j}^n$, $i,j = 1,2$, if we choose not to drop them once sharing is initialized (for the reasons described in §2.2 above). In that case, the scale of $\kappa^n_{i,j}$ is as in (2.5). The basic model is included by simply having $\kappa^n_{1,2} = \kappa^n_{2,1} = 0$. To summarize, we consider $\kappa^n_{i,j} = O(n)$, but without specifying their exact scale.

Sharing with pool 2 helping class 1 is allowed when first $\tilde{D}^n_{1,2} > k^n_{1,2}$, but because we use the shifting thresholds, a class-1 customer will be assigned to pool 2 only when $\tilde{D}^n_{1,2} > k^n_{1,2} \vee \kappa^n_{1,2}$. If $\kappa^n_{1,2} \to \infty$, then $\tilde{D}^n_{1,2} \to \infty$ as $n \to \infty$. Hence, we redefine the queue-difference process, hereafter referred to simply as *difference process*, by subtracting $\kappa^n_{1,2}$ from Q^n_1, i.e.,

$$D^n_{1,2}(t) \equiv (Q^n_1(t) - \kappa^n_{1,2}) - r_{1,2}Q^n_2(t), \quad t \ge 0. \tag{2.7}$$

We now apply FQR using the process $D^n_{1,2}$ in (2.7): if $D^n_{1,2}(t) > 0$, then every newly available agent (in either pool) takes his new customer from the head of the class-1 queue. If $D^n(t) \le 0$, then every newly available agent takes his new customer from the head of his own queue. As before, the sharing is terminated altogether at time t if either $Q^n_1(t) = 0$ or if $\tilde{D}^n_{2,1}(t) = k^n_{2,1}$. (Note that we use the original queue-difference process $\tilde{D}^n_{2,1}$ and threshold $k^n_{2,1}$.) Because of the one-way sharing rule, sharing in the opposite direction, with pool 1 helping class 2, can start only after $Z^n_{1,2} = 0$. Once the sharing has been terminated, the startup procedure for sharing is as specified above, essentially the same in each direction.

2.4. *The MS-HT ED Regime.* For a Markovian I model, having one service pool, one customer class and customer abandonment, i.e., the $M/M/m+M$ model (also called the Erlang-A model), three different MSHT limiting regimes were identified in [14]: If the system is asymptotically overloaded, then it is called the *efficiency-driven* (ED) limiting regime; if the system is asymptotically critically loaded, then it is called the *quality-and-efficiency-driven* (QED) limiting regime; if the system is asymptotically underloaded, then it is called the *quality-driven* (QD) limiting regime. These same cases without abandonment had been specified by [19]. For one class and one pool, it is natural to let n be the total number of servers ($m_n = n$ for all n). Then the regimes are determined by the limit $(1-\rho^n)\sqrt{n} \to \beta$ as $n \to \infty$, where $\rho^n \equiv \lambda^n/n\mu$ is the traffic intensity in model n. The regimes (i) ED, (ii) QED, and (iii) QD then occur, respectively, if the limit holds with (i) $\beta = -\infty$, (ii) $-\infty < \beta < \infty$, and (iii) $\beta = +\infty$.

Let

$$\rho^n_i \equiv \frac{\lambda^n_i}{\mu_{i,i} m^n_i}, \quad \text{and} \quad \rho_i \equiv \lim_{n\to\infty} \rho^n_i = \frac{\lambda_i}{\mu_{i,i} m_i}, \quad i = 1, 2. \tag{2.8}$$

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:

$$q_i^a \equiv \frac{(\lambda_i - \mu_{i,i} m_i)^+}{\theta_i} \quad \text{and} \quad s_i^a \equiv \left(m_i - \frac{\lambda_i}{\mu_{i,i}} \right)^+, \quad i = 1, 2, \tag{2.9}$$

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 (2.9) see Theorem 2.3 in [49], especially equation (2.19), and §5.1 in [37]. See also §6 in [39].

2.5. *Conventions About Notation.* We use the usual $\mathbb{R}$, $\mathbb{Z}$ and $\mathbb{N}$ 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 $\mathcal{D}_k(I)$. Since we will be considering continuous limits, the topology is equivalent to uniform convergence on compact subintervals of I. 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., $e(t) \equiv t$, $t \in I$. The function $0 \in \mathcal{D}$ will be denoted simply by 0, when the context is clear, or by $0e$. Let $\Rightarrow$ denote convergence in distribution.

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

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

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

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

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

Let $Y \equiv \{Y(t) : t \ge 0\}$ 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* (SB), i.e., if

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

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 $\|Y\|_t/f(t) \Rightarrow 0 \quad \text{as} \quad t \to \infty$. If $f(t) \equiv 1$, then $Y = O_P(1)$ if it is SB, 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 $\mathbb{N}$, i.e., $f : \mathbb{N} \to [0, \infty)$.

For a sequence $\{Y^n : n \geq 1\}$ (of stochastic processes, random variables or real numbers) we denote its fluid-scaled version by $\bar{Y}^n \equiv Y^n/n$. The fluid limit of stochastic processes $\bar{Y}^n$ is denoted by $\bar{Y}$. The diffusion-scaled sequence of stochastic processes, centered about their fluid limit, is denoted by $\hat{Y}^n \equiv (Y^n - n\bar{Y})/\sqrt{n}$, and its limit by $\hat{Y}$. We let $\breve{Y}^n \equiv Y^n/\sqrt{n}$ be the $\sqrt{n}$-scaled processes without the centering about the fluid limit.

3. The Main Assumptions. We now specify the three main assumptions: Assumptions 1, 2 and 3 below. *These three assumptions are assumed to hold throughout the paper.*

First, we have the three assumptions already made, (2.3), (2.5) and (2.6). (Here we do not require (2.1).) Our first new assumption is on the asymptotic behavior of the rates; it specifies the essential form of the overload. For the statement, recall the definitions in (2.3), (2.6) and (2.9), which describe the asymptotic behavior of the rates.

Assumption 1. (*system overload, with class* 1 *more overloaded*)

The rates in the overload are such that the limiting rates satisfy

(1) $\theta_1(q_1^a - \kappa_{1,2}) > \mu_{1,2}s_2^a$.
(2) $q_1^a - \kappa_{1,2} > rq_2^a$.

Condition (1) in Assumption 1 ensures that class 1 is asymptotically overloaded, even after receiving help from pool 2. To see why, first observe that, since $s_2^a \geq 0$, $q_1^a > \kappa_{1,2} \geq 0$, 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 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 §7. Note that Condition (1) in Assumption 1 is slightly stronger than Condition (I) of Assumption A in [39]. because here there is a strong inequality instead of a weak inequality.

Condition (2) in Assumption 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 $Z^n_{2,1}$ should remain at 0.

We now expand upon the centering constants. Given Assumption 1, we can simplify the notation, dropping the subscripts from $\kappa^n_{1,2}$ and $\kappa_{1,2}$.

Assumption 2. (*centering constants*)

For the sequence $\{\kappa^n : n \in \mathbb{N}\}$ of centering constants, we require that

(1) $\kappa^n \geq 0$ *for all* n *and* $\kappa^n/n \to \kappa$, *where* $0 \leq \kappa < \infty$.
(2) *If* $\kappa = 0$, *then in addition we require that* $\kappa^n \to c_1$ *and* $\kappa^n/\log n \to c_2$ *as* $n \to \infty$, *where* $0 \leq c_i \leq \infty$ *for* $i = 1, 2$.

In Assumption 2, the first condition is the standard scaling for the centering constants. If $\kappa = 0$, then we have FQR after sharing has been activated by passing the thresholds; if $\kappa \neq 0$, then we have shifted FQR after sharing has been activated by passing the thresholds. From the perspective of the centering constants alone, it would suffice to consider $\kappa^n = n\kappa$. However, we have imposed additional conditions for the case $\kappa = 0$. We did this so that we could consider the FQR-T control with the original thresholds retained. As discussed in §2.3, we want those thresholds to be $o(n)$ but large compared to $O(\sqrt{n})$; e.g., we might have $\kappa^n = n^p$ for $1/2 < p < 1$. The regularity conditions involving scaling by $\log n$ is for results in §7 showing that the idleness is at most $O(\log n)$.

Our third assumption is about the initial conditions. For the initial conditions, we assume that the overload, whose asymptotic character is specified by Assumption 1, has begun some time in the past and is ongoing. In addition, sharing with pool 2 allowed to help class 1 has been activated by having the threshold $k^n_{1,2}$ exceeded by the queue difference process $\tilde{D}^n_{1,2}$ and is in process. Thus actual sharing is being controlled by the difference process $D^n_{1,2}$ in (2.7). The initial time 0 might be the time that $D^n_{1,2}$ first becomes strictly positive, and a class 1 customer is sent to pool 2, but we allow more general initial conditions.

We require that a fluid-scale limit exists at time 0, where the limit $x(0)$ satisfies the initial conditions required for the existence of a unique solution to the ODE, established in [39]. The ODE and the FSTP will be reviewed here in §5. Specifically, Assumption 3 refers to the set $\mathbb{A}$ defined in (5.16) and expressed in (5.22). We will be explaining Assumption 3 in the next two sections. For the statement, recall the definition of the six-dimensional fluid-scaled process $\bar{X}_6^n$ in (2.4) and let $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ be the associated three-dimensional process. (In §4 we show that it suffices to consider $\bar{X}^n$.) We also need to separately specify initial conditions for the queue-difference processes in (2.7).

Let

$$B^n \equiv \{Z_{1,1}^n(0) = m_1^n, \quad Z_{2,1}^n(0) = 0, \quad Z_{1,2}^n(0) + Z_{2,2}^n(0) = m_2^n\}. \tag{3.1}$$

Assumption 3. (*initial conditions*) *As* $n \to \infty$,

$$P(B^n) \to 1, \quad D_{1,2}^n(0) \Rightarrow L \quad \textit{and} \quad \bar{X}^n(0) \Rightarrow x(0) \in \mathbb{A},$$

where B^n *is defined in (3.1),* $D_{1,2}^n$ *is defined in (2.7),* L *is a proper random variable and* $x(0)$ *is a deterministic element of* $\mathbb{R}_3$.

If the system is initialized not in $\mathbb{A}$, then other fluid models hold during the initial period before $\mathbb{A}$ is hit; See §8 in [39]. In this paper we concentrate on time intervals on which the averaging principle is operating.

4. Representation of X_6^n. The statements of our asymptotic results are easier to understand if we first exhibit the representation of X_6^n that we will use in our proof.

4.1. *Starting with Rate-*1 *Poisson Processes.* Let $A_i^n(t)$ count the number of class-i customer arrivals, let $S_{i,j}^n(t)$ count the number of service completions of class-i customers by agents in pool j, and let $U_i^n(t)$ count the number of class-i customers to abandon from queue, all in model n during the time interval $[0,t]$. the fundamental evolution equations for the queue lengths are:

$$Q_i^n(t) = A_i^n(t) - S_{i,1}^n(t) - S_{i,2}^n(t) - U_i^n(t), \quad t \geq 0, \quad i = 1, 2, \tag{4.1}$$

where the processes $S_{i,j}^n(t)$ depend on the service processes $Z_{i,j}^n(t)$ and the routing rule.

Following common practice, as reviewed in §2 of [36], 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 as

$$\begin{aligned} A_i^n(t) &\equiv N_i^a(\lambda_i^n t), \\ S_{i,j}^n(t) &\equiv N_{i,j}^s\left(\mu_{i,j}\int_0^t Z_{i,j}^n(s)\,ds\right), \\ U_i^n(t) &\equiv N_i^u\left(\theta_i\int_0^t Q_i^n(s)\,ds\right), \quad t \geq 0, \end{aligned} \tag{4.2}$$

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.

Within this framework, the evolution of the four service processes $Z_{i,j}^n$, and thus X^n in (2.4), depends on the routing and the state of the service processes. Suppose that one-way sharing has been activated with pool 2 allowed to help class 1. Thus we have previously had $D_{1,2}^n(t) > 0$ and $Z_{2,1}^n(t) = 0$. Since then, we assume that the other retained threshold $k_{2,1}^n$ has not been crossed, so that $Z_{2,1}^n$ has remained at 0. At the present (later) time, we need to know whether or not $D_{1,2}^n(t) > 0$. If $D_{1,2}^n(t) > 0$, then each newly available server should take a customer from the head of queue 1. However, if $D_{1,2}^n(t) \leq 0$, then a pool-2 server will only take a customer from class 2.

4.2. *Simplification via SSC.* However, since the system is assumed to be overloaded, it is reasonable to expect that the idleness processes in the two service pools are asymptotically negligible in diffusion (and thus in fluid) scale. That means that $Z_{1,1}^n(t) + Z_{2,1}^n(t) \approx m_2^n$ and $Z_{2,2}^n(t) + Z_{1,2}^n(t) \approx m_2^n$ for all $t > 0$, provided that those approximations hold at $t = 0$. Also, since we assume that class 1 is more overloaded than class 2, it is reasonable to expect that $Z_{1,2}^n$ becomes positive before the threshold $k_{2,1}^n$ is crossed (for large n), so that $Z_{2,1}^n(t) = 0$, at least on some initial interval $[0, \tau]$, $\tau > 0$. If that is true, then $Z_{1,1}^n(t) \approx m_1^n$ and $Z_{2,2}^n(t) \approx m_2^n - Z_{1,2}^n(t)$, $t \in [0, \tau]$. The approximation signs will be replaced with equality with both diffusion and fluid scaling, producing a SSC result. Specifically, the dimension of the service process reduces from four to one in the limit with diffusion scaling. That will be proved in Theorem 7.1 below.

We now state a result which will allow us to represent the system in a relatively simple form, building on the SSC for the service process just explained (and which will be proved in §7). Recall that X_6^n has been defined in §2.3, the assumptions in §3 are in force, d_{J_1} denotes the standard Skorohod J_1 metric and $\breve{Y}^n \equiv Y^n/\sqrt{n}$ for any $Y^n \in \mathcal{D}_k$.

THEOREM 4.1. (*Representation via SSC*) *As* $n \to \infty$, $d_{J_1}(\breve{X}_6^n, \breve{X}^{n,*}) \Rightarrow 0$ *in* $\mathcal{D}_6$, *where* $X^{n,*} \equiv X_6^n \equiv (Q_1^n, Q_2^n, Z_{1,1}^n, Z_{2,1}^n, Z_{1,2}^n, Z_{2,2}^n)$ *under the extra condition that* $Z_{1,1}^n = m_1^n$, $Z_{2,1}^n = 0$ *and* $Z_{1,2}^n + Z_{2,2}^n = m_2^n$, *with* $X^n \equiv$

$(Q_1^n, Q_2^n, Z_{1,2}^n)$ *being represented via*

$$
\begin{aligned}
Z_{1,2}^n(t) &\equiv Z_{1,2}^n(0) + \int_0^t 1_{\{D_{1,2}^n(s)>0\}}\, dS_{2,2}^n(t) - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{1,2}^n(t) \\
&= Z_{1,2}^n(0) + N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(m_2^n - Z_{1,2}^n)(s)\, ds\right) \\
&\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}} Z_{1,2}^n(s)\, ds\right), \quad t \ge 0,
\end{aligned} \tag{4.3}
$$

$$
\begin{aligned}
Q_1^n(t) &\equiv Q_1^n(0) + A_1^n(t) - \int_0^t 1_{\{D_{1,2}^n(s)>0\}}\, dS^n(t) \\
&\quad - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{1,1}^n(t) - U_1^n(t) \\
&= Q_1^n(0) + N_1^a(\lambda_1^n t) - N_{1,1}^s(\mu_{1,1} Z_{1,1}^n t) \\
&\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}} Z_{1,2}^n(s))\, ds\right) \\
&\quad - N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)>0\}}(m_2^n - Z_{1,2}^n(s))\, ds\right) \\
&\quad - N_1^u\left(\theta_1\int_0^t Q_1^n(s)\, ds\right), \quad t \ge 0,
\end{aligned} \tag{4.4}
$$

$$
\begin{aligned}
Q_2^n(t) &\equiv Q_2^n(0) + A_2^n(t) - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{2,2}^n(t) \\
&\quad - \int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}\, dS_{1,2}^n(t) - U_2^n(t) \quad t \ge 0 \\
&= Q_2^n(0) + N_2^a(\lambda_2^n t) \\
&\quad - N_{2,2}^s\left(\mu_{2,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}}(m_2^n - Z_{1,2}^n(s))\, ds\right) \\
&\quad - N_{1,2}^s\left(\mu_{1,2}\int_0^t 1_{\{D_{1,2}^n(s)\le 0\}} Z_{1,2}^n(s)\, ds\right) \\
&\quad - N_2^u\left(\theta_2\int_0^t Q_2^n(s)\, ds\right), \quad t \ge 0.
\end{aligned} \tag{4.5}
$$

With a slight abuse of notation, henceforth we use $X^n \equiv (Q_1^n, Q_2^n, Z_{1,2}^n)$ to refer to both its direct representation in $\mathcal{D}_3$ and (by virtue of Theorem 4.1) the essentially three-dimensional process $X^{n,*}$ in $\mathcal{D}_6$.

Theorem 4.1 is achieved as a corollary of Theorem 7.1, which will be stated in §7. Without it, we could not write the representation (4.3)-(4.5). In fact, if we do not know that $Z_{2,1}^n$ is asymptotically negligible, then the evolution of X_6^n becomes intractable. Specifically, the system may oscillate between different directions of sharing, with $Z^n_{1,2}$ being positive at some instances, and $Z^n_{2,1}$ being positive at other instances. The system may also get "stuck" with $Z^n_{2,1}(t) > 0$ and $Z^n_{1,2}(t) = 0$ for all $t > t_0$, for some $t_0 > 0$, even though we want to have sharing in the other direction. (See Lemma 7.2 below. If at some $t_0 \geq 0$ we have that $z_{2,1}(t_0) > 0$ then $z_{2,1}(t) > 0$ for all $t > t_0$, where $z_{2,1}$ is the fluid limit of $\bar{Z}^n_{2,1}$.) These situations are ruled out by Theorems 7.1 and 4.1.

4.3. *Simplification via Martingales.* We now obtain further simplification using the familiar martingale representation, again see [36]. Consider the representation of X^n in (4.3) - (4.5) above, and let

$$\begin{aligned} M_i^{n,a}(t) &\equiv N_i^a(\lambda_i^n t) - \lambda_i^n t, \\ M_i^{n,u}(t) &\equiv N_i^u\left(\theta_i \int_0^t Q_i^n(s))\, ds\right) - \theta_i \int_0^t Q_i^n(s)\, ds, \\ M_{i,2}^{n,s}(t) &\equiv N_{i,2}^s(J_{i,2}^n(t)) - J_{i,2}^n(t), \end{aligned} \tag{4.6}$$

where $J^n_{i,2}(t)$ are the compensators of the Poisson-processes $N^s_{i,2}(t)$ in (4.3)-(4.5), $i = 1, 2$, e.g.,

$$J^n_{1,2}(t) \equiv \mu_{1,2} \int_0^t 1_{\{D^n_{1,2}(s)<0\}} Z^n_{1,2}(s)\, ds.$$

The quantities in (4.6) can be shown to be martingales (with respect to an appropriate filtration); See [36]. However, we will not use any martingale property, and call those terms martingales for convenience.

The following lemma follows easily from the FSLLN for Poisson processes and the $\mathcal{C}$-tightness to be established in Lemma 8.1:

Lemma 4.1. *(fluid limit for the martingale terms) As $n \to \infty$,*

$$n^{-1}(M_1^{n,a}, M_2^{n,a}, M_1^{n,u}, M_2^{n,u}, M_{1,2}^{n,s}, M_{2,2}^{n,s}) \Rightarrow 0 \quad \textit{in } \mathcal{D}_6.$$

Proof. By Lemma 8.1, the sequence $\{\bar{X}^n_6 : n \geq 1\}$ is tight in $\mathcal{D}$. Thus any subsequence has a convergent subsequence. By the proof of Lemma 8.1, the sequences $\{J^n_{i,j}/n\}$ are also $\mathcal{C}$-tight, so that $\{J^n_{i,j}/n\}$, $i = 1, 2$, all converge along a converging subsequence as well. Consider a converging subsequence $\{X^n\}$ and its limit $\bar{X}$, which is continuous by Lemma 8.1. 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 [48]. In this case, the limit of each fluid-scaled martingale is the zero function $0e \in \mathcal{D}$, regardless of the converging subsequence we consider, and is thus unique. Hence we have completed the proof. ∎

We can thus obtain an alternative martingale representation for $\bar{X}^n$. In particular, we can let

$$\bar{M}^n \equiv \bar{X}^n - \bar{C}^n, \tag{4.7}$$

where $\bar{X}^n$ is defined in (4.3)-(4.5) and, with an abuse of notation, $\bar{C}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ for

$$\begin{aligned} \bar{Z}_{1,2}^n(t) &\equiv \bar{Z}_{1,2}^n(0) + \mu_{2,2} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} (\bar{m}_2^n - \bar{Z}_{1,2}^n(s))\, ds \\ &\quad - \mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s) \le 0\}} \bar{Z}_{1,2}^n(s)\, ds, \\ \bar{Q}_1^n(t) &\equiv \bar{Q}_1^n(0) + \bar{\lambda}_1^n t - \bar{m}_1^n t - \mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} \bar{Z}_{1,2}^n(s))\, ds \\ &\quad - \mu_{2,2} \int_0^t 1_{\{D_{1,2}^n(s) > 0\}} (\bar{m}_2^n - \bar{Z}_{1,2}^n(s))\, ds - \theta_1 \int_0^t \bar{Q}_1^n(s)\, ds, \\ \bar{Q}_2^n(t) &\equiv \bar{Q}_2^n(0) + \bar{\lambda}_2^n t - \mu_{2,2} \int_0^t 1_{\{D_{1,2}^n(s) \le 0\}} (\bar{m}_2^n - \bar{Z}_{1,2}^n(s))\, ds \\ &\quad - \mu_{1,2} \int_0^t 1_{\{D_{1,2}^n(s) \le 0\}} \bar{Z}_{1,2}^n(s)\, ds - \theta_2 \int_0^t \bar{Q}_2^n(s))\, ds. \end{aligned} \tag{4.8}$$

(We have used the same notation $(\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ in the definition of the different procsses $\bar{X}^n$ in (4.3)-(4.5) and $\bar{C}^n$ in (4.8) above. The following result shows that this anomaly causes no problem. Recall that d_{J_1} denotes the standard J_1 metric.

Theorem 4.2. *As $n \to \infty$, $\bar{M}^n \Rightarrow 0$, so that $d_{J_1}(\bar{X}^n, \bar{C}^n) \Rightarrow 0$ in $\mathcal{D}_3$ as $n \to \infty$, where $\bar{X}^n$ is defined in (4.3)-(4.5) and $\bar{C}^n$ is defined in (4.8).*

Proof. Since the weak limit of the centered fluid-scaled Poisson processes in (4.6) is the (continuous) 0 function, the sum of any two or more of those processes also converges to $0 \equiv 0e$ in $\mathcal{D}$, by the continuity of addition at continuous limits, and is therefore $o_P(1)$. Hence we get $\bar{M}^n \Rightarrow 0$ as $n \to \infty$ directly from Lemma 4.1, from which the remaining convergence follows directly. ∎

As a consequence of Theorem 4.2, henceforth we can focus on $\bar{C}^n$ in (4.8) instead of $\bar{X}^n$ in (4.3)-(4.5). We will do so, but redefining $\bar{X}^n$: We let $\bar{X}^n \equiv \bar{C}^n$; i.e., henceforth we let $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ in (4.8).

Theorem 4.2 reduces the expression of $\bar{X}^n$ to the random rates of the Poisson processes, and reveals the basic structure of the limiting ODE in

(5.13). Due to Theorem 4.1, the representation in (4.8) is equivalent to the representation of the six-dimensional process $\bar{X}_6^n$, for X_6^n in (2.4). Hence, proving that $\bar{X}^n$ converges to a unique deterministic limit, will imply the convergence of $\bar{X}_6^n$ to a limit in a three-dimensional hyperplane of $\mathcal{D}_6$, which is homeomorphic to $\mathcal{D}_3$. It is thus enough to work with the three-dimensional process in (4.8). Given Theorems 4.1 and 4.2, we will show that

$$\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n) \Rightarrow x \equiv (q_1, q_2, z_{1,2}) \quad \text{in} \quad \mathcal{D}_3([0,\delta]) \quad \text{as} \quad n \to \infty$$

for some $\delta > 0$, where x is a deterministic element of $\mathcal{C}_3$, with $x(t) \in \mathbb{A}$ for all $t \in [0, \delta]$.

5. The FTSP and the ODE. Even though Theorems 4.1 and 4.2 allow us to consider only the three-dimensional process X^n in (4.8), we still must cope with the indicator functions in the integrands in (4.8), which appear because of the FQR routing. Thus, the key to a successful analysis of X^n is understanding the behavior of the stochastic queue-difference process $D_{1,2}^n \equiv (Q_1^n - \kappa^n) - r_{1,2}Q_2^n$ in (2.7) when some, but not all, type-2 servers are helping class-1 customers, and the system is overloaded in the sense of Assumption 1.

In [39] we presented and analyzed a three dimensional ODE (which we refer to simply as "the ODE" since it is the only ODE under consideration). This ODE was conjectured to arise as the limit of the fluid-scaled version of X^n in (4.3)-(4.5). In this paper we will prove that conjecture. Specifically, we will show that $\bar{X}^n$ indeed converges weakly to the solution of that three-dimensional ODE, so that the fluid limit of $\bar{X}^n$ and the solution to the ODE coincide. However, *the ODE is well defined and its solution exists as an element of* $\mathcal{C}_3$, regardless of any convergence results.

Since an understanding of the ODE, its state space and its solution is required in order to characterize the fluid limit, we begin by defining the ODE (motivated by the sequence $\bar{X}^n$). In doing so, we will be reviewing [39]; see [39] for a complete analysis of the ODE. Recall that the ODE is driven by a stochastic process, whose local steady-state distributions govern the evolution of the solution to the ODE. We thus start by defining the driving process, which we call the FTSP. To understand the FTSP, we need to better understand the queue-difference process.

5.1. *The Drift Rates of the Queue-Difference Processes.* In this subsection we specify the transition rates of the queue-difference process $\{D_{1,2}^n(t) : t \geq 0\}$ in (2.7) at any time t_0 conditional on $X^n(t_0) = \Gamma^n$, where sharing is taking place; i.e., we consider the transition rates of the process

$$D^n \equiv D^n(\Gamma^n) \equiv \{D^n(\Gamma^n, t) : t \geq t_0\} \equiv \{D_{1,2}^n(X^n(t_0), t) : t \geq t_0\} \tag{5.1}$$

at time t_0 conditional on $X^n(t_0) = \Gamma^n$, where Γ^n is a deterministic state, under the assumption that sharing is taking place. (We will explain when sharing will be taking place in the following subsections.) The initial difference at time t_0 is $D^n_{1,2}(X^n(t_0), t_0) = Q^n_1(t_0) - \kappa^n - r_{1,2}Q^n_2(t_0)$, where $(Q^n_1(t_0), Q^n_2(t_0))$ is part of $X^n_6(t_0)$. To be well defined, the state Γ^n should be for the full CTMC X^n_6. The transition rates are independent of time t_0 for any given process state Γ^n. However, because of §4, it suffices to focus on the three-dimensional process $\bar{X}^n$. In other words, we can think of Γ^n as a state of X^n, i.e., a vector in $\mathbb{N}^2 \times [0, m^n_2]$. Thus the transition rates in (5.2)-(5.5) below, under this simplifying assumption, are asymptotically correct with $o(n)$ terms as $n \to \infty$ (which we omit).

To simplify analysis, we will work with an integer state space. Thus we assume that the shifting thresholds $\kappa^n_{1,2}$ in (2.7) are integers and that $r_{1,2}$ is rational, in particular, $r_{1,2} = j/k$ for positive integers j and k. We then look at queue differences measured in units of $1/k$. Hence, we have transitions of $\pm j$ and $\pm k$ instead of the original values of ± 1 and $\pm r$.

When $D^n(\Gamma^n, t_0) = m \le 0$, let the transition rates be $\lambda^{(j)}_-(n, m, \Gamma^n)$, $\lambda^{(k)}_-(n, m, \Gamma^n)$, $\mu^{(j)}_-(n, m, \Gamma^n)$ and $\mu^{(k)}_-(n, m, \Gamma^n)$ for transitions of $+j$, $+k$, $-j$ and $-k$, respectively. When $D^n(\Gamma^n, t_0) = m > 0$, let the transition rates be $\lambda^{(j)}_+(n, m, \Gamma^n)$, $\lambda^{(k)}_+(n, m, \Gamma^n)$, $\mu^{(j)}_+(n, m, \Gamma^n)$ and $\mu^{(k)}_+(n, m, \Gamma^n)$ for transitions of $+j$, $+k$, $-j$ and $-k$, respectively.

First, for $D^n(\Gamma^n, t_0) = m \le 0$ with $\Gamma^n \equiv (Q^n_1, Q^n_2, Z^n_{1,2})$, the upward rates are

$$
\begin{aligned}
&\lambda^{(k)}_-(n, m, \Gamma^n) \equiv \lambda^n_1, \quad \text{and} \\
&\lambda^{(j)}_-(n, m, \Gamma^n) \equiv \mu_{1,2}Z^n_{1,2} + \mu_{2,2}(m^n_2 - Z^n_{1,2}) + \theta_2 Q^n_2,
\end{aligned}
\tag{5.2}
$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 queue, caused by a type-2 agent service completion (of either customer type) or by a class-2 customer abandonment. Similarly, the downward rates are

$$
\mu^{(k)}_-(n, m, \Gamma^n) \equiv \mu_{1,1}m^n_1 + \theta_1 Q^n_1 \quad \text{and} \quad \mu^{(j)}_-(n, m, \Gamma^n) \equiv \lambda^n_2,
\tag{5.3}
$$

corresponding, first, to a departure from the class-1 customer queue, caused by a class-1 agent service completion or by a class-1 customer abandonment, and, second, to a class-2 arrival.

Next, for $D^n(\Gamma^n, t_0) = m \in (0, \infty)$, we have upward rates

$$
\lambda^{(k)}_+(n, m, \Gamma^n) \equiv \lambda^n_1 \quad \text{and} \quad \lambda^{(j)}_+(n, m, \Gamma^n) \equiv \theta_2 Q^n_2,
\tag{5.4}
$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 customer queue caused by a class-2 customer abandonment. The downward rates are

$$\begin{aligned} \mu_+^{(k)}(n,m,\Gamma^n) &\equiv \mu_{1,1}m_1^n + \mu_{1,2}Z_{1,2}^n + \mu_{2,2}(m_2^n - Z_{1,2}^n) + \theta_1 Q_1^n \quad \text{and} \\ \mu_+^{(j)}(n,m,\Gamma^n) &\equiv \lambda_2^n, \end{aligned} \tag{5.5}$$

corresponding, first, to a departure from the class-1 customer queue, caused by (i) a type-1 agent service completion, (ii) a type-2 agent service completion (of either customer type), or (iii) by a class-1 customer abandonment and, second, to a class-2 arrival.

Using these transition rates, we can define the *drift rates* for $D^n(X^n(t),t) \equiv D^n(\Gamma^n,t)$, conditional upon $X^n(t) = \Gamma^n$. Let these drift rates in the regions $(0,\infty)$ and $(-\infty,0]$ be denoted by $\delta_+^n(X^n(t))$ and $\delta_-^n(X^n(t))$, respectively, Combining (5.20) and (5.2)-(5.5), we obtain

$$\begin{aligned} \delta_+^n(X^n(t)) \equiv\; & j[\lambda_1^n - \mu_{1,1}m_1^n + (\mu_{2,2}-\mu_{1,2})Z_{1,2}^n(t) - \mu_{2,2}m_2^n(t) - \theta_1 Q_1^n(t)] \\ & - k[\lambda_2^n - \theta_2 Q_2^n(t)], \\ \delta_-^n(X^n(t)) \equiv\; & j[\lambda_1^n - \mu_{1,1}m_1^n - \theta_1 Q_1^n(t)] \\ & - k[\lambda_2^n + (\mu_{2,2}-\mu_{1,2})Z_{1,2}^n(t) - \mu_{2,2}m_2^n - \theta_2 Q_2^n(t)]. \end{aligned} \tag{5.6}$$

In order to have sharing, we will want to have $\delta_+^n(\Gamma^n) < 0 < \delta_-^n(\Gamma^n)$.

5.2. *The FSTP.* 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

$$D_e^n(\Gamma^n,s) \equiv D_{1,2}^n(t_0 + s/n), \quad s \geq 0. \tag{5.7}$$

where 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 sharing will occur (or will occur eventually for n large enough). Since we divide s in (5.7) by n, we are effectively dividing the rates by n. 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 HT limits. See [46] for a previous limit using time expansion. We will explain the limit in detail in §5.6 below.

With that in mind, we see that the FTSP should have the same state space as $D_{1,2}^n$. When we relate the FTSP to the expanded queue-difference process in §5.6 below, we will also relate the initial differences, which so far are unspecified here. Since we already converted to an integer state space, the FTSP will be a continuous-time Markov chain (CTMC) on $\mathbb{Z}$. With that convention, the FTSP $\{D(\gamma, s) : s \geq 0\}$ has transition rates among the integers determined at any time s (in the newly introduced "infinitesimal" time scale) by both its current state $D(\gamma, s) \equiv m$ and the vector γ. The vector γ is a possible state of the fluid model $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$ at some time t, where averaging may take place. Thus $\gamma \in [0, \infty)^2 \times [0, m_2]$. Specifically, γ can be any vector in the subset $\mathbb{A}$ defined in (5.16) below.

Given the current state m, we let the rates of the FTSP D as a function of γ be the limit of the rates of $D^n(\Gamma^n, \cdot)$ divided by n, where the rates of $D^n(\Gamma^n, \cdot)$ are themselves a function of the current state $D^n(\Gamma^n, 0) = m$ with $\Gamma^n/n \to \gamma$ as $n \to \infty$. Since $\Gamma^n/n \to \gamma$ as $n \to \infty$, there will be sharing in all systems for all n sufficiently large. (For the corresponding rates of the queue-difference process $D^n(\Gamma^n, \cdot)$ itself, see (5.2)-(5.5).)

Since the drift rates of $D^n(\Gamma^n, t)$ in (5.6) are linear functions of the state $X^n(t)$, we have

$$\delta_+^n(X^n(t)) \Rightarrow \delta_+(\bar{X}(t)) \quad \text{and} \quad \delta_-^n(X^n(t)) \Rightarrow \delta_-(\bar{X}(t)) \tag{5.8}$$

whenever $\bar{X}^n(t) \Rightarrow \bar{X}(t)$ in $\mathbb{R}$, which we will have (for all t along a convergent subsequence, because along that subsequence we have $\bar{X}^n \Rightarrow \bar{X}$ in $\mathcal{D}_3$ as a consequence of tightness).

Directly, we let the FTSP $\{D(\gamma, s) : s \geq 0\}$ be a CTMC with transition rates $\lambda_-^{(j)}(m, \gamma)$, $\lambda_-^{(k)}(m, \gamma)$, $\mu_-^{(j)}(m, \gamma)$ and $\mu_-^{(k)}(m, \gamma)$ for transitions of $+j$, $+k$, $-j$ and $-k$, respectively, when $D(\gamma, s) = m \leq 0$. Similarly, let the transition rates be $\lambda_+^{(j)}(m, \gamma)$, $\lambda_+^{(k)}(m, \gamma)$, $\mu_+^{(j)}(m, \gamma)$ and $\mu_+^{(k)}(m, \gamma)$ for transitions of $+j$, $+k$, $-j$ and $-k$, respectively, when $D(\gamma, s) = m > 0$.

Paralleling the definitions in (5.2)-(5.5), we define the transition rates for $D(\gamma)$ as follows: First, for $D(\gamma, s) = m \in (-\infty, 0]$ with $\gamma \equiv (q_1, q_2, z_{1,2})$, the upward rates are

$$\begin{aligned} &\lambda_-^{(k)}(m, \gamma) \equiv \lambda_1, \quad \text{and} \\ &\lambda_-^{(j)}(m, \gamma) \equiv \mu_{1,2} z_{1,2} + \mu_{2,2}(m_2 - z_{1,2}) + \theta_2 q_2. \end{aligned} \tag{5.9}$$

Similarly, the downward rates are

$$\mu_-^{(k)}(m, \gamma) \equiv \mu_{1,1} m_1 + \theta_1 q_1 \quad \text{and} \quad \mu_-^{(j)}(m, \gamma) \equiv \lambda_2 \tag{5.10}$$

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

$$\lambda_+^{(k)}(m, \gamma) \equiv \lambda_1 \quad \text{and} \quad \lambda_+^{(j)}(m, \gamma) \equiv \theta_2 q_2. \tag{5.11}$$

The downward rates are

$$\begin{aligned}\mu_+^{(k)}(m,\gamma) &\equiv \mu_{1,1}m_1 + \mu_{1,2}z_{1,2} + \mu_{2,2}(m_2 - z_{1,2}) + \theta_1 q_1 \quad \text{and}\\ \mu_+^{(j)}(m,\gamma) &\equiv \lambda_2.\end{aligned} \tag{5.12}$$

5.3. *The ODE.* We can now present the three-dimensional ODE in terms of the FTSP D. 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

$$\begin{aligned}\dot{q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(x(t))\,[z_{1,2}(t)\mu_{1,2} + z_{2,2}(t)\mu_{2,2}] - \theta_1 q_1(t)\\ \dot{q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(x(t)))\,[z_{2,2}(t)\mu_{2,2} + z_{1,2}(t)\mu_{1,2}] - \theta_2 q_2(t)\\ \dot{z}_{1,2}(t) &\equiv \pi_{1,2}(x(t))z_{2,2}(t)\mu_{2,2} - (1 - \pi_{1,2}(x(t)))z_{1,2}(t)\mu_{1,2},\end{aligned} \tag{5.13}$$

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 $\gamma = x(t)$.

Equivalently, we have the following integral representation of the ODE in (5.13):

$$\begin{aligned}z_{1,2}(t) &\equiv z_{1,2}(0) + \mu_{2,2}\int_0^t \pi_{1,2}(x(s))(m_2 - z_{1,2}(s))\,ds\\ &\qquad - \mu_{1,2}\int_0^t (1 - \pi_{1,2}(x(s)))z_{1,2}(s)\,ds,\\ q_1(t) &\equiv q_1(0) + \lambda_1 t - m_1 t - \mu_{1,2}\int_0^t \pi_{1,2}(x(s))z_{1,2}(s))\,ds\\ &\qquad - \mu_{2,2}\int_0^t \pi_{1,2}(x(s))(m_2 - z_{1,2}(s))\,ds - \theta_1\int_0^t q_1(s)\,ds,\\ q_2(t) &\equiv q_2(0) + \lambda_2 t - \mu_{2,2}\int_0^t (1 - \pi_{1,2}(x(s)))(m_2 - z_{1,2}(s))\,ds\\ &\qquad - \mu_{1,2}\int_0^t (1 - \pi_{1,2}(x(s)))z_{1,2}(s)\,ds - \theta_2\int_0^t q_2(s))\,ds.\end{aligned} \tag{5.14}$$

The integral representation is closely related to the integral representation of $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$ in (4.8); $\bar{X}^n$ has been replaced by x and the indicators $1_{\{D_{1,2}^n(s)>0\}}$ have been replaced by $\pi_{1,2}(x(s))$.

Since $\gamma = x(t)$, the relevant FTSP at time t depends on the solution of the ODE at time t, $x(t)$. Since the right side of the ODE has $\pi_{1,2}(x(t))$, the evolution of the ODE beyond t in turn depends on the FTSP, specifically, upon the steady-state distribution of that FTSP. Given $x(t)$ for some $t > 0$, we can determine the FTSP $\{D(x(t), s) : s \ge 0\}$. Given that FTSP, we can determine the steady-state quantity $\pi_{1,2}(x(t))$. Then $\pi_{1,2}(x(t))$ appears on the right side of the ODE in (5.13), determining the future of the ODE. We provided an efficient algorithm to solve this ODE coupled with the FTSP in [39]. The efficiency is based on the *quasi-birth-and-death* (QBD) process structure discussed in §5.5.

5.4. *The State Space of the ODE.* Since the ODE in (5.13) is driven by the family of FTSP $D(\gamma,\cdot)$ (just as the stochastic systems are driven by the process $D^n_{1,2}$), we divide the state space of the fluid limit according to the relation that holds between q_1 and q_2, and the behavior of the FTSP in the different regions.

Denote by $\mathbb{S}$ the state space of the ODE. That is, $\mathbb{S} \equiv [0,\infty)^2 \times [0,m_2] \equiv \{\gamma \equiv (q_1,q_2,z_{1,2})\}$, and let

$$\mathbb{S}^{\mathbf{b}} \equiv \{q_1 - rq_2 = \kappa\}, \quad \mathbb{S}^+ \equiv \{q_1 - rq_2 > \kappa\}, \quad \mathbb{S}^- \equiv \{q_1 - rq_2 < \kappa\}, \tag{5.15}$$

with $\mathbb{S} = \mathbb{S}^b \cup \mathbb{S}^+ \cup \mathbb{S}^-$.

The "boundary" subset $\mathbb{S}^b$ is a hyperplane in the state space $\mathbb{S}$, and is therefore a closed subset. It is the subset of $\mathbb{S}$ in which the AP is taking place, and the function $\pi_{1,2}$ can assume its full range of values, $0 \le \pi_{1,2}(\gamma) \le 1$, $\gamma \in \mathbb{S}^b$.

The region $\mathbb{S}^+$ "above the boundary" is an open subset of $\mathbb{S}$. For all $\gamma \in \mathbb{S}^+$, $\pi_{1,2}(\gamma) = 1$. The region $\mathbb{S}^-$ below the boundary is also an open subset of $\mathbb{S}$. For all $\gamma \in \mathbb{S}^-$, $\pi_{1,2}(\gamma) = 0$.

Let $\mathbb{A} \subset \mathbb{S}^b$ be the set in which $D(x,\cdot)$ is positive recurrent. We have $0 < \pi_{1,2}(\gamma) < 1$ if and only if $\gamma \in \mathbb{A}$. Thus, for each $\gamma \in \mathbb{S}^b$, we define

$$\mathbb{A} \equiv \{\gamma \in \mathbb{S}^b : 0 < \pi_{1,2}(\gamma) < 1\}. \tag{5.16}$$

5.5. *The Fundamental QBD structure.* Characterizing the set $\mathbb{A}$ in (5.16) is essential to our analysis. Our analysis is simplified by exploiting matrix geometric methods, as in [29]. In particular, we represent the integer-valued FTSP $D \equiv \{D(\gamma,s) : s \ge 0\}$ constructed above as a homogeneous continuous-time QBD, as in Definition 1.3.1 and §6.4 of [29]. To do so, we reorder the states appropriately. We order the states so that the infinitesimal generator matrix Q can be written in block-tridiagonal form, as in Definition 1.3.1 and (6.19) of [29] (imitating the shape of a generator matrix of a birth-and-death process). In particular, for each three-dimensional state γ, we write

$$Q \equiv Q(\gamma) \equiv \begin{pmatrix} B & A_0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \tag{5.17}$$

where the four component submatrices B, A_0, A_1 and A_2 are all $2m \times 2m$ submatrices for $m \equiv \max\{j, k\}$ (and are also functions of γ). These $2m \times 2m$ matrices B, A_0, A_1 and A_2 in turn can be written in block-triangular form composed of four $m \times m$ submatrices, i.e.,

$$B \equiv \begin{pmatrix} A_1^+ & B_\mu \\ B_\lambda & A_1^- \end{pmatrix} \qquad \text{and} \qquad A_i \equiv \begin{pmatrix} A_i^+ & 0 \\ 0 & A_i^- \end{pmatrix} \tag{5.18}$$

for $i = 0, 1, 2$. (All these matrices are also functions of γ.)

To achieve this representation, we need to re-order the states into levels. The main idea is to represent transitions above 0 and below 0 within common blocks. Let $L(n)$ denote level n, $n = 0, 1, 2, \dots$ We assign original states $\phi(n)$ to positive integers n according to the mapping:

$$\begin{aligned} \phi(2nm + i) &\equiv nm + i \quad \text{and} \\ \phi((2n+1)m + i) &\equiv -nm - i + 1, \quad 1 \le i \le m. \end{aligned} \tag{5.19}$$

Then we order the states in levels as follows

$$\begin{aligned} L(0) &\equiv \{1, 2, 3, 4, \dots m, 0, -1, -2, \dots, -(m-1)\}, \\ L(1) &\equiv \{m+1, m+2, \dots, 2m, -m, -(m+1), \dots, -(2m-1)\}, \quad \dots \end{aligned}$$

With this representation, the generator-matrix Q can be written in the form (5.17) above, where A_1 groups all the transitions within a level, A_0 groups the transitions from level $L(n)$ to level $L(n+1)$ and A_2 groups all transitions from level $L(n)$ to level $L(n-1)$. Matrix B groups the transitions within the boundary level $L(0)$, and is thus different than A_1. An example is given in §5.5.

QBD's having a generator matrix Q of the form (5.17)-(5.18) will be repeatedly constructed in our proofs. We thus refer to the QBD structure, represented by the generator matrix Q as specified by (5.18) as the *fundamental QBD.*

To determine when the AP holds, we need to be able to determine when the FTSP D is positive recurrent. Fortunately, QBD theory allows us to determine that easily for each γ, as explained in §4.3 of [39] and summarized below.

Let δ_+ and δ_- be the drift of the QBD in the positive and negative region, respectively (see §4.3 in [39]. See [29] for the general theory); i.e., let

$$\begin{aligned} \delta_+(\gamma) &\equiv j\left(\lambda_+^{(j)}(\gamma) - \mu_+^{(j)}(\gamma)\right) + k\left(\lambda_+^{(k)}(\gamma) - \mu_+^{(k)}(\gamma)\right), \\ \delta_-(\gamma) &\equiv j\left(\lambda_-^{(j)}(\gamma) - \mu_-^{(j)}(\gamma)\right) + k\left(\lambda_-^{(k)}(\gamma) - \mu_-^{(k)}(\gamma)\right). \end{aligned} \tag{5.20}$$

By our construction of the rates above, it holds that $\delta_-(\gamma) > \delta_+(\gamma)$ for every $\gamma \in \mathbb{S}$. The following is Theorem 4.1 in [39]:

THEOREM 5.1. *The QBD representing the FTSP $\{D(\gamma, s) : s \geq 0\}$ is positive recurrent if and only if*

$$\delta_-(\gamma) > 0 > \delta_+(\gamma). \tag{5.21}$$

For every $\gamma \in \mathbb{R}_3$, the set $\mathbb{A}$ in (5.16) where the AP is operating, is the same set in which (5.21) holds, i.e.,

$$\mathbb{A} = \{\gamma \in \mathbb{S}^b : \delta_-(\gamma) > 0 > \delta_+(\gamma)\}. \tag{5.22}$$

From the continuity of the QBD drift-rates in (5.20), if follows that $\mathbb{A}$ is an open and connected subset of $\mathbb{S}^b$. Hence, $\mathbb{A}$ can be regarded as an open connected subset of $\mathbb{R}_2^+$ (since $\mathbb{S}^b$ is homoeomorphic to $\mathbb{R}^+ \times [0, m_2]$). Our proofs (here and in [39]) rely on the fact that if $x(s) \in \mathbb{A}$, then for some $h > 0$, $x(u) \in \mathbb{A}$, $0 < u < h$. In particular, if $x(0) \in \mathbb{A}$, then there exists a $\delta > 0$ such that $\{x(t) : 0 \leq t < \delta\} \subset \mathbb{A}$. The following is Theorem 5.2 in [39]:

THEOREM 5.2. *If $x(0) \in \mathbb{A}$, then there exists a unique solution $x \in \mathcal{C}_3([0, \delta))$ to the fluid ODE (5.13) for some $\delta > 0$.*

We will initially work on an interval $[0, \delta)$, $\delta > 0$, over which we can guarantee that the AP and Theorem 5.2 hold. After the convergence is established, this interval can be extended, typically all the way to ∞; see §7 in [39]. However, the extension of the initial interval $[0, \delta)$ depends only on the solution to the ODE. Thus, it suffices to prove the convergence over $[0, \delta)$ no matter how small δ is. We will characterize a $\delta > 0$ in §8.3. For the rest of the discussion, assume that $\delta > 0$ is known.

5.6. *The FTSP Arising as a Limit.* We now present some results in which the FTSP $D \equiv \{D(\gamma, s) : s \geq 0\}$ arises as a limit. These results connect the queue difference process $D^n \equiv \{D^n_{1,2}(t) : t \geq 0\}$ defined in (2.7) and (5.1) and the time-expanded queue-difference processes D^n_e in (5.7) to the FTSP defined above. These results help explain the main theorem.

We first formalize the separation of time scales using the time-expanded queue-difference processes D^n_e defined in (5.7). The following result "explains" the AP, but does not complete the proof of the FWLLN. We prove this theorem in Appendix A.

THEOREM 5.3. *If $\Gamma^n/n \to \gamma$ and $D^n(\Gamma^n, 0) \Rightarrow D(\gamma, 0)$ in $\mathbb{R}$ as $n \to \infty$, where $\gamma \in \mathbb{A}$, then*

$$\{D^n_e(\Gamma^n, s) : s \geq 0\} \Rightarrow \{D(\gamma, s) : s \geq 0\} \quad in \quad \mathcal{D} \quad as \quad n \to \infty, \tag{5.23}$$

where D^n_e is the time-expanded queue-difference process in (5.7) and D is the FTSP; i.e., we have convergence of the sequence of non-Markov processes $\{D^n_e : n \geq 1\}$ to a limiting time-homogeneous CTMC.

Of course, we are actually interested in the queue-difference processes. We will obtain the following result in Corollary 8.5. Recall the definition of stochastic boundedness (SB) in §2.5.

THEOREM 5.4. *Over an appropriate interval, $[0, \delta)$, the sequence of stochastic processes $\{\{D^n_{1,2}(t) : 0 \leq t \leq \delta\} : n \geq 1\}$ is SB in $\mathcal{D}$, so that the sequence of random variables $\{D^n_{1,2}(t) : n \geq 1\}$ is SB in $\mathbb{R}$ for each t, $0 \leq t \leq \delta$.*

Nevertheless, one implication of Theorem 5.3 is that, as n increases, $D^n_{1,2}$ fluctuates "too much" in the neighborhood of every point $t \in [0, \delta)$ for the sequence of stochastic processes $\{\{D^n_{1,2}(t) : 0 \leq t \leq \delta\} : n \geq 1\}$ to be $\mathcal{D}$-tight. If the sequence were tight, then it would have a convergent subsequence. If $D^n_{1,2}$ were to converge on $[0, \delta)$ to a process in $\mathcal{D}$ along that subsequence, then the limiting process must have at most finitely many discontinuities exceeding any constant $\epsilon > 0$, see e.g., Lemma 1 on p. 122 of [8]. However, for every neighborhood of any $t \in [0, \delta]$, there would necessarily be infinitely many jumps of size 1 in the limit as $n \to \infty$. Moreover, every t would have to be a discontinuity point of the limit, but there can be only countably many discontinuities. Hence, the limit process could not be an element of $\mathcal{D}$. Hence tightness does not hold.

However, we do obtain a proper limit for the sequence of random variables $\{D^n_{1,2}(t) : n \geq 1\}$ in $\mathbb{R}$ for each fixed t by exploiting the AP. After we prove Theorem 6.1, we will establish the following pointwise AP result, which helps explain the AP. See [47] for a similar result. We prove this theorem in Appendix A after proving Theorem 5.3.

THEOREM 5.5. (*pointwise AP*) *Fix* $t \in [0, \delta)$. *As* $n \to \infty$, $D^n_{1,2}(t) \Rightarrow D(x(t), \infty)$ *in* $\mathbb{R}$ *as* $n \to \infty$, *where* $x(t)$ *is the solution to the ODE at time* t *and* $D(x(t), \infty)$ *has the limiting steady-state distribution of the FTSP* $D(\gamma, s)$ *for* $\gamma = x(t)$.

REMARK 5.1. Even though the limit of $\bar{X}^n$ turns out to be deterministic, Theorems 5.3 and 5.5 imply that the process $D^n_{1,2}$ does not become deterministic as $n \to \infty$. Given Theorems 5.3 and 5.5, we see that indeed *the deterministic ODE is driven by a stochastic process.* More precisely, the evolution of the (deterministic) solution to the ODE over $[0, \delta)$ is governed by a stochastic process, although the ODE describing that evolution is itself deterministic, depending on the time-dependent steady-state distribution of the FTSP's.

The limiting ODE and its solution are deterministic because two kinds of averaging phenomena taking place simultaneously: The first is the typical strong-law type of averaging, which is achieved by the spatial fluid scaling. The second, more interesting one, is the AP, providing instantaneous long-run averaging through the separation of time scales in the fluid limit.

As an immediate consequence of Theorem 5.4, we obtain the following SSC result.

COROLLARY 5.1. (*SSC of the queue process*) *As* $n \to \infty$,

$$c_n^{-1}((Q^n_1 - \kappa^n) - r_{1,2}Q^n_2) \Rightarrow 0 \quad \text{in} \quad \mathcal{D}([0, \delta))$$

for any sequence $\{c_n : n \ge 1\}$ *satisfying* $c_n \to \infty$ *as* $n \to \infty$.

Corollary 5.1 shows that the two-dimensional scaled queue process is effectively a one-dimensional process as $n \to \infty$. Combining Theorem 4.1 and Corollary 5.1 gives the following SSC result, which reduces the dimension of the process from the original six dimension, to only two when we consider the fluid-scaled or diffusion-scaled versions of the process X^n_6 in (2.4). In particular, asymptotically, the six-dimensional process $\bar{X}^n_6 \in \mathcal{D}_6$ actually exists in a two-dimensional hyperplane of $\mathcal{D}_6$, which is homeomorphic to $\mathcal{D}_2$ over the interval $[0, \delta)$. For $\mathcal{D}_3 \equiv \{(a_1, a_2, a_3) : a_1, a_2, a_3 \in \mathcal{D}\}$, $\bar{X}^n_3$ is asymptotically an element of the two-dimensional hyperplane $\{(a_1, r_{1,2}a_1 + \kappa, a_3) : a_1, a_3 \in \mathcal{D}\}$ of $\mathcal{D}_3$.

Recall that for a sequence of processes $\{Y^n\}$ in $\mathcal{D}$, $\check{Y}^n \equiv Y^n/\sqrt{n}$.

THEOREM 5.6. (*Complete SSC*) *As* $n \to \infty$, $d_{J_1}(\breve{X}_6^n, \breve{X}_2^n) \Rightarrow 0$ *in* $\mathcal{D}_6([0,\delta))$, *where* $X_2^n \equiv (Q_1^n, r_{1,2}Q_1^n + \kappa^n, Z_{1,2}^n)$.

REMARK 5.2. The SSC result in Theorem 4.1 is stated for $\mathcal{D}_6 \equiv \mathcal{D}_6([0,\infty))$, while the SSC in Corollary 5.1, and thus also Theorem 5.6, holds on $\mathcal{D}_6([0,\delta))$. However, the SSC result in Corollary 5.1 and Theorem 5.6 can be extended as long as the solution to the ODE is in $\mathbb{A}$, since the SSC of the queue process is implied by the AP. (This will become clear in the proofs.) As we mentioned above, the solution to the ODE is typically in $\mathbb{A}$ for all $t \geq 0$; see [39]

6. The FWLLN. 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

$$X_6^n \equiv (Q_i^n, Z_{i,j}^n) \in \mathcal{D}_6 \quad \text{and} \quad Y_8^n \equiv (A_i^n, S_{i,j}^n, U_i^n) \in \mathcal{D}_8, \quad i,j = 1,2. \tag{6.1}$$

For the FWLLN, we focus on the scaled vector process

$$(\bar{X}_6^n, \bar{Y}_8^n) \equiv n^{-1}(X_6^n, Y_8^n), \tag{6.2}$$

for X_6^n and Y_8^n in (6.1). Recall that Assumptions 1-3 are in force.

THEOREM 6.1. (FWLLN) *There exists* $\delta > 0$ *such that*

$$(\bar{X}_6^n, \bar{Y}_8^n) \Rightarrow (x, y) \quad in \quad \mathcal{D}_{14}([0,\delta)) \quad as \quad n \to \infty, \tag{6.3}$$

where (x, y) *is a deterministic element of* $\mathcal{C}_{14}([0,\delta))$ *with*

$$x \equiv (q_i, z_{i,j}) \quad and \quad y \equiv (a_i, s_{i,j}, u_i), \quad i = 1,2; j = 1,2; \tag{6.4}$$

$z_{2,1} = s_{2,1} = m_1 - z_{1,1} = m_2 - z_{2,2} - z_{1,2} = 0e$ *and* $(q_1, q_2, z_{1,2})$ *being the unique solution to the three-dimensional ODE in* (5.13). *The remaining limit function* y *is defined in terms of* x:

$$\begin{aligned} a_i(t) &\equiv \lambda_i t, \quad s_{i,j}(t) \equiv \mu_{i,j}\int_0^t z_{i,j}(s)\,ds, \\ u_i(t) &\equiv \theta_i \int_0^t q_i(s)\,ds \quad for \quad t \geq 0, \quad i = 1,2; \quad j = 1,2. \end{aligned} \tag{6.5}$$

The time interval $[0,\delta)$ *can be expanded to the largest interval (typically* $[0,\infty)$*) such that there exists a unique solution to the ODE in* (5.13).

Theorem 6.1 established convergence over some interval $[0, \delta)$. Theorem 6.1 concludes by stating that the interval can be extended whenever the solution to the ODE can be extended. Ensuring convergence over $[0, \delta)$ will usually imply convergence over an interval $[0, T)$, for some $T \gg \delta$, often even $T = \infty$. First, the convergence over $[0, \delta)$ implies that the SSC results in the next section, §7, hold globally - see the explanation right above Lemma 7.3. Second, once the convergence is established, and the unique solution to the ODE (5.13) is known to exist (Theorem 5.2 in [39]), we can use the results in Section 7 of [39], to infer whether we can extend the convergence to the whole halfline $[0, \infty)$ *by analyzing the limiting ODE itself, and not the stochastic sequence* X^n. In particular, the solution to the ODE (5.13) can be extended to the entire halfline $[0, \infty)$ by showing that $x(t) \in \mathbb{A}$ for all $t \geq 0$. Often, this can be done without even solving the ODE; see Theorem 5.4 and §7 in [39].

By Theorem 5.6, it is enough to present the fluid limit of $(\bar{Q}^n_1, \bar{Z}^n_{1,2})$, since each queue determines the other in the limit. Nevertheless, in Theorem 6.1 we presented the fluid limit for both queues. We did so, because the three-dimensional framework applies in other regions. For example, in [39] we analyzed that same ODE in all three regions. More importantly, even in our settings, when Assumption 3 holds and the solution is in $\mathbb{A}$ over $[0, \delta)$, it is helpful to solve the fluid equations without explicitly forcing the SSC relation between the queues. Having the solution satisfying $q_1(t) - r_{1,2}q_2(t) = \kappa$ strongly indicates that the numerical solution to the fluid ODE is correct; See the last paragraph in §9.2 in [39].

Most of the rest of this paper is devoted to the proof of Theorem 6.1. Most proofs of supporting theorems and lemmas appear in the Appendix (in order of appearance in the main paper). The final §9 establishes a WLLN for the stationary distributions.

7. SSC for the Service Process. In this section we establish state-space collapse (SSC) for the service process $Z^n \equiv (Z^n_{1,1}, Z^n_{1,2}, Z^n_{2,1}, Z^n_{2,2})$; i.e., we show that we can consider the process $(m^n_1, Z^n_{1,2}, 0, m^n_2 - Z^n_{1,2})$ instead of Z^n in diffusion scale (and thus, in fluid scale). Thus, the relevant dimension of the stochastic service process is one instead of four. We accomplish this goal by showing that $Z^n_{2,1}$ is asymptotically null and that the idleness in each pool is asymptotically negligible in diffusion scale (in preparation for a future FCLT refinement of the FWLLN here).

Unlike our main convergence result - Theorem 6.1 - which is proved on an initial interval, the SSC of the service process holds globally on $[0, \infty)$ for FQR-T, given Assumptions 1-3. However, here we do not yet show that a limit of $\bar{Z}^n_{1,2}$ as $n \to \infty$ exists. We only show that, when analyzing Z^n, it is sufficient to consider $Z^n_{1,2}$, prove that its fluid-scaled and diffusion-scaled versions converge and then characterize the limits. That will be done for the fluid-scaled case in the next section (and the Appendix).

Here is the SSC result to be established in this section. Note that it directly implies Theorem 4.1.

THEOREM 7.1. (*global SSC of the service process*) *As* $n \to \infty$,

$$n^{-1/2}(m^n_1 - Z^n_{1,1} - Z^n_{2,1},\ Z^n_{2,1},\ m^n_2 - Z^n_{1,2} - Z^n_{2,2}) \Rightarrow (0,0,0) \quad \text{in } \mathcal{D}_3.$$

Let $I^n_1 \equiv m^n_1 - Z^n_{1,1} - Z^n_{2,1}$ and $I^n_2 \equiv m^n_2 - Z^n_{1,2} - Z^n_{2,2}$ be the idleness processes in service pools 1 and 2, respectively, and let

$$\bar{I}^n_j \equiv I^n_j/n \quad \text{and} \quad \hat{I}^n_j \equiv I^n_j/\sqrt{n}, \quad j = 1,2. \tag{7.1}$$

Theorem 7.1 will be proved in two steps. First, we show that $Z^n_{2,1} \Rightarrow 0$; second, we show that $\hat{I}^n_1$ and $\hat{I}^n_2$ are asymptotically negligible. By the first step, $I^n_1 = m^n_1 - Z^n_{1,1} + o_P(1)$, so that we can disregard the $o_P(1)$ term in the second step.

So far, we know only that the initial state converges by Assumption 3. We do not yet have convergence results for any of the stochastic processes we consider. Hence, the results in this section will be established by (i) determining bounding stochastic processes (using sample-path stochastic order) for which the limits are known or easy to establish, and (ii) using extreme-value theory for the bounding processes to justify our claims. The bounding processes established in step (i) will have a QBD form (or an $M/M/1$ form, which can also be viewed as a trivial QBD). Hence we start by establishing extreme-value limits for homogeneous QBD processes.

7.1. *Extreme-Value Limits for QBD Processes.* We are unaware of any established extreme-value limits for QBD processes, so 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. The proof of this theorem, like most others appears in the Appendix.

THEOREM 7.2. (*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*

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

Both the statement and the proof of Theorem 7.2 are complicated by the discreteness of the integer-valued process $\mathcal{L}$. The proof is also somewhat complicated by the continuity of time. It is well known that the stationary distribution of the QBD level is asymptotically geometric, e.g., see §9.1 in [29]. Hence, we are unambiguously in the light-tailed case, but we do not have the conventional convergence in law to the Gumbel distribution if we subtract by $c\log t$ instead of divide. Indeed, there do not exist scaling functions $a(t)$ and $b(t)$ such that $a(t)(\|\mathcal{L}\|_t - b(t))$ converges in law to a proper limit as $t \to \infty$; see Sections 1.5 and 1.7 of [30]. Even though the conventional extreme-value limit cannot hold, Theorem 7.2 evidently is not in best possible form. First, we should have $\|\mathcal{L}\|_t/\log t \Rightarrow c$ for a specific constant c (which is easy to identify); second, we should have tightness of the family $\{\|\mathcal{L}\|_t - c\log t : t \geq 1\}$ for that same constant c; e.g., see Example C.2.6 of [1] and Problem 4.2 of [4], but our weaker implication of such results suffices for the application here and has a relatively simple proof; see §B.

7.2. *Basic Stochastic-Order Bounds.* As we mentioned before, the proofs will involve stochastic-order bounds, using sample-path stochastic order, involving coupling; see [45], Ch. 4 of [31] and §2.6 of [35]. We briefly discuss those bounds for a sequence of stochastic processes $\{Y^n : n \in \mathbb{N}\}$. 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.

The stochastic-order bounds will be of the form

$$Y^n(t) = Y^n(0) + \sum_{i=1}^{k} N_i\left(\int_0^t J_i^n(s)\ ds\right), \quad t \geq 0, \tag{7.2}$$

where N_i, $i = 1, 2, \ldots, k$, denote independent rate-1 Poisson processes, and J_i^n is a stochastic process that serves as a random time change of the Poisson process N_i. If we are concerned with the fluid limit of Y^n, then we next divide both sides of (7.2) by n, subtract and then add back J_i^n to get the representation

$$\begin{aligned}\bar{Y}^n(t) \equiv Y^n(t)/n &= \bar{Y}^n(0) + n^{-1}\int_0^t J_i^n(s)\, ds \\ &\quad + n^{-1}\sum_{i=1}^k \left[N_i\left(\int_0^t J_i^n(s)\, ds\right) - \int_0^t J_i^n(s)\, ds\right].\end{aligned} \tag{7.3}$$

The third step is to apply a version of the continuous mapping theorem to (7.3) (The purpose of the bounds is to be able to use the continuous mapping theorem, which can not be used on X^n.) However, to avoid unnecessary repetitions, we will not write the second step (7.3) and write only the representation as in (7.2), with the understanding that the continuous mapping theorem is actually applied to the version of $\bar{Y}^n$ in (7.3).

We now construct lower and upper stochastic-order bounds for the queues, that will be repeatedly used in following proofs, including in the proof of the AP. Throughout, N_i^a, $N_{i,j}^s$ and N_i^u, $i,j = 1,2$, denote independent rate-1 Poisson processes.

We start with the bound $X_a^n \equiv (Q_{1,a}^n, Q_{2,a}^n, Z_a^n)$ in which $Q_{1,a}^n \geq_{st} Q_1^n$, $Q_{2,a}^n \leq_{st} Q_2^n$ and $Z_a^n \leq_{st} Z_{1,2}^n$. For later use, we will consider the evolution of $\{X_a^n(t) : t \geq y\}$ for any $y \geq 0$. To construct $\{X_a^n(t) : t \geq y\}$ for a fixed $y \geq 0$, we initialize with $X_a^n(y) = X^n(y)$, and act as if all newly available pool-2 servers (after time y) take their next customers from the head of pool 2, even if $Q_{2,a}^n(t) \leq 0$ (we allow the queues to become negative), so that queue 1 is served by pool-1 servers only. Then, for any $y \geq 0$ and $t \geq y$, $X_a^n(t)$ can be represented via

$$\begin{aligned}Q_{1,a}^n(t) &= Q_{1,a}^n(y) + N_1^a(\lambda_1^n t) - N_{1,1}^s(\mu_{1,1} m_1^n t) \\ &\quad - N_1^u\left(\theta_1 \int_0^t (Q_{1,a}^n(s) \vee 0)\, ds\right), \\ Q_{2,a}^n(t) &= Q_{2,a}^n(y) + N_2^a(\lambda_2^n t) - N_{1,2}^s\left(\mu_{1,2}\int_0^t Z_a^n(s)\, ds\right) \\ &\quad - N_{2,2}^s\left(\mu_{2,2}\int_0^t (m_2^n - Z_a^n(s))\, ds\right) \\ &\quad - N_2^u\left(\theta_2 \int_0^t (Q_{2,a}^n(s) \vee 0)\, ds\right), \\ Z_a^n(t) &= Z_a^n(y) - N_{1,2}^s\left(\mu_{1,2}\int_0^t Z_a^n(s)\, ds\right).\end{aligned} \tag{7.4}$$

Observe that Z_a^n is non-increasing, and will eventually reach 0. By our construction, $Z_a^n(y) = Z_{1,2}^n(y)$, where $Z_{1,2}^n(y)$ is the number of pool-2 servers helping class-1 customers. Starting at time y, every server takes his new customers from queue 2, so that the downward drift of $Q^n_{2,a}$ may become negative. Since we have no reflection, $Q^n_{2,a}$ itself may become negative, and if the downward drift is greater than the upward one, it will drift to $-\infty$ as $t \to \infty$. However, the above bounds will be used to bound X^n on small intervals $[y, y+\epsilon)$, over which they will be meaningful. Note that the operators inside the integrants of N^u_i ensure that there is no abandonment when $Q^n_{i,a} < 0$, $i = 1, 2$.

Next, we construct the bounding system $X^n_b \equiv (Q^n_{1,b}, Q^n_{2,b}, Z^n_b)$, having $Q^n_{1,b} \le_{st} Q^n_1$, $Q^n_{2,b} \ge_{st} Q^n_2$ and $Z^n_b \ge_{st} Z^n_{1,2}$. Once again, for each $y \ge 0$, we consider the evolution the process $\{X^n_b(t) : t \ge y\}$. First, we initialize with $X^n_b(y) = X^n(y)$, $n \ge 1$. We now act as if every newly available server at time $t \ge y$ takes his next customer from queue 1, even if $Q^n_{1,b}(t) \le 0$. (Once again, we allow the queues to become negative, although in this case, $Q^n_{2,b}(t) \ge 0$ for all t and n.) Then, for any fixed $y \ge 0$ and $t \ge y$, X^n_b can be represented via

$$
\begin{aligned}
Q^n_{1,b}(t) &= Q^n_{1,b}(y) + N^a_1(\lambda^n_1 t) - N^s_{1,1}(\mu_{1,1} m^n_1 t) - N^s_{1,2}\left(\mu_{1,2}\int_0^t Z^n_b(s)\right) \\
&\quad - N^u_1\left(\theta_1 \int_0^t (Q^n_{1,b}(s) \vee 0)\, ds\right), \\
Q^n_{2,b} &= Q^n_{2,b}(y) + N^a_2(\lambda^n_2 t) - N^u_2\left(\theta_2 \int_0^t Q^n_{2,b}(s)\, ds\right), \\
Z^n_b(t) &= Z^n_b(y) + N^s_{2,2}\left(\mu_{2,2}\int_0^t (m^n_2 - Z^n_b(s))\, ds\right),
\end{aligned}
\tag{7.5}
$$

Observe that Z^n_b is nondecreasing, and will eventually reach m^n_2. Thus, the downwards drift of $Q^n_{1,b}$ might eventually become larger than the upwards drift, which means that $Q^n_{1,b}$ may drift to $-\infty$ (as $t \to \infty$). Again, these bounds will be used over short intervals over which they will be meaningful.

By a simple application of the continuous mapping theorem we can prove the next lemma:

LEMMA 7.1. *For $y \ge 0$ consider the processes $\{X^n_a(t) : t \ge y\}$ in (7.4) and $\{X^n_b(t) : t \ge y\}$ in (7.5), for which the following holds for all $n \ge 1$:*

$$(-Q^n_{1,a}, Q^n_{2,a}, Z^n_a) \le_{st} (-Q^n_1, Q^n_2, Z^n_{1,2}) \le_{st} (-Q^n_{1,b}, Q^n_{2,b}, Z^n_b).$$

Also assume that $\bar{X}^n_a(y) \equiv X^n_a(y)/n \Rightarrow X_a(y)$ and $\bar{X}^n_b(y) \Rightarrow X_b(y)$ in $\mathbb{R}$ as $n \to \infty$. Then $\{\bar{X}^n_a(t) : t \ge y\} \Rightarrow \{X_a(t) : t \ge y\}$ and $\{\bar{X}^n_b(t) : t \ge y\} \Rightarrow \{X_b(t) : t \ge y\}$ in $\mathcal{D}_3$ as $n \to \infty$, where X_a and X_b are defined as follows:

For $t \geq y$, $X_a(t) \equiv (Q_{1,a}(t), Q_{2,a}(t), Z_a(t))$ satisfies the following integral equation

$$
\begin{aligned}
Q_{1,a}(t) &= Q_{1,a}(y) + \lambda_1 t - \mu_{1,1} m_1 t - \theta_1 \int_0^t (Q_{1,a}(s) \vee 0)\; ds, \\
Q_{2,a}(t) &= Q_{2,a}(y) + \lambda_2 t - \mu_{1,2} \int_0^t Z_a(s)\; ds - \mu_{2,2} \int_0^t (m_2 - Z_a(s))\; ds \\
&\quad - \theta_2 \int_0^t (Q_{2,a}(s) \vee 0)\; ds, \\
Z_a(t) &= Z_a(y) + \mu_{2,2} m_2 t - \mu_{2,2} \int_0^t Z_a(s)\; ds,
\end{aligned}
\tag{7.6}
$$

and $X_b(t) \equiv (Q_{1,b}(t), Q_{2,b}(t), Z_b(t))$ satisfies the integral equation

$$
\begin{aligned}
Q_{1,b}(t) &= Q_{1,b}(y) + \lambda_1 t - \mu_{1,1} m_1 t - \mu_{1,2} \int_0^t Z_b(s)\; ds \\
&\quad - \theta_1 \int_0^t (Q_{1,b}(s) \vee 0)\; ds, \\
Q_{2,b}(t) &= Q_{2,b}(y) + \lambda_2 t - \theta_2 \int_0^t Q_{2,b}(s)\; ds, \\
Z_b(t) &= Z_b(y) + \mu_{2,2} m_2 t - \mu_{2,2} \int_0^t Z_b(s)\; ds,
\end{aligned}
\tag{7.7}
$$

Proof. By the continuous mapping theorem, applied to the integral representation, Theorem 4.1 in [36], $\bar{Z}_a^n \equiv Z_a^n/n$ and $\bar{Z}_b^n \equiv Z_b^n/n$ converge to the processes Z_a and Z_b with continuous sample paths. We can then apply Theorem 4.1 in [36] again, to conclude that the fluid-scaled queue lengths, $\bar{Q}_{i,a}^n \equiv Q_{i,a}^n/n$ and $\bar{Q}_{i,b}^n \equiv Q_{i,b}^n/n$, $i = 1, 2$, converge as well. (Note that $h(s) \equiv \theta(s \vee 0)$ is Lipschitz continuous, as required for the integral representation to be continuous.) ∎

Note that the condition $\bar{X}_a^n(y) \Rightarrow X_a(y)$ and $\bar{X}_b^n(y) \Rightarrow X_b(y)$ in $\mathbb{R}$ as $n \to \infty$ holds for $y = 0$ with $X_a(0) = X_b(0) = x(0)$, where $x(0)$ is deterministic, by Assumption 3 and our construction (since we take $X_a^n(0) = X_b^n(0) = X^n(0)$). In that case, and whenever $X_a(y)$ and $X_b(y)$ are deterministic, the limits X_a and X_b are deterministic functions. Indeed, we anticipate that the limits X_a and X_b will be deterministic, but we use the more general form in our proof of Lemma 8.11, exploiting convergence along subsequences, where we do not yet know that the limit is deterministic.

7.3. *The $Z_{2,1}^n$ Process.* We now treat $Z_{2,1}^n$, proving that it is asymptotically globally (for all $t \geq 0$) null in distribution. This conclusion for $Z_{2,1}^n$ holds without any scaling.

THEOREM 7.3. *(global one-way sharing)* $Z^n_{2,1} \Rightarrow 0$ *in* $\mathcal{D}$ *as* $n \to \infty$.

The proof of Theorem 7.3 relies on three lemmas, which we state now. The proofs of these lemmas and Theorem 7.3 appear in Appendix B. The first lemma proves a special case which implies Theorem 7.3. The other two lemmas prove a local version of the theorem, i.e., that $\|Z^n_{2,1}\|_\tau \Rightarrow 0$ as $n \to \infty$ for some $\tau > 0$. In the proof of Theorem 7.3 we extend the local result to the full halfline $[0, \infty)$.

Our first lemma treats the simplest case.

LEMMA 7.2. *If* $z_{1,2}(0) > 0$*, then, for all* $T > 0$*,* $P(\inf_{0 \le t \le T} \bar{Z}^n_{1,2}(t) > 0) \to 1$ *as* $n \to \infty$*. As a consequence,* $Z^n_{2,1} \Rightarrow 0$ *as* $n \to \infty$*.*

Given Lemma 7.2, 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: The SSC statement for $Z^n_{2,1}$ will first be proved locally on an interval $[0, \tau]$, for some $\tau > 0$. Then, we can use later results, proving that $\bar{Z}^n_{1,2}(t) \Rightarrow z_{1,2}(t)$ as $n \to \infty$ on $[0, \delta]$ for some $\delta \le \tau$, to extended the local SSC statement to a global one. That is, our proof follows three steps: (1) We first prove that $\|Z^n_{2,1}\|_\tau \Rightarrow 0$, for some $\tau > 0$. (2) For some δ satisfying $0 < \delta \le \tau$, we can use the local result established in the first step, to prove Theorem 6.1, and deduce that the *deterministic* fluid limit $z_{1,2}(t)$ of $\bar{Z}^n_{1,2}(t)$ exists over $[0, \delta]$. (3) Finally, we show that $z_{1,2}(t_0) > 0$ for some t_0, $0 < t_0 < \delta \le \tau$, so that Lemma 7.2 can be applied to extend the local statement in step (1) to a global one. *We emphasize at the outset that the extension to a global statement is not circular, since the convergence of the* $\bar{Z}^n_{1,2}$ *process over* $[0, \delta]$ *(established in Theorem 6.1) uses only the local SSC result (since we take* $\delta \le \tau$*).*

The next two lemmas establish step (1) described above, namely that $Z^n_{2,1} \Rightarrow 0$ on an interval $[0, \tau]$.

LEMMA 7.3. *If either (i)* $\kappa > 0$ *or (ii)* $r_{1,2} > r_{2,1}$ *and* $q_1(0) > 0$*, then there exists* τ*,* $0 < \tau \le \infty$*, such that*

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

so that $\|Z^n_{2,1}\|_\tau \Rightarrow 0$ *as* $n \to \infty$*.*

The proof of Lemma 7.3 relies on a fluid argument. That fluid reasoning fails when $\kappa = 0$ and $r_{2,1} = r_{1,2} \equiv r$ or when $\kappa = 0$ and $q_1(0) = 0$, since then $q_1(0) - r_{1,2}q_2(0) = q_1(0) - r_{2,1}q_2(0)$. In these cases we will rely on the threshold $k^n_{2,1}$, and construct a finer sample-path stochastic-order bound for the stochastic system.

When we consider the stochastic sequence $\{X^n\}$, we need to have $rQ^n_2(t) - Q^n_1(t) > k^n_{2,1}$ in order to have sharing, with pool 1 helping class 2. It is thus clear that we need to consider the stochastic fluctuations of the weighted queue-length processes $D^n_{2,1}$, and show that the probability of the threshold $k^n_{2,1}$ being crossed over an initial interval $[0, \tau]$ converges to 0 as $n \to \infty$. Arguments relying solely on the fluid-scaled processes (which are of order $O_P(n)$) are too crude, and cannot reveal whether $k^n_{2,1}$ is exceeded on an interval, since $k^n_{2,1}$ is taken to be $o(n)$. We treat that case in the next lemma by appealing to the extreme-value result established in Theorem 7.2.

REMARK 7.1. Recall that the two initial thresholds $k^n_{1,2}$ and $k^n_{2,1}$ are designed to prevent sharing when the two classes are not overloaded, and are thus chosen to satisfy $k^n_{i,j}/\sqrt{n} \to \infty$ as $n \to \infty$. Once sharing starts, with pool 2 helping queue 1, $k^n_{1,2}$ may be dropped (unless shifted-FQR is employed, in which case $k^n_{1,2} = \kappa^n = O(n)$), but $k^n_{2,1}$ is kept, in order to prevent sharing in the other direction. In the proof of the next lemma, Lemma 7.4, we will see that when sharing is taking place, it is enough to have $k^n_{2,1}/\log n \to \infty$ as $n \to \infty$. This suggests that, once sharing starts, we can replace the original threshold $k^n_{2,1}$, with a new and smaller threshold, which satisfies $k^n_{2,1}/\log n \to \infty$ as $n \to \infty$.

In the next lemma we treat the cases not treated in Lemma 7.3. In addition to $z_{1,2}(0) = 0$, we assume that $\kappa = 0$ and that $q_1(0) - r_{2,1}q_2(0) = 0$. This latter assumption implies that either $q_1(0) = 0$ (so that $q_2(0) = 0$ as well), or, if $q_1(0) > 0$, then necessarily $r_{1,2} = r_{2,1}$.

LEMMA 7.4. *Assume that $\kappa = 0$ and that $k^n_{2,1}/\log n \to \infty$ as $n \to \infty$. Also assume that $q_1(0) - r_{2,1}q_2(0) = 0$ (where $r_{2,1}$ is a rational number). Then there exists τ, $0 < \tau \le \infty$, such that*

$$\lim_{n\to\infty} P\left(\sup_{t\in[0,\tau]} D^n_{2,1}(t) < k^n_{2,1}\right) = 1.$$

Hence, $\|Z^n_{2,1}\|_\tau \Rightarrow 0$ as $n \to \infty$.

Lemmas 7.3 and 7.4 prove that, for some $\tau > 0$, $\|Z^n_{2,1}\|_\tau \Rightarrow 0$ as $n \to \infty$. We will use this local result in the proof of Theorem 6.1, which shows that, for some $0 < \delta \le \tau$, $\{\bar{X}^n(t) : 0 \le t \le \delta\} \Rightarrow \{x(t) : 0 \le t \le \delta\}$, where

x is deterministic. In particular, $\bar{Z}^n_{1,2}(t) \Rightarrow z_{1,2}(t)$ over $[0, \delta]$, where $z_{1,2}(t)$ is deterministic. Recall that Theorem 6.1 relies only on the local version of Theorem 7.3 established already.

REMARK 7.2. The conclusion of Lemma 7.2 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 $t_0 \geq 0$ $Z^n_{1,2}(t_0) > \epsilon n$, then $Z^n_{1,2}(t)$ is very likely not to reach 0 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 $k^n_{1,2}$ and $k^n_{2,1}$ 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 $s^n_{i,j}$, $i \neq j$, such that pool 2 is allowed to start helping class 1 at time t if $D^n_{2,1} > k^n_{2,1}$ and $Z^n_{1,2}(t) < s^n_{1,2}$, and similarly in the other direction.

We do not analyze either of these modified controls in this paper. We observe that a global result stating that $Z^n_{2,1} \Rightarrow 0$ as $n \to \infty$ will be much harder to show, because we cannot use the reasoning in Lemma 7.2. Specifically, showing that $Z^n_{1,2}$ becomes positive in fluid scale and never empties, does not imply that $Z^n_{2,1} \Rightarrow 0$, since sharing may be allowed at time t even if $Z^n_{1,2}(t) > 0$. Nevertheless, Lemmas 7.3 and 7.4 still hold, so that $Z^n_{2,1}(t) = 0$ for all $t \in [0, \tau)$ for some $\tau > 0$ and all n large enough. Since the convergence to the fluid limit in Theorem 6.1 is initially established for an interval $[0, \delta)$, we can decrease δ if necessary, so that $\delta \leq \tau$. Once convergence of the fluid limit to its stationary point is established (using the results in §7 of [39]), we have that the fluid cannot leave $\mathbb{A}$, and $z_{2,1}$ is guaranteed to remain zero throughout.

7.4. *The Idleness Processes.* We next address the two idleness processes. We will use the standard concept of stochastic boundedness, extended to stochastic processes, which was defined in §2.5.

THEOREM 7.4. *For $j = 1, 2$, $I^n_j / \log n$ is SB, which implies that $\hat{I}^n_j \Rightarrow 0$ as $n \to \infty$.*

REMARK 7.3. The proof Theorem 7.4 uses the result in the previous subsection, namely that $Z^n_{2,1} \Rightarrow 0$ as $n \to \infty$. Hence, the statement of the theorem should first be shown to hold on $[0, \tau]$, for τ in Lemmas 7.3 and 7.4. Once the local result is shown to hold, it is used to prove Theorem 6.1, so that the convergence of $\bar{X}^n$ to the deterministic fluid limit x is established over an interval $[0, \delta]$, for some $0 < \delta \leq \tau$. In the proof of Theorem 7.3 this was shown to imply that $Z_{2,1}^n \Rightarrow 0$ as $n \to \infty$ over the entire halfline $[0, \infty)$. We can thus extend the proof of Theorem 7.4 to the entire halfline as well. For that reason, the statement of the theorem refers to the global result and its proof also assumes that $Z_{2,1}^n$ is asymptotically null globally.

8. Proof of the Main Theorem. We now come to the proof of Theorem 6.1. There are eight subsections here. In §8.1 we establish tightness. In §8.2 we establish explicit stochastic bounds on all the processes, which control the total rate of transitions. In §8.3 we identify an interval $[0, \delta)$ over which the frozen difference processes are positive recurrent, asymptotically. In §8.4 we state a continuity result for QBD's that we will apply. In §8.5 we establish stochastic-process bounds. In §8.6 we establish bounds for the integrals over small subintervals. In §8.7 we complete the proof of Theorem 6.1, exploiting the preparation in the previous subsections. The string of inequalities in (5.37) in Appendix E.5 shows what is needed. Finally, in §A.2 we prove Theorem 5.5. Most of the proofs for this section appear in Appendix E.

8.1. *Tightness.* We start by establishing tightness.

LEMMA 8.1. *The sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in (6.2) is $\mathcal{C}$-tight in $\mathcal{D}_{14}$.*

For background on tightness, see [8, 36, 48]. 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 [8]. Thus it suffices to verify conditions (6.3) and (6.4) of Theorem 11.6.3 of [48]. Hence, it suffices to prove SB of the sequence of stochastic processes evaluated at time 0 and appropriately control the oscillations, using the modulus of continuity on $\mathcal{C}$. We obtain the stochastic boundedness at time 0 immediately from Assumption 3 in §3. We show that we can control the oscillations in our proof of Lemma 8.1. The resulting tightness implies that the sequence of stochastic processes is SB. We give an alternative proof of SB in §8.2, which yields explicit bounds on the limit processes.

Since the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in (6.2) is $\mathcal{C}$-tight by Lemma 8.1, every subsequence has a further subsequence which converges to a continuous limit. We conclude this section by applying the modulus-of-continuity inequalities established in the proof of Lemma 8.1 to deduce additional smoothness properties of the limits of all converging subsequence.

LEMMA 8.2. *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*

$$w(\bar{X}_i, \zeta, T) \leq c\zeta \quad w.p.1 \tag{8.1}$$

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

In closing this subsection, we remark that we cannot employ these bounds on the modulus of continuity to directly deduce that the limit $(\bar{X}_6, \bar{Y}_8)$ is either differentiable or deterministic. For example, a nonlinear piecewise-linear function with bounded slope is Lipschitz continuous without being differentiable, and the random function At, $t \geq 0$, where A is a bounded (non-deterministic) random variable satisfies (8.1) without itself being deterministic.

The $\mathcal{C}$-tightness result in Lemma 8.1 implies that every subsequence of the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in (6.1) has a further converging subsequence in $\mathcal{D}_{14}$, whose limit is in the function space $\mathcal{C}_{14}$. However, by Theorem 7.1, it suffices to focus on $\bar{X}^n$ in $\mathcal{D}_3$, where the limits of the subsequences will be in $\mathcal{C}_3$. To establish the convergence of the sequence $\bar{X}^n$, we must show that every converging subsequence converges to the same (unique) limit. We thus need to characterize the limit of any converging subsequence, show that it is deterministic and that it satisfies the ODE (5.13) of Theorem 6.1. The existence and uniqueness of the solution to the ODE over an interval $[0, \delta)$, for some $\delta > 0$, is stated in Theorem 5.2. This δ can be increased as long as the solution x to the limiting ODE 5.13 remains in $\mathbb{A}$. In this section we will characterize an initial interval $[0, \delta]$ for which the solution is ensured to be in $\mathbb{A}$. Since we will be using the results of §7, we can decrease δ if necessary, so that $\delta \leq \tau$, for τ defined in Lemmas 7.3 and 7.4.

8.2. *Explicit Stochastic Bounds.* In this section we establish some explicit stochastic bounds on the sequence $\{(\bar{X}_6^n, \bar{Y}_8^n) : n \geq 1\}$ in (6.1) and (6.2). These bounds complement the material in §8.1 and will be used to control the transition rates of the queue-difference stochastic processes $D_{1,2}^n$.

To treat $\bar{Y}_8^n$, we use the inequalities

$$\begin{aligned} S_{i,j}^n(t) &\leq N_{i,j}^s \left(\mu_{i,j} m_j^n t\right), \\ Q_i^n(t) &\leq Q_i^n(0) + A_i^n(t), \\ U_i^n(t) &\leq N_i^u \left(\theta_i [Q_i^n(0)t + A_i^n(t)t]\right), \quad t \geq 0. \end{aligned} \tag{8.2}$$

We apply the FWLLN for the Poisson process with (8.2) and Assumption 3 to obtain the following lemma.

LEMMA 8.3. *$\bar{Y}_8^n \le \bar{Y}_{bd}^n$, where $\bar{Y}_{bd}^n \Rightarrow y_{bd}$ in $\mathcal{D}_8$, with*

$$y_{bd}(t) \equiv \Big(\lambda_1 t, \lambda_2 t, \mu_{1,1} m_1 t, 0, \mu_{1,2} m_2 t, \mu_{2,2} m_2 t, \theta_1[q_1(0)t + \lambda_1 t^2], \theta_2[q_2(0)t + \lambda_2 t^2]\Big) \quad in \quad \mathbb{R}^8. \tag{8.3}$$

We now turn to $\bar{X}_6^n$. Since $\bar{Z}_{i,j}^n \le n^{-1} m_j^n \to m_j$ as $n \to \infty$, the agent occupancy processes $\bar{Z}_{i,j}^n$ present no problem. Let $Q_\Sigma^n \equiv Q_1^n + Q_2^n$ be the stochastic process representing the total number of customers waiting in queue in our stochastic model indexed by n. It is easy to see that we can bound Q_Σ^n above stochastically by Q_{bd}^n, where Q_{bd}^n is defined to be the number in system in an $M/M/\infty$ model with arrival rate $\lambda^n \equiv \lambda_1^n + \lambda_2^n$ and individual service rate $\theta \equiv \theta_1 \wedge \theta_2 \equiv \min\{\theta_1, \theta_2\}$. The upper bound is created by simply removing all the servers in the original model, and only allowing departure by abandonment.

For the following comparison result we use the same sample-path stochastic-order construction as in §7.

LEMMA 8.4. *If $Q_\Sigma^n(0) \le_{st} Q_{bd}^n(0)$ in $\mathbb{R}$, then $Q_\Sigma^n \le_{st} Q_{bd}^n$ in $\mathcal{D}$.*

It is well known that, if $Q_{bd}^n(0) = 0$, then $Q_{bd}^n(t)$ has a Poisson distribution with a finite mean for each $t \ge 0$. Moreover, it is easy to establish a FSLLN and a FWLLN for Q_{bd}^n; we state the FWLLN.

LEMMA 8.5. *If $\bar{Q}_{bd}^n(0) \Rightarrow q_{bd}(0)$ in $\mathbb{R}$ w.p.1, where $q_{bd}(0)$ is deterministic, then we have the FWLLN*

$$\bar{Q}_{bd}^n \Rightarrow q_{bd} \quad in \quad \mathcal{D} \quad as \quad n \to \infty, \tag{8.4}$$

where q_{bd} evolves deterministically according to the ODE $\dot{q}_{bd}(t) = \lambda - \theta q_{bd}(t)$, starting at $q_{bd}(0)$. Thus

$$q_{bd}(t) \le q_{bd}^* \equiv q_{bd}(0) \vee (\lambda/\theta) \quad for\ all \quad t \ge 0. \tag{8.5}$$

Proof. Let N^a and N^s be independent rate-1 Poisson processes. Then,

$$Q_{bd}^n(t) = Q_{bd}^n(0) + N^a(\lambda^n t) - N^s\left(\theta \int_0^t Q_{bd}^n(s)\, ds\right).$$

Applying the continuous mapping theorem for the integral representation, Theorem 4.1 in [36], we have that $\bar{Q}_{bd}^n \Rightarrow q_{bd}$ in $\mathcal{D}$ as $n \to \infty$, where q_{bd}

satisfies the ODE in the statement of the lemma. The solution to this ODE is easily seen to be $q(t) = \lambda/\theta + (q(0) - \lambda/\theta)e^{-\theta t}$, from which (8.5) follows. ∎

Lemma 8.5 implies that the sequence $\{\bar{Q}^n_{bd} : n \geq 1\}$ is C-tight in $\mathcal{D}$ whenever there is convergence of the initial conditions. Together with Lemma 8.4, that implies the following result.

COROLLARY 8.1. *The sequence* $\{\bar{Q}^n_{\Sigma} : n \geq 1\}$ *is SB in* $\mathcal{D}$*. For each* $t > 0$*, the limit of any converging subsequence of* $\{\|\bar{Q}^n_{\Sigma}\|_t : n \geq 1\}$*, where* $n \to \infty$*, is almost surely contained in the bounded interval* $[0, (q_1(0)+q_2(0)) \vee (\lambda/\theta)]$.

Proof. We use Assumption 3 to ensure that there is convergence of the initial conditions: $\bar{X}^n(0) \Rightarrow x(0)$ in $\mathbb{R}_6$ as $n \to \infty$, where $x(0)$ is deterministic. We can then let the initial conditions in Lemma 8.5 be $q_{bd}(0) \equiv q_1(0)+q_2(0)$. Hence, we get

$$\bar{Q}^n_{bd} \Rightarrow q_{bd} \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \quad \text{for} \quad q_{bd}(0) \equiv q_1(0) + q_2(0).$$

That FWLLN for $\bar{Q}^n_{bd}$ implies that $\{\bar{Q}^n_{bd}\}$ is SB, which in turn implies that $\{\bar{Q}^n_{\Sigma}\}$ is SB. Moreover, we get the final conclusion of Corollary 8.1. ∎

We now have the following strengthening of the SB conclusion that can be deduced from Lemma 8.1.

COROLLARY 8.2. *The sequence* $\{(\bar{X}^n_6, \bar{Y}^n_8) : n \geq 1\}$ *in* (6.1) *and* (6.2) *is SB in* $\mathcal{D}_{14}$*. For each* $t > 0$*, the limit of any convergent subsequence of the sequence* $\{\|(\bar{X}^n_6, \bar{Y}^n_8)\|_t : n \geq 1\}$ *is contained in a compact subset of* $\mathbb{R}_{14}$.

We also want to control the changes in the queue-length processes over intervals. For that purpose, let $T^n(t)$ be the total number of transitions of the process $(\bar{X}^n_6, \bar{Y}^n_8)$ in the time interval $(0, t]$.

LEMMA 8.6. *For* $0 \leq t < t + u$ *with* $u > 0$,

(8.6)
$$\sup_{t \leq s \leq t+u} \{|Q^n_1(s) - Q^n_1(t)| + |Q^n_2(s) - Q^n_2(t)|\} \leq T^n(t+u) - T^n(t) \leq_{st} T^n_b(u),$$

where $\{T^n_b(t) : t \geq 0\}$ *is a Poisson process with rate* c_n , $c_n/n \to c$*, with*

(8.7)
$$c \equiv \lambda_1 + \lambda_2 + \mu_{1,1} m_1 + (\mu_{1,2} \vee \mu_{2,2}) m_2 + (\theta_1 \vee \theta_2) \left((q_1(0) + q_2(0)) \vee \left(\frac{\lambda_1 + \lambda_2}{\theta_1 \vee \theta_2} \right) \right).$$

As a consequence, $n^{-1} T^n_b \Rightarrow T_b$ *in* D *as* $n \to \infty$*, where* $T_b(t) \equiv ct$*,* $t \geq 0$*, for* c *in* (8.7)*. Thus, for any* $(t, u, \tilde{c}, \epsilon)$ *with* $0 \leq t < t + u$*,* $\tilde{c} > c$ *and* $\epsilon > 0$*,* *there exists* $n_0 \equiv n_0(t, u, \tilde{c}, \epsilon)$ *such that*

$$P(T^n(t+u) - T^n(t) > \tilde{c}nu) \le \epsilon \quad \textit{for all} \quad n \ge n_0. \tag{8.8}$$

Proof. Apply Lemma 8.3 to bound the rate of arrivals and service completions. Apply Corollary 8.2 to bound the total queue content, then multiply by $\theta_1 \vee \theta_2$ to bound the rate of abandonments. ∎

8.3. *Positive Recurrence of the Frozen Difference Process.* We defined the transition rates of the queue-difference process in (5.1). We assumed that $X^n(t_0) = \Gamma^n$ where Γ^n is some fixed deterministic state where sharing is taking place, and specified the transition rates at time t_0. We now consider the *constant-rate QBD* with those transition rates. We also extend the definition by letting Γ^n be a random variable, where it is understood that Γ^n only determines the constant transition rates, and does not otherwise affect the future evolution of the stochastic process. Let $D^n_f(\Gamma^n) \equiv \{D^n_f(\Gamma^n, t) : t \ge 0\}$ denote this process. (Since t_0 plays no role in (5.1), we take it to be 0.) We use the subscript f because we refer to this constant-rate QBD as the *frozen queue-difference process*, thinking of the constant transition rates being achieved because the state has been frozen at the state Γ^n. (As in §5.1 the now-constant transition rates in (5.2)-(5.5) are asymptotically correct as $n \to \infty$ with extra $o(n)$ terms, which we omit.)

We will frequently apply this constant-rate QBD with Γ^n being a state of some process, such as $X^n(t)$. We then write $D^n_f(X^n(t)) \equiv \{D^n_f(X^n(t), s) : s \ge 0\}$, where it is understood that $D^n_f(X^n(t)) \stackrel{\mathrm{d}}{=} D^n_f(\Gamma^n)$ under the condition that $\Gamma^n \stackrel{\mathrm{d}}{=} X^n(t)$.

It is important that this frozen difference process $D^n_f(\Gamma^n)$ can be directly identified with a version of the FSTP, because both are QBD's 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 and m_j, and transformed time. In order to express the relationship, we indicate the dependence upon the arrival rates and number of servers. In particular,

$$\{D^n_f(\lambda^n_i, m^n_j, \Gamma^n, s) : s \ge 0\} \stackrel{\mathrm{d}}{=} \{D(\lambda^n_i/n, m^n_j/n, \Gamma_n/n, ns) : s \ge 0\}, \tag{8.9}$$

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

$$D(\lambda^n_i/n, m^n_j/n, \Gamma_n/n, 0) \equiv D^n_f(\lambda^n_i, m^n_j, \Gamma^n, 0) \equiv Q^n_1(0) - r_{1,2}Q^n_2(0), \tag{8.10}$$

where (Q^n_1, Q^n_2) is part of the state Γ^n. This can be checked by verifying that the constant transition rates are indeed identical for the two processes, referring to (5.2)-(5.5) and (5.9)-(5.12). 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 MS-HT scaling in (2.3), 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 $\Gamma_n/n \to \gamma$. Of course, (8.9) should not be surprising, because we defined the FTSP in terms of the queue-difference process by a limit that asymptotically reverses (8.9): The transition rates of $D(\gamma)$ were defined to be the limit of the transition rates of $D^n(\Gamma_n)/n$ when $\Gamma/n \to \gamma$.

Since the process $D_f^n(X^n(t_0), t)$ has the same QBD structure as the FTSP D, a version of Theorem 5.1 holds, i.e., for a given fixed $X^n(t_0)$, the frozen difference process $\{D_f^n(X^n(t_0), t) : t \ge 0\}$ is positive recurrent if and only if

$$\delta_+^n(X^n(t_0)) < 0 < \delta_-^n(X^n(t_0)). \tag{8.11}$$

In this subsection we find a $\xi > 0$, such that the frozen process $D_f^n(X^n(t), \cdot)$ is positive recurrent for all $t \in [0, \xi)$ with probability converging to 1 as $n \to \infty$. We do not actually use this result in the following, but the result is interesting and the proof illustrates the technique we will use in a relatively simple setting.

For $\xi > 0$ and $\eta > 0$, let $B_n(\xi, \eta)$ be the following subset of the underlying probability space:

$$B_n(\xi, \eta) \equiv \{\sup_{t \in [0,\xi]} \delta_+^n(X^n(t)) < -\eta \quad \text{and} \quad \inf_{t \in [0,\delta]} \delta_-^n(X^n(t)) > \eta\}. \tag{8.12}$$

On $B_n(\xi, \eta)$, the process $\{D_f^n(X^n(t), s) : s \ge 0\}$ is positive recurrent for all $t \in [0, \xi]$.

LEMMA 8.7. *There exist $\xi > 0$ and $\eta > 0$ such that $P(B_n(\xi, \eta)) \to 1$ as $n \to \infty$, where $B_n(\xi, \eta)$ is the subset in (8.12), on which the process $\{D_f^n(X^n(t), s) : s \ge 0\}$ is positive recurrent for all $t \in [0, \xi]$.*

8.4. *Continuity of the FTSP QBD.* In the remaining proof, we will ultimately reduce everything down to the behavior of the FTSP QBD D. First, we intend to analyze the inhomogeneous queue-difference processes $D^n(\Gamma^n)$ in terms of associated homogeneous (constant-rate) processes $D_f^n(\Gamma^n)$ introduced in §8.3, obtained by freezing the transition rates at the transition rates in the initial state Γ^n. In (8.9) above, 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. In the following subsections, 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, we will exploit a continuity property possessed by this family of QBD processes. We will be applying this to the FTSP D.

To set the stage, we review basic properties of the QBD process. From the transition rates defined in (5.9)-(5.12), we see that there are only 8 different transition rates overall. The generator Q in (5.17) is based on the four basic $2m \times 2m$ matrices B, A_0, A_1, and A_2, involving the 8 transition rates. By Theorem 6.4.1 and Lemma 6.4.3 of [29], 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 [29]; See [39] for applications in our settings. If the drift condition (5.21) holds, then the spectral radius of R is strictly less than 1 and the QBD is positive recurrent (Corollary 6.2.4 of [29]). As a consequence, we have $\sum_{n=0}^{\infty} R^n = (I - R)^{-1}$. Also, by Lemma 6.3.1 of [29], the boundary probability vector α_0 is the unique solution to the system $\alpha_0(B + RA_2) = 0$ and $\alpha \mathbf{1} = \alpha_0 (I - R)^{-1} \mathbf{1} = 1$.

Like any irreducible positive recurrent CTMC, the positive recurrent QBD is regenerative, with successive visits to any state constituting an embedded renewal process. As usual for QBD's (see [29]), 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 τ be the return time and let N be the number of Poisson transitions (with specified Poisson rate). Because of the QBD structure, the return time τ has a moment generating function (mgf) $\phi_\tau(\theta) \equiv E[e^{\theta \tau}]$, for which there exists a critical value $\theta^* > 0$ such that $\phi_\tau(\theta) < \infty$ for $\theta < \theta^*$ and $\phi_\tau(\theta) = \infty$ for $\theta > \theta^*$, while 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 $z > z^*$.

Moreover, 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 (5.17)). 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 [29]. Given $G(z)$, it is not difficult to write an expression for the generating function $\psi_{N^n}(z)$, just as in the familiar BD case; e.g., see §4.3 of [29].

We will be interested in the *cumulative process*

$$C(t) \equiv \int_0^t (f(D(s)) - E[f(D(\infty))])\, ds \quad t \geq 0, \tag{8.13}$$

for the special function $f(x) \equiv 1_{\{x \geq 0\}}$. Cumulative processes associated with regenerative processes obey CLT's and FCLT's, 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 [4] and [15]. From [9], we have the following CLT with a Berry-Esseen bound on the rate of convergence (stated in continuous time, unlike [9]): For any bounded measurable function f, there exists t_0 such that

$$|E[f(C(t))/\sqrt{t}] - E[f(N(0,\sigma^2))]| \leq \frac{K}{\sqrt{t}} \quad \text{for all} \quad t > t_0, \tag{8.14}$$

where

$$\sigma^2 \equiv E\left[\left(\int_0^\tau f(D(s)) - E[f(D(\infty))]\, ds\right)^2\right], \tag{8.15}$$

again assuming for this definition that $D(0) = s^*$. The constant K depends on the function f 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 (8.14) is an indicator function. Hence, we have the simple domination:

$$\int_0^\tau |f(D(s)|\, ds = \int_0^\tau f(D(s))\, ds \leq \tau \quad \text{w.p.1} \tag{8.16}$$

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 QBD's.

LEMMA 8.8. (*continuity of QBD's*) *Consider a sequence of irreducible, positive recurrent QBD's having the structure of the fundamental QBD in* §*5.5, with generator matrices* $\{Q_n : n \geq 1\}$ *of the form* (5.17)*. If* $Q_n \to Q$ *as* $n \to \infty$*, where the positive-recurrence drift condition* (5.21) *holds for* Q*, then there exists* n_0 *such that the positive-recurrence drift condition* (5.21) *holds for* Q_n *for* $n \geq n_0$*. For* $n \geq n_0$*, 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* (8.14) *and* (8.15)*, 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 [39], building on Theorem 2.3 in [20]. The continuity of σ^2 follows from the explicit representation in (8.15) above (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 mgf's and gf's) and then for convergence of all desired moments, exploiting (8.16) and the mgf of τ to get the required uniform integrability. Finally, we get the continuity of K from [9] 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 [9]. However, we do not get an explicit expression for the constants K. ∎

We use the continuity of the steady-state distribution α in (5.33) in §E.5. In addition, we use the following corollary to Lemma 8.8 in (5.32) in §E.5.

COROLLARY 8.3. *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 QBD's, where* (5.21) *holds for* (λ_i, m_j, γ)*, then for all* $\epsilon > 0$ *there exist* t_0 *and* n_0 *such that*

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

for all $t \geq t_0$ *and* $n \geq n_0$*.*

Proof. First apply Lemma 8.8 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 $n \geq n_0$. 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 (8.14), for any M, there exists t_0 such that for all $t \geq t_0$,

$$P\left(|\frac{1}{t}\int_0^t 1_{\{D(\lambda_i^n, m_j^n, \gamma_n, s)>0\}}\, ds - P(D(\lambda_i^n, m_j^n, \gamma_n, \infty) > 0)| > \frac{M}{\sqrt{t}}\right) < P(|N(0, \sigma^2(\lambda_i^n, m_j^n, \gamma_n))| > M) + \frac{K(\lambda_i^n, m_j^n, \gamma_n)}{\sqrt{t}}. \tag{8.17}$$

Next, choose M so that $P(|N(0,\sigma^2(\lambda_i, m_j, \gamma))| > M) < \epsilon/2$. Then, invoking Lemma 8.8, 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 (8.17) 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. ∎

8.5. *Process Bounds.* Our next step is to find a $\xi > 0$ for which we can uniformly bound the frozen difference processes $\{D_f^n(X^n(t), \cdot)\}$ and the queue-difference processes $\{D_{1,2}^n(t)\}$ for all $t \in [0, \xi]$, with two QBD's - one from above and the other from below. We thus translate the uniformity of the bounds on the drifts, established in Lemma 8.7, to a uniformity of bounds on the family of process $\{D_f^n(X^n(t), \cdot)\}$ for $t \in [0, \xi]$. Having two bounding QBD's will eventually allow us to use a sandwiching argument. Now, instead of sample path stochastic order, we use rate order, denoted by $X_1 \leq_r X_2$, by which we mean that, from every integer state and for every possible state that can be reached from that state in a single transition, both (i) the transition rates up in CTMC X_1 are less than or equal to the corresponding transition rates up in CTMC X_2, and (ii) the transition rates down in CTMC X_1 are greater than or equal to the corresponding transition rates down in CTMC X_2.

LEMMA 8.9. *For any $\epsilon > 0$, there exist states $x_m, x_M \in \mathbb{A}$ and random vectors X_m^n, X_M^n with $\|x_m - x(0)\| < \epsilon$, $\|x_M - x(0)\| < \epsilon$, $n^{-1}X_m^n \Rightarrow x_m$ and $n^{-1}X_M^n \Rightarrow x_M$ as $n \to \infty$. Moreover, there exist $\xi > 0$, $\eta > 0$ and a sequence of sets $\{B_n(\xi, \eta) : n \geq 1\}$ in the underlying probability space with $P(B_n(\xi, \eta)) \to 1$ as $n \to \infty$, such that, for $0 \leq t \leq \xi$, the frozen-difference processes associated with X_m^n and X_M^n, defined as in (8.9), provide lower and upper bounds in rate order, i.e.,*

$$\begin{aligned} D_f^n(X_m^n, \cdot) &\leq_r D_f^n(X^n(t), \cdot) \leq_r D_f^n(X_M^n, \cdot), \\ D_f^n(X_m^n, \cdot) &\leq_r D_{1,2}^n(t) \leq_r D_f^n(X_M^n, \cdot), \end{aligned} \tag{8.18}$$

where the bounding processes $D_f^n(X_M^n, \cdot)$ and $D_f^n(X_m^n, \cdot)$, and thus also the interior processes $D_f^n(X^n(t), \cdot)$, satisfy (8.12) on $B_n(\xi, \eta)$, $n \geq 1$, and are thus positive recurrent.

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. However, there is no gap when $r_{1,2} = 1$ because then $j = k = 1$.

COROLLARY 8.4. *Let $\zeta \equiv (j \vee k) - 1$. Under the conditions of Lemma 8.9, there exist $\xi > 0$ and $\eta > 0$, random vectors X_M^n and X_m^n, and a sequence of sets $\{B_n(\xi,\eta) : n \ge 1\}$ in the underlying probability space with $P(B_n(\xi,\eta)) \to 1$ as $n \to \infty$, such that, whenever*

$$\begin{aligned} D_f^n(X_m^n,0) - \zeta \quad &\le_{st} \quad D_f^n(X^n(0),0) \le_{st} D_f^n(X_M^n,0) + \zeta, \\ D_f^n(X_m^n,0) - \zeta \quad &\le_{st} \quad D_{1,2}^n(0) \le_{st} D_f^n(X_M^n,0) + \zeta, \end{aligned} \tag{8.19}$$

in $\mathbb{R}$,

$$\begin{aligned} D_f^n(X_m^n,\cdot) - \zeta \quad &\le_{st} \quad D_f^n(X^n(t),\cdot) \le_{st} D_f^n(X_M^n,\cdot) + \zeta, \\ D_f^n(X_m^n,t) - \zeta \quad &\le_{st} \quad D_{1,2}^n(t) \le_{st} D_f^n(X_M^n,t) + \zeta, \end{aligned} \tag{8.20}$$

in $\mathcal{D}([0,\xi])$), where the bounding processes $D_f^n(X_M^n,\cdot)$ and $D_f^n(X_m^n,\cdot)$, and thus also $D_f^n(X^n(t),\cdot)$, satisfy (8.12) on $B_n(\xi,\eta)$, $n \ge 1$, and are thus positive recurrent.

Proof. We can do the standard sample path construction: Provided that the processes are on the same side of state 0 in the CTMC 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 CTMC representation (which involves the matrix B for the QBD). When the upper process is above 0 and the lower process is at or below 0, 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, e.g., whenever the processes are ordered $D_f^n(X_M^n,t) \le D_f^n(X_m^n,t)$, no further discrepancies can be introduced. ∎

As an immediate corollary to Corollary 8.4, we can deduce stochastic boundedness (SB) as $n \to \infty$. The following corollary implies Theorem 5.4.

COROLLARY 8.5. *For $n \geq 1$, let $\mathcal{S}^n$ be the set of all processes $\{D^n_{1,2}(t) : 0 \leq t \leq \xi\}$ and $\{D^n_f(X^n(t), s) : 0 \leq s \leq \xi\}$ for $0 \leq t \leq \xi$ with ξ from Corollary 8.4. (The sets $\mathcal{S}^n$ form an uncountably infinite subset of the space $\mathcal{D}([0,\xi])$.) Suppose that condition (8.19) is satisfied. Then the sequence $\{\mathcal{S}^n : n \geq 1\}$ is SB. Consequently, the sequence of processes $\{\{D^n_{1,2}(t) : 0 \leq t \leq \xi\} : n \geq 1\}$ is SB in $\mathcal{D}([0,\xi])$, so that the sequence $\{D^n_{1,2}(t) : n \geq 1\}$ is SB in $\mathbb{R}$ for each t with $0 \leq t \leq \xi$.*

Proof. By letting $n \to \infty$ in Corollary 8.5, we are able to exploit the stochastic order bound in (8.20), where the bounds are positive recurrent, satisfying (8.12). ∎

We will later show that the conclusions of Corollary 8.5 hold when ξ is replaced by δ, where $[0,\delta)$ is the interval over which there exists a unique solution to the ODE in $\mathbb{A}$. Together with Theorem 5.3, Corollary 8.5 proves that the sequence of processes $\{\{D^n_{1,2}(t) : 0 \leq t \leq \xi\} : n \geq 1\}$ is SB but not tight in $\mathcal{D}([0,\xi])$; the oscillations are too rapid.

8.6. *Special Construction to Bound the Integrals.* The comparisons in Lemma 8.9 and Corollary 8.4 are important, but they are not directly adequate for our purpose. The sample-path stochastic order bound works fine for the special case of $r_{1,2} = 1$, but not more generally, because of the gap ζ. However, we now show that an actual gap will only be present rarely, if we choose the interval length ξ small enough and n big enough. We use the construction in the previous section, 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 ξ suitably small.

However, we must specify the initial conditions for all the difference processes under consideration. Consistent with Assumption 3, we assume that

$$D^n_{1,2}(0) = D^n_f(X^n_m, 0) = D^n_f(X^n_M, 0) = D^n_f(X^n(t), 0) \tag{8.21}$$

for all t, $0 \leq t \leq \xi$, and $D^n_{1,2}(0) \Rightarrow L$ as $n \to \infty$, where L is a proper random variable.

LEMMA 8.10. *Assume that condition (8.21) holds. For any $\epsilon > 0$, there exist states $x_m, x_M \in \mathbb{A}$ and random vectors X^n_m, X^n_M with $\|x_m - x(0)\| < \epsilon$, $\|x_M - x(0)\| < \epsilon$, $n^{-1}X^n_m \Rightarrow x_m$ and $n^{-1}X^n_M \Rightarrow x_M$ as $n \to \infty$. In addition, there exist $\xi > 0$ and $\eta > 0$, frozen-difference QBD processes $\{D^n_f(X^n_M, s) : s \geq 0\}$ and $\{D^n_f(X^n_m, s) : s \geq 0\}$, associated with the random vectors X^n_m and X^n_M above, defined as in (8.9), and a sequence of sets $\{B_n(\xi,\eta) : n \geq 1\}$ in the underlying probability space with $P(B_n(\xi,\eta)) \to 1$ as $n \to \infty$, such* *that, on the set* $B_n(\xi, \eta)$,

$$\begin{aligned} \delta_+^n(X_m^n) &< -\eta \quad \textit{and} \quad \delta_-^n(X_m^n) > \eta, \\ \delta_+^n(X_M^n) &< -\eta \quad \textit{and} \quad \delta_-^n(X_M^n) > \eta \end{aligned} \tag{8.22}$$

(*so that the bounding processes* $D_f^n(X_m^n, \cdot)$ *and* $D_f^n(X_M^n, \cdot)$, *and thus also* $D_f^n(X^n(t), \cdot)$, *are positive recurrent*) *and, for* $0 \le t \le \xi$, (*also on* $B_n(\xi, \eta)$)

$$\begin{aligned} \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_m^n,s)>0\}}\, ds - \epsilon &\le \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X^n(t),s)>0\}}\, ds \\ &\le \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_M^n,s)>0\}}\, ds + \epsilon \\ \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_m^n,s)>0\}}\, ds - \epsilon &\le \frac{1}{\xi}\int_0^\xi 1_{\{D_{1,2}^n(s)>0\}}\, ds \\ &\le \frac{1}{\xi}\int_0^\xi 1_{\{D_f^n(X_M^n,s)>0\}}\, ds + \epsilon. \end{aligned} \tag{8.23}$$

8.7. *Proof of Theorem 6.1.* By the tightness established in Lemma 8.1, we know that every subsequence of $\{\bar{X}^n : n \in \mathbb{N}\}$ has a further subsequence converging weakly in $\mathcal{D}_3$. We will be considering a converging subsequence with limit $\bar{X}$, but without changing the indexing notation. (We understand that n runs through a subsequence.) It suffices to show that the limit $\bar{X}$ is deterministic and satisfies the ODE in (5.13) or, equivalently, the integral representation in (5.14).

By Theorems 4.1 and 4.2, which draws on §7, it suffices to focus on the integral representation for $\bar{X}^n$ in (4.8). Many of the terms converge directly to their counterparts in (5.14) because of the assumed MS-HT 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. However, these integral terms are easily seen to be tight as well, as a conseuence of the tightness of the sequences $\{\bar{Z}_{i,j}^n : n \ge 1\}$ established in §8.1. 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 integral representation in (4.8) converge to the system

$$
\begin{aligned}
\bar{Z}_{1,2}(t) &= z_{1,2}(0) + \mu_{2,2}\bar{I}_{z,1}(t) - \mu_{1,2}\bar{I}_{z,2}(t)\\
\bar{Q}_1(t) &= q_1(0) + \bar{\lambda}_1 t - \bar{m}_1 t - \mu_{1,2}\bar{I}_{q,1,1}(t)\\
&\qquad - \mu_{2,2}\bar{I}_{q,1,2}(t) - \theta_1 \int_0^t \bar{Q}_1(s)\, ds,\\
\bar{Q}_2(t) &= q_2(0) + \bar{\lambda}_2 t - \mu_{2,2}\bar{I}_{q,2,1}(t)\\
&\qquad - \mu_{1,2}\bar{I}_{q,2,2}(t) - \theta_2 \int_0^t \bar{Q}_2(s))\, ds.
\end{aligned} \tag{8.24}
$$

We have exploited the assumed convergence of the initial conditions in Assumption 3 to replace $\bar{X}(0)$ by $x(0)$ in (8.24). In more detail, for one integral term we have

$$
\{\{\int_0^t 1_{\{D^n_{1,2}(s)>0\}}\bar{Z}^n_{1,2}(s)\, ds : t \geq 0\} : n \geq 1\} \Rightarrow \{\bar{I}_{q,1,2}(t) : t \geq 0\} \quad \text{in} \quad \mathcal{D}
$$

through the final converging subsequence.

At this point, it suffices to identify the limit of each integral term with the corresponding term in the integral representation in (5.14). That will uniquely characterize the limit over an initial interval $[0, \delta)$ because, by Theorem 5.2, there exists a unique solution to the ODE over an initial interval $[0, \delta)$. Since each of these integrals can be treated in essentially the same way, we henceforth focus only on the term $\bar{I}_{q,1,2}(t)$. Thus, it suffices to show that

$$
\bar{I}_{q,1,2}(t) = \int_0^t \pi_{1,2}(\bar{X}(s))\bar{Z}_{1,2}(s))\, ds
$$

for each t. (It suffices to look at only any one t.) From a differential perspective, it suffices to show that

$$
\bar{I}_{q,1,2}(t+\xi) - \bar{I}_{q,1,2}(t) = \pi_{1,2}(\bar{X}(t))\bar{Z}_{1,2}(t)\xi + o(\xi) \quad \text{as} \quad \xi \to 0.
$$

We achieve that goal by applying Lemma 8.11 below. ∎

Recall that $[0, \delta)$ is the interval where the ODE has a unique solution. It is initially reduced to satisfy the requirements of §7, but then can be increased once a smaller interval has been treated. However, here we reduce δ again if necessary, so that $\delta < \xi$ for ξ in Lemmas 8.7, 8.9 and 8.10. After Lemma 8.11 and Theorem 6.1 have been proved for this reduced δ, δ can be further increased to the point where the existence of a unique solution to the ODE has been determined. Below we will be introducing a new ξ less than this new δ.

LEMMA 8.11. (convergence of the integral terms) *For any $\epsilon > 0$ and t with $0 \leq t < \delta$, with δ specified above, there exists $\xi \equiv \xi(\epsilon, \delta, t)$ with $0 < \xi < \delta - t$ and n_0 such that*

$$P\left(|\frac{1}{\xi}\int_t^{t+\xi} 1_{\{D^n_{1,2}(s)>0\}}\bar{Z}^n_{1,2}(s)\, ds - \pi_{1,2}(\bar{X}(t))\bar{Z}_{1,2}(t)| > \epsilon\right) < \epsilon \tag{8.25}$$

for all $n \geq n_0$.

9. WLLN for the Stationary Distributions. Under the shifted FQR-T control, the six-dimensional stochastic process $X^n \equiv (Q^n_i, Z^n_{i,j}; i, j = 1, 2)$ is a an irreducible CTMC for each n. Equivalently, the associated fluid-scaled processes $\bar{X}^n \equiv n^{-1}X^n$ is an irreducible CTMC, with the states replaced by vectors of integers divided by n. Hence $\bar{X}^n$ has a unique steady-state (limiting and stationary) distribution $\bar{X}^n(\infty)$ for each n. We now prove that $\bar{X}^n(\infty) \Rightarrow x^*$, where x^* is the unique stationary point in $\mathbb{S}$, characterized in Theorem 6 and its two corollaries in [39].

THEOREM 9.1. (*WLLN for the stationary distributions*) *As $n \to \infty$, $\bar{X}^n(\infty) \Rightarrow x^*$, where x^* is the unique fluid stationary point given in Corollary 2 of* [39].

We apply two lemmas, which we establish in §**??**. Let $\bar{X}^{n*}_6$ be the stationary version of $\bar{X}^n_6$ created by initializing with the stationary distribution $\bar{X}^n(\infty)$ for each n.

LEMMA 9.1. *The sequence of stationary versions $\{\bar{X}^{n*}_6 : n \geq 1\}$ is $\mathcal{C}$-tight in $\mathcal{D}_6$.*

We now consider the limit of a convergent subsequence of $\{\bar{X}^{n*}_6 : n \geq 1\}$; let one such limit be denoted by $X^* \equiv (Q^*_i, Z^*_{i,j}; i, j = 1, 2)$. In general, at this stage of the reasoning, we do not know that X^* is deterministic. We first show that $P(X^*(0) \in \mathbb{S}) = 1$.

LEMMA 9.2. *As $n \to \infty$, $P(\bar{X}^n(t) \in \mathbb{S} \quad \text{for all} \quad t) \to 1$, so that $P(X^*(0) \in \mathbb{S}) = 1$ for each $t \geq 0$.*

Proof of Theorem 9.1. We apply Lemma 9.1 to obtain a convergent subsequence; let its limit be $X^* \equiv (Q^*_i, Z^*_{i,j}; i, j = 1, 2)$. We apply Lemma 9.2 to conclude that $P(X^* \in \mathbb{S}) = 1$. For each sample point of $X^*(0)$, we apply Theorem 6.1 to conclude that $X^* \equiv (Q^*_1, Q^*_2, Z^*_{1,2})$ must evolve as a three-dimensional ODE, except the initial values $(Q^*_1(0), Q^*_2(0), Z^*_{1,2}(0))$ in $\mathbb{R}^3$, which may in general be random. Since the converging stochastic processes $\bar{X}^n$ in $\mathcal{D}$ are stationary stochastic processes, so is the limiting stochastic process X^*. Hence, each initial value $(Q_1^*(0), Q_2^*(0), Z_{1,2}^*(0))$ corresponding to one sample point in the underlying probability space must be a stationary point of the fluid model. However, by Theorem 6 of [39] and its corollaries, there is a unique stationary point for the fluid model (ODE), depending on the model parameters. Hence, we must have $P(X^* = x^*) = 1$. Since this argument applies to each convergent subsequence, the limit of each convergent subsequence must be x^*. Hence, the full sequence must converge to x^*, which completes the proof. ∎

Acknowledgments. This research is part of the first author's doctoral dissertation in the IEOR Department at Columbia University. The research was supported by NSF grants DMI-0457095 and CMMI 0948190.

APPENDIX

There are six appendices here, providing important supporting material. The material is presented in the order of the associated material in the main paper. Appendix A contains the proofs of Theorems 5.3 and 5.5 in §5. Appendix B contains the proofs for theorems and lemmas establishing SSC for the service processes in §7. Appendices C and D contain supplementary material for §7. In particular, Appendix C displays the bounding QBD used in the proof of Lemma 7.4, while Appendix D provides more on the idleness processes, going beyond Theorem 7.4 in §7.

Appendix E contains the proofs for the theorems and lemmas completing the proof of Theorem 6.1 in §8. Appendix E has four subsections, corresponding to the subsections of §8 where the results are located. Appendix **??** contains the proofs for the two lemmas used in proving the WLLN for the stationary distributions.

APPENDIX A: REMAINING PROOFS IN SECTION 5

In this section we provide the two remaining proof in §5: We prove Theorems 5.3 and 5.5.

A.1. Proof of Theorem 5.3. We first establish the claimed convergence of processes in (5.23). For any $\gamma \in \mathbb{A}$, the limiting FTSP $\{D(\gamma, s) : s \geq 0\}$ is a CTMC with bounded constant transition rates, as specified in §5.2. (In this section we view the FTSP as a CTMC rather than as a QBD process.) Hence, the FTSP can make only finitely many transitions in any bounded interval. Moreover, there are only four possible transitions from any state, and there are only two possible forms for these transitions, depending upon whether $D(\gamma, s) > 0$ or $D(\gamma, s) \leq 0$. Thus, the FTSP is a well-defined random element of $\mathcal{D}$. In this framework of integer-valued processes, convergence in $\mathcal{D}$ is equivalent to convergence of the finite-dimensional distributions (fidi's).

The converging processes $\{D_e^n(\Gamma^n, s) : s \geq 0\}$ defined in (5.7) are more complicated, having time-dependent transition rates, but they have essentially the same structure. For each n and s, these processes also have only four possible transitions from any state, and there are only two possible forms for these transitions, depending upon whether $D_e^n(\Gamma^n, s) > 0$ or $D_e^n(\Gamma^n, s) \leq 0$. By assumption, the initial conditions converge. Since $\Gamma^n/n \to \gamma$ as $n \to \infty$, and because of the special time scaling in (5.7), we have uniform convergence of the time-varying transition rates of $D_e^n(\Gamma^n, s) > 0$ to the constant transition rates of the FTSP over the interval $[0, t]$. Hence, we have convergence of the fidi's, and thus convergence in $\mathcal{D}$.

We now elaborate on the way this last step can be formalized. That can be done cleanly using a uniformization framework, as in Theorem 3.1 of [32], in which all transitions of $\{D_e^n(\Gamma^n, s) : s \geq 0\}$ are generated from a single Poisson process with constant rate. However, there is a complication, because in general the transition rates are not unbounded above. One approach to this problem is to use adaptive uniformization as in [34] and references cited therein. However, by Corollary 8.1, the scaled total queue content $n^{-1}Q_{\Sigma}^n$ is stochastically bounded above by a process $n^{-1}Q_{bd}^n$, which converges in law to the deterministic finite bound $q_{bd}(t) \leq q_{bd}^*$ given in (8.5). Hence, $D_e^n(\Gamma^n, \cdot)$ is asymptotically equivalent to a process with uniformly bounded transition rates. (For a direct stochastic bound on the number of transitions over a subinterval, see Lemma 8.6.) Hence, without loss of generality, we work with the asymptotically equivalent processes that do have uniformly bounded transition rates. However, we do not introduce new notation; instead we simply act as if $n^{-1}Q_{\Sigma}^n$ is bounded above and the transition rates of $\{D_e^n(\Gamma^n, s) : s \geq 0\}$ are bounded above. Hence, we just apply standard uniformization.

Given the Poisson process with a fixed rate, which exceeds the transition rate out of any state, all potential transitions are the transition epochs of the Poisson process. The actual transitions at the transition epochs of the Poisson process occur according to a discrete-time Markov chain (DTMC). However, in our nonstationary context, the DTMC is nonstationary as well. In particular, as in [32], we can express the time-dependent transition function as

$$
\begin{aligned}
P_{i,j}^{(n)}(t) &\equiv P(D_e^n(\Gamma^n, t) = j | D_e^n(\Gamma^n, 0) = i) \\
&= \sum_{k=0}^{\infty} \frac{e^{-\eta t}(\eta t)^k}{k!} \int \cdots \int_{0 \leq s_1 < s_2 < \cdots < s_k \leq t} (\prod_{l=1}^{k} P_{\eta}^{(n)}(s_l))_{i,j} \frac{k!}{t^k} \, ds_1 \cdots ds_k, \qquad (1.1)
\end{aligned}
$$

where η is an upper bound on the total transition rate out of each state for all $n \geq 1$, and $P_{\eta}^{(n)}(s) \equiv I + Q^{(n)}(s)/\eta$ is the discrete-time markov chain transition matrix at time s, based on the infinitesimal generator matrix $Q^{(n)}(s)$ at time s.

Thus, for any given time interval $[0, t]$ and $\epsilon > 0$, we can find an integer ν such that the total number of transitions of all of the processes $\{D_e^n(\Gamma^n, s) : s \geq 0\}$ over $[0, t]$ is at most ν with probability $1 - \epsilon$. This will apply to all processes under discussion. Moreover, the occurrence of those ν transitions is distributed over $[0, t]$ according to ν i.i.d. uniformly random variables, using the classical property of the Poisson process. We can thus take the number ν and the locations of the transitions as fixed, independent of n. We are then left with the product of ν DTMC transition matrices at time-varying locations, as shown in (1.1). These transition matrices here are infinite matrices, but each has at most 5 positive entries in each row. For any given ν and initial state, we can only reach a finite number of states. So, at this point, these transition matrices actually are equivalent to finite matrices. Moreover, these transition matrices converge to the common limiting transition matrix corresponding to the FTSP, uniformly. Hence, we can uniformly bound the difference between the product of these ν matrices and the corresponding product for the FTSP, independent of their time-varying locations. In that way, we can bound the total error by an arbitrarily small quantity by choosing first ν and then n to be suitably large. ∎

A.2. Proof of Theorem 5.5. By Corollary 8.5, the sequence of random variables $\{D_{1,2}^n(t) : n \geq 1\}$ is SB. Since SB is equivalent to tightness in $\mathbb{R}$, every subsequence has a converging subsequence. We now apply Theorem 6.1 to show that every such converging subsequence must converge to the random variable $D(x(t), \infty)$, which has the steady-state distribution of the FTSP D determined by the fluid state $x(t)$ at time t. (We do not apply Theorem 5.5 in the proof of Theorem 6.1.) That implies that the entire sequence must converge to that same limit and completes the proof.

To characterize the limit of a convergent subsequence, we exploit the continuity of, first, $x(t)$ and, second, the distribution of $D(x(t), \infty)$, exploiting Lemma 8.8. With these properties, we obtain the following lemma, which relates the FTSP at finite times to its steady-state distribution. To express the distance between probability distributions on $\mathbb{R}$, corresponding to convergence in distribution, we use the Lévy metric, defined for any two cdf's F_1 and F_2 by

$$\mathcal{L}(F_1, F_2) \equiv \inf \{\epsilon > 0 : F_1(x - \epsilon) - \epsilon \leq F_2(x) \leq F_1(x + \epsilon) + \epsilon \,\text{forall}\, x\}.$$

For random variables X_1 and X_2, $\mathcal{L}(X_1, X_2)$ denotes the Lévy distance between their probability distributions.

LEMMA A.1. *For any t_0 with $0 \leq t_0 < \delta$, where δ is chosen to ensure that the ODE has a unique solution x with $x(t) \in \mathbb{A}$ for all $t \in [0, \delta)$, and any $\epsilon > 0$, there exist s_0 and $\zeta > 0$ such that $t_0 + \zeta < \delta$ and*

$$\mathcal{L}(D(x(t_0), \infty), D(x(t), s)) \leq \epsilon \,\textit{for all}\, s \geq s_0 \,\textit{and}\, t \in (t_0 - \zeta, t_0 + \zeta). \tag{1.2}$$

Proof. As stated above, Lemma 8.8 establishes continuity in $x(t)$ of the distributions of the steady-state variables $D(x(t), \infty)$ of the FTSP D. Since

$x(t)$ is continuous in t, we get continuity of the distribution of $D(x(t), \infty)$ in the time argument t. Thus, for any $\epsilon > 0$ we can find $\zeta_1 > 0$, such that

$$\mathcal{L}(D(x(t_0), \infty), D(x(t), \infty)) \leq \epsilon/2 \quad \text{for all} \quad t \in [t_0 - \zeta_1, t_0 + \zeta_1]. \tag{1.3}$$

Lemma 8.8 also establishes continuity for the distribution of the return time to a fixed regeneration state, which was argued to have a finite mgf in §8.4. Let $\tau(t)$ denote the return time to the regeneration state s^* of the process $D(x(t), \cdot)$. Now, the existence of a finite mgf of the return time $\tau(t)$ to a "small set" (for background and definition see, e.g., [33], [27]) is equivalent to exponential ergodicity of the Markov process; See Theorem 2.5 in [27]. Moreover, in a countable state space, every set is small, and in particular our chosen regeneration state s^*. Hence, $\tau(t)$ is the return time to the small set $\{s^*\}$ having a finite mgf, which implies that $D(x(t), \cdot)$ is exponentially ergodic:

$$\mathcal{L}(D(x(t), s), D(x(t), \infty)) \leq \beta(t) e^{-\rho(t)s}$$

or some $\beta(t), \rho(t) > 0$.

However, for our purposes, we need to show that the bounds on the convergence rates $\beta(t)$ and $\rho(t)$ are themselves bounded uniformly in t in the neighborhood of t_0. Specifically, we need to find $\beta_0 < \infty$ and $\rho_0 > 0$, such that, for some ζ_2 and for all $t \in [t_0 - \zeta_2, t_0 + \zeta_2] \subset [0, \delta)$

$$\mathcal{L}(D(x(t), s), D(x(t), \infty)) \leq \beta_0 e^{-\rho_0 s} \tag{1.4}$$

in $\mathbb{R}$. We prove that result in Lemma A.2 below. Then by 1.4, we can take s_0 large enough, such that

$$\mathcal{L}(D(x(t), \infty), D(x(t), s)) \leq \epsilon/2$$

for all $s \geq s_0$ and for all $t \in [t_0 - \zeta_2, t_0 + \zeta_2]$. This, together with (1.3) implies the claim of the lemma for $\zeta \equiv \zeta_1 \wedge \zeta_2$. ∎

Next, by Theorem 6.1, $\bar{X}^n \Rightarrow x$ in $D([0, \delta))$ as $n \to \infty$, where x is a deterministic continuous function with $x(t) \in \mathbb{A}$ for all $t \in [0, \delta)$. Then we can apply Theorem 5.3, just proved above, to obtain

$$D^n_{1,2}(t + s_0/n) = D^n_e(X^n(t), s_0) \Rightarrow D(x(t), s_0) \quad \text{as} \quad n \to \infty. \tag{1.5}$$

From the proof of Theorem 5.3 we can conclude the convergence is uniform for t in a neighborhood of t_0. Hence we can apply Lemma A.1 to conclude that there exists n_0 such that

$$\mathcal{L}(D(x(t_0), \infty), D^n_{1,2}(t + s_0/n)) \leq 2\epsilon \,\text{for all}\, t \in (t_0 - \zeta, t_0 + \zeta), \tag{1.6}$$

provided that $n \geq n_0$. Hence, the limit of the convergent subsequence of $\{D^n_{1,2}(t_0)\}$ must be $D(x(t_0), \infty)$, as claimed. ∎

To finish the proof, we now show that (1.4) indeed holds for the family of processes $\{D(x(t), \cdot) : t \in [t_0 - \zeta_2, t_0 + \zeta_2]\}$, for some $\zeta_2 > 0$.

LEMMA A.2. (*uniform bounds on convergence rate to stationarity*) *Fix $t_0 \in (0, \delta)$, for δ in Theorem 6.1. Then there exist $\zeta_2 > 0$ and constants $\beta_0 < \infty$ and $\rho_0 > 0$, such that (1.4) holds for all $t \in [t_0 - \zeta_2, t_0 + \zeta_2] \subset (0, \delta)$. If $t_0 = 0$, then the statement is true on an interval $[0, \zeta_2]$.*

Proof. As before, we 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., $s^* = 0$. As explained in the proof of Lemma A.1, the return time of $\tau(t)$ has a finite mgf for all $t \in [0, \delta)$, which imply that each QBD $D(x(t), \cdot)$ is exponentially ergodic, $t \in [0, \delta)$.

Consider the infinitesimal generator matrix $Q(t)$ of the process $D(x(t), \cdot)$. In a countable state space, every compact set is "small" (for background and definition see, e.g., [33] and [27]). In particular, $\{0\}$ is a small set. Moreover, by Theorem 2.5 in [27], the existence of a finite mgf for the hitting time of the small set $\{0\}$ is equivalent to the exponential drift condition on the generator, Condition (**V4**) in [27]:

$$Q(t)V \leq -c_t V + d_t \mathbf{1}_{\{0\}}, \tag{1.7}$$

where c_t and d_t are strictly positive constants, and $V : \mathbb{Z} \to [1, \infty]$ is a Lyapunov function.

Consider the time $t_0 \in (0, \delta)$. Then (1.7) holds at t_0 with constants c_{t_0} and d_{t_0}. Since $c_{t_0} > 0$ we can decrease it such that (1.7) holds with strict inequality and the new c_{t_0} is still strictly positive. We increase d_{t_0} appropriately, such that $\sum_j q_{0,j}(t)V(j) < -c_{t_0}V(0) + d_{t_0}\mathbf{1}_{\{\mathbf{0}\}}$. The continuity of $Q(t)$ on $(0, \delta)$ as a function of t (Lemma 8.8) implies that there exist $\zeta_2 > 0$ and two positive constants c_0 and d_0, such that 1.7 holds for all $t \in [t_0 - \zeta_2, t_0 + \zeta_2]$ with the same constants c_0 and d_0. However, this is still not sufficient to conclude that the bounds in (1.4) are the same for all $t \in [t_0 - \zeta_2, t_0 + \zeta_2]$; see Theorem 1.1 in [7] (for discrete-time Markov chains).

Let $P^t_{i,j}(s)$ denote the transition probabilities of the CTMC $\{D(x(t), s) : s \geq 0\}$. (Here, the subscript t denotes the fixed time t we are considering and s is the fast time scale i.e., the time argument of the QBD: $P^t_{i,j}(s) \equiv P(D(x(t), s) = j \mid D(x(t), 0) = i)$.) For the CTMC, we can establish uniform bounds on the convergence rates to stationarity by showing that $\{0\}$ is a $(u, \alpha) - small$ set as in pg. 3 of [42]. That is, for some $\alpha > 0$ and a time

$u > 0$, and for a probability measure φ on the state space $\mathbb{Z}$, the small set $\{0\}$ satisfies the following "minorization condition":

$$P^t_{0,j}(u) \geq \alpha\varphi(j), \quad j \in \mathbb{Z}. \tag{1.8}$$

In particular, we need to show that (1.8) holds for all $t \in [t_0 - \zeta_2, t_0 + \zeta_2]$ with the same α (but φ is allowed to change with t). This step, together with the uniform bounds c_0 and d_0 in (1.7) established above, will be shown to be sufficient to conclude the proof.

Hence, it is left to show that (1.8) holds for all $t \in [t_0 - \zeta_2, t_0 + \zeta_2]$ 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 $\varphi(j) \equiv P^t_{0,j}(u)$. With this definition of φ we can take any $\alpha \leq 1$ in (1.8). As in the discrete-time case in [7] (the *strong aperiodicity condition* (A3) in [7] is irrelevant in continuous time), the bounds on the convergence rates in (1.4) depend explicitly on α in the minorization condition (1.8), the bounds in the drift condition (1.7) and the Lyapunov function V in (1.7). This can be justified by uniformization, but can also be justified directly for continuous-time processes, for example, from the expressions in Theorem 3 and Corollary 4 in [42].

The uniform bounds on the rate of convergence to steady state established above by applying [42] 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 A.1. A minor variant of Lemma 8.11 (proved in the same way) establishes the weaker limit for local averages:

$$\lim_{\xi \downarrow 0} \lim_{n\to\infty} \frac{1}{\xi} \int_t^{t+\xi} 1_{\{D^n_{1,2}(s) \leq k\}} ds \to P(D(x(t), \infty) \leq k) \quad \text{for all} \quad k, \tag{1.9}$$

but (1.9) and tightness alone are evidently insufficient to establish Theorem 5.5.

APPENDIX B: REMAINING PROOFS IN SECTION 7

Proof of Theorem 7.2. 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. Since 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 $\mathcal{L}$ above w.p.1 by another process $\mathcal{L}_b$ that is continuous real-valued with an exponential tail and which inherits the regenerative structure of $\mathcal{L}$.

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

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

where X is an exponential random variable distributed as X_i that is independent of $\mathcal{L}(t)$. We now consider the stationary version of $\mathcal{L}$, which makes $\mathcal{L}_b$ 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 $P(\mathcal{L}_b(t) > x) \sim e^{-x/c}$ as $x \to \infty$, because the first and third terms become asymptotically negligible as $x \to \infty$. (We choose c to make $\mathcal{L}(t)$ asymptotically negligible compared to X.)

It now remains to establish the conventional extreme-value limit for the bounding process $\mathcal{L}_b$. 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 [43]. One is based on stationary processes, while 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 [30]. The regenerative structure implies that the mixing condition in [30] is satisfied; see Section 4 of [43].

However, the classical theory in [30] and the analysis in [43] applies to sequences of random variables as opposed to continuous-time processes. In general, the established results for stationary sequences in [30] do not extend to stationary continuous-time processes. That is demonstrated by extreme-value limits for positive recurrent diffusion processes in [10, 12]. Proposition 3.1, Corollary 3.2 and Theorem 3.7 of [10] 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 $\mathcal{L}_b(t) = \mathcal{L}_d(N(t))$, where $\mathcal{L}_d(n)$ is the embedded discrete-time process associated with $\mathcal{L}_b$. 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 $\mathcal{L}_b$. 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 [43]. In fact, there is a substantial literature on extremes with a random index, e.g., see Proposition 4.20 and (4.53) of [41] and also [44]. Hence, for the QBD we can initially work in discrete time, to be consistent with [30, 43]. 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 $\mathcal{L}_b$, but our process $\mathcal{L}_b$ is actually not a stationary process. So it is natural to apply the second approach based on cycle maxima, which is given in [43, 3] and Section VI.4 of [4]. We would get the same extreme-value limit for the given version of $\mathcal{L}_b$ 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 [43] has connected the two approaches (see p. 380 of [43]), showing that all the versions of the regenerative process have the same extreme-value limit. Hence, the given version of the process $\mathcal{L}_b$ 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. ∎

REMARK B.1. (an alternative proof) An alternative proof of Theorem 7.2 would be based on a direct demonstration that the cycle maximum of $\mathcal{L}$ has a geometric tail. That alternative reasoning has the advantage that it applies directly in continuous time; see [3] and Section VI.4 of [4]. However, we are unaware of such a result in the literature. Evidently, it can be derived from the known behavior of the first passage times between levels. By Theorem 8.2.2 of [29], the probability of moving from level 0 to level $k+1$ before returning to level 0 is asymptotically geometric as $k \to \infty$. However, the return to level 0 may not be in the same phase as the initial phase. Hence, we must consider the random evolution within level 0 until we either hit the initial phase or leave level 0, and then the random number of those returns until we do return to level 0 in the same phase as the initial phase. Evidently that will not alter the geometric tail, but that remains to be shown.

In fact, if we show that the cycle maximum has a geometric tail, then we need not construct the bounding process $\mathcal{L}_b$. Instead, we can directly apply the extreme-value theorem for regenerative processes with geometric tail, Theorem 6 in [2] or Problem 4.2 on p. 185 of [4], from which our conclusion would follow. In particular, it is well known that the maximum queue length over a busy cycle in an $M/M/1$ is asymptotically geometric. We can thus use Theorem 6, and, more directly, the example on p. 112 in [2], for the extreme-value bound for the $M/M/1$ queue-length process, which we apply in the proof of Theorem 7.4.

Proof of Lemma 7.2. By Assumption 3, the condition $z_{1,2}(0) > 0$ implies that $P(Z^n_{1,2}(0) > 0) \to 1$ as $n \to \infty$. Clearly, for every $n \geq 1$, $Z^n_{1,2}$ is stochastically bounded from below, in sample-path stochastic order, by a process Z^n_b which has $Z^n_b(0) = Z^n_{1,2}(0)$, has only departures and no new arrivals, i.e., $Z^n_{1,2} \geq_{st} Z^n_b$ for all $n \geq 1$ and $t \geq 0$, where

$$Z^n_b(t) = Z^n_b(0) - N^s_{1,2}\left(\mu_{1,2}\int_0^t Z^n_b(s)\, ds\right),$$

with $N^s_{1,2}$ being a rate-1 Poisson process.

Given the FSLLN for the Poisson process $N^s_{1,2}$, by applying the continuous mapping theorem, we have $Z^n_b/n \Rightarrow z_b$ in $\mathcal{D}$, as $n \to \infty$, where

$$z_b(t) = z_b(0) - \mu_{1,2}\int_0^t z_b(s)\, ds, \quad t \geq 0.$$

It follows that $z_b(t) \geq z_b(0)e^{-\mu_{1,2}t}$, so that $z_b(t) > 0$ for all $t \geq 0$. Thus $P(\inf_{0\leq s\leq t} Z^n_b(s) > 0) \to 1$ as $n \to \infty$. The stochastic order bound implies that the same is true for $Z^n_{1,2}$, which proves the first claim of the lemma. The second claim that $Z^n_{2,1} \Rightarrow 0$ as $n \to \infty$ follows from the first together with the one-way sharing rule. ∎

Proof of Lemma 7.3. When either of the conditions (i) or (ii) holds, $d_{2,1}(0) < 0$, where $d_{2,1}(t) \equiv r_{2,1}q_2(t) - q_1(t)$, $t \geq 0$. Under condition (i), by Assumption 3, $-d_{2,1}(0) \geq d_{1,2}(0) \equiv q_1(0) - r_{1,2}q_2(0) = \kappa$. If $\kappa = 0$ and Condition (ii) holds, then $d_{2,1}(0) < r_{1,2}q_2(0) - q_1(0) = \kappa = 0$.

We will construct a sample-path stochastic-order bound from above for $D^n_{2,1}$, and show that this bounding process is asymptotically strictly negative on an interval $[0,\tau]$, for some $\tau > 0$. To stochastically bound $D^n_{2,1}$, we consider a sequence of systems $\{X^n_b : n \geq 1\}$ in (7.5) initialized at time 0 with $X^n_b(0) \equiv X^n(0)$, $n \geq 1$. Thus, $Q^n_{i,b}(0) = Q^n_i(0)$, and both service pools start full with only their own customers. (Recall that we are considering the case $Z^n_{1,2}(0) = 0$ for all n large enough.)

Let $D^n_b \equiv r_{2,1} Q^n_{2,b} - Q^n_{1,b}$ be the weighted difference process in X^n_b. By construction, $Q^n_{1,b} \leq_{st} Q^n_1$ and $Q^n_{2,b} \geq_{st} Q^n_2$, so that $D^n_b \geq_{st} D^n_{2,1}$. Now, as was shown in §7.2, $\bar{X}^n_b \Rightarrow x_b$ as $n \to \infty$, for x_b in (7.7). Hence, $\bar{D}^n_b \equiv D^n_b/n \Rightarrow d_b \equiv r_{2,1} q_{2,b} - q_{1,b}$ as $n \to \infty$, with $d_b(0) < 0$.

The limit process $q_{1,b}(t)$ may eventually become negative as t increases, at which point it becomes meaningless as a stochastic-order bound for q_1. However, the continuity of $q_{1,b}$, together with the initial condition, $q_{1,b}(0) > 0$, implies that we can find a time $\tau_1 > 0$, such that $q_{1,b}(t) > 0$ for all $t \in [0, \tau_1]$. Similarly, the continuity of d_b implies that there exists $\tau_b > 0$, where $\tau_b \equiv \inf\{t \geq 0 : d_b(t) = 0\}$. Then, for $\tau^n_b \equiv \inf\{t \geq 0 : D^n_b(t) \geq 0\}$, by applying a version of Theorem 13.6.4 in [48], the continuous mapping theorem gives $\tau^n_b \Rightarrow \tau_b$. Now, for $\tau^n \equiv \inf\{t \geq 0 : D^n_{2,1}(t) \geq 0\}$ we have that $\tau^n \geq_{st} \tau^n_b$. Taking $\tau \equiv \tau_1 \wedge \tau_b$ gives the first claim of the statement.

The second claim of the statement follows from the first, together with the initial condition in Assumption 3, namely, that $Z^n_{2,1}(0) = 0$ for all n. ∎

Proof of Lemma 7.4. We will prove the lemma by constructing a QBD process that serves as a stochastic-order bound for the process $D^n_{2,1}$ over some interval $[0,\tau]$. The claims will then follow from an application of the extreme-value limit in Theorem 7.2. As a first step, we define the following processes:

For $s \geq 0$, let $X^n_*(s) \equiv (Q^n_{1,a}(s), Q^n_{2,a}(s), Z^n_b(s))$, where $Q^n_{i,a}$, $i = 1,2$, are defined in (7.4) and Z^n_b is defined in (7.5). For a fixed $s > 0$ and a fixed $X^n_*(s)$, define the following processes:

$$
\begin{aligned}
Q^n_{1,*}(X^n_*(s), t) &= Q^n_{1,a}(0) + N^a_1(\lambda^n_1 t) - N^s_{1,1}(\mu_{1,1} m^n_1 t) - N^s_{1,2}(\mu_{1,2} Z^n_b(s) t) \\
&\quad - N^u_1\left(\theta_1 (Q^n_{1,a}(s) \vee 0) t\right), \\
Q^n_{2,*}(X^n_*(s), t) &= Q^n_{2,a}(0) + N^a_2(\lambda^n_2 t) - N^s_{2,2}(\mu_{2,2}(m^n_2 - Z^n_b(s)) t) \\
&\quad - N^u_2\left(\theta_2 (Q^n_{2,a}(s) \vee 0) t\right),
\end{aligned}
$$

where, as before, N^a_i, $N^s_{i,j}$ and N^u_i, $i,j = 1,2$, are independent rate-1 Poisson processes. Then the process

$$
D^n_*(X^n_*(s), t) \equiv r_{2,1} Q^n_{2,*}(X^n_*(s), t) - (Q^n_{1,*}(X^n_*(s), t) - \kappa^n) - \inf_{0 \leq u \leq t} D^n_*(X^n_*(s), u)
$$

conditional on $X_*^n(s)$, is a continuous-time Markov chain as a function of the time argument t. (That is because X_*^n is constructed independently of D_*^n.) The key observation here is that the conditional process D_*^n (given $X_*^n(s)$), can be analyzed as a QBD, just as in §4 of [39]. In particular, if $r_{2,1} = j/k$, where j, k are positive integers with no common divisors, then the process $\tilde{D}_*^n \equiv jQ_{2,*}^n - kQ_{1,*}^n$ is a CTMC with state space in the nonnegative integers, and can be represented as a QBD. Moreover, the process $\tilde{D}_*^n$ is positive recurrent if and only if D_*^n is.

Our next objective is to replace the family of processes $\{\tilde{D}_*^n(X_*^n(s), t) : t \geq 0\}$ (there is a different process for each $X_*^n(s)$) with one positive-recurrent QBD which will bound $D_{2,1}^n$ from above over an entire interval $[0, \tau]$, for some $\tau > 0$, and then translate the scaling by n in X_*^n to a scaling by n of the time argument t. More specifically, we continue the proof in two steps: in the first step we find a positive recurrent QBD $D_*^n(X_m^n, t)$, such that $D_*^n(X_m^n, \cdot) \geq_{st} D_*^n(X_*^n(s), \cdot)$ for all $s \in [0, \tau]$. In the second step, the bounding process $D_*^n(X_m^n, \cdot)$ is shown to be equal in distribution to a rate-1 QBD on the interval $[0, a_n\tau]$, for some $\{a_n\}$ such that $a_n/n \to 1$ as $n \to \infty$. The second step allows us to employ Theorem 7.2 and show that the probability that the threshold $k_{2,1}^n$ is crossed over $[0, \tau]$ converges to 0 as $n \to \infty$.

However, before we find a QBD that uniformly bounds all the processes $D_*^n(X_*^n(s), \cdot)$, for all $s \in [0, \tau]$, we need to find all $s \geq 0$ for which $D_*^n(X_*^n(s), \cdot)$ is positive recurrent. That will allow us to characterize τ. As mentioned above, D_*^n is positive recurrent if and only if $\tilde{D}_*^n$ is positive recurrent. We thus analyze the family of processes $\{\{\tilde{D}_*^n(X_*^n(s), t) : t \geq 0\} : s \geq 0\}$. (For every fixed $s \geq 0$ and $X_*^n(s)$ we have a whole process $\tilde{D}_*^n$ with time argument t.)

Given $X_*^n(s)$, the process $\{\tilde{D}_*^n(X_*^n(s), t) : t \geq 0\}$ has upward jumps of size j with rate $\hat{\lambda}_j(X_*^n(s)) \equiv \lambda_2^n$, and downward jumps of size j (away from the boundary) with rate $\hat{\mu}_j(X_*^n(s)) \equiv \mu_{2,2}(m_2^n - Z_b^n(s)) + \theta_2 Q_{2,a}^n(s)$. It has upward jumps of size k with rate $\hat{\lambda}_k(X_*^n(s)) \equiv \mu_{1,1}m_1^n + \mu_{1,2}Z_b^n(s) + \theta_1 Q_{1,a}^n(s)$, and downwards jumps of size k (away from the boundary) with rate $\hat{\mu}_k(X_*^n(s)) \equiv \lambda_1^n$. Now, by Theorem 7.2.3 in [29], for a given $X_*^n(s)$, $\tilde{D}_*^n(X_*^n(s), \cdot)$ is positive recurrent if and only if $\tilde{\delta}_*(X_*^n(s)) < 0$, where

$$\tilde{\delta}_*(X_*^n(s)) \equiv j(\hat{\lambda}_j(X_*^n(s)) - \hat{\mu}_j(X_*^n(s))) + k(\hat{\lambda}_k(X_*^n(s)) - \hat{\mu}_k(X_*^n(s))).$$

Since $\bar{X}_*^n \equiv X_*^n/n \Rightarrow x_* \equiv (q_{1,a}, q_{2,a}, z_b)$, for z_b in (7.7) and $q_{i,a}$, $i = 1, 2$ in (7.6), we can define for every $s \geq 0$ the functions $\hat{\lambda}_j(x_*(s))$, $\hat{\mu}_j(x_*(s))$, $\hat{\lambda}_k(x_*(s))$ and $\hat{\mu}_k(x_*(s))$ to be the limits of $\hat{\lambda}_j(X_*^n(s))/n$, $\hat{\mu}_j(X_*^n(s))/n$, $\hat{\lambda}_k(X_*^n(s))/n$ and $\hat{\mu}_k(X_*^n(s))/n$, respectively, as $n \to \infty$.

By the linearity of $\tilde{\delta}_*$ and the continuity of the addition mapping when the limits are continuous, e.g. Theorem 12.7.1 in [48], we have that $\tilde{\delta}_*(X^n_*(s))/n \Rightarrow \tilde{\delta}_*(x_*(s))$, where

$$\tilde{\delta}_*(x_*(s)) \equiv j(\hat{\lambda}_j(x_*(s)) - \hat{\mu}_j(x_*(s))) + k(\hat{\lambda}_k(x_*(s)) - \hat{\mu}_k(x_*(s))).$$

Note that, by our construction of X^n_*, $x(0) = x_*(0)$ (that is because $X^n_a(0) = X^n_b(0) = X^n(0)$ for all $n \geq 1$). It is easy to see that, if $r_{2,1} = r_{1,2}$ (recall also that $z_{1,2}(0) = 0$), then $\tilde{\delta}_*(x_*(0)) = -\delta_-(x(0))$ for $\delta_-(x(0))$ in (5.20). Since, by Assumption 3, $\delta_-(x(0)) > 0$, it holds that $\tilde{\delta}_*(x_*(0)) < 0$.

If $r_{2,1} < r_{1,2}$, then necessarily $q_1(0) = q_2(0) = 0$ (see the explanation before the statement of the lemma). In that case we have that $\tilde{\delta}_*(x_*(0)) = j\theta_1(\lambda_1 - \mu_{1,1}m_1) + k(\lambda_2 - \mu_{2,2}m_2)$, so that $\tilde{\delta}_*(x_*(0)) < 0$ if and only if $\theta_1(\lambda_1 - \mu_{1,1}m_1) + r_{2,1}(\lambda_2 - \mu_{2,2}m_2) < 0$. To see that this inequality must hold, observe that with $q_1(0) = q_2(0) = z_{1,2}(0) = 0$, and by Assumption 3,

$$\delta_-(x(0)) = \theta_1(\lambda_1 - \mu_{1,1}m_1) - r_{1,2}(\lambda_2 - \mu_{2,2}m_2) > 0,$$

which implies that $\lambda_2 > \mu_{2,2}m_2$, since by Assumption 1, $q^a_1 \equiv \lambda_1 - \mu_{1,1}m_1 > 0$. It follows from the latter inequality and the fact that $r_{2,1} < r_{1,2}$, that $\tilde{\delta}_*(x_*(0)) < 0$. To summarize, $\tilde{\delta}_*(x_*(0)) < 0$ in both cases considered in the statement of the lemma.

Since x_* and $\tilde{\delta}_*(x_*)$ are continuous functions, we can find $\tau > 0$ such that $\sup_{s\in[0,\tau]} \tilde{\delta}_*(x_*(s)) < 0$. Hence, there exists $\eta_1 > 0$ such that

$$P\left(\sup_{s\in[0,\tau]} \tilde{\delta}^n_*(X^n_*(s)) < -\eta_1\right) \to 1 \quad \text{as} \quad n \to \infty.$$

That is, for some $\tau > 0$ there exists a sequence of sets $\{B^n : n \in \mathbb{N}\}$ satisfying $P(B^n) \to 1$ as $n \to \infty$, such that the process $\{D^n_*(X^n_*(s), t) : t \geq 0\}$ is positive recurrent for all $s \in [0, \tau]$ and for every sample path of X^n_* contained in B^n.

We now construct a single bounding QBD process that bounds $\tilde{D}^n_*(X^n_*(s), \cdot)$ for all $s \in [0, \tau]$. For that purpose, let $X^n_m \equiv (Q^n_{1,m}, Q^n_{2,m}, Z^n_m)$, where

$$Q^n_{1,m} \equiv \|Q^n_{1,a}\|_\tau, \quad Q^n_{2,m} \equiv \inf_{0\leq t\leq\tau} Q^n_{2,a}(t) \quad \text{and} \quad Z^n_m \equiv \|Z^n_b\|_\tau.$$

Applying the continuous mapping theorem for the supremum function, e.g., Theorem 12.11.7 in [48], we have that $\bar{X}^n_m \equiv X^n_m/n \Rightarrow x_m \equiv (q_{1,m}, q_{2,m}, z_m)$, with $q_{1,m} \equiv \|q_{1,a}\|_\tau$, $q_{2,m} \equiv \inf_{0\leq t\leq\tau} q_{2,a}(t)$ and $z_m \equiv \|z_b\|_\tau$.

Let $D_*(t) \equiv r_{2,1}Q_{2,*}(t) - Q_{1,*}(t) - \inf_{0\leq u\leq t} D_*(u)$, where

$$Q_{1,*}(t) = q_{1,a}(0) + N^a_1(\lambda_1 t) - N^s_{1,1}(\mu_{1,1}m_1 t) - N^s_{1,2}(\mu_{1,2}z_m t) - N^u_1(\theta_1 q_{1,m} t),$$
$$Q_{2,*}(t) = q_{2,a}(0) + N^a_2(\lambda_2 t) - N^s_{2,2}(\mu_{2,2}(m_2 - z_m)t) - N^u_2(\theta_2 q_{2,m} t).$$

By our choice of x_m, the QBD D_* is positive recurrent. Observe that for every sequence of sample paths $\{X^n_m : n \in \mathbb{N}\}$, the scaling in $D^n_*(X^n_m, \cdot)$ is equivalent to scaling time by a factor of order $O(n)$ in D_*. That is, for every $T > 0$, and every sample path of X^n_m contained in the sets B^n defined above $\{D^n_*(X^n_m, t) : 0 \le t \le T\} \stackrel{d}{=} \{D_*(a_n t) : 0 \le t \le T\}$, with $a_n/n \to 1$ as $n \to \infty$.

Let $M_*(t) \equiv \sup_{s\in[0,t]} D_*(s)$ denote the running maximum of the positive recurrent QBD D_*. It follows from Theorem 7.2 that there exists $c > 0$ such that

$$\lim_{n\to\infty} P(\|D^n_{2,1}\|_\tau / \log n > c) \le \lim_{n\to\infty} P(\|M^*\|_{a_n\tau}/\log n > c) = 0.$$

The claim of the lemma then follows from the assumption that $k^n_{2,1}/\log n \to \infty$ as $n \to \infty$. ∎

Proof of Theorem 7.3. By Lemma 7.2, we only need to consider the case $z_{1,2}(0) = 0$. By Lemmas 7.3 and 7.4, there exists $\tau > 0$ such that

$$\lim_{n\to\infty} P\left(\|D^n_{2,1}\|_\tau < k^n_{2,1}\right) = 1.$$

Hence, the claim of the theorem will follow from Lemma 7.2 and Theorem 6.1 if we show that for some t_0 satisfying $0 < t_0 \le \delta \le \tau$ it holds that $z_{1,2}(t_0) > 0$, where $z_{1,2}$ is the (deterministic) fluid limit of $\bar{Z}^n_{1,2}$ as $n \to \infty$ (shown to exist in the proof of Theorem 6.1 on $[0, \delta]$). We will actually show a somewhat stronger result, namely, that for any $0 < \epsilon \le \delta$ there exists $t_0 < \epsilon$ such that $z_{1,2}(t_0) > 0$. We prove that by assuming the contradictory statement: for some $0 < \epsilon \le \delta$ and for all $t \in [0, \epsilon]$, $z_{1,2}(t) = 0$.

Since, by our contradictory assumption, $z_{1,2}(t) = 0$ over $[0, \epsilon]$, we have that $Z^n_{1,2} = o_P(n)$. Recall also that $Z^n_{2,1} = o_P(1)$ over $[0, \epsilon]$ (since $\epsilon \le \tau$, and τ is chosen according to Lemmas 7.3 and 7.4). Define the processes

$$L^n_1 \equiv Q^n_1 + Z^n_{1,1} + Z^n_{1,2} - m^n_1 \quad \text{and} \quad L^n_2 \equiv Q^n_2 + Z^n_{2,1} + Z^n_{2,2} - m^n_2, \tag{2.1}$$

representing the excess number in system for each class. Note that $(L^n_i)^+ = Q^n_i$, $i = 1, 2$. Then,

$$\begin{aligned} L^n_i(t) = L^n_i(0) + N^a_i(\lambda^n_i t) - N^s_{i,i}\left(\mu_{i,i}\int_0^t (L^n_i(s)\wedge 0)\, ds\right) \\ - N^u_i\left(\theta_i \int_0^t (L^n_i(s) \vee 0)\, ds\right) + o_P(n), \quad i = 1, 2 \end{aligned} \tag{2.2}$$

for $0 \le t \le \delta$ as $n \to \infty$, where N^a_i, $N^s_{i,i}$ and N^u_i are independent rate-1 Poisson processes. The $o_P(n)$ terms are replacing the (random-time changed)

Poisson processes related to $Z^n_{1,2}$ and $Z^n_{2,1}$, which can be disregarded when we consider the fluid limits of (2.2).

Letting $\bar{L}^n_i \equiv L^n_i/n$, $i = 1, 2$, and applying the continuous mapping theorem for the integral representation function in (2.2), Theorem 4.1 in [36], (see also Theorem 7.1 and its proof in [36]), we have that $(\bar{L}^n_1, \bar{L}^n_2) \Rightarrow (\bar{L}_1, \bar{L}_2)$ as $n \to \infty$, where, for $i = 1, 2$,

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

so that

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

(We denote the fluid limit of $\bar{L}^n_i$ by $\bar{L}_i$, $i = 1, 2$, instead of our usual lower-case letters notation in order to avoid confusion.)

It is easy to see that $q_i = (\bar{L}_i(t))^+$, $i = 1, 2$, where q_i is the fluid limit of $\bar{Q}^n_i$. Now, by Assumption 3, both pools are full at time 0, so that $L_i(0) \geq 0$. Moreover, for $i = 1, 2$, $\bar{L}^e_i \equiv (\lambda_i - \mu_{i,i})/\theta_i$ is an equilibrium point of the ODE $\bar{L}'_i$, in the sense that, if $\bar{L}_i(t_0) = \bar{L}^e_i$, then $\bar{L}_i(t) = \bar{L}^e_i$ for all $t \geq t_0$. (That is, $\bar{L}^e_i$ 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 $\bar{L}_i(0) < \bar{L}^e_i$, and strictly decreasing if $\bar{L}_i(0) > \bar{L}^e_i$, $i = 1, 2$.

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, $i = 1, 2$, and write

$$\begin{aligned} q_1(t) &= q_1(0) - (\lambda_1 - \mu_{1,1} m_1)t - \theta_1 \int_0^t q_1(s)\, ds, \\ q_2(t) &= q_2(0) - (\lambda_2 - \mu_{2,2} m_2)t - \theta_2 \int_0^t q_2(s)\, ds, \quad t \in [0, \epsilon], \end{aligned}$$

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

$$\begin{aligned} d_{1,2}(t) &= q^a_1 + (q_1(0) - q^a_1)e^{-\theta_1 t} - r\left(q^a_2 + (q_2(0) - q^a_2)e^{-\theta_2 t}\right) \\ &= (q^a_1 - r q^a_2) + (q_1(0) - q^a_1)e^{-\theta_1 t} - r(q_2(0) - q^a_2)e^{-\theta_2 t}. \end{aligned} \tag{2.3}$$

and $d_{1,2}(0) = \kappa$.

It is easy to see that

$$d'_{1,2}(t) \equiv \frac{d}{dt} d_{1,2}(t) = -\theta_1(q_1(0) - q^a_1)e^{-\theta_1 t} + r\theta_2(q_2(0) - q^a_2)e^{-\theta_2 t}.$$

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 (5.20) and the assumption $z_{1,2}(0) = 0$, that $d'_{1,2}(0) = \delta_-(x(0))$. By Assumption 3, $x(0) \in \mathbb{A}$, so that $d'_{1,2}(0) > 0$, and $d_{1,2}$ is strictly increasing at 0. Now, since $d_{1,2}(0) = \kappa$, we can find $t_1 \in (0, \epsilon]$, such that $d_{1,2}(t) > \kappa$ for all $0 < t < t_1$. This implies that $P(\inf_{0<t\leq t_1} D^n_{1,2}(t) > 0) \to 1$ as $n \to \infty$.

It follows from the representation of $Z^n_{1,2}$ in (4.3) that for any $t \in [0, t_1]$,

$$\bar{Z}^n_{1,2}(t) = \frac{N^s_{2,2}\left(\mu_{2,2}m^n_2 t\right)}{n} + o_P(1). \tag{2.4}$$

The $o_P(1)$ term follows from our assumption that $\bar{Z}^n_{1,2}(t) \Rightarrow 0$ as $n \to \infty$. However, by the FSLLN for Poisson processes, the fluid limit $z_{1,2}$ of $\bar{Z}^n$ in 2.4 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 $z(t) = 0$ 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 $q_1(t) - rq_2(t) > \kappa$ 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. ∎

Proof of Theorem 7.4. We will start working with the processes L^n_1 and L^n_2 defined in (2.1) (but recall that, by Theorem 7.3 $Z^n_{2,1} \Rightarrow 0$, and in particular $\hat{Z}^n_{2,1} \Rightarrow 0$). For each $n \geq 1$, we will bound the two-dimensional process (L^n_1, L^n_2) below in sample-path stochastic order by another two-dimensional process $(L^n_{1,b}, L^n_{2,b})$.

We construct the lower-bound process $(L^n_{1,b}, L^n_{2,b})$ by increasing the departure rates in both processes L^n_1 and L^n_2, making it so that each goes down at least as fast, regardless of the state of the other. 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^n_1 at κ^n, where $\kappa^n \geq 0$; we place the reflecting upper barrier on L^n_2 at 0. With the upper barrier at κ^n, the departure rate of L^n_1 is bounded above by $\mu_{1,1}m^n_1 + \theta_1\kappa^n + \mu_{1,2}Z^n_{1,2}(t)$, based on assuming that pool 1 is fully busy serving class 1 (since $\mu_{2,1}Z^n_{2,1}(t) = o_p(1)$ we ignore it), that L^n_1 is at its upper barrier, and that $Z^n_{1,2}(t)$ agents from pool 2 are currently busy serving class 1 in the original system. Second, with the upper barrier at 0, the departure rate of L^n_2 is bounded above by $\mu_{2,2}m^n_2 - \mu_{1,2}Z^n_{1,2}(t)$, based on assuming that pool 2 is fully busy with $Z^n_{1,2}(t)$ agents from pool 2 currently busy serving class 1, and that L^n_2 is at its upper barrier 0. Thus, we give $L^n_{1,b}$ and $L^n_{2,b}$ these bounding rates at all times.

Of course, as constructed, the evolution of $(L^n_{1,b}, L^n_{2,b})$ depends on the process $Z^n_{1,2}$ 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

$$U^n \equiv \mu_{2,2}(L^n_1 - \kappa^n) + \mu_{1,2}L^n_2 \quad \text{and} \quad U^n_b \equiv \mu_{2,2}(L^n_{1,b} - \kappa^n) + \mu_{1,2}L^n_{2,b}. \tag{2.5}$$

By the established sample-path stochastic order $(L^n_1, L^n_2) \geq_{st} (L^n_{1,b}, L^n_{2,b})$ and the monotonicity of the linear map in (2.5), we get the associated sample-path stochastic order $U^n \geq_{st} U^n_b$. Moreover, the stochastic process U^n_b is independent of the process $Z^n_{1,2}$, because of the particular linear combination we have chosen for the one-dimensional processes U^n and U^n_b in (2.5). We have chosen that linear combination so that the number of pool-2 agents working on class 1 does not matter.

Now observe that the lower-bound stochastic process U^n_b is a BD process on the set of all integers in $(-\infty, 0]$. The BD process will have both constant birth rate $\lambda^n_b = \mu_{2,2}\lambda^n_1 + \mu_{1,2}\lambda^n_2$ and by the definitions above, the stochastic process U^n_b has death rate

$$\begin{aligned} \mu^n_b &\equiv \mu_{2,2}(\mu_{1,1}m^n_1 + \theta_1\kappa^n + \mu_{1,2}Z^n_{1,2}(t)) \\ &\qquad + \mu_{1,2}(\mu_{2,2}m^n_2 - \mu_{2,2}Z^n_{1,2}(t))) \\ &= \mu_{2,2}(\mu_{1,1}m^n_1 + \theta_1\kappa^n) + \mu_{1,2}\mu_{2,2}m^n_2. \end{aligned} \tag{2.6}$$

As a consequence, for each $n \geq 1$, the drift in U^n_b is

$$\begin{aligned} \delta^n_b &\equiv \lambda^n_b - \mu^n_b = \mu_{2,2}(\lambda^n_1 - m^n_1\mu_{1,1} - \theta_1\kappa^n) \\ &\qquad + \mu_{1,2}(\lambda^n_2 - m^n_2\mu_{2,2}). \end{aligned} \tag{2.7}$$

Hence, after scaling, we get $\delta^n_b/n \to \delta$, where

$$\delta_b \equiv \mu_{2,2}(\lambda_1 - m_1\mu_{1,1} - \theta_1\kappa) + \mu_{1,2}(\lambda_2 - m_2\mu_{2,2}) > 0, \tag{2.8}$$

with the inequality following from Assumption 1.

Now we observe that $-U^n_b$ is equivalent to the number in system in a stable $M/M/1$ queueing model with traffic intensity $\rho^n_* \to \rho_* < 1$. Let Q_* be the number-in-system process in an $M/M/1$ system having arrival rate equal to $\lambda_* \equiv \mu_{2,2}(m_1\mu_{1,1} + \theta_1\kappa) + \mu_{1,2}m_2\mu_{2,2}$, 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^n_b is tantamount to accelerating time by a factor of order $O(n)$ in Q_*. That is, $\{-U^n_b(t) : t \geq 0\}$ can be represented as $\{Q_*(c_n t) : t \geq 0\}$, where $c_n/n \to 1$ as $n \to \infty$.

Let $M_*(t) \equiv \|Q_*\|_t$. We can now apply the extreme-value result in Theorem 7.2 for the $M/M/1$ queue above (since an $M/M/1$ is trivially a QBD) to conclude that $M_*(t) = O_P(\log(t))$. This implies that $U_b^n/\log(n)$ is SB.

From the way that the reflecting upper barriers were constructed, we know at the outset that $L_{1,b}^n(t) \le \kappa^n$ and $L_{2,b}^n(t) \le 0$. Hence, we must have both $(\kappa^n - L_{1,b}^n)^+$ and $(-L_{2,b}^n)^+$ nonnegative. Combining this observation with the result that $(U_b^n)/\log n$ is SB, we deduce first that both $(\kappa^n - L_{1,b}^n)^+/\log n$ and $(-L_{2,b}^n)^+/\log n$ are SB, so that both $I_1^n/\log n$ and $I_2^n/\log n$ are SB as well. ∎

APPENDIX C: THE BOUNDING QBD IN LEMMA 7.4

In this section we add some more supporting detail to §7. In particular, we now describe how to present the process $\tilde{D}_*^n \equiv jQ_*^n - kQ_*^n$ in the proof of Lemma 7.4 as a QBD for each n. To that end, let $m \equiv j \vee k$. We divide the state space $\mathbb{N} \equiv \{0, 1, 2, \dots\}$ into level of size m: Denoting level i by $L(i)$, we have

$$\begin{aligned} L(0) &= (0, 1, \dots, m-1) \\ L(1) &= (m, m+1, \dots, 2m-1) \quad \text{etc.} \end{aligned}$$

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

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

(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

$$B^{(n)} = \begin{pmatrix} -\sigma^n & 0 & \hat{\lambda}_2^n \\ \hat{\mu}_\Sigma^n & -\sigma^n & 0 \\ \hat{\mu}_\Sigma^n & 0 & -\sigma^n \end{pmatrix}, \qquad A_0^{(n)} = \begin{pmatrix} \hat{\lambda}_3^n & 0 & 0 \\ \hat{\lambda}_2^n & \hat{\lambda}_3^n & 0 \\ 0 & \hat{\lambda}_2^n & \hat{\lambda}_3^n \end{pmatrix},$$

$$A_1^{(n)} = \begin{pmatrix} -\sigma^n & 0 & \hat{\lambda}_2^n \\ 0 & -\sigma^n & 0 \\ \hat{\mu}_2^n & 0 & -\sigma^n \end{pmatrix}, \qquad A_2^{(n)} = \begin{pmatrix} \hat{\mu}_3^n & \hat{\mu}_2^n & 0 \\ 0 & \hat{\mu}_3^n & \hat{\mu}_2^n \\ 0 & 0 & \hat{\mu}_2^n \end{pmatrix},$$

where $\hat{\mu}^n_{\Sigma} \equiv \hat{\mu}^n_3 + \hat{\mu}^n_2$ and $\sigma^n \equiv \hat{\mu}^n_{\Sigma} + \hat{\lambda}^n_2 + \hat{\lambda}^n_3$.

Let $A^{(n)} \equiv A^{(n)}_0 + A^{(n)}_1 + A^{(n)}_2$. 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 [29], the QBD is positive recurrent if and only if

$$\nu A^{(n)}_0 \mathbf{1} < \nu A^{(n)}_2 \mathbf{1},$$

where $\mathbf{1}$ is the vector of all 1's. This translates to the stability condition given in the proof of Lemma 7.4.

APPENDIX D: MORE ON THE IDLENESS PROCESSES

In this section we present additional results about the idleness processes, going beyond Theorem 7.4. We treat pools 1 and 2 in the following subsections.

D.1. The Idleness Process in Pool 1.

We now show how to analyze the idleness in pool 1 without paying attention to what happens in pool 2. This provides a more elementary derivation of the results for I^n_1 in Theorem 7.4.

We start by showing that Q^n_1 is never "too much" below κ^n if κ^n is large enough, where "large enough" in our setting is $\kappa^n/\log(n) \to \infty$ as $n \to \infty$. Since the thresholds in FQR-T are of order greater than $O(\sqrt{n})$, this includes the case in which the thresholds are kept throughout (i.e., they are not dropped once they are crossed, so that $\kappa^n = k^n_{1,2}$), and the case in which κ^n is the centering constant used in shifted FQR-T, where $\kappa^n/n \to \kappa > 0$.

For $t \in \mathbb{R}^+$, let $\lfloor t \rfloor$ be the integer part of t, i.e., the largest integer smaller than t. Let

$$\rho_* \equiv \frac{\mu_{1,1} m_1 + \theta_1 \kappa}{\lambda_1} < 1, \tag{4.1}$$

where the inequality follows from Assumption 1.

We define the difference-process

$$E^n_1 \equiv \kappa^n - Q^n_1. \tag{4.2}$$

We will focus on the positive part: $(E^n_1)^+(t) \equiv \max\{E^n_1(t), 0\}$.

Lemma D.1. *If $\kappa^n/\log(n) \to \infty$ as $n \to \infty$, then $(E^n_1)^+/\log(n)$ is SB.*

Proof. To prove the statement, we will use a stochastic bound argument for Q_1^n. Specifically, we will bound Q_1^n from below in sample-path stochastic order by the queue-length process of an $M/M/m_1^n/\kappa^n + M$ system having a finite buffer of size κ^n, arrival rate λ_1^n, service rate $\mu_{1,1}$ and abandonment rate θ_1. This stochastic-order lower bound for Q_1^n allows us to consider the service process in pool 1 alone, ignoring pool 2. The idea is that Q_1^n is the smallest possible (stochastically), when there are always available servers in pool 2 to ensure that queue 1 never goes above κ^n. In that case, Q_1^n is equivalent to the queue-length process in the $M/M/m_1^n/\kappa^n + M$ model.

In the bounding system, every arriving customer who finds κ^n customers waiting in queue is blocked and lost. Let Q_b^n and Z_b^n (the subscript b is for blocking) denote the number of customers in queue and the number of customers in service, respectively, in the $M/M/m_1^n/\kappa^n + M$ system. Let $\bar{Q}_b^n$ and $\bar{Z}_b^n$ denote the associated sequence of fluid-scaled processes. Also let the initial condition be $Q_b^n(0) = \min\{\kappa^n, Q_1^n(0)\}$ and $Z_b^n(0) = Z_{1,2}^n(0)$ for all n. From the definition of $Q_b^n(0)$ and Assumption 3, we see that $Q_b^n(0) = \kappa^n$ for all n. Hence, $\bar{Q}_b^n(0) \to \kappa$ and $\bar{Z}_b^n(0) \to z_b(0) = z_{1,2}(0)$ as $n \to \infty$.

We can bound the process Q_1^n from below by Q_b^n in the sense of sample-path stochastic order; i.e., for each n, it is possible to construct stochastic processes $\tilde{Q}_b^n$ and $\tilde{Q}_1^n$ on a common probability space, with $\tilde{Q}_b^n$ having the same distribution as Q_b^n, $\tilde{Q}_1^n$ having the same distribution as Q_1^n, and every sample path of $\tilde{Q}_b^n$ lies below the corresponding sample path of $\tilde{Q}_1^n$. The stochastic bound is constructed directly by generating the same arrival processes to both systems. We let departures from service coincide in both systems whenever $Z_b^n = Z_{1,1}^n$. Similarly, we let abandonments from Q_1^n coincide with abandonments from Q_b^n whenever both queues are equal. The argument follows the reasonings in Theorems 6 and 9 in [45].

As explained above, $Q_b^n(0) = \kappa^n$ for all n. Consider the (nonnegative) difference process $E_b^n \equiv \kappa^n - Q_b^n$. Similar to our construction of the bounding process above, we can bound E_b^n from above, in sample-path stochastic order, by an $M/M/1$ system having arrival rate $\mu_{1,1}m_1^n + \theta_1\kappa^n$ and service rate λ_1^n, i.e., denoting sample-path stochastic order by $\le_{st}$, for each n and for all $t \ge 0$, we have

(4.3)
$$E_b^n(t) \le_{st} Q_*^n(t) = N_*^a\Big((\mu_{1,1}m_1^n + \theta_1\kappa^n)t\Big) - N_*^s\left(\lambda^n \int_0^t 1_{\{Q_*^n(s)>0\}}\, ds\right),$$

where N_*^a and N_*^s are two independent rate-1 Poisson processes, and Q_*^n is the number-in-system process in the n^{th} $M/M/1$ system (customers in queue and in service).

Let Q_* be the number-in-system process in an $M/M/1$ system having arrival rate equal to $\mu_{1,1}m_1 + \theta_1\kappa$ and service rate λ_1, so that ρ_* in (4.1) is the traffic intensity to Q_*, and $\rho_* < 1$. Observe that the effect of increasing the size of the $M/M/m_1^n/\kappa^n + M$ system and its arrival rate (by increasing m_1^n, κ^n and λ_1^n) is tantamount to accelerating time by a factor of order $O(n)$ in Q_*. That is, $\{E_b^n(t) : t \geq 0\}$ is stochastically bounded from above (in sample-path stochastic order) by $\{Q_*(c_n t) : t \geq 0\}$, where $c_n/n \to 1$ as $n \to \infty$, for every $t \geq 0$. We can now apply extreme-value theory for the $M/M/1$ queue. In particular, if we let $M_*(t) \equiv \max\{Q_*(s) : 0 \leq s < t\}$, then $\|E_b^n\|_t$ is bounded from above, in the sample-path stochastic-order sense, by the process $M_*(c_n t)$.

Since the queue length is discrete, with a geometric stationary distribution, a standard extreme-value limit does not exist. Nevertheless, we can bound the lim sup above; in particular, it follows from Theorem 6 in [2] and the example following it, (see also Problem 4.2 pg. 185 of [4]), that, for $c = [(\mu_{1,1}m_1 - \theta_1\kappa)(1-\rho_*)]^{-1} > 0$,

$$\begin{aligned}
&\lim_{x\to\infty} \limsup_{t\to\infty} P(M_*(t) - a\log(t) + b(t) > x) \\
&= 1 - \lim_{x\to\infty} \liminf_{t\to\infty} P(M_*(t) - a\log(t) + b(t) \leq x) \\
&\leq 1 - \lim_{x\to\infty} e^{-\rho_*^{x-1}/c} = 0,
\end{aligned}$$

where

$$a \equiv \frac{1}{-\log(\rho_*)}, \qquad b(t) \equiv \frac{\log(t) - \log\lfloor t\rfloor - \log(1-\rho_*)}{-\log(\rho_*)}$$

and $b(t) \to -\log(1-\rho_*)/\log(\rho_*)$ as $t \to \infty$. The last inequality is the result in [2].

Hence, $M_*(t) = O_P(\log(t))$. Since $\|E_b^n\|_T$ is stochastically smaller than $M_*(c_n T)$, where $c_n/n \to 1$, we have that $\|E_b^n\|_T/\log(n)$ is stochastically bounded for all $T > 0$. The desired result then follows from the fact that $(E_1^n)^+$ is itself stochastically smaller than E_b^n. ∎

From the fact that $(E_1^n)^+$, is at most of order $O_P(\log(n))$ when $\kappa^n/\log(n) \to \infty$, we deduce that, asymptotically, there are always customers waiting in the class-1 queue. The following corollary is immediate:

COROLLARY D.1. *Under the conditions of Lemma D.1, for any $T > 0$,*

$$\lim_{n\to\infty} P\left(\inf_{0\leq t\leq T} Q_1^n(t) > 0\right) = 1, \quad \textit{so that} \quad \lim_{n\to\infty} P\left(\sup_{0\leq t\leq T} I_1^n(t) > 0\right) = 0.$$

We now treat the case in which $\kappa^n/\log(n) \to c$, where $c < \infty$, which is the only other case with $\kappa \geq 0$ by virtue of 2. Since the order of size of the thresholds in FQR-T is greater than $O(\sqrt{n})$, we are mainly concerned with the case in which the thresholds are dropped once they are crossed, and FQR is employed. That is, the main case is $\kappa^n = 0$ for all n.

PROPOSITION D.1. *If $\kappa^n/\log(n) \to c$, where $0 \leq c < \infty$, then $I_1^n/\log(n)$ is SB.*

Proof. The proof is similar to the proof of Lemma D.1. If we prove the result for any bounded sequence, then the result will follow trivially for any unbounded sequence. We thus assume that $0 \leq \kappa^n \leq M < \infty$. We use the same sample-path stochastic-order $M/M/1$-bound Q_*^n in (4.3) to bound I_1^n, only now we replace κ^n with M in the representation (4.3). Since M becomes negligible relative to the scaling by n as n increases, the traffic intensity for the process Q_*, defined in the proof of Lemma D.1, is $\rho_* = \mu_{1,1} m_1/\lambda_1$, so that $\rho_* < 1$ by Assumption 1. Hence, the bound M_* in the proof of Lemma D.1, applies to I_1^n. ∎

We can combine Corollary D.1 and Proposition D.1. To that end, we define the process

$$L_1^n \equiv Q_1^n + Z_{1,1}^n - m_1^n. \tag{4.4}$$

Observe that $(L_1^n)^+ \equiv Q_1^n$ and $(L_1^n)^- \equiv I_1^n$, so that $I_1^n \leq (\kappa^n - L_1^n)^+$ w.p. 1.

COROLLARY D.2. *The sequence $(\kappa^n - L_1^n)^+/\log(n)$ is SB. Hence, $I_1^n/\log(n)$ is SB.*

D.2. The Idleness Process in Pool 2. We now turn to the pool-2 idleness process. We establish a stronger property away from the time origin.

PROPOSITION D.2. *For all ϵ and T satisfying $0 < \epsilon < T < \infty$,*

$$P\Big(\sup_{\epsilon \leq t \leq T} I_2^n(t) > 0\Big) \to 0 \quad as \quad n \to \infty.$$

Proof. Much of the argument here repeats the proof of Theorem 7.4. For the first statement, we will create a stochastic lower bound and show that it satisfies the statement. We will exploit a linear combination of processes associated with the two queues. For that purpose, we define the process

$$L_2^n \equiv Q_2^n + Z_{1,2}^n + Z_{2,2}^n - m_2^n, \tag{4.5}$$

representing the excess number in system for class 2. Then let U^n be the linear combination of the processes L_i^n, $i = 1, 2$, defined in (4.4) and (4.5):

$$U^n \equiv \mu_{2,2}(L_1^n - \kappa^n) + \mu_{1,2} L_2^n. \tag{4.6}$$

As we will explain below, this provides a one-dimensional view that can be regarded as independent of the customer assignments for pool 2.

Because of our FQR (or shifted FQR) routing rule, $L_1^n(t) > \kappa^n$ implies that $L_2^n(t) \geq 0$. If $U^n(t) > 0$, then necessarily we must have either $L_1^n(t) > \kappa^n$ or $L_2^n(t) > 0$, and so either $Q_1^n(t) > \kappa^n$ or $Q_2^n(t) > 0$. If either of those events holds, then necessarily we must have $I_2^n(t) = 0$. Hence, we will show that $P(B^n) \to 1$ as $n \to \infty$, where $B^n \equiv \{\sup_{\epsilon \leq t \leq T} U^n(t) > 0\}$.

Just as in the proof of Lemma D.1, we will bound the process U^n in (4.6) below in sample-path stochastic order by another process, U_b^n, a one-dimensional birth-and-death (BD) process. As a first step, we give U_b^n the same Poisson arrival processes as the original system has. Thus, U_b^n has constant birth rate $\lambda_b^n \equiv \mu_{2,2}\lambda_1^n + \mu_{1,2}\lambda_2^n$.

We next bound the pair of processes (L_1^n, L_2^n) below in sample-path stochastic order by another two-dimensional process $(L_{1,b}^n, L_{2,b}^n)$. We construct the lower-bound process$(L_{1,b}^n, L_{2,b}^n)$ 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 other. 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_1^n at $\kappa^n + \epsilon_1 n$; we place the reflecting upper barrier on L_2^n at $\epsilon_1 n$. With the upper barrier at $\epsilon_1 n$, the departure rate of L_1^n is bounded above by $\mu_{1,1} m_1^n + \theta_1 \kappa^n + \theta_1 \epsilon_1 n + \mu_{1,2} Z_{1,2}^n(t)$, based on assuming that pool 1 is fully busy serving class 1, that L_1^n 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. Second, with the upper barrier at $\epsilon_1 n$, the departure rate of L_2^n is bounded above by $\mu_{2,2} m_2^n + \theta_2 \epsilon_1 n - \mu_{1,2} Z_{1,2}^n(t)$, 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 is at its upper barrier $\epsilon_1 n$. Thus, we give $L_{1,b}^n$ and $L_{2,b}^n$ these bounding rates at all times

Of course, 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. However, we can avoid this difficulty by looking at the special linear combination in (2.5); i.e., we define the associated process

$$U_b^n \equiv \mu_{2,2}(L_{1,b}^n - \kappa^n) + \mu_{1,2} L_{2,b}^n. \tag{4.7}$$

By the sample-path stochastic order $(L_1^n, L_2^n) \geq_{st} (L_{1,b}^n, L_{2,b}^n)$, we get the associated sample-path stochastic order $U^n \geq_{st} U_b^n$. Moreover, the stochastic process U_b^n is independent of the process $Z_{1,2}^n$, because of the particular linear combination we have chosen for the one-dimensional processes U^n and U_b^n in (2.5) and (4.7). We have chosen that linear combination so that the number of pool-2 agents working on class 1 does not matter.

Now observe that the lower-bound stochastic process U_b^n is a BD process on the set of all integers in $(-\infty, (\mu_{2,2}+\mu_{1,2})\epsilon_1 n]$. The BD process will have both constant birth rate λ_b^n defined above and constant death rate μ_b^n. The important point is that we will choose ϵ_1 so small that the constant drift $\delta^n \equiv \lambda_b^n - \mu_b^n$ is strictly positive for all suitably large n. To achieve the positive drift below, we will rely heavily on the overload assumption, 1.

By the definitions above, the stochastic process U_b^n has death rate

$$
\begin{aligned}
\mu_b^n &\equiv \mu_{2,2}(\mu_{1,1}m_1^n + \theta_1\kappa^n + \theta_1\epsilon_1 n + \mu_{1,2}Z_{1,2}^n(t)) \\
&\qquad +\mu_{1,2}(\mu_{2,2}m_2^n + \theta_2\epsilon_1 n - \mu_{2,2}Z_{1,2}^n(t))) \\
&= \mu_{2,2}(\mu_{1,1}m_1^n + \theta_1\kappa^n) + \mu_{1,2}\mu_{2,2}m_2^n + (\mu_{2,2}\theta_1 + \mu_{1,2}\theta_2)\epsilon_1 n. \qquad (4.8)
\end{aligned}
$$

As a consequence, for each $n \geq 1$, the drift in U_b^n is

$$
\begin{aligned}
\delta_b^n &\equiv \lambda_b^n - \mu_b^n = \mu_{2,2}(\lambda_1^n - m_1^n\mu_{1,1} - \theta_1\kappa^n) \\
&\qquad +\mu_{1,2}(\lambda_2^n - m_2^n\mu_{2,2}) + (\mu_{2,2}\theta_1 + \mu_{1,2}\theta_2)\epsilon_1 n. \qquad (4.9)
\end{aligned}
$$

Hence, after scaling, we get $\delta_b^n/n \to \delta$, where

$$\delta_b \equiv \mu_{2,2}(\lambda_1 - m_1\mu_{1,1} - \theta_1\kappa) + \mu_{1,2}(\lambda_2 - m_2\mu_{2,2}) + (\mu_{2,2}\theta_1 + \mu_{1,2}\theta_2)\epsilon_1. \qquad (4.10)$$

By Assumption 1, we see that we would have $\delta_b > 0$ if $\epsilon_1 = 0$. However, because of the strict inequality in Assumption 1, we can always choose ϵ_1 sufficiently small, so that $\delta_b > 0$, and we do that.

Now we can establish a FWLLN for U_b^n. Such a FWLLN is elementary since the BD process has constant birth and death rates with positive drift. After exploiting the fact that we start at $L_1^n(0) = \kappa^n$ and $L_2^n(0) = 0$, so that $U_b^n(0) = U^n(0) = 0$, we see that

$$\bar{U}_b^n \Rightarrow u_b \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \qquad (4.11)$$

where

$$\bar{U}_b^n(t) \equiv U_b^n(t)/n \quad \text{and} \quad u_b(t) \equiv \delta_b t \wedge \epsilon_1 \quad \text{for} \quad t \geq 0, \qquad (4.12)$$

with $u_b(0) = 0$.

As a consequence, we deduce that, for any ϵ and T with $0 < \epsilon < T < \infty$,

$$P(\inf_{\epsilon \leq t \leq T} U^n(t) > 0) \to 1 \quad \text{as} \quad n \to \infty. \qquad (4.13)$$

Next, we recall that on the subset in the underlying probability space for which $\inf_{\epsilon \le t \le T} U^n(t) > 0$, we must have, for each t, that either $Q_1^n(t) > \kappa^n$ or $Q_2^n(t) > 0$. However, either one of these inequalities implies that $I_2^n(t) = 0$. Thus the idleness must be 0 throughout the interval $[\epsilon, T]$. Hence we have established the proposition. ∎

APPENDIX E: REMAINING PROOFS IN SECTION 8

E.1. Remaining Proofs in §8.1.

Proof of Lemma 8.1. For background on tightness, see [8, 36, 48]. 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 $\mathcal{D}_k$, 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 [8]. Thus it suffices to verify conditions (6.3) and (6.4) of Theorem 11.6.3 of [48]. Hence, it suffices to prove SB of the sequence of stochastic processes evaluated at time 0 and appropriately control the oscillations, using the modulus of continuity on $\mathcal{C}$. We obtain the stochastic boundedness at time 0 immediately from Assumption 3 in §3. We show that we can control the oscillations below. The resulting tightness implies that the sequence of stochastic processes is SB. We give an alternative proof of SB in §8.2, which yields explicit bounds on the limit processes.

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

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

Using the representations (4.2)-(4.5), for $t_2 > t_1 \ge 0$ we have

$$\begin{aligned}|\bar{Q}_1^n(t_2) - \bar{Q}_1^n(t_1)| \le \frac{A_1^n(t_2) - A_1^n(t_1)}{n} &+ \frac{\int_{t_1}^{t_2} 1_{\{D^n(s) > 0\}}\, dS^n(s)}{n} \\ &+ \frac{\int_{t_1}^{t_2} 1_{\{D^n(s) \le 0\}}\, dS_{1,1}^n}{n} + \frac{U_1^n(t_2) - U_1^n(t_1)}{n},\end{aligned}$$

Hence, for any $\zeta > 0$ and $T > 0$,

$$\begin{aligned} w(Q_1^n/n, \zeta, T) \quad &\le \quad w(A_1^n/n, \zeta, T) + w(S^n/n, \zeta, T) + w(S_{1,1}^n/n, \zeta, T) \\ &\quad + w(U_1^n/n, \zeta, T). \end{aligned} \tag{5.1}$$

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

(4.2),

$$w(S^n_{i,j}/n,\zeta,T) \le w(N^s_{i,j}(n\mu_{i,j}m_j\cdot)/n,\zeta,T) \le w(N^s_{i,j}(n\cdot)/n,c\zeta,T) \tag{5.2}$$

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

$$\begin{aligned} Q^n_i(t) &\le Q^n_i(0) + A^n_i(t), \\ |U^n_i(t_2) - U^n_i(t_1)| &= |N_i(\theta_i \int_{t_1}^{t_2} Q^n_i(s)\,ds| \\ &\le |N_i(n\theta(\bar{Q}^n_i(0) + \bar{A}^n_i(T)(t_2-t_1))|. \end{aligned} \tag{5.3}$$

Let $q_{bd} \equiv 2(q_i(0)+T)$, where $\bar{Q}^n_i(0) \Rightarrow q_i(0)$ by Assumption 3, and let B_n be the following subset of the underlying probability space:

$$B_n \equiv \{\bar{Q}^n_i(0) + \bar{A}^n_i(T) \le q_{bd}\}.$$

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

$$w(U^n_i/n,\zeta,T) \le w(N^u_i(nq_{bd}\cdot)/n,\zeta,T) \le w(N^u_i(n\cdot)/n,c\zeta,T) \tag{5.4}$$

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 $n \ge n_0$, $P(B_n) > 1 - \eta/2$ and on B_n

$$\begin{aligned} w(Q^n_i/n,\zeta,T) &\le w(N^a_i(n\cdot)/n,c\zeta,T) + 2\sum_{i=1}^{2}\sum_{j=1}^{2} w(N^s_{i,j}(n\cdot)/n,c\zeta,T) \\ &\quad + w(N^u_i(n\cdot)/n,c\zeta,T). \end{aligned} \tag{5.5}$$

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

$$P(w(Q^n_i/n,\zeta,T) \ge \epsilon) \le \eta \quad \text{for all} \quad n \ge n_0.$$

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

We now turn to the sequence $\{\bar{Z}^n_{1,2}\}$. Let $A^n_{1,2}(t)$ 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 $\bar{A}^n_{1,2} \equiv A^n_{1,2}/n$ and $\bar{S}^n_{1,2}(t) \equiv S^n_{1,2}(t)/n$, for $S^n_{1,2}(t)$ in (4.2). Since

$$Z^n_{1,2}(t) = Z^n_{1,2}(0) + A^n_{1,2}(t) - S^n_{1,2}(t),$$

we have

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

However, for A^n_1 in (4.2),

$$A^n_{1,2}(t_2) - A^n_{1,2}(t_1) \le A^n_1(t_2) - A^n_1(t_1).$$

Since $\bar{A}^n_1 \Rightarrow \lambda_1 e$ in $\mathcal{D}$, the sequence $\{\bar{A}^n_1\}$ is tight. Together with (5.2), that implies that the sequence $\{\bar{Z}^n_{1,2}\}$ is tight as well. Finally, we observe that the tightness of $\{\bar{Y}^n_8\}$ follows from (5.2), (5.4) and the convergence of $\bar{A}^n_i$. ∎

Proof of Lemma 8.2. Apply the bounds on the modulus of continuity involving Poisson processes in the proof of Lemma 8.1 above. For a Poisson process N, let $\hat{N}^n \equiv \sqrt{n}(\bar{N}^n - e)$, where $\bar{N}^n(t) \equiv N(nt)/n$, $t \ge 0$. By the triangle inequality, for each n, ζ, and T,

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

Since, $w(x, \zeta, T)$ is a continuous function of x for each fixed ζ and T, we can apply this bound with the inequalities in the proof of Lemma 8.1 to deduce (8.1). ∎

E.2. Remaining Proof in §8.3.

Proof of Lemma 8.7. Consider the drift rates of the QBD-version of D^n_f in (5.6), and observe that, by the linearity of the drift expressions and Assumption 3, $\delta^n_+(X^n(0))/n \Rightarrow \delta_+(x(0))$ and $\delta^n_-(X^n(0))/n \Rightarrow \delta_-(x(0))$ for δ_+ and δ_- in (5.20). Also by Assumption 3, $x(0) \in \mathbb{A}$ so that (5.21) holds. This implies that there exists $\eta > 0$ such that

$$\lim_{n\to\infty} P(\delta^n_+(X^n(0)) < -\eta \quad \text{and} \quad \delta^n_-(X^n(0)) > \eta) = 1,$$

i.e., (8.11) holds at $t = 0$ with probability converging to 1 as $n \to \infty$.

To prove the lemma, we bound the drifts in (5.6). We do that by bounding the change in the components of $X^n(t)$ in a short interval after time 0. To do that, we use the stochastic-order bounds in (7.4)-(7.5). Recall the rather special ordering obtained there:

$$(5.6) \qquad (-Q^n_{1,a}, Q^n_{2,a}, Z^n_a) \le_{st} (-Q^n_1, Q^n_2, Z^n_{1,2}) \le_{st} (-Q^n_{1,b}, Q^n_{2,b}, Z^n_b).$$

In particular, we will find two processes X^n_+ and X^n_- in $\mathcal{D}$, such that

$$(5.7) \qquad \delta^n_+(X^n(t)) \le_{st} \delta^n_+(X^n_+(t)), \qquad \delta^n_-(X^n(t)) \ge_{st} \delta^n_-(X^n_-(t))$$

and, for some $\delta > 0$ and $\eta > 0$,

$$\lim_{n\to\infty} P\left(\sup_{t\in[0,\xi]} \delta_+^n(X_+^n(t)) < -\eta \quad \text{and} \quad \inf_{t\in[0,\xi]} \delta_-^n(X_-^n(0)) > \eta\right) = 1. \tag{5.8}$$

To construct the processes X_+^n and X_-^n with these properties, we use the bounding processes X_a^n and X_b^n in (7.4) and (7.5) (appearing again in (5.6)). Specifically, we let

$$X_+^n \equiv (Q_{1,a}^n, Q_{2,b}^n, Z_+^n) \quad \text{and} \quad X_-^n \equiv (Q_{1,b}^n, Q_{2,a}^n, Z_-^n), \tag{5.9}$$

respectively, where $Z_+^n = Z_b^n$ if $\mu_{2,2} \geq \mu_{1,2}$, and $Z_+^n = Z_a^n$ otherwise. $Z_-^n = Z_a^n$ if $\mu_{2,2} \geq \mu_{1,2}$, and $Z_-^n = Z_b^n$ otherwise. As a consequence, for each $t \geq 0$, the drifts satisfy

$$\begin{aligned}
\delta_+^n(X_+^n(t)) &\equiv j[\lambda_1^n - \mu_{1,1}m_1^n - (\mu_{1,2} - \mu_{2,2})Z_+^n(t) - \mu_{2,2}m_2^n - \theta_1 Q_{1,b}^n(t)] \\
&\quad - k[\lambda_2^n - \theta_2 Q_{2,a}^n(t)], \\
\delta_-^n(X_-^n(t)) &\equiv j[\lambda_1^n - \mu_{1,1}m_1^n - \theta_1 Q_{1,a}^n(t)] \\
&\quad - k[\lambda_2^n - (\mu_{1,2} - \mu_{2,2})Z_-^n(t) - \mu_{2,2}m_2^n - \theta_2 Q_{2,b}^n(t)].
\end{aligned} \tag{5.10}$$

We have directly defined the processes in (5.9) to ensure that the inequalities in (5.7) are satisfied.

Assume that $X_+^n(0) = X_-^n(0) = X^n(0)$. By Assumption 3, $\bar{X}^n(0) \Rightarrow x(0)$ as $n \to \infty$, so that the condition in Lemma 7.1 holds at $t = 0$. Hence, by Lemma 7.1, $\bar{X}_+^n \Rightarrow x_+ \equiv (q_{1,b}, q_{2,a}, z_+)$, where $z_+ = z_a$ if $\mu_{2,2} \geq \mu_{1,2}$ and $z_+ = z_b$ otherwise. Also, $\bar{X}_-^n \Rightarrow x_- \equiv (q_{1,a}, q_{2,b}, z_-)$, where $z_- = z_b$ if $\mu_{2,2} \geq \mu_{1,2}$ and $z_- = z_a$ otherwise. Hence, by the linearity of the functions δ_+^n and δ_-^n,

$$\delta_+^n(X_+^n)/n \Rightarrow \delta_+(x_+) \quad \text{and} \quad \delta_-^n(X_-^n)/n \Rightarrow \delta_-(x_-) \quad \text{in } \mathcal{D} \text{ as } n \to \infty. \tag{5.11}$$

Since $x_+(0) = x_-(0) = x(0) \in \mathbb{A}$, and by the continuity of $\delta_+(\cdot)$ and $\delta_-(\cdot)$, we can find $\xi > 0$ and $\eta > 0$, such that $\delta_+(x_+(t)) < -\eta$ and $\delta_-(x_-(t)) > \eta$. for all $t \in [0, \xi]$. That implies that we have (5.8). Together with (5.7), that concludes the proof. ∎

E.3. Remaining Proof in §8.5.

Proof of Lemma 8.9. We can apply essentially the same reasoning as in the proof of Lemma 8.7. We only need to change the order. Now we aim to achieve:

$$\begin{aligned} &\delta^n_+(X^n_m(t)) \le \delta^n_+(X^n(t)) \le \delta^n_+(X^n_M(t)), \quad \text{and} \\ &\delta^n_-(X^n_m(t)) \le \delta^n_-(X^n(t)) \le \delta^n_-(X^n_M(t)) \end{aligned} \tag{5.12}$$

instead of (5.7). Moreover, we will do so such that the two bounding QBD's are positive recurrent over some interval $[0, \xi]$ on the sets B_n where $P(B_n) \to 1$ as $n \to \infty$. In other words, we will use random vectors X^n_M and X^n_m instead of full processes.

We again use the stochastic-order bounds in (7.4)-(7.5), with the ordering in (5.6). To construct X^n_M, let

$$X^n_{M^+} \equiv (Q^n_{1,M}, Q^n_{2,M}, Z^n_{M^+}) \quad \text{and} \quad X^n_{M^-} \equiv (Q^n_{1,M}, Q^n_{2,M}, Z^n_{M^-}), \tag{5.13}$$

where

$$\begin{aligned} Q^n_{1,M} &\equiv \inf_{0\le t\le\xi} Q^n_{1,b}(t) \vee 0, \quad Q^n_{2,M} \equiv \|Q^n_{2,b}\|_\xi, \\ Z^n_{M^+} &\equiv \inf_{0\le t\le\xi} Z^n_+(t), \qquad\quad Z^n_{M^-} \equiv \|Z^n_-\|_\xi, \end{aligned} \tag{5.14}$$

with $Z^n_+(t) \equiv Z^n_b$ and $Z^n_-(t) \equiv Z^n_a$ if $\mu_{2,2} \ge \mu_{1,2}$, and $Z^n_+(t) \equiv Z^n_a$ and $Z^n_-(t) \equiv Z^n_b$ otherwise. Note that we can regard $\{X^n_{M^+}(\xi) : \xi \ge 0\}$ as a stochastic process as a function of ξ, but we work with the final value $X^n_{M^+} \equiv X^n_{M^+}(\xi)$, and similarly for $X^n_{M^-}$. Let $\{D^n_f(X^n_M, s) : s \ge 0\}$ have the rates determined by $X^n_{M^-}$ when $D^n_f(X^n_M, s) \le 0$, and the rates determined by $X^n_{M^+}$ when $D^n_f(X^n_M, s) > 0$.

We do a similar construction for X^n_m. Let

$$X^n_{m^+} \equiv (Q^n_{1,m}, Q^n_{2,m}, Z^n_{m^+}) \quad \text{and} \quad X^n_{m^-} \equiv (Q^n_{1,m}, Q^n_{2,m}, Z^n_{m^-}), \tag{5.15}$$

where

$$\begin{aligned} Q^n_{1,m} &\equiv \|Q^n_{1,a}\|_\xi, \quad Q^n_{2,m} \equiv \inf_{0\le t\le\xi} Q^n_{2,b}(t) \vee 0, \\ Z^n_{m^+} &\equiv \|Z^n_+\|_\xi, \qquad Z^n_{m^-} \equiv \inf_{0\le t\le\xi} Z^n_-(t). \end{aligned} \tag{5.16}$$

with $Z^n_+(t) \equiv Z^n_a$ and $Z^n_-(t) \equiv Z^n_b$ if $\mu_{2,2} \ge \mu_{1,2}$, and $Z^n_+(t) \equiv Z^n_b$ and $Z^n_-(t) \equiv Z^n_a$ otherwise (the reverse of what is done in (5.14)). Let $\{D^n_f(X^n_m, s) : s \ge 0\}$ have the rates from $X^n_{m^-}$ when $D^n_f(X^n_m, s) \le 0$, and the rates from $X^n_{m^+}$ when $D^n_f(X^n_m, s) > 0$. By this construction, we achieve the ordering in (8.18). We cover the rates of $D^n_{1,2}(t)$ too because we can make the identification: the rates of $D^n_{1,2}(t)$ given $X^n(t)$ coincide with the rates of $D^n_f(X^n(t), \cdot)$.

It remains to find a ξ such that both the processes $\{D^n_f(X^n_m, s) : s \ge 0\}$ and $\{D^n_f(X^n_M, s) : s \ge 0\}$ are positive recurrent. To do so, we will use a minor modification of the reasoning in the final step of the proof of Lemma 8.7. We use Lemma 7.1, which concludes that the bounding processes as functions of ξ have fluid limits. By Lemma 7.1, we can conclude that $\bar{X}^n_{m^+} \equiv n^{-1}X^n_{m^+} \Rightarrow x^+_m$, $\bar{X}^n_{m^-} \equiv n^{-1}X^n_{m^-} \Rightarrow x^-_m$, $\bar{X}^n_{M^+} \equiv n^{-1}X^n_{M^+} \Rightarrow x^+_M$ and $\bar{X}^n_{M^-} \equiv n^{-1}X^n_{M^-} \Rightarrow x^-_M$ in $\mathcal{D}$, where x_{m^+}, x^-_m, x^+_M and x^-_M are all continuous with $x^+_m(0) = x_m-(0) = x^+_M(0) = x^-_M = x(0) \in \mathbb{A}$. Hence, we can find ξ' such that $x_m(\xi) \in \mathbb{A}$ and $x_M(\xi) \in \mathbb{A}$ for all $\xi \in [0, \xi']$. Hence, we can choose ξ such that the constant vectors $x_m \equiv x_m(\xi)$ and $x_M \equiv x_M(\xi)$ both arbitrarily close to $x(0)$.

Finally, we use the linearity of the drift function to deduce the positive recurrence of the processes depending upon n. As in (5.11), we have

$$\begin{aligned} \delta^n_-(X^n_{m^-})/n &\Rightarrow \delta_-(x^-_m), \quad \delta^n_+(X^n_{m^+})/n \Rightarrow \delta_+(x^+_m), \\ \delta^n_-(X^n_{M^-})/n &\Rightarrow \delta_-(x^-_M), \quad \text{and} \quad \delta^n_+(X^n_{M^+})/n \Rightarrow \delta_+(x^+_M). \end{aligned} \tag{5.17}$$

As a consequence, we can deduce the conclusion of the lemma. ▪

E.4. Remaining Proof in §8.6.

Proof of Lemma 8.10. We start with the processes $D^n_f(X^n_m, \cdot)$ and $D^n_f(X^n_M, \cdot)$ already constructed in §§8.5 and E.3, with the understanding that the interval length ξ will in general need to be redefined, now depending on ϵ. Since the initial state has been frozen in $D^n_f(X^n_m, \cdot)$, $D^n_f(X^n_M, \cdot)$ and $D^n_f(X^n(t), \cdot)$, these three processes are stationary CTMC's (have stationary transition rates), but $D^n_{1,2}(t)$ is not. In the following we construct modified versions of these processes, but so as not to alter their individual distributions. For the following, we regard all the QBD processes as CTMC's and use the natural order on the integer state space (instead of the special order in the QBD structure).

As in the proof of Theorem 5.3, we can apply uniformization. As explained there, without loss of generality, we can regard the transition rates in $D^n_{1,2}$ as being uniformly bounded. Thus, for for all n suitably large, and for each process under consideration, we can generate all potential transitions from constant-rate Poisson processes. Because of the scaling by $O(n)$ in (2.3), the Poisson processes for model n can be given rate αn, $n \geq 1$, for some positive constant α. The constant α is chosen so that the rate αn exceeds the maximum total transition rate out of any state for any of the processes for each $n \geq 1$. Then the actual transitions of the process are governed by a DTMC. The Poisson process generates potential transitions. When there is not a real transition, that is captured in the DTMC by a transition from that state back to itself. By choosing the Poisson transition rate sufficiently large, for every state in the state space, there is positive probability of a one-step transition immediately back to that same state. Hence, the DTMC is aperiodic as well as ireducible and positive recurrent. Note that the Poisson process captures the scaling by n.

For the new construction, we use a regenerative approach, using the regenerative structure discussed in §8.4. Provided that the QBD's $D_f^n(X_M^n, \cdot)$ and $D_f^n(X_m^n, \cdot)$ are positive recurrent, which will hold on $B_n(\xi, \eta)$. By virtue of the construction in §8.5, successive visits to any fixed state constitute regenerative cycles for these stationary CTMC's with constant transition rates. It is convenient to let the regenerative state, denoted by s^*, be contained in the boundary of the QBD.

We use the common initial state, say s^*. For simplicity, we initially assume that

$$D_f^n(X_m^n, 0) = D_f^n(X_M^n, 0) = D_f^n(X^n(t), 0) = D_{1,2}^n(0) = s^*, \tag{5.18}$$

but we will later show that this initial condition is not needed; e.g., it can be replaced by the convergence condition imposed in Assumption 3.

For the new construction, we couple all four processes; i.e., we start by constructing *all* the processes together, starting in their common initial state, based on the rate order established in (8.18). That means that we use a single Poisson process with rate αn to generate potential transitions for all the processes under consideration. We match the actual transitions as much as possible in order to keep the processes evolving together as much as possible. We will choose ξ to ensure that the transition probabilities differ by only a negligible amount, so the processes will only rarely have different transitions during a single regenerative cycle. Even though we cannot achieve full sample path stochastic order for the stochastic processes over the full time interval, we can keep all the processes together over each regenerative cycle, with high probability. (Recall that the number of transitions in each regenerative cycle is of order $O(1)$, but the transitions are occurring at rate $O(n)$, so we are *not* succeeding in keeping the process paths identical over positive time intervals, but that is not needed. Because we are concerned with the integrals in (8.23), it suffices to have the *proportion* of time that the paths are identical be large. Also recall that the inequalities in (8.23) need not hold w.p.1; we are only claiming that the probability that they hold should converge to 1 as $n \to \infty$.)

Our general idea is to construct an alternating renewal process for each n, which involves a sequence $\{(U_{1,k}^n, U_{2,k}^n) : k \geq 1\}$ of i.i.d pairs of nonnegative random variables, $U_{1,k}^n$ and $U_{2,k}^n$. These variables measure times in the full process and so will be $O(1/n)$. The first random variable $U_{1,k}^n$ is the geometric random sum of the cycle lengths of all the regenerative cycles where the processes all coincide, while the second interval $U^n_{2,k}$ is a subsequent interval on which the processes do not necessarily coincide. The second interval ends when all processes are in the regenerative state together. We then repeat the construction. We will make the first interval $U^n_{1,k}$ much longer than the second interval $U^n_{2,k}$, ensuring that the proportion of time that the processes all agree is arbitrarily close to 1 (falling within the ϵ gaps in (8.23)). The cycles will have $O(1)$ transitions, but since the transitions occur according to the Poisson process at rate αn, the cycle lengths are asymptotically negligible, making the limiting proportions all that matters.

With the general strategy laid out, it now remains to show that we can make the first intervals $U^n_{1,k}$ suitably long and make the second intervals $U^n_{2,k}$ relatively short. The construction is more complicated for the second interval $U^n_{2,k}$. The second interval is made up of two parts. The first part of $U^n_{2,k}$ is the exceptional cycle on which the processes first disagree. The second part of $U^n_{2,k}$ starts at the end of that exceptional cycle, where the upper process is in the regenerative state, but in general the other processes are not. At that point, we change the construction. We use *independent* Poisson processes, all with rate αn, to generate the transitions in the four processes. This second part ends when all the processes are simultaneously together in the regenerative state. We start over after the second interval ends, i.e., afterwards we again use a single Poisson process to generate the transitions of all processes, starting when they are all together in the regenerative state, and so forth. In this way we produce the alternating renewal process structure.

We do a careful analysis to ensure that the second random variable $U^n_{2,k}$ is appropriately controlled, independent of ξ, and then we choose ξ suitably small to make the first interval relatively long, so that the long-run proportion of time that the process is in the second interval, which is

$$\frac{E[U^n_{2,k}]}{E[U^n_{1,k}] + E[U^n_{2,k}]}, \tag{5.19}$$

is as small as desired. In fact, our construction will make $E[U^n_{1,k}] \uparrow \infty$ as $\xi \uparrow \infty$, while $E[U^n_{2,k}] \downarrow 0$ as $\xi \uparrow \infty$. Since the Poisson rate αn produces a time scaling of order $O(n)$, the cycles are occurring more rapidly as $n \to \infty$. In that way we can achieve the inequalities in (8.23) with probabiity converging to 1 as $n \to \infty$. Since we are working with indicator functions in (8.23), in computing the bound we allow the worst case, in which the indicator functions differ by 1 throughout the second interval.

We now present the details. Let the random number of transitions in a regenerative cycle for the upper bound process $D_f^n(X_M^n,\cdot)$ be N^n. Since the events are occurring at rate of order $O(n)$, we can use a version of the time-expanded queue-difference process for $D_f^n(X_M^n,\cdot)$, as in (5.7). By Theorem 5.3, we have $N^n \Rightarrow N$ as $n \to \infty$, where N is the corresponding random number of transitions during a regenerative cycle for the FTSP $D(x_M,\cdot)$, using the same designated regenerative state, where $\bar{X}_M^n \Rightarrow x_M$ as $n \to \infty$, as in §8.5. Moreover, because of the special QBD structure we also have additional regularity properties.

Let p_n be the probability mass function of N^n, i.e., $p_n(k) \equiv P(N^n = k)$. As in §8.4, From the convergence $N^n \Rightarrow N$ and the QBD structure of all processes, we know that p_n has a proper generating function (gf) $\psi_{N^n}(z) \equiv E\left[z^{N^n}\right]$. Combining the QBD and gf structure, we can conclude that there is an integer k_0 such that we can bound the probabilities $p_n(k)$ above and below by

$$C_L\tilde{q}^k \le p_n(k) \le C_U q^k \quad \text{for all} \quad k \ge k_0, \tag{5.20}$$

for positive constants C_L, C_U, $\tilde{q}$ and q with $0 < \tilde{q} < q < 1$, independent of n for n suitably large. That implies associated uniform integrability, from which we obtain associated convergence of means: $E[N^n] \to E[N]$ as $n \to \infty$, and higher moments as well if desired.

We now focus on the event, say A_n, that any of the processes ever differ from the upper bound process over a regenerative cycle of the n^{th} upper bound process. In addition to the upper bound process, it suffices to consider only the lower bound process, because the rate order implies that we can construct the processes so that the lower bound process will differ from the upper bound process at some transition whenever any of the other intermediate processes do, i.e., whenever the other processes do; i.e., whenever $D_f^n(X^n(t),\cdot)$ or $D_{1,2}^n(\cdot)$ do.

Both the upper and lower bound processes are constant rate CTMC's, with common rates in the two regions $(-\infty,0]$ and $(0,\infty)$. Thus there are only two different cases to consider: the two processes are either both in the upper region or both in the lower region. To simplify the analysis, it is convenient to modify the construction of the two processes $D_f^n(X_m^n,\cdot)$ and $D_f^n(X_M^n,\cdot)$ in order to make the probability that the two processes differ at any transition be the same in both regions for all ξ and n, and thus the same for all transitions for all ξ and n. That can be done by adjusting the bounds, while still keeping the rate order and the asymptotic properties as $\xi \downarrow 0$. (For each n, we can make the difference in the total transition rate in each region the maximum of what it was originally in each of the two regions. Clearly, the maximum difference also converges to 0 as $\xi \downarrow 0$.) That allows us to totally decouple the probability of a different transition at each transition epoch from the evolution of the processes, and thus simplifies calculations of bounds.

With that modified construction in place, let $W_i^n = 1$ if the lower bound process $D_f^n(X_m^n, \cdot)$ makes a different transition from the upper bound process $D_f^n(X_M^n, \cdot)$ at the i^{th} transition of the Poisson process, given that has not happened so far. Given our revised construction above, we can assume that the sequence $\{W_i^n : i \geq 1\}$ is a sequence of i.i.d random variables with $P(W_i^n = 1) = \phi_n$, where $\phi_n \to \phi$ as $n \to \infty$ and $\phi \downarrow 0$ as $\xi \downarrow 0$. To see why, recall that, by Lemma 7.1, $\bar{X}_M^n \Rightarrow x_M$ and $\bar{X}_m^n \Rightarrow x_m$ in $\mathcal{D}_6$ as $n \to \infty$, where $x_M(0) = x_m(0) = (x(0), x(0))$. Hence, by taking ξ small enough and n large enough, we can make $\bar{X}_M^n$ and $\bar{X}_m^n$ arbitrarily close for all $t \in [0, \xi]$. Consequently, the probability that any of the processes differ at step $k \geq 1$ during a regenerative cycle, depends on the number of transitions during a regenerative cycle being at least k. Hence,

$$
\begin{aligned}
P(A_n) \equiv P(\text{any processes differ}) &= \sum_{k=1}^{\infty} \phi_n (1-\phi_n)^{k-1} \sum_{j=k}^{\infty} p_n(j) \\
&\leq \sum_{k=1}^{k_0} \phi_n (1-\phi_n)^{k-1} + \sum_{k=k_0+1}^{\infty} \phi_n (1-\phi_n)^{k-1} \sum_{j=k}^{\infty} C_U q^j \\
&= \phi_n \left(\sum_{k=1}^{k_0} (1-\phi_n)^{k-1} + \frac{C_U q}{1-q} \sum_{k=k_0+1}^{\infty} [(1-\phi_n) q]^{k-1} \right) \\
&\leq C_1 \phi
\end{aligned} \tag{5.21}
$$

for a new constant C_1, provided that $(1-\phi)q < 1$ and n is suitably large. The condition $(1-\phi)q < 1$ holds since $q < 1$, so that the overall probability $P(A_n)$ can be made arbitrarily small, by making ϕ small enough by choosing ξ suitably small and n suitably large.

The first interval $U_{1,k}^n$ is the random sum of $V_{1,k}^n$ i.i.d. exponential random variables, each with mean $1/n\alpha$ (corresponding to the Poisson process with rate $n\alpha$), where $V_{1,k}^n$ is the geometric random sum, with mean $1/P(A_n)$, of the numbers of transitions in the successive cycles, in which no transitions disagree. We now give an expression for a lower bound for the means:

$$
E[V_{1,k}^n] = \frac{E[N^n]}{P(A_n)} \geq \frac{C_2 E[N]}{\phi} \quad \text{for all suitably large} \quad n, \tag{5.22}
$$

where $C_2 < 1/C_1$ for C_1 in (5.21). We obtain the lower bound in (5.22) by applying the convergence of the means $E[N^n] \to E[N]$ as $n \to \infty$, indicated above. Thus,

$$E[U_{1,k}^n] \geq \frac{C_2 E[N]}{\phi n \alpha} \quad \text{for all suitably large} \quad n, \tag{5.23}$$

as well. The main point is that we can make these means in (5.22) and (5.23) large in the relevant scale by making ϕ suitably small, which we can achieve by the proper choice of ξ.

We now want to show that $V_{2,k}^n$, the number of transitions of the Poisson process with rate $n\alpha$ in the second interval $U_{2,k}^n$, can be suitably controlled. To go with (5.22), it suffices to show that $V_{2,k}^n$ is SB as $n \to \infty$. Equivalently, it suffices to show that $nU_{2,k}^n$ is SB as $n \to \infty$. We will consider the two parts of this second interval in turn.

First consider the exceptional cycle. Let N_e^n be the random number of transitions in an exceptional regenerative cycle for the upper bound process. First, N_e^n is not distributed the same as N^n, because longer cycles are more likely to become exceptional cycles than shorter ones, because they generate more opportunities for a difference. Nevertheless, we can bound $E[N_e^n]$ above. To do so, we need to bound $P(A_n)$ below, instead of above as in (5.21). We can do so by using the lower bound for the probabilities $p_n(k) \equiv P(N^n = k)$ in (5.20).

We can now bound the mean $E[N_e^n]$ above for all n suitably large. In particular,

$$E[N_e^n] = E[N^n | A_n] = \frac{E[N^n; A_n]}{P(A_n)}. \tag{5.24}$$

We start with the numerator of (5.24):

$$\begin{aligned}
E[N^n; A_n] &= \sum_{k=1}^{\infty} \sum_{j=1}^{k} kP(N = k; \text{processes first differ at transition } j) \\
&= \sum_{k=1}^{\infty} \sum_{j=1}^{k} kP(N^n = k)\phi_n (1-\phi_n)^{j-1} = \sum_{k=1}^{\infty} kp_n(k)\phi_n \frac{1-(1-\phi_n)^k}{\phi_n} \\
&= E[N^n] - (1-\phi_n) \sum_{k=1}^{\infty} kp_n(k)(1-\phi_n)^{k-1} = E[N^n] - z_n \frac{d}{dz}\psi_{N^n}(z_n),
\end{aligned}$$

where $z_n \equiv (1-\phi_n)$.

Note that, by Abel's Lemma (Lemma 5.1 pg. 64 in [24]), $\psi_{N^n}(z_n)$ and, consequently, $\frac{d}{dz_n}\psi_{N^n}(z_n)$ are continuous from the left at $z_n = 1$. Also, $z_n \to 1$ (from the left) as $\phi_n \to 0$. Hence, the numerator of (5.24) converges to 0 as $\phi_n \to 0$. We next show that the rate of convergence to 0 is the same as that of the denominator of (5.24), so that (5.24) is bounded from above by a constant. By (5.21) and Fubini's theorem,

$$\begin{aligned} P(A_n) &= \sum_{k=1}^{\infty} \phi_n (1-\phi_n)^{k-1} \sum_{j=k}^{\infty} p_n(j) = \sum_{j=1}^{\infty} p_n(j) \sum_{k=1}^{j} \phi_n (1-\phi_n)^{k-1} \\ &= \sum_{j=1}^{\infty} p_n(j) [1-(1-\phi_n)^j] = 1 - \psi_{N^n}(z_n). \end{aligned}$$

Applying L'Hôpital's rule and Abel's lemma, we see that the limit of (5.24) as $\phi_n \to 0$ (by taking n to infinity and then ξ to zero) is bounded from above by a constant. Specifically,

$$\lim_{z_n \uparrow 1} \frac{\frac{d}{dz_n}\psi_{N^n}(z_n) + z_n \frac{d^2}{d^2 z_n}\psi_{N^n}(z_n)}{\frac{d}{dz_n}\psi_{N^n}(z_n)} = \frac{E[N^n] + E[(N^n)^2]}{E[N^n]} \le C_3$$

for some constant C_3. (Recall that $E[N^n] \to E[N]$ and $E[(N^n)^2] \to E[N^2]$ as $n \to \infty$ by (5.20).)

For the next step, we will also want to bound the tail probabilities of N_e^n. By a minor variation of the argument in (5.24), we can show they are bounded by a random variable with a geometric tail. If $k_1 \ge k_0$, then

$$\begin{aligned} P(N_e^n \ge k_1) &= \frac{P(N^n \ge k_1; A_n)}{P(A_n)} = \frac{\sum_{k=k_1}^{\infty} \sum_{j=1}^{k} \phi_n (1-\phi_n)^{j-1} p_n(k)}{\sum_{k=1}^{\infty} \sum_{j=1}^{k} \phi_n (1-\phi_n)^{j-1} p_n(k)} \\ &\le \frac{\sum_{k=k_1}^{\infty} [1-(1-\phi_n)^k] C_U q^k}{\sum_{k=k_0}^{\infty} [1-(1-\phi_n)^k] C_L \tilde{q}^k} \le C_4 [(1-\phi)q]^{k_1} \qquad (5.25) \end{aligned}$$

for a new constant C_4 (depending upon k_0), provided that ϕ is close enough to 0, which can be ensured by making ξ small, and that n is suitably large.

We now are ready to treat the second part of the second interval $U_{2,k}^n$, focusing on the number of transitions $V_{2,k}^n$. Our main idea now is to let the four processes evolve independently with the transitions generated by independent Poisson processes. Thus, to be concrete, let $V_{2,k}^n$ refer specifically to the number of transitions in the Poisson process generating the upper bound process $D_f^n(X_M^n, \cdot)$. To understand the essential point, we first consider the relatively simple case in which there are four independent versions of $D_f^n(X_M^n, \cdot)$ starting together in the regenerative state. But now we generate the vector-valued four-tuple of processes together using the superposition of four independent Poisson processes, which is a Poisson process with rate $4\alpha n$. At each transition epoch of this Poisson process, we let the transition correspond to each of the four individual processes independently with probability $1/4$. We thus construct the 4 independent versions together. We can thus focus on the vector-valued discrete-time Markov chain representing the transitions of all 4 processes, but each of these transitions corresponds to only one of the four Poisson processes, and the four processes remain independent. Now let N_c^n be the total number of transitions of this Poisson process with rate $4\alpha n$ before the interval ends with all four processes together again in the regenerative state s^*.

Now observe that the intervals between successive visits of all four processes to this regenerative state constitute a renewal process. In the long run, each process will be in the regenerative state a proportion $\pi^n(s*)$ of the time, for $0 < \pi^n(s*) < 1$; i.e., $\pi^n(s*)$ is the steady-state probability of the regenerative state, say s^*, i.e., $\pi^n(s*) = P(D_f^n(X_M^n, \infty) = s^*)$, with $1/\pi^n(s*)$ being the mean interval between successive visits to s^*. Consequently, in the long run, the four copies will all be in the state s^* together a proportion $\pi^n(s*)^4$ of the time. Since successive return times to s^* form a renewal process, the mean time between successive returns of all four copies of the upper bound process $D_f^n(X_M^n, \cdot)$ to s^* is $1/\pi^n(s*)^4$ for each n.

By (i) the convergence of $\bar{X}_M^n \Rightarrow x_M$, (ii) the convergence of the transition rates of $\{D_f^n(X_M^n, s) : s \geq 0\}$ defined in (5.2)-(5.5) to the transition rates of the FTSP $\{D(x_M, s) : s \geq 0\}$ defined in (5.9)-(5.12) as $n \to \infty$, which is justified by (8.9) and the following discussion, and (iii) Lemma 8.8, we deduce that $\pi^n(s^*) \to \pi(s^*)$ as $n \to \infty$, where $\pi(s^*)$ is the steady-state probability of the FTSP, i.e., $\pi(s^*) = P(D(x_M, \infty) = s^*)$. Hence, for this special initial condition, we have established the bound $E[N_c^n] \leq C_7/\pi(s^*)^4 < \infty$ for $C_7 > 1$ for all n suitably large (depending on our choice of C_7).

Of course, we do not actually have four copies of the upper bound process and the four processes we do have are not all starting in the regenerative state. Hence we have to do more. There is a further complication, because the process $D_{1,2}^n$ is *not* a constant-rate CTMC. However, we circumvent this difficulty by treating *all* the independent processes under consideration as independent copies of the upper bound process $D_f^n(X_M^n, \cdot)$, but with different initial conditions. (This addresses the first difficulty.) In particular, we generate four independent copies of $D_f^n(X_M^n, \cdot)$ with the given initial conditions at the end of the exceptional cycle. And, together with the three processes that are not actually the upper-bound process, we also generate the other process using that *same* Poisson process. Hence three of the four independent Poisson processes will be used to generate two processes each. We do those pairwise constructions as before, aiming to keep the two processes as close together as possible, for each of the three pairs of processes. We have already described how to analyze the probability of a difference occurring over successive transitions, which can be (and will be) made negligible.

We will succeed in using the four independent copies of $D_f^n(X_M^n,\cdot)$ constructed as above if none of the three independent versions $D_f^n(X_M^n,\cdot)$ serving for other processes make a different transition from the original process over the interval under consideration. Since we will be showing that the total interval is SB, the probability of a different transition here can be made arbitrarily small as well. We will thus do the construction until the four processes meet again in the regenerative state, but in doing so, we also keep track of whether or not any of the interior processes make any different transitions. If there were no differences in transitions for the interior processes, then the cycle has ended when all the processes first reach the regenerative state at the same transition epoch.

For the moment, assume that no differences occur between the three original processes and the version of $D_f^n(X_M^n,\cdot)$. Hence, we now focus on the different initial conditions actually holding at the end of an exceptional cycle. To facilitate having these four independent copies of $D_f^n(X_M^n,\cdot)$ with different initial conditions reach the regenerative together as soon as possible, we couple each process with the upper-bound process as soon as the two processes are ever in the same state. From that hitting time forward, we let both processes be the upper bound process, generated by its Poisson process. This leaves the distribution of the individual processes unchanged. We now proceed until all three independent copies of $D_f^n(X_M^n,\cdot)$ have coupled with the upper-bound process $D_f^n(X_M^n,\cdot)$ and the upper-bound process (and thus all four) processes have reached the regenerative state.

We can bound this expected number of transitions until the four processes reach the regeneration state together if we can bound the first hitting time of s^*. That is so, because we can bound the expected number of transitions for all four independent processes to reach the regeneration state together, if at transition k all four processes have visited state s^* at least once in the last k transitions. That makes the other three discrete-time processes distributed as index shifted versions of the upper-bound DTMC.

e now want to bound the first passage time to s^* for each of the processes not starting in s^*. he first passage time can be controlled provided the initial condition can be controlled. e thus control the separation between the processes that can occur during the rest of the exceptional cycle, after the first non-identical transition. After the first non-identical transition, we focus on the upper bound process. We say that the exceptional cycle ends when the upper bound process next hits the regenerative state. However, because of the non-identical transitions, the other processes typically will not hit the regenerative state at that same transition epoch. It is evident that, as long as the processes stay together on the same side of 0, the probability of a second different transition during the cycle will be negligible. However, we lose control when the processes are on different sides of state 0. Fortunately, it suffices to use a crude bound on the maximum possible separation of the processes during the exceptional cycle. We can suppose that the maximum possible separation is achieved at each transition over the entire cycle. The worst case would have the separation increase by $K \equiv 2(j \vee k)$ at every transition. (The two processes would have a transition at the same time going the maximum possible distance away from each other.) Hence, since the total number of transitions of the upper bound process in the exceptional cycle after the initial non-identical transition is N_e^n, then the other processes are in a state within KN_e^n states of the regenerative state, where the upper bound process $D_f^n(X_M^n, \cdot)$ will be at the end of the exceptional cycle. In (5.25) we have shown that this random bound on the initial difference has a geometric tail, so that the probability of large differences are controlled. Since the first passage time (number of transitions) from any fixed state to s^* has a generating function, the number of transitions until all the processes have hit s^* is SB. Consequently, N_c^n is SB.

We now specify what we do if there are differences within the period considered above. If there were any differences (an event of small probability), then we repeat the construction for the second part of the second interval using four independent versions of $D_f^n(X_M^n, \cdot)$ until the four processes are again together in the regenerative state. This second try will produce a number of transitions $N_{c,2}^n$ different from $N_{c,1}^n \equiv N_c^n$ in the first try, but actually somewhat more favorable (tending to be smaller) because the initial conditions are more favorable, with three of the four processes likely to be starting in the regenerative state and the interior process differing at most by the gap ζ, by virtue of Corollary 8.4. (By the independence of the pairs, two or more differences will be asymptotically negligible compared to a single difference.) So, if the second try is needed, we will be able to control $N_{c,2}^n$ just as we can control $N_{c,1}^n$.

However, even the second try may be unsuccessful, because again we may find that one or more of the three processes makes a transition different from its representation by $D_f^n(X_M^n, \cdot)$. Thus we may possibly need to repeat the second-try construction an indefinite number of times until we get all four processes together in the regenerative state. However, these successive repetitions will be independent copies of the second try, each with the same initial conditions, yielding numbers of transitions again distributed as $N_{c,2}^n$. Thus we can represent $V_{2,k}^n$ as the sum of N_c^n and an independent geometric random sum of i.i.d. random variables distributed as N_c^n, where the geometric probability can be made very small by choosing ξ small enough. Thus we can control all of $V^n_{2,k}$ if we can control N^n_c, assuming that all four processes are four independent copies of the upper-bound process $D^n_f(X^n_M, \cdot)$, but with different initial conditions.

The final task is to show that the special initial conditions imposed in (5.18) are actually not needed. Instead, we have the assumed condition (8.21) with $D^n_{1,2}(0) \Rightarrow L$, where L is a proper random variable. We now replace this assumed convergence in distribution by convergence w.p.1 by applying the Skorohod representation theorem. We thus write $D^n_{1,2}(0) \to L$, without using special notation to denote these alternative versions having the same probability law as the orginal versions. However, since these random variables are integer valued, we have for each underlying sample point ω that $D^n_{1,2}(0) \equiv D^n_{1,2}(0, \omega) = L(\omega) \equiv L$ for all sufficiently large n. Thus, we can condition on ω and thus the value of L and regard that state as the initial state s^* in (5.18) (for all sufficiently large n). Since X^n_6 is a CTMC the evolution after the initial state, given that initial state is independent of the inital state. Hence this construction is justified. ∎

E.5. Remaining Proof in §8.7.

Proof of Lemma 8.11. First, let $\delta > 0$, $\epsilon > 0$ and t with $0 < t < \delta$ be given, where the δ is chosen so that $\delta < \xi$ for ξ in Lemmas 8.7, 8.9 and 8.10. Below we will be introducing a new ξ less than this δ.

We start by observing that versions of Lemmas 8.9 and 8.10 hold on an interval $[t, t+\xi]$, where $\xi \equiv \xi(t)$ satisfies $0 < \xi < \delta - t$. Before, we started with the convergence $\bar{X}^n(0) \Rightarrow x(0)$ in $\mathbb{R}^3$ at time 0 based on Assumption 3. Now, instead, we base the convergence $\bar{X}^n(t) \Rightarrow \bar{X}(t)$ at time t on the convergence we have along the converging subsequence. Since the processes are Markov processes, we can construct the processes after time t, given only the value of $X^n(t)$, independently of what happens on $[0, t]$. We apply Lemma 8.7 to deduce that $P(\bar{X}(t) \in \mathbb{A}) = 1$ (which is justified by our choice of δ).

We now indicate how the proofs of Lemmas 8.9 and 8.10 need to be modified, proceeding forward after time t. Let $X^{n,\xi}_M \equiv (X^{n,\xi}_{M^+}, X^{n,\xi}_{M^-})$ be defined similar to X^n_M in (5.13) and $X^{n,\xi}_m \equiv (X^{n,\xi}_{m^+}, X^{n,\xi}_{m^-})$ be defined similar to and X^n_m in (5.15), but with supremum and infimum taken over the interval $[t, t+\xi]$ (instead of over the interval $[0, \xi]$ as before (where the constants ξ need not be the same for each t; i.e., $\xi \equiv \xi(t)$). Recall that the associated bounding quantities are constructed from separate processes related to X^n only through their distributions. These too do not depend on the evolution of X^n after time t.

Reasoning as before, by virtue of Lemma 7.1, the limits $x^{\xi}_{M} \equiv (x^{\xi}_{M^+}, x^{\xi}_{M^-})$ and $x^{\xi}_{m} \equiv (x^{\xi}_{m^+}, x^{\xi}_{m^-})$ of $\bar{X}^{n,\xi}_{M}$ and $\bar{X}^{n,\xi}_{m}$ exist. (Since $\bar{X}(t)$ so far is a random variable, so are x^{ξ}_{M} and x^{ξ}_{m}. However, we can regard $\bar{X}(t)$ as a constant by conditioning upon it, without affecting the evolution after time t, because of the Markov property.) In particular, Applying the continuous mapping theorem for the supremum, Theorem 12.11.7 in [48], we have that $X^{n,\xi}_{M^+}/n \Rightarrow x^{\xi}_{M^+} \equiv (q^{\xi}_{1,M}, q^{\xi}_{2,M}, z^{\xi}_{M^+})$ and $X^{n,\xi}_{M^-}/n \Rightarrow x^{\xi}_{M^-} \equiv (q^{\xi}_{1,M}, q^{\xi}_{2,M}, z^{\xi}_{M^-})$ as $n \to \infty$, where

$$(5.26)\qquad \begin{aligned} q^{\xi}_{1,M} &\equiv \inf_{t \le s \le t+\xi} q^{\xi}_{1}(s) \vee 0, \\ q^{\xi}_{2,M} &\equiv \sup_{t \le s \le t+\xi} q_{2}(s), \\ z_{M^+} &\equiv \begin{cases} \inf_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \\ \sup_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \end{cases} \\ z_{M^-} &\equiv \begin{cases} \inf_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \\ \sup_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \end{cases} \end{aligned}$$

Similarly, $X^{n,\xi}_{m^+}/n \Rightarrow x^{\xi}_{m^+} \equiv (q^{\xi}_{1,m}, q^{\xi}_{2,m}, z^{\xi}_{m^+})$ and $X^{n,\xi}_{m^-}/n \Rightarrow x^{\xi}_{m^-} \equiv (q^{\xi}_{1,m}, q^{\xi}_{2,m}, z^{\xi}_{m^-})$ as $n \to \infty$, with

$$(5.27)\qquad \begin{aligned} q^{\xi}_{1,m} &\equiv \sup_{t \le s \le t+\xi} q_{1}(s), \\ q^{\xi}_{2,m} &\equiv \inf_{t \le s \le t+\xi} q_{2}(s) \vee 0, \\ z_{m^+} &\equiv \begin{cases} \inf_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \\ \sup_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \end{cases} \\ z_{m^-} &\equiv \begin{cases} \inf_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \le \mu_{2,2}, \\ \sup_{t \le s \le t+\xi} z_{1,2}(s) & \mu_{1,2} \ge \mu_{2,2}, \end{cases} \end{aligned}$$

The two bounding frozen difference processes are $\{D^{n}_{f}(X^{n,\xi}_{M}, s) : s \ge t\}$ and $\{D^{n}_{f}(X^{n,\xi}_{m}, s) : s \ge t\}$. As a consequence of this construction, we can conclude that there exists $\xi > 0$ and an integer n_1 such that the drift rates of these bounding processes satisfy both the inequalities in (8.12) in order for them to be positive recurrent and the rate order in (8.18) with probability at least $1 - \epsilon/6$ for all $n \ge n_1$.

We next apply Lemma 8.10 to conclude that there exists a new ξ, taken no bigger than the one created so far, such that the following variants of the integral inequalities in (8.23) hold with probability at least $1 - \epsilon/6$ as well:

$$(5.28)\qquad \begin{aligned} \frac{1}{\xi}\int_{t}^{t+\xi} 1_{\{D^{n}_{f}(X^{n,\xi}_{m}, s)\}}\, ds - \frac{\epsilon}{6m_2} &\le \frac{1}{\xi}\int_{t}^{t+\xi} 1_{\{D^{n}_{1,2}(s) > 0\}}\, ds \\ &\le \frac{1}{\xi}\int_{t}^{t+\xi} 1_{\{D^{n}_{f}(X^{n,\xi}_{M}, s) > 0\}}\, ds + \frac{\epsilon}{6m_2}. \end{aligned}$$

(We divide by m_2 because we will be multiplying by $z_{1,2}(t)$.)

We now represent the bounding frozen queue-difference processes directly in terms of the FTSP, using the relation (8.9):

(5.29)
$$\{D_f^n(\lambda_i^n, m_j^n, X_m^{n,\xi}, t+s) : s \ge 0\} \stackrel{\mathrm{d}}{=} \{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, t+sn) : s \ge 0\}$$
$$\{D_f^n(\lambda_i^n, m_j^n, X_M^{n,\xi}, t+s) : s \ge 0\} \stackrel{\mathrm{d}}{=} \{D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, t+sn) : s \ge 0\}.$$

Upon making a change of variables, the bounding integrals in (5.28) become

(5.30)
$$\frac{1}{\xi}\int_t^{t+\xi} 1_{\{D_f^n(\lambda_i^n, m_j^n, X_m^{n,\xi}, s)>0\}}\, ds \stackrel{\mathrm{d}}{=} \frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s>0\}}\, ds$$
$$\frac{1}{\xi}\int_t^{t+\xi} 1_{\{D_f^n(\lambda_i^n, m_j^n, X_M^{n,\xi}, s)>0\}}\, ds \stackrel{\mathrm{d}}{=} \frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s)>0\}}\, ds.$$

For each integer k, we have the iterated limits

$$\begin{aligned}
&\lim_{n\to\infty}\lim_{s\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s) = k) \\
&\quad = \lim_{s\to\infty}\lim_{n\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s) = k), \\
&\lim_{n\to\infty}\lim_{s\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, s) = k) \\
&\quad = \lim_{s\to\infty}\lim_{n\to\infty} P(D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, s) = k),
\end{aligned} \tag{5.31}$$

where the first limit is $P(D(x_m^\xi, \infty) = k) \equiv P(D(\lambda_i, m_j, x_m^\xi, \infty) = k)$, while the second is $P(D(x_M^\xi, \infty) = k) \equiv P(D(\lambda_i, m_j, x_M^\xi, \infty) = k)$.

By Corollary 8.3, we also have the associated double limit for the averages over intervals of length $O(n)$ as $n \to \infty$

(5.32)
$$\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s)>0\}}\, ds \Rightarrow P(D(\lambda_i, m_j, x_m^\xi, \infty) > 0) \equiv \pi_{1,2}(x_m^\xi),$$
$$\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, s)>0\}}\, ds \Rightarrow P(D(\lambda_i, m_j, x_M^\xi, \infty) > 0) \equiv \pi_{1,2}(x_M^\xi).$$

(It is significant that for each t we have different x_m^ξ and x_M^ξ. Recall that we are now considering a fixed t.)

Invoking Lemma 8.8, choose ξ less than or equal to the previous value of ξ such that

$$|\pi_{1,2}(x_m^\xi) - \pi_{1,2}(\bar{X}(t))| \le \frac{\epsilon}{6m_2}. \tag{5.33}$$

For that ξ, applying (5.32), choose $n_2 \geq n_1$ such that

(5.34)
$$P\left(|\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_m^{n,\xi}/n, s)>0\}}\, ds - \pi_{1,2}(x_m)| > \frac{\epsilon}{6m_2}\right) < \frac{\epsilon}{6}$$
$$\text{and}\quad P\left(|\frac{1}{n\xi}\int_t^{t+n\xi} 1_{\{D(\lambda_i^n/n, m_j^n/n, X_M^{n,\xi}/n, s)>0\}}\, ds - \pi_{1,2}(x_M)| > \frac{\epsilon}{6m_2}\right) < \frac{\epsilon}{6}$$

for all $n \geq n_2$.

We now use the convergence 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 ξ less than or equal to the previous value and $n_3 \geq n_2$ such that

$$P(\sup_{u: t\leq u\leq t+\xi} \{|\bar{X}^n(u) - \bar{X}(t)|\} > \epsilon/6) < \epsilon/6 \quad \text{for all} \quad n \geq n_3. \tag{5.35}$$

For the current proof, we will use the consequence

$$P(\sup_{u: t\leq u\leq t+\xi} \{|\bar{Z}_{1,2}^n(u) - \bar{Z}_{1,2}(t)|\} > \epsilon/6) < \epsilon/6 \quad \text{for all} \quad n \geq n_3. \tag{5.36}$$

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 afterwards) holding with probability at least $1 - \epsilon$ (counting $\epsilon/6$ once each for (5.26), (5.27), (5.28), (5.36) and twice for (5.34)):

(5.37)

$$
\begin{aligned}
&(a)\ \int_t^{t+\xi} 1_{\{D^n_{1,2}(s)>0\}} \bar{Z}^n_{1,2}(s)\ ds \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \int_t^{t+\xi} 1_{\{D^n_{1,2}(s)>0\}}\ ds \\
&(b) \qquad \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \left(\int_t^{t+\xi} 1_{\{D^n_f(\lambda^n_i, m^n_j, X^n_M, s)>0\}}\ ds + \frac{\epsilon\xi}{6m_2}\right) \\
&(c) \qquad \stackrel{\mathrm{d}}{=} \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \left(\int_0^{\xi} 1_{\{D(\lambda^n_i/n, m^n_j/n, X^{n,\xi}_M/n, t+sn)>0\}}\ ds + \frac{\epsilon\xi}{6m_2}\right) \\
&(d) \qquad \stackrel{\mathrm{d}}{=} \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \xi \left(\frac{1}{n\xi}\int_0^{n\xi} 1_{\{D(\lambda^n_i/n, m^n_j/n, X^{n,\xi}_M/n, t+s)>0\}}\ ds + \frac{\epsilon}{6m_2}\right) \\
&(e) \qquad \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \xi \left(\pi_{1,2}(x^{\xi}_M) + \frac{2\epsilon}{6m_2}\right) \\
&(f) \qquad \le \left(\bar{Z}_{1,2}(t) + \frac{\epsilon}{6}\right) \xi \left(\pi_{1,2}(\bar{X}(t)) + \frac{3\epsilon}{6m_2}\right) \\
&(g) \qquad \le \bar{Z}_{1,2}(t)\pi_{1,2}(\bar{X}(t))\xi + \frac{\pi_{1,2}(\bar{X}(t))}{6}\epsilon\xi + \frac{1}{2}\epsilon\xi + \frac{\xi\epsilon^2}{12m_2} \\
&(h) \qquad \le \bar{Z}_{1,2}(t)\pi_{1,2}(\bar{X}(t))\xi + \frac{3}{4}\epsilon\xi \\
&\qquad\quad \le (\bar{Z}_{1,2}(t)\pi_{1,2}(\bar{X}(t)) + \epsilon)\xi \quad \text{for all} \quad n \ge n_0 \equiv n_3.
\end{aligned}
$$

We now explain the steps in (5.37): First, for (a) we replace $\bar{Z}^n_{1,2}(s)$ by $\bar{Z}_{1,2}(t)$ for $t \le s \le t + \xi$ by applying (5.36). For (b), we apply Lemma 8.10. For (c), we use the alternative representation in terms of the FTSP in (5.29). For (d), we use the change of variables in (5.30). For (e), we use (5.34), exploiting the convergence in (5.32). For (f), we use (5.33). Step (g) is simple algebra, exploiting $\bar{Z}_{1,2}(t) \le m_2$. Step (h) is more algebra, exploiting $\pi_{1,2}(\bar{X}(t)) \le 1$, and $\epsilon \le 1 \wedge m_2$. That completes the proof of the lemma. ∎

APPENDIX F: PROOFS FOR SECTION 9

Proof of Lemma 9.1. First, we apply the stochastic bounds in §8.2 to show that the the sequence of steady-state random vectors $\{\bar{X}^n(\infty) : n \ge 1\}$ is tight in $\mathbb{R}_6$. For each i, these bounds bound $\bar{Q}^n_i$ above by stochastic processes that converge to proper steady-state distributions as $t \to \infty$ and converge to fluid limits as $n \to \infty$. Hence, the family of random variables $\{\bar{Q}^n_i(t) : t \ge 0, n \ge 0\}$ is SB in $\mathbb{R}$ for $i = 1, 2$. As a consequence, the associated sequences of steady-state distributions $\{\bar{Q}^n_i(\infty) : n \ge 1\}$ are tight in $\mathbb{R}$. Since $\bar{Z}^n_{i,j} \le m^n_j/n \to m_j$, the families of random variables $\{\bar{Z}^n_{i,j}(t) : t \ge 0, n \ge 0\}$ is SB in $\mathbb{R}$ as well. Since tightness of the marginals implies tightness of vectors, the sequence of steady-state random vectors $\{\bar{X}^n(\infty) : n \geq 1\}$ is tight in $\mathbb{R}_6$. Finally, to treat the rest of the processes in $\mathcal{D}_6$, we can use the same proof as for Lemma 8.1. ∎

Proof of Lemma 9.1. Apply Theorem 6.1, using special initial conditions, so that $\bar{X}^n(0) \Rightarrow x^* \in \mathbb{S}$ as $n \to \infty$. Apply Section 7 to deduce global SSC, which implies that $P(\bar{X}^n(t) \in \mathbb{S} \quad \text{for all} \quad t) \to 1$ as $n \to \infty$. As a consequence, $P(\bar{X}^n(\infty) \in \mathbb{S} \quad \text{for all} \quad t) \to 1$ as $n \to \infty$, where here we regard $\bar{X}^n(\infty)$ as a random variable with the limiting distribution as $t \to \infty$. However, $\bar{X}^n(\infty) \stackrel{\mathrm{d}}{=} \bar{X}^{n*}(0)$ for each n. ∎

References.

[1] Aldous, D. A. (1989). *Probability Approximations via the Poisson Clumping Heuristic*, Springer, New York.

[2] Anderson, C.W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Prob.* **7** 99–113.

[3] Asmussen, S. (1998). Extreme value theory for queues via cycle maxima. *Extremes* **1** 137–168.

[4] Assmusen S. (2003). *Applied Probability and Queues*, 2nd ed., Springer, New York.

[5] Bassamboo, A., Harrison, J. M. and 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.

[6] Bassamboo, A., Harrison, J. M. and Zeevi, A. (2006). Dynamic and control of a large call center: asymptotic analysisof an LP-based method. *Oper. Res.* **54** (3) 419–435.

[7] Baxendale, P. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Prob.*, **15** (1) 700–738.

[8] Billingsley, P. (1999) *Convergence of Probability Measures*, 2nd ed. Wiley, New York. MR1700749

[9] Bolthausen, E. (1980). The Berry-Esseen theorem for functionals of discrete Markov chains. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **54** 59–73.

[10] Borkovec, M. and Klüppelberg, C. (1998). Extremal behavior of diffusion models in finance. *Extremes.* **1**, 47–80.

[11] Coffman, E. G., Puhalskii, A. A. and Reiman, M. I. (1995). Polling systems with zero switchover times: a heavy-traffic averaging principle. *Annals of Applied Probability* **5** 681–719.

[12] Davis, R. A. (1982). Maximum and minimum of one-dimensional diffusions. *Stochastic Processes and their Applications.* **13** 1–9.

[13] Ethier, S. N. and Kurtz, T. G. (1985). *Markov Processes: Characterization and Convergence*. Wiley, New York. MR838085

[14] Garnett, O., Mandelbaum, A. and Reiman, M. I. (2002). Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** (3) 208–227.

[15] Glynn, P. W. and Whitt, W. (1993). Limit theorems for cumulative processes. *Stoch. Proc. Appl.* **47** 299–314.

[16] Gurvich, I. and Whitt, W. (2009a). Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* **34** (2) 363–396.

[17] Gurvich, I. and Whitt, W. (2009b). Scheduling flexible servers with convex delay

costs in many-server service systems. *Manufacturing Service Oper. Management* **11** (2) 237–253.

[18] Gurvich, I. and Whitt, W. (2010). Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* **58** (2) 316–328.

[19] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** (3) 567–588.

[20] He, Q. (1993). Differentiability of the matrices R and G in the matrix analytic method. *Stochastic Models* **11** (1) 123–132.

[21] Hunt, P.J. and Kurtz, T. G. (1994). Large loss networks. *Stochastic Processes and their Applications* **53** 363–378.

[22] Kamae, T., Krengel, U. and O'Brien, G. L. (1977). Stochastic inequalities on partially ordered spaces. *Ann. Prob.* **5** 899–912.

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

[24] Karlin S. and Taylor, H.M. (1975). *A First Course in Stochastic Processes* Academic Press, New York.

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

[26] Khasminskii, R. Z. and Yin, G. (2004). On averaging principles: an asymptotic expansion approach. *SIAM J. Math. Anal.* **35** 1534–1560.

[27] Kontoyiannis, I. and Meyn, S.P. (2003). Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Prob.*, **13** (1) 304–362.

[28] Kurtz, T. G. (1992). Averaging for martingale problems and stochastic approximations, *Applied Stochastic Analysis, Proc.* US-French Workshop, Lecture Notes in Control and Information Sciences **177** Springer, Berlin, 186–209.

[29] Latouche G. and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling,* Siam and ASA, Philadelphia.

[30] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes* Springer-Verlag, New York.

[31] Lindvall, T. (1992). *Lectures on the Coupling Method*, Wiley, New York.

[32] Massey, W. A. and Whitt, W. (1998). Uniform acceleration expansions for Markov chains with time-varying rates. *Ann. Appl. Prob.* **8** (4) 1130–1155.

[33] Meyn, S. P. and Tweedie, R, L. (1993). *Markov Chains and Stochastic Stability*, Springer, New York.

[34] van Moorsel, A. P. A. and Sanders, W. H. (1994). Adaptive uniformization. *Stochastic Models* **10** (3) 619–647.

[35] Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models with Risks*, Wiley, New York.

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

[37] Perry, O. and Whitt, W. (2009a). Responding to unexpected overloads in large-scale service systems. *Management Sci.*, **55** (8) 1353–1367.

[38] Perry, O. and Whitt, W. (2009b). A fluid approximation for service systems responding to unexpected overloads. Submitted to *Oper. Res.* Available at: http://www.columbia.edu/∼ww2040/allpapers.html

[39] Perry, O. and Whitt, W. (2009c). An ODE for an overloaded X model involving a stochastic averaging principle. working paper, Columbia University, NY. (Revision of "Transient and stability analysis of the many-server heavy-traffic fluid limit for the overloaded X call-center model.") Available at: http://www.columbia.edu/∼ww2040/allpapers.html

[40] Perry, O., W. Whitt. 2010. Gaussian approximations for an overloaded X model via an averaging principle. Working paper, Columbia University. Available at: http://www.columbia.edu/∼ww2040/allpapers.html

[41] Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*, Springer-Verlag, New York.

[42] Roberts, G. O. and Rosenthal, J. S. (1996). Quantative bounds for convergence rates of continuous time Markov processes. *Electronic Journal of Applied Probability*, **1** (9) 1–21.

[43] Rootzén, H. (1988). Maxima and exceedances of stationary Markov chains. *Advances in Applied Probability.* **20** (2) 371–390.

[44] Silvestrov, D. S. and Teugels, J. L. (1998). Limit theorems for extremes with random sample size. *Adv. Appl. Prob.* **30** 777–806.

[45] Whitt, W. (1981). Comparing counting processes and queues. *Adv. Appl. Prob.* **13** (1) 207–220.

[46] Whitt, W. (1984). Departures from a queue with many busy servers. *Math. Oper. Res.* **9** (4) 534–544.

[47] 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

[48] Whitt, W. (2002). *Stochastic-Process Limits*, Springer, New York.

[49] Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50** (10) 1449–1461.

CWI, Science Park 123, 1098 XG Amsterdam,
The Netherlands,
E-mail: o.perry@cwi.nl

Department of Industrial Engineering and Operations Research,
Columbia University New York, New York 10027-6699,
E-mail: ww2040@columbia.edu