Stochastic Systems
2011, Vol. 1, 17–66
DOI: 10.1214/10-SSY009

AN ODE FOR AN OVERLOADED X MODEL INVOLVING A STOCHASTIC AVERAGING PRINCIPLE

By Ohad Perry* and Ward Whitt†

CWI and Columbia University

We study an ordinary differential equation (ODE) arising as the many-server heavy-traffic fluid limit of a sequence of overloaded Markovian queueing models with two customer classes and two service pools. The system, known as the X model in the call-center literature, operates under 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 unanticipated overload. Each pool serves only its own class until a threshold is exceeded; then one-way sharing is activated with all customer-server assignments then driving the two queues toward a fixed ratio. For large systems, that fixed ratio is achieved approximately. The ODE describes system performance during an overload. The control is driven by a queue-difference stochastic process, which operates in a faster time scale than the queueing processes themselves, thus achieving a time-dependent steady state instantaneously in the limit. As a result, for the ODE, the driving process is replaced by its long-run average behavior at each instant of time; i.e., the ODE involves a heavy-traffic averaging principle (AP).

1. Introduction. We study an *ordinary differential equation* (ODE) that arises as the *many-server heavy-traffic* (MS-HT) fluid limit of a sequence of overloaded Markovian X queueing models under the *fixed-queue-ratio-with-thresholds* (FQR-T) control. The ODE is especially interesting, because it involves a heavy-traffic *averaging principle* (AP).

The system consists of two large service pools that are designed to operate independently, but can help each other when one of the pools, or both, encounter an unexpected overload, manifested by an instantaneous shift in the arrival rates. We assume that the time that the arrival rates shift and the values of the new arrival rates are not known when the overload

Received July 2010.

*Centrum Wiskunde & Informatica (CWI).

†Department of Industrial Engineering and Operations Research, Columbia University.

AMS 2000 subject classifications: Primary 60K25; secondary 60K30, 60F17, 90B15, 90B22, 37C75, 93D05.

Keywords and phrases: Many-server queues, averaging principle, separation of time scales, heavy traffic, deterministic fluid approximation, quasi-birth-death processes, ordinary differential equations, overload control.

occurs. We want the control to automatically detect the overload. The FQR-T control is designed to prevent sharing of customers (i.e., sending customers to be served at the other-class service pool) when sharing is not needed, and automatically activate sharing when the system becomes overloaded due to a sudden shift in the arrival rates. Others have also considered many-server systems with unknown arrival rates [2, 20], but our overload control problem addresses a different issue. Here the service pools are intended to operate independently except during the overload incident.

This paper is the third in a series of five papers. First, in [16] we initiated study of this overload-control problem and proposed the FQR-T control; see [16] for a discussion of related literature. We used a heuristic stationary fluid approximation to analyze the performance of the system during an overload. For the fluid model, we derived the optimal ratio control parameters when a convex holding cost is charged to the two queues during the overload incident (considering the long-run average cost during the overload incident, assuming that the overload condition lasts sufficiently long). Within that framework, we showed with simulations that FQR-T outperforms the best fixed allocation of servers, even when the new arrival rates are known. The stationary point of the fluid model was derived using a heuristic flow-balance argument, which equates the rate of flow into the system to the rate of flow out of the system, when the system is in steady state.

Second, in [19] we applied the heavy-traffic AP as an engineering principle in order to justify the ODE considered here to describe the transient fluid approximation of the X system under FQR-T after an overload has begun. For the overload problem, it is natural to go beyond the steady-state performance during an extended overload incident considered in [16] to develop a good approximation for the transient performance. We observed that the FQR-T control is driven by a queue-difference stochastic process, which operates in a faster time scale than the queueing processes themselves, so that it should achieve a time-dependent steady state instantaneously in the MS-HT limit, i.e., as the scale (arrival rate and number of servers) increases; see §3.1. We argued heuristically that the ODE should arise as the limit of a properly-scaled sequence of overloaded X-model systems, provided that the driving process is replaced by its long-run average behavior at each instant of time.

As reported in [16, 19], we performed simulation to justify the approximations. The close agreement between simulations of large-scale X queueing systems with the FQR-T control and the results of the numerical algorithm for solving the ODE demonstrated that the ODE considered here is indeed properly defined, that the numerical algorithm is effective and that the fluid

model approximation is effective for approximating the performance of the queueing system; see also Figures 3–4, where the numerical solution to the ODE is presented together with simulated sample paths of a large stochastic system. However, it still remained to state and prove theorems mathematically justifying the results.

The present paper and the next two provide that mathematical support. The present paper establishes important properties of the ODE introduced in [19]. The fourth and fifth papers establish limits. In particular, in [17] we prove that the fluid approximation, as a deterministic function of time, arises as the MS-HT limit of a sequence of X models; i.e., we prove a *functional weak law of large numbers* (FWLLN). This FWLLN is based on the AP; see [5, 9] for previous examples. In [18] we prove the corresponding *functional central limit theorem* (FCLT) that describes the stochastic fluctuations about the deterministic fluid path.

We prove convergence to the ODE in [17] by the standard two-step procedure, described in Ethier and Kurtz [6]: (i) establishing tightness and (ii) uniquely characterizing the limit process. The tightness argument follows familiar lines, but characterizing the limit process turns out to be challenging. Indeed, characterizing the limit process depends on the results here. Thus, the present paper provides a crucial ingredient for the limits established in [17, 18].

The AP makes the ODE unconventional. The AP creates a singularity region, causing the ODE not to be continuous in its full state space. Hence, classical results of ODE theory, such as those establishing existence, uniqueness and stability of solutions, cannot be applied directly. Moreover, existing algorithms for numerically solving ODE's cannot be applied directly either, since the solution to the ODE requires that the time-dependent steady state of the *fast-time-scale process* (FTSP) be computed at each instant. Nevertheless, we establish the existence of a unique solution to the ODE, show that there exists a unique stationary point; and show that the fluid process converges to its stationary point as time evolves. Moreover, we show that the convergence to stationarity is exponentially fast. The key is a careful analysis of the FTSP, which we represent as a *quasi-birth-and-death* (QBD) process. Finally, we provide a numerical algorithm for solving the ODE based on the matrix-geometric method [11].

Here is how the rest of this paper is organized: The next two sections provide background, intending to help the reader understand the motivation for the rigorous development that begins here in §4. In §2 we elaborate on the X queueing model and the FQR-T control; that primarily is a review of [16]. In §3 we provide a brief overview of the MS-HT scaling,

a heuristic explanation of the AP and a statement of the MS-HT limit; that primarily is a brief account of [17]. The reader should refer to [16, 17, 19] for additional details.

In §4 we introduce the ODE that we study in subsequent sections. In §5 we state out main result, establishing the existence of a unique solution. In §6 we establish properties of the FTSP, which depends on the state of the ODE, and whose steady-state distribution influences the evolution of the ODE. In §7 we define the state space of the ODE, and prove the main theorem about existence of a unique solution. In §8 we establish the existence of a unique stationary point and show that the fluid solution converges to that stationary point as time evolves. In §9 we prove that a solution converges to stationarity exponentially fast. In §10 we provide conditions for global state space collapse, i.e., for having the AP operate for all $t \geq 0$. In §11 we develop an algorithm to numerically solve the ODE (given an initial condition), based on the theory developed in the previous sections. We conclude in §12 with one postponed proof.

Additional material appears in an appendix, available from the authors' web pages. There we analyze the system with an underloaded initial state, and show that the approximating fluid models then lead to our main ODE in finite time. We elaborate on the algorithm and give two more numerical examples. We also provide a few omitted proofs. Finally, we mention remaining open problems.

2. The motivating queueing system. This section reviews the highlights of [16], starting with a definition of the original X queueing model, for which the ODE serves as an approximation.

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

This X model was introduced to study two large systems that are designed to operate independently under normal loads, but can help each other in 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 point of time, the arrival rates change, with at least one increasing. 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 $\lambda_i > m_i\mu_{i,i}$ 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 after the overload has occurred, so that we start with the new arrival rate pair (λ_1, λ_2).

2.2. *The FQR-T control for the original queueing model.* 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 $r_{2,1}$. (Ways to choose these parameters are discussed in [16, 19].) 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 long as $\tilde{D}_{1,2}(t) \le k_{1,2}$ and $\tilde{D}_{2,1}(t) \le k_{2,1}$ 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 condition of *one-way sharing*: we allow sharing in only one direction at any time. Thus, in the example above, where sharing is done with pool 2 helping class 1, we do not later allow pool 1 to

help class 2 until there are no more pool-2 agents serving class-1 customers. Sharing is initiated with pool 1 helping class 2 when $\tilde{D}_{2,1}(t) > k_{2,1}$ and there are no pool-2 agents serving class-1 customers. And similarly in the other direction.

In simulation experiments, we found that it may be advantageous to relax one-way sharing in order to prevent the system from becoming stuck with sharing in the wrong direction for a long time when sharing is needed in the opposite direction. That could arise if two different overload incidents occur in rapid succession. In addition, even with well-chosen thresholds to activate sharing, it is possible that one-way sharing would get initiated due to stochastic fluctuations under normal loading. If, after such an event, an overload occurs in the opposite direction, then it might not be possible to activate new sharing in the desired direction. (See Remark 8.1 of [17] for further discussion.) We found that the problem posed by one-way sharing can be substantially reduced by incorporating additional lower thresholds, such that one-way sharing with pool 2 helping class 1 is no longer enforced when the number of class-1 customers served by pool 2 falls below the threshold. However, it is not the purpose of the present paper to investigate detailed implementation of the FQR-T control. Hence, here we simply assume that one-way sharing is enforced.

Once sharing is initialized, the control makes the overloaded X model operate as an overloaded N model, and 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 [16, 19].

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, $i, j = 1, 2$. With the assumptions on the X system and the FQR-T control, the six-dimensional stochastic process $(Q_i(t), Z_{i,j}(t); i, j = 1, 2)$ describing the overloaded system becomes a *continuous-time Markov chain* (CTMC) (with stationary transition rates).

In addition to the thresholds $k_{1,2}$ and $k_{2,1}$, discussed above, the model also includes shifting constants $\kappa_{1,2}$ and $\kappa_{2,1}$. The shifting constants may be introduced after the threshold is dropped, because it may be dictated by the optimal ratio function in [16]. If class 1 is overloaded, then *shifted FQR-T* centers about $\kappa_{1,2}$ instead at about zero. Then every server takes his new customer from the head of queue 1 if $\tilde{D}_{i,j}(t) > \kappa_{1,2}$. Otherwise, it takes the new customer from the head of its own class queue. With shifted FQR-T, we aim to achieve $Q_1(t) \approx r_{1,2}Q_2(t) + \kappa_{1,2}$. We can think of FQR-T as the

special case of shifted FQR-T with $\kappa_{1,2} = 0$.

The beauty of the control is that, for large-scale service systems, FQR-T and shifted FQR-T tend to achieve their purpose; i.e., they keep the two queues approximately in fixed relation.

3. The many-server heavy-traffic fluid limit. In this section we briefly describe the convergence of the sequence of stochastic systems to the fluid limit, as established in [17]. That limit, and the ODE studied here, only describe the performance of the the system after an overload incident has occurred, and only during that overload incident. The analysis describes the performance of the FQR control that is activated after the threshold is crossed.

Since we consider an arbitrary single overload incident, without loss of generality, we assume that class 1 is overloaded, and is more overloaded than class 2, so that class 1 receives help from service-pool 2. (Class 2 may also be overloaded, but less than class 1, so that pool 2 should be serving some class-1 customers.) We also assume that one-way sharing is enforced. As a consequence, the system is behaving like an overloaded N model with a shifted FQR control.

3.1. *Many-server heavy-traffic (MS-HT) scaling.* To develop the fluid limit in [17], we consider a sequence of X systems, indexed by n (denoted by superscript), with arrival rates and number of servers growing proportionally to n, i.e.,

$$\bar{\lambda}_i^n \equiv \frac{\lambda_i^n}{n} \to \lambda_i \quad \text{and} \quad \bar{m}_i^n \equiv \frac{m_i^n}{n} \to m_i \quad \text{as} \quad n \to \infty, \tag{3.1}$$

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

$$\bar{Q}_i^n(t) \equiv \frac{Q_i^n(t)}{n} \quad \text{and} \quad \bar{Z}_{i,j}^n(t) \equiv \frac{Z_{i,j}^n(t)}{n}, \quad i,j = 1,2, \quad t \geq 0. \tag{3.2}$$

For each system n, there are threshold $k_{1,2}^n$ and $k_{2,1}^n$, scaled so that

$$\frac{k_{i,j}^n}{n} \to 0 \quad \text{and} \quad \frac{k_{i,j}^n}{\sqrt{n}} \to \infty \quad \text{as} \quad n \to \infty, \quad i,j = 1,2. \tag{3.3}$$

The first scaling by n is chosen to make the thresholds asymptotically negligible in MS-HT fluid scaling, so they detect overloads immediately when they occur (asymptotically). 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 reasonably well in large systems.

There are also the shifting constants $\kappa_{i,j}^n$ discussed in §2.2, but we do not specify their scale. If sharing is taking place, then at some time it was activated by sending the first class-1 customer to service pool 2. We thus need only consider $\kappa_{1,2}^n$ and the weighted-difference process $\tilde{D}_{1,2}^n(t) \equiv Q_1^n(t) - r_{1,2}^* Q_2^n(t)$. Note that if $\kappa_{1,2}^n \to \infty$, then $\tilde{D}_{1,2}^n \to \infty$ as $n \to \infty$. Hence, we redefine the difference process. Let

$$D^n(t) \equiv (Q_1^n(t) - \kappa^n) - rQ_2^n(t), \quad t \geq 0, \tag{3.4}$$

where $\kappa \equiv \kappa_{1,2}$ and $r \equiv r_{1,2}^*$.

With the new definition in (3.4), we allow κ^n to be of any order less than or equal to $O(n)$; in particular, we assume that $\kappa^n/n \to \kappa$ for $0 \leq \kappa < \infty$. There are two principle cases: $\kappa = 0$ and $\kappa > 0$. The first case produces FQR (after sharing has began); the second case produces shifted FQR (shifted by the constant κ^n).

With the new process D^n in (3.4), we can apply the same FQR routing rule for both the FQR and shifted FQR cases: if $D^n(t) > 0$, then every newly available agent (in either pool) takes his new customer from the head of the class-1 queue. If $D^n(t) \leq 0$, then every newly available agent takes his new customer from the head of his own queue.

3.2. *A heuristic view of the AP.* The AP is concerned with the system behavior when sharing is taking place; i.e., when some, but not all, of the pool 2 agents are serving class 1. That takes place when $Q_1 = rQ_2 + \kappa$. In that situation, it can be shown that the queue-difference process D^n in (3.4) is an order $O(1)$ process, without any spatial scaling, i.e., for each t, the sequence of unscaled random variables $\{D^n(t) : n \geq 1\}$ turns out to be stochastically bounded (or tight) in $\mathbb{R}$. That implies that D^n operates in a time scale that is different from the other processes Q_i^n and $Z_{1,2}^n$, which are scaled by dividing by n in (3.2). With the MS-HT scaling in (3.1), in order for the two queues to change significantly (in a relative sense, which is captured by the scaling in (3.2)), there needs to be $O(n)$ arrivals and departures from the queues. In contrast, the difference process D^n can never go far from 0, because it has drift pointing towards 0 from both above and below. Thus, the difference process oscillates more and more rapidly about 0 as n increases. Thus, over short time intervals in which X^n remains nearly unchanged for large n, the process D^n moves rapidly in its state space, nearly achieving a local steady state. As n increases, the speed of the difference process increases, so that

in the limit, it achieves a steady state instantaneously. That steady state is a local steady state, because it depends on $x(t)$, the fluid limit x at time t.

To formalize this separation of time scales, we define a family of *time-expanded* difference processes: for each $n \geq 1$ and $t \geq 0$, let

$$D_t^n(s) \equiv D^n(t + s/n), \quad s \geq 0. \tag{3.5}$$

Dividing s by n in (3.5) allows us to examine what is happening right after time t in the faster time scale. For each t, a different process D_t^n is defined. For every $t \geq 0$ and $s > 0$, the time increment $[t, t + s/n)$ becomes infinitesimal in the limit. Theorem 4.3 in [17] proves that, for each $t \geq 0$,

$$D_t^n \equiv \{D_t^n(s) : s \geq 0\} \Rightarrow D_t(s) \equiv \{D_t(s) : s \geq 0\} \quad \text{as } n \to \infty \tag{3.6}$$

where the limit $D_t \equiv \{D_t(s) : s \geq 0\}$ is the FTSP, and convergence is in the space $\mathcal{D}$ of right continuous functions with left limits.

For each n, the control depends on whether or not $D^n(t) > 0$. In turn, the limiting ODE depends on the corresponding steady-state probability of the FTSP,

$$\pi_{1,2}(x(t)) \equiv \lim_{s \to \infty} P(D_t(s) > 0) \tag{3.7}$$

which depends on x because the distribution of $\{D_t(s) : s \geq 0\}$ depends on the value of $x(t) \in \mathbb{R}^3$.

In this paper, we directly define the FTSP D_t and its steady-state probability $\pi_{1,2}$ in §6. The limit provides important motivation.

3.3. *The FWLLN.* The main result in [17] is the FWLLN. We now briefly summarize the main part of the statement, without providing all conditions. The limit is for the six-dimensional scaled process $\bar{X}_6^n \equiv (\bar{Q}_i^n, \bar{Z}_{i,j}^n)$, where $\bar{Q}_i^n$ and $\bar{Z}_{i,j}^n$ are defined in (3.2). Let $\mathcal{D}_6(I)$ be the usual space of right-continuous $\mathbb{R}^6$ valued functions on the interval I with left limits everywhere except the left endpoint. Let $\Rightarrow$ denote convergence in distribution. The FWLLN in Theorem 4.1 of [17] concludes, under regularity conditions (including the initial convergence $\bar{X}_6^n(0) \Rightarrow x(0)$ in $\mathbb{R}^6$), that

$$\bar{X}_6^n \Rightarrow x \quad \text{in} \quad \mathcal{D}_6([0, \infty)) \quad \text{as} \quad n \to \infty, \tag{3.8}$$

where $x \equiv (q_i, z_{i,j})$ is a continuous deterministic element of $\mathcal{D}_6([0, \infty))$.

Throughout, the limit x in (3.8) is effectively three dimensional because $z_{1,1}(t) = m_1$, $z_{2,1}(t) = 0$ and $z_{1,2}(t) + z_{2,2}(t) = m_2$ for all t. Hence, the limit can be considered to be of the form $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$. We characterize the limit x in [17] as the solution to the ODE considered in this paper.

4. The ODE. We now specify the ODE, which is the main subject of this paper. The rigorous development starts here.

We introduce the ODE to describe the evolution of the system state of a deterministic fluid model, which is approximating the performance of the stochastic system with FQR-T during an overload incident. The deterministic fluid model has two classes, with class-i input arriving at rate λ_i. There are two service pools with service pool j having capacity m_j. The state at time t is $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$, where $q_i(t)$ is the content of the class-i queue and $z_{i,j}(t)$ is the amount of service pool j occupied by serving class i at time t. Since the service pools are always full in the setting we consider, fluid enters service only from the queue and only when service is completed. In the fluid model, we stipulate that a *proportion* $\pi_{1,2}(x(t))$ of the newly available capacity in pool 2 at time t, i.e., of $z_{1,2}(t)\mu_{1,2} + (m_2 - z_{1,2}(t))\mu_{2,2}$, is allocated to class 1, while the remaining proportion $1 - \pi_{1,2}(x(t))$ is allocated to class 2. The proportion $\pi_{1,2}(x(t))$ is a function of $x(t)$, the state of the fluid model (solution to the ODE) at time t. That function $\pi_{1,2}(x(t))$ is rigorously defined and characterized in §6 below.

For *understanding* why the ODE has the form it does, it is helpful to recall that the fluid model is approximating the queueing system after an overload incident has occurred. We are considering the case in which class 1 is overloaded and more so than class 2. Moreover, we are considering the system after sharing has been initiated with pool 2 starting to help class 1. Both service pools are fully busy and some of pool-2 is serving class 1, so that $z_{1,1}(t) = m_1$, $z_{2,1}(t) = 0$ and $z_{1,2}(t)+z_{2,2}(t) = m_2$. As a consequence, we only need consider $z_{1,2}$ among the four $z_{i,j}$ variables. Hence, as in the statement of the FWLLN above, the ODE is three-dimensional. The associated state space is $\mathbb{S} \equiv [0,\infty)^2 \times [0, m_2]$.

The ODE characterizes the evolution of the fluid model described above. Consistent with the description above, we *define* the transient behavior of the fluid model by the ODE

$$\dot{x}(t) \equiv (\dot{q}_1(t), \dot{q}_2(t), \dot{z}_{1,2}(t)) = \Psi(x(t)) \equiv \Psi(q_1(t), q_2(t), z_{1,2}(t)), \tag{4.1}$$

$t \geq 0$, where $\Psi : [0,\infty)^2 \times [0, m_2] \to \mathbb{R}^3$ can be displayed via

$$\begin{aligned}
\dot{q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(x(t))\left[z_{1,2}(t)\mu_{1,2} + (m_2 - z_{1,2}(t))\mu_{2,2}\right] - \theta_1 q_1(t) \\
\dot{q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(x(t)))\left[(m_2 - z_{1,2}(t))\mu_{2,2} + z_{1,2}(t)\mu_{1,2}\right] - \theta_2 q_2(t) \\
\dot{z}_{1,2}(t) &\equiv \pi_{1,2}(x(t))(m_2 - z_{1,2}(t))\mu_{2,2} - (1 - \pi_{1,2}(x(t)))z_{1,2}(t)\mu_{1,2},
\end{aligned} \tag{4.2}$$

with $\pi_{1,2} : [0,\infty)^2 \times [0, m_2] \to [0, 1]$ defined by §6 below when $q_1 - rq_2 = \kappa$, $\pi_{1,2}(x) \equiv 1$ when $q_1 - rq_2 > \kappa$ and $\pi_{1,2}(x) \equiv 0$ when $q_1 - rq_2 < \kappa$.

We also consider the associated *initial value problem* (IVP)

$$\dot{x}(t) = \Psi(x(t)), \quad x(0) = w_0 \tag{4.3}$$

for $\Psi(x)$ in (4.1)–(4.2).

It is significant that the ODE in (4.1) is *autonomous*, i.e., that Ψ in (4.1) is a function of x and not of t. (It is a function of t only through $x(t)$; we do not have the more general form $\dot{x}(t) = \Psi(x(t), t)$.)

The important function $\pi_{1,2}$ has been defined informally by (3.7) in §3.2 above, referring to the MS-HT limit. The function $\pi_{1,2}$ will be defined here in a rigorous self-contained way in §6. The MS-HT limit is then proved in [17], building on this paper.

5. The main result. The state space $\mathbb{S}$ is a subset of $\mathbb{R}^3$ with the boundary constraints: $q_1 \geq 0$, $q_2 \geq 0$ and $0 \leq z_{1,2}(t) \leq m_2$. The differential equation for $z_{1,2}$ prevents its boundary states 0 and m_2 from being active, because $\dot{z}_{1,2}(t) = \pi_{1,2}(x(t))m_2\mu_{2,2} \geq 0$ when $z_{1,2}(t) = 0$ and $\dot{z}_{1,2}(t) = -(1-\pi_{1,2}(x(t))m_2\mu_{1,2} \leq 0$ when $z_{1,2}(t) = m_2$. However, the queue-length constraints can alter the evolution. In general, we can have $\dot{q}_i(t) < 0$ when $q_i(t) = 0$, which we understand as leaving $q_i(t)$ fixed at 0. However, we are primarily interested in overloaded cases, in which these boundaries are not reached. Under Assumption A below, we can consider the ODE without constraints.

Recall that the shifting constant satisfies $\kappa \geq 0$. We consider the *restricted state space* $S \equiv [\kappa, \infty) \times [0, \infty) \times [0, m_2]$. We thus avoid the transient region in which $q_1 < rq_2 + \kappa$ with $q_2 = 0$, where $\dot{q}_1(t) > 0$ and $\dot{q}_2(t) < 0$, but q_2 remains at 0 while q_1 increases to the shifting constant κ. The restricted state space, with $q_1 \geq \kappa$ is shown to be the space of the fluid limit of the system in [17]. We will also show in Theorem 5.1 below that the ODE cannot leave this restricted state space.

It is convenient to specify the conditions on the model parameters in terms of the steady-state formulas for the queues in isolation. For that purpose, let q_i^a be the length of fluid-queue i and let s_i^a be the amount of spare service capacity in service-pool i, in steady state, when there is no sharing, $i = 1, 2$. The quantities q_i^a and s_i^a are well known, since they are the steady state quantities of the fluid model for the Erlang-A model $(M/M/m_i + M)$ with arrival-rate λ_i, service-rate $\mu_{i,i}$ and abandonment-rate θ_i; see Theorem 2.3 in [23], especially equation (2.19), and §5.1 in [16]. In particular,

$$q_i^a \equiv \frac{(\lambda_i - \mu_{i,i}m_i)^+}{\theta_i} \quad \text{and} \quad s_i^a \equiv \left(m_i - \frac{\lambda_i}{\mu_{i,i}}\right)^+, \quad i = 1, 2, \tag{5.1}$$

where $(x)^+ \equiv \max\{x, 0\}$. It is easy to see that $q_i^a s_i^a = 0$, $i = 1, 2$. We thus make the following assumption, which is *assumed to hold henceforth.*

ASSUMPTION A.

(I) The model parameters satisfy $\theta_1(q_1^a - \kappa) \geq \mu_{1,2} s_2^a$.
(II) The initial conditions satisfy $x(0) \in \mathbb{S} \equiv [\kappa, \infty) \times [0, \infty) \times [0, m_2]$.

We now explain these assumptions. Clearly, a sufficient condition for both pools to be overloaded is to have $s_1^a = s_2^a = 0$, i.e., to have no spare service capacity in either pool in their individual steady states. However, if $s_2^a > 0$, both pools can still be overloaded, provided that enough class-1 fluid is processed in pool 2. To have the solution be eventually in $\mathbb{S}$, we require that $\theta_1(q_1^a - \kappa) \geq \mu_{1,2} s_2^a$. This condition ensures that service pool 2 is also full of fluid when sharing is taking place, i.e., $z_{1,2}(t) + z_{2,2}(t) = m_2$ for all $t \geq 0$ (assuming that pool 2 is full at time 0). To see why, note that when service-pool 2 has spare service capacity ($s_2^a > 0$), sharing will be activated if $q_1^a > \kappa$, because $q_2^a = 0$. Now, the maximum amount of class-1 fluid that pool 2 can process, while still processing all of the class-2 fluid (so that q_2 is kept at zero), is $\mu_{1,2} s_2^a$. hence, $\mu_{1,2} s_2^a$ is the minimal amount of class-1 fluid that should flow to pool 2. On the other hand, $\theta_1 q_1^a = \lambda_1 - \mu_{1,1} m_1$ is equal to the "extra" class-1 fluid input, i.e., all the class-1 fluid that pool 1 cannot process. Some of this "extra" class-1 fluid might abandon (if $q_1 > 0$). The minimal amount of class-1 fluid that abandons is $\theta_1 \kappa$ (but κ can be equal to zero).

We thus require that all the class-1 fluid, *that is not served in pool* 1, minus the minimal amount of class-1 fluid that abandons, is larger than $\mu_{1,2} s_2^a$. With this requirement, pool 2 is assured to be full, assuming that it is initialized full. (If pool 2 is not initialized full, then it will fill up after some finite time period; see the appendix.)

REMARK 5.1 (class 1 need not be more overloaded than class 2). In this paper we are interested in analyzing the ODE (4.2) as given. Hence, in Assumption A we do not assume that class 1 is more overloaded than class 2; i.e., we do not require that $q_1^a - \kappa \geq r q_2^a$. This extra assumption is not needed for our results for the specified ODE. In contrast, this assumption is needed in order to show that the ODE holds as the fluid limit, with class 1 receiving help; see Assumption 3.1 in [17].

We exploit Assumption A to show that the boundaries of $\mathbb{S}$ in $\mathbb{R}^3$ play no role.

THEOREM 5.1. *$x(t) \in \mathbb{S}$ for all $t \geq 0$.*

We give the proof in §8.4 after the necessary tools have been developed. Our main result establishes the existence of a unique solution.

THEOREM 5.2 (existence and uniqueness). *For any $w_0 \in \mathbb{S}$, there exists a unique function $x : [0, \infty) \to \mathbb{S}$ such that, (i) for all $t \geq 0$, there exist $\delta(t) > 0$ such that x is right-differentiable at t, differentiable on $(t, t + \delta(t))$ and satisfies the IVP (4.3) based on the ODE (4.1) over $[t, t + \delta(t))$ with initial value $x(t)$, and (ii) x is continuous and differentiable almost everywhere.*

Theorem 5.2 has two parts: First, there is (i) establishing the local existence and uniqueness of a conventional differentiable solution on each interval $[t, t+\delta(t))$, for which it suffices to consider a single t, e.g., $t = 0$. Second, there is (ii) justifying an overall continuous solution.

We prove Theorem 5.2 in the next two sections. The proof is tied to the characterization of $\pi_{1,2}$ in (4.2) and (3.7), and thus the FTSP D_t. We need to determine conditions for the FTSP D_t to be positive recurrent, so that the AP holds, and then calculate its steady-state distribution in order to find $\pi_{1,2}$. Moreover, we need to establish topological properties of the function $\pi_{1,2}$, such as continuity and differentiability. We discuss the FTSP D_t next.

6. The fast-time-scale process. In this section we define the function $\pi_{1,2}$, which is a crucial component of the ODE in (4.2). The value $\pi_{1,2}(x)$ depends on the state $x \in \mathbb{S}$ of the ODE. The value $\pi_{1,2}(x)$ is a steady-state probability of the fast-time-scale process (FTSP) D_t, which also depends on the state x of the ODE. Below we will first define the FTSP and then we will characterize $\pi_{1,2}$.

For understanding, it is helpful to recall §3.2, where we indicated that the FTSP D_t arises as the limit of D_t^n without any scaling (see (3.6)), where D_t^n is the time-expanded difference process defined in (3.5) associated with the queue-difference stochastic process $D^n \equiv (Q_1^n - \kappa^n) - rQ_2^n$ in (3.4). Since there is no scaling of space, the state space for the FTSP D_t is the countable lattice $\{\pm j \pm kr : j, k \in \mathbb{Z}\}$ in $\mathbb{R}$. To see this, first observe from (3.4) that D^n has state space $\{\pm j \pm kr - \kappa^n : j, k \in \mathbb{Z}\}$. Next, because of the subtraction in (3.5), D_t^n has state space $\{\pm j \pm kr : j, k \in \mathbb{Z}\}$. Finally, because of the convergence in (3.6), the FTSP D_t has this same state space. This limiting behavior motivates what we do below, but here what is given below can be taken as the definition.

6.1. *The fast-time-scale CTMC.* We fix a time t and assume that we are given the value $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$. In order to simplify the analysis

we assume that r is rational. That clearly is without any practical loss of generality. Specifically, we assume that $r = j/k$ for some positive integers j and k without any common factors. We then multiply the process by k, so that all transitions can be expressed as $\pm j$ or $\pm k$ in the state space $\mathbb{Z}$. In that case, the FTSP $D_t \equiv \{D_t(s) : s \geq 0\}$ becomes a CTMC.

Let $\lambda_+^{(j)}(x(t))$, $\lambda_+^{(k)}(x(t))$, $\mu_+^{(j)}(x(t))$ and $\mu_+^{(k)}(x(t))$ be the transition rates of the FTSP D_t for transitions of $+j$, $+k$, $-j$ and $-k$, respectively, when $D_t(s) > 0$. Similarly, we define the transitions when $D_t(s) \leq 0$: $\lambda_-^{(j)}(x(t))$, $\lambda_-^{(k)}(x(t))$, $\mu_-^{(j)}(x(t))$ and $\mu_-^{(k)}(x(t))$. (We remark that these rates are the limits of the rates of D_t^n as $n \to \infty$ with $\bar{X}^n(t) \Rightarrow x(t)$.)

First, for $D_t(s) \in (-\infty, 0]$, the upward rates are

$$\begin{aligned} &\lambda_-^{(k)}(x(t)) = \lambda_1, \\ &\lambda_-^{(j)}(x(t)) = \mu_{1,2} z_{1,2}(t) + \mu_{2,2}(m_2 - z_{1,2}(t)) + \theta_2 q_2(t), \end{aligned} \tag{6.1}$$

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

$$\mu_-^{(k)}(x(t)) = \mu_{1,1} z_{1,1}(t) + \theta_1 q_1(t), \qquad \mu_-^{(j)}(x(t)) = \lambda_2, \tag{6.2}$$

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 $D_t(s) \in (0, \infty)$, we have upward rates

$$\lambda_+^{(k)}(x(t)) = \lambda_1, \qquad \lambda_+^{(j)}(x(t)) = \theta_2 q_2(t), \tag{6.3}$$

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

$$\begin{aligned} &\mu_+^{(k)}(x(t)) = \mu_{1,1} z_{1,1}(t) + \mu_{1,2} z_{1,2}(t) + \mu_{2,2}(m_2 - z_{1,2}(t)) + \theta_1 q_1(t), \\ &\mu_+^{(j)}(x(t)) = \lambda_2, \end{aligned} \tag{6.4}$$

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.

6.2. *Representing the FTSP D_t as a QBD.* Further analysis is simplified by exploiting matrix geometric methods, as in [11]. In particular, we represent the integer-valued CTMC $D_t \equiv \{D_t(s) : s \geq 0\}$ just constructed as a homogeneous continuous-time *quasi-birth-and-death* (QBD) process, as in Definition 1.3.1 and §6.4 of [11]. In passing, note that the special case $r = 1$ is especially tractable, because then the QBD process reduces to an ordinary *birth-and-death* (BD) process.

To represent D_t as a QBD process, we must re-order 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 [11] (imitating the shape of a generator matrix of a BD process). In particular, we write

$$Q \equiv \begin{pmatrix} B & A_0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \tag{6.5}$$

where the four component submatrices B, A_0, A_1 and A_2 are all $2m \times 2m$ submatrices for $m \equiv \max\{j, k\}$. In particular, 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.,

$$B \equiv \begin{pmatrix} A_1^+ & B_\mu \\ B_\lambda & A_1^- \end{pmatrix} \quad \text{and} \quad A_i \equiv \begin{pmatrix} A_i^+ & 0 \\ 0 & A_i^- \end{pmatrix} \tag{6.6}$$

for $i = 0, 1, 2$. (All matrices are also functions of $x(t)$ because for each t we have a different generator matrix corresponding to the FTSP D_t.)

To achieve this representation, we need to re-order the states into levels. The main idea is to represent transitions of D_t above 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:

(6.7)
$$\phi(2nm + i) \equiv nm + i \quad \text{and} \quad \phi((2n+1)m + i) \equiv -nm - i + 1, \quad 1 \leq i \leq m.$$

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 (6.5) 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.

To illustrate, consider an example with $r = 0.8$, so that we can choose $j = 4$ and $k = 5$, yielding $m = 5$. The states are ordered in levels as follows

$$\begin{aligned} L(0) &= \{1,2,3,4,5,0,-1,-2,-3,-4\}, \\ L(1) &= \{6,7,8,9,10,-5,-6,-7,-8,-9\}, \\ L(2) &= \{11,12,13,14,15,-10,-11,-12,-13,-14\}, \quad \ldots \end{aligned}$$

Then the submatrices B_μ, B_λ, A_i^+ and A_i^-, which form the block matrices B and A_i, $i = 0,1,2$, have the form in (6.12) below, where

$$\sigma_+ = \lambda_+^{(5)} + \lambda_+^{(4)} + \mu_+^{(5)} + \mu_+^{(4)} \quad \text{and} \quad \sigma_- = \lambda_-^{(5)} + \lambda_-^{(4)} + \mu_-^{(5)} + \mu_-^{(4)}. \tag{6.8}$$

(We solve a full numerical example with these matrices in §11.3.)

Henceforth, we refer to D_t interchangeably as the QBD or the FTSP.

6.3. *Positive recurrence.* We show that positive recurrence depends only on the constant drift rates in the two regions, as one would expect. Let δ_+ and δ_- be the drift in the positive and negative region, respectively; i.e., let

$$\begin{aligned} \delta_+(x(t)) &\equiv j\left(\lambda_+^{(j)}(x(t)) - \mu_+^{(j)}(x(t))\right) + k\left(\lambda_+^{(k)}(x(t)) - \mu_+^{(k)}(x(t))\right) \\ \delta_-(x(t)) &\equiv j\left(\lambda_-^{(j)}(x(t)) - \mu_-^{(j)}(x(t))\right) + k\left(\lambda_-^{(k)}(x(t)) - \mu_-^{(k)}(x(t))\right). \end{aligned} \tag{6.9}$$

THEOREM 6.1. *The QBD D_t is positive recurrent if and only if*

$$\delta_-(x(t)) > 0 > \delta_+(x(t)). \tag{6.10}$$

PROOF. We employ the theory in §7 of [11], modified for the continuous-time QBD. We first construct the aggregate matrices $A \equiv A_0 + A_1 + A_2$, $A^+ \equiv A_0^+ + A_1^+ + A_2^+$ and $A^- \equiv A_0^- + A_1^- + A_2^-$. We then observe that the aggregate matrix A is reducible, so we need to consider the component matrices A^+ and A^-, which both are irreducible CTMC infinitesimal generators in their own right. Let ν^+ and ν^- be the unique stationary probability vectors of A^+ and A^-, respectively, e.g., with $\nu^+ A^+ = 0$ and $\nu^+ \mathbf{1} = \mathbf{1}$. The theory concludes that our QBD is positive recurrent if and only if

$$\nu^+ A_0^+ \mathbf{1} < \nu^+ A_2^+ \mathbf{1} \quad \text{and} \quad \nu^- A_0^- \mathbf{1} < \nu^- A_2^- \mathbf{1}. \tag{6.11}$$

In our application it is easy to see that both ν^+ and ν^- are the uniform probability vector, attaching probability $1/m$ to each of the m states, from which the conclusion follows directly. □

(6.12)

$$B_\mu = \begin{pmatrix} 0 & 0 & 0 & \mu_+^{(4)} & \mu_+^{(5)} \\ 0 & 0 & \mu_+^{(4)} & \mu_+^{(5)} & 0 \\ 0 & \mu_+^{(4)} & \mu_+^{(5)} & 0 & 0 \\ \mu_+^{(4)} & \mu_+^{(5)} & 0 & 0 & 0 \\ \mu_+^{(5)} & 0 & 0 & 0 & 0 \end{pmatrix} \qquad B_\lambda = \begin{pmatrix} 0 & 0 & 0 & \lambda_-^{(4)} & \lambda_-^{(5)} \\ 0 & 0 & \lambda_-^{(4)} & \lambda_-^{(5)} & 0 \\ 0 & \lambda_-^{(4)} & \lambda_-^{(5)} & 0 & 0 \\ \lambda_-^{(4)} & \lambda_-^{(5)} & 0 & 0 & 0 \\ \lambda_-^{(5)} & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_0^+ = \begin{pmatrix} \lambda_+^{(5)} & 0 & 0 & 0 & 0 \\ \lambda_+^{(4)} & \lambda_+^{(5)} & 0 & 0 & 0 \\ 0 & \lambda_+^{(4)} & \lambda_+^{(5)} & 0 & 0 \\ 0 & 0 & \lambda_+^{(4)} & \lambda_+^{(5)} & 0 \\ 0 & 0 & 0 & \lambda_+^{(4)} & \lambda_+^{(5)} \end{pmatrix} \qquad A_0^- = \begin{pmatrix} \mu_-^{(5)} & 0 & 0 & 0 & 0 \\ \mu_-^{(4)} & \mu_-^{(5)} & 0 & 0 & 0 \\ 0 & \mu_-^{(4)} & \mu_-^{(5)} & 0 & 0 \\ 0 & 0 & \mu_-^{(4)} & \mu_-^{(5)} & 0 \\ 0 & 0 & 0 & \mu_-^{(4)} & \mu_-^{(5)} \end{pmatrix}$$

$$A_1^+ = \begin{pmatrix} -\sigma_+ & 0 & 0 & 0 & \lambda_+^{(4)} \\ 0 & -\sigma_+ & 0 & 0 & 0 \\ 0 & 0 & -\sigma_+ & 0 & 0 \\ 0 & 0 & 0 & -\sigma_+ & 0 \\ \mu_+^{(4)} & 0 & 0 & 0 & -\sigma_+ \end{pmatrix} \qquad A_1^- = \begin{pmatrix} -\sigma_- & 0 & 0 & 0 & \mu_-^{(4)} \\ 0 & -\sigma_- & 0 & 0 & 0 \\ 0 & 0 & -\sigma_- & 0 & 0 \\ 0 & 0 & 0 & -\sigma_- & 0 \\ \lambda_-^{(4)} & 0 & 0 & 0 & -\sigma_- \end{pmatrix}$$

$$A_2^+ = \begin{pmatrix} \mu_+^{(5)} & \mu_+^{(4)} & 0 & 0 & 0 \\ 0 & \mu_+^{(5)} & \mu_+^{(4)} & 0 & 0 \\ 0 & 0 & \mu_+^{(5)} & \mu_+^{(4)} & 0 \\ 0 & 0 & 0 & \mu_+^{(5)} & \mu_+^{(4)} \\ 0 & 0 & 0 & 0 & \mu_+^{(5)} \end{pmatrix} \qquad A_2^- = \begin{pmatrix} \lambda_-^{(5)} & \lambda_-^{(4)} & 0 & 0 & 0 \\ 0 & \lambda_-^{(5)} & \lambda_-^{(4)} & 0 & 0 \\ 0 & 0 & \lambda_-^{(5)} & \lambda_-^{(4)} & 0 \\ 0 & 0 & 0 & \lambda_-^{(5)} & \lambda_-^{(4)} \\ 0 & 0 & 0 & 0 & \lambda_-^{(5)} \end{pmatrix}$$

The alternative cases are simplified by the following relation:

$$\begin{aligned} \delta_-(x(t)) - \delta_+(x(t)) &= (j+k)(\mu_{1,2} z_{1,2} + (m_2 - z_{1,2})\mu_{2,2}) \\ &> (j+k) m_2 (\mu_{1,2} \wedge \mu_{2,2}) > 0. \end{aligned} \tag{6.13}$$

Hence there are only two cases in which the drift does not point inward: (i) $\delta_+(x(t)) \geq 0$ and $\delta_-(x(t)) > 0$, (ii) $\delta_-(x(t)) \leq 0$ and $\delta_+(x(t)) < 0$. In both cases the behavior is unambiguous: In case (i), clearly $\pi_{1,2}(x(t)) = 1$; in case (ii), clearly $\pi_{1,2}(x(t)) = 0$.

6.4. *Computing* $\pi_{1,2}$. When the QBD is positive recurrent, the stationary vector of the QBD can be expressed as $\alpha \equiv \{\alpha_n : n \geq 0\} \equiv \{\alpha_{n,j} : n \geq 0, 1 \leq j \leq m\}$, where $\alpha_n \equiv (\alpha_n^+, \alpha_n^-)$ for each n, with α_n^+ and α_n^- both being $1 \times m$ vectors. Then the desired probability $\pi_{1,2}$ can be expressed as

$$\pi_{1,2} = \sum_{n=0}^{\infty} \sum_{j=1}^{m} \alpha_{n,j}^+ = \sum_{n=0}^{\infty} \alpha_n^+ \mathbf{1} = \sum_{n=0}^{\infty} \alpha_n \mathbf{1}_+, \tag{6.14}$$

where $\mathbf{1}$ denotes a column vector with all entries 1 of the right dimension (here $m \times 1$), while $\mathbf{1}_+$ represents a $2m \times 1$ column vector, with m $1's$ followed by m $0's$.

By Theorem 6.4.1 and Lemma 6.4.3 of [11], the steady-state distribution has the matrix-geometric form

$$\alpha_n = \alpha_0 R^n, \tag{6.15}$$

where 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 [11] (see §11 below). Since the matrices A_0, A_1 and A_2 have the block-diagonal form in (6.6), so does R, with submatrices R^+ and R^-.

Since the spectral radius of the rate matrix R is strictly less than 1 (Corollary 6.2.4 of [11]), the sum of powers of R is finite, yielding

$$\sum_{n=0}^{\infty} R^n = (I - R)^{-1}.$$

Also, by Lemma 6.3.1 of [11], the boundary probability vector α_0 in (6.15) is the unique solution to the system

$$\alpha_0(B + RA_2) = 0 \quad \text{and} \quad \alpha \mathbf{1} = \alpha_0 (I - R)^{-1} \mathbf{1} = 1. \tag{6.16}$$

Finally, given the above, and using (6.14), we see that the desired quantity $\pi_{1,2}$ can be represented as

$$\pi_{1,2} = \alpha_0 (I - R)^{-1} \mathbf{1}_+. \tag{6.17}$$

For further analysis, it is convenient to have alternative representations for $\pi_{1,2}(x)$. Let the vector $\mathbf{1}$ have the appropriate dimension in (6.19) below.

THEOREM 6.2 (alternative representations for $\pi_{1,2}$). *Assume that $\delta_+(x) < 0 < \delta_-(x)$, so that the QBD is positive recurrent at x. (a) For $r = 1$,*

$$\pi_{1,2}(x) = \frac{\delta_-(x)}{\delta_-(x) - \delta_+(x)}. \tag{6.18}$$

(b) For rational r, we have the sub-block representation

$$\pi_{1,2}(x) = \frac{\alpha_0^+(x)(I - R^+(x))^{-1} \mathbf{1}}{\alpha_0^+(x)(I - R^+(x))^{-1} \mathbf{1} + \alpha_0^-(x)(I - R^-(x))^{-1} \mathbf{1}}, \tag{6.19}$$

where we choose $\alpha_0(x)$ to satisfy $\alpha_0(B(x) + R(x)A_2(x)) = 0$, renormalize to $\alpha_0(x)\mathbf{1} = 1$, which corresponds to multiplying the original $\alpha_0(x)$ by a constant, decompose $\alpha_0(x)$ consistent with the blocks as $\alpha_0(x) = (\alpha_0^+(x), \alpha_0^-(x))$.

PROOF. (a) When $r = 1$, the FTSP $D_t \equiv D_t(x)$ evolves as an $M/M/1$ queue in each of the regions $D_t > 0$ and $D_t \le 0$. Thus, we can look at the system at the successive times at which D_t transitions from state 0 to state 1, and then again from state 1 to state 0. That construction produces an alternating renewal process of occupation times in each region, where these occupation times are distributed as the busy periods of the corresponding $M/M/1$ queues. Hence, $\pi_{1,2}(x)$ can be expressed as

$$\pi_{1,2}(x) = \frac{E[T^+(x)]}{E[T^+(x)] + E[T^-(x)]}, \tag{6.20}$$

where $T^+(x)$ is the busy period of the $M/M/1$ queue in the upper region, while $T^-(x)$ is the busy period of the $M/M/1$ queue in the lower region. By the definition of $\mathbb{A}$, these mean busy periods are finite in each region. In particular,

$$E[T^\pm(x)] = \frac{1}{\mu^\pm(x)(1-\rho^\pm(x))} = \frac{1}{\mu^\pm(x) - \lambda^\pm(x)} = \frac{1}{|\delta_\pm(x)|}, \tag{6.21}$$

where $\rho^\pm(x) \equiv \lambda^\pm(x)/\mu^\pm(x)$, $\lambda^+(x)$ and $\mu^+(x)$ are the constant drift rates up (away from the boundary) and down (toward the boundary) in the upper region in (6.3) and (6.4), depending on state x, while $\lambda^-(x)$ and $\mu^-(x)$ are the constant drift rates down (away from the boundary) and up (toward the boundary) in the lower region in (6.1) and (6.2); e.g., $\lambda^-(x) \equiv \mu_-^{(j)}(x) + \mu_-^{(k)}(x)$ with $j = k = 1$ from (6.2).

(b) We first observe that we can reason as in the case $r = 1$, using a regenerative argument. We can let the regeneration times be successive transitions from one specific QBD state in level 0 with $D_t \le 0$ to a specific state in level 1 where $D_t > 0$. The intervals between successive transitions will be i.i.d. random variables with finite mean. Hence, we can represent $\pi_{1,2}(x)$ just as in (6.20), but where now $T^+(x)$ is the total occupation time in the upper region with $D_t(s) > 0$ during a regeneration cycle, while $T^-(x)$ is the total occupation time in the lower region with $D_t(s) \le 0$ during a regeneration cycle. Each of these occupation times can be broken up into first passage times. For example, $T^+(x)$ is the sum of first passage times from some state at level 0 to some other state in level 1 where $D_t(s) > 0$. The regenerative cycle will end when the starting and ending states within levels 0 and 1 are the designated pair associated with the specified regeneration time. The successive pairs (i, j) of starting and ending states within the levels 0 and 1 evolve according to a positive-recurrent finite-state discrete-time Markov chain.

Paralleling that regenerative argument, we can work with the QBD matrices, as in (6.17), but now using an alternative representation. Since $\mathbf{1}_+ + \mathbf{1}_- = \mathbf{1}$, where all column vectors are $2m \times 1$, we can apply the second equation in (6.16) to write

$$\pi_{1,2} = \frac{\alpha_0 (I - R)^{-1} \mathbf{1}_+}{\alpha_0 (I - R)^{-1} \mathbf{1}_+ + \alpha_0 (I - R)^{-1} \mathbf{1}_-}.$$

Then we can choose α_0 to satisfy $\alpha_0(B + RA_2) = 0$, renormalize to $\alpha_0 \mathbf{1} = 1$ (which corresponds to multiplying the original α_0 by a constant), decompose α_0 consistent with the blocks, letting $\alpha_0 = (\alpha_0^+, \alpha_0^-)$, to obtain (6.19). □

With the QBD representation, we can determine when the FTSP D_t is positive recurrent, for a given $x(t)$, using (6.10), and then numerically calculate $\pi_{1,2}$. That allows us to numerically solve the ODE (4.1) in §11. We will also use the representations (6.17), (6.18), (6.19) and other QBD properties to deduce topological properties of $\pi_{1,2}$.

7. Existence and uniqueness of solutions. This section is devoted to proving Theorem 5.2. For the local existence and uniqueness in Theorem 5.2 (i), we will show that the function Ψ in (4.2) is locally Lipschitz continuous in Theorem 7.1 below. That allows us to apply the classical Picard-Lindelöf theorem to deduce the desired existence and uniqueness of solutions to the IVP (4.3); see Theorem 2.2 of Teschl [21] or Theorem 3.1 in [10]. Afterwards, in §7.4 we establish the global properties in Theorem 5.2 (ii).

7.1. *Subsets of the state space.* The analysis is complicated because the function Ψ in (4.1) and (4.2) is not continuous, let alone Lipschitz continuous, on all of the state space of the ODE, i.e., on $\mathbb{S} \equiv [\kappa, \infty) \times [0, \infty) \times [0, m_2] \equiv \{(q_1, q_2, z_{1,2})\}$. However, we can obtain the required local Lipschitz continuity in appropriate neighborhoods about each point in $\mathbb{S}$, but specifying these neighborhoods requires care. When care is taken, we can eventually construct a continuous solution to the ODE, which is differentiable a.e. with respect to Lebesgue measure.

We first divide the state space $\mathbb{S}$ into three regions:

$$\mathbb{S}^{\mathbf{b}} \equiv \{q_1 - rq_2 = \kappa\}, \quad \mathbb{S}^+ \equiv \{q_1 - rq_2 > \kappa\}, \quad \mathbb{S}^- \equiv \{q_1 - rq_2 < \kappa\}, \tag{7.1}$$

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 $\mathbb{S}$ where the AP

is taking place. In $\mathbb{S}^b$ the function $\pi_{1,2}$ can assume its full range of values, $0 \le \pi_{1,2}(x) \le 1$.

The regions $\mathbb{S}^+$ and $\mathbb{S}^-$ are open subsets of $\mathbb{S}$. For all $x \in \mathbb{S}^+$, $\pi_{1,2}(x) = 1$; for all $x \in \mathbb{S}^-$, $\pi_{1,2}(x) = 0$. In order for $\mathbb{S}^-$ to be a proper subspace of $\mathbb{S}$, both service pools must be constantly full. Thus, if $x \in \mathbb{S}^-$, then $z_{1,1} = m_1$ and $z_{1,2} + z_{2,2} = m_2$, but q_1 and q_2 are allowed to be equal to zero.

To analyze Ψ on $\mathbb{S}^b$, we exploit properties of the QBD introduced in §6. We partition $\mathbb{S}^b$ into three subsets, depending on the drift rates in (6.9). Let $\mathbb{A}$ be the set of all $x \in \mathbb{S}^b$ for which the QBD is positive recurrent, as given in (6.10); i.e., let

$$\mathbb{A} \equiv \{x \in \mathbb{S}^b \mid \delta_-(x) > 0 > \delta_+(x)\}. \tag{7.2}$$

Let the other two subsets be

$$\mathbb{A}^+ \equiv \{x \in \mathbb{S}^b \mid \delta_+(x) \ge 0\} \quad \text{and} \quad \mathbb{A}^- \equiv \{x \in \mathbb{S}^b \mid \delta_-(x) \le 0\}. \tag{7.3}$$

By the relation (6.13), there are no other alternatives; i.e., $\mathbb{S}^b = \mathbb{A} \cup \mathbb{A}^+ \cup \mathbb{A}^-$. From §6, we know that $\pi_{1,2}(x) = 1$ in $\mathbb{A}^+$, while $\pi_{1,2}(x) = 0$ in $\mathbb{A}^-$.

7.2. *Local Lipschitz continuity.* We now are ready to establish the local Lipschitz continuity. Here we need an unconventional setting, because we will be changing the reference set.

Definition 7.1 (local Lipschitz continuity). A function $f : \Omega_2 \to \mathbb{R}^m$, where $\Omega_1 \subseteq \Omega_2$ with Ω_2 a connected subset in $\mathbb{R}^n$, is locally Lipschitz continuous on Ω_1 within Ω_2 if, for every $v_0 \in \Omega_1$, there exists a neighborhood $U \subseteq \Omega_2$ of v_0 such that f restricted to U is Lipschitz continuous; i.e., there exists a constant $K \equiv K(U)$ such that $\|f(v_1) - f(v_2)\| \le K\|v_1 - v_2\|$ for every $v_1, v_2 \in U$.

The complexity of Definition 7.1 occurs because we are envisioning, not a single application of the Picard-Lindelof Theorem, but instead different applications over different regions. When we consider initial states in the set $\mathbb{A}$, we are regarding $\mathbb{A}$ as a two-dimensional subset; we are regarding $\mathbb{A}$ as a subset of $\mathbb{R}^2$. In the other cases, the subset is three-dimensional.

To elaborate, let $\mathbb{S}^2$ be the subset of all $x \equiv (x_1, x_2, x_3)$ in $\mathbb{R}^3$ such that $x_1 - rx_2 = \kappa$. Equivalently, we can let $\mathbb{S}^2$ be the set of all (x_2, x_3) in $\mathbb{R}^2$, which is just $\mathbb{R}^2$ itself, with the understanding that we separately define x_1 in terms of x_2 via $x_1 = rx_2 + \kappa$. In this sense $\mathbb{S}^2$ can directly be identified with $\mathbb{R}^2$. It is a Banach space, as required for the Picard-Lindelof Theorem. Moreover, it is easy to see that $\mathbb{A}$ is an open subset of $\mathbb{S}^b$ and thus of the Banach

space $\mathbb{S}^2$. When we consider the set $\mathbb{A}$, we are requiring that $q_1 = rq_2 + \kappa$ hold for all time (over the short interval we are considering). Exploiting this perspective, we will show that the IVP starting in $\mathbb{A}$ has a solution that remains in $\mathbb{A}$ over a positive interval. A variant of that same conclusion will be deduced for every initial point in $\mathbb{S}$, but for the other statements the region is three-dimensional.

THEOREM 7.1. *The function Ψ in (4.2) is locally Lipschitz continuous on $\mathbb{A}$ within $\mathbb{S}^b$, on $\mathbb{S}^+$ within $\mathbb{S}^+$, on $\mathbb{S}^-$ within $\mathbb{S}^-$, on $\mathbb{A}^+$ within $\mathbb{S}^b \cup \mathbb{S}^+$ and on $\mathbb{A}^-$ within $\mathbb{S}^b \cup \mathbb{S}^-$.*

We prove Theorem 7.1 in §12, drawing heavily upon the properties of $\pi_{1,2}$ established in §6. In particular, we use the fact that $\pi_{1,2}(x)$ is continuous in $\mathbb{S}^b$. In order to ensure that the ODE starting at a point in the set $\mathbb{A}$ within $\mathbb{S}^b \subset \mathbb{S}^2$ remains in the set $\mathbb{A}$, we apply the FTSP studied in §6. The FTSP has a proper steady state distribution at a state x, with $0 < \pi_{1,2}(x) < 1$, if and only if $x \in \mathbb{A}$, which requires that $q_1 = rq_2+\kappa$. Since $\pi_{1,2}(x)$ is continuous for $x \in \mathbb{A}$, that steady-state probability can change only smoothly, all of which takes place within the set $\mathbb{A}$. This property of the FTSP implies that the solution to the ODE, starting within $\mathbb{A}$ will remain in $\mathbb{A}$ for a short interval of time. For any x outside $\mathbb{A}$, we necessarily have either $\pi_{1,2}(x) = 0$ or $\pi_{1,2}(x) = 1$.

7.3. *Proof of Theorem 5.2 (i).*

PROOF. Note that all of $\mathbb{S}$ is covered by the five cases in Theorem 7.1; i.e., every point in $\mathbb{S}$ belongs to one (and only one) of the five sets $\mathbb{A}$, $\mathbb{S}^+$, $\mathbb{S}^-$, $\mathbb{A}^+$ and $\mathbb{A}^-$. In each of the five cases, we can apply Theorem 7.1 to conclude that Ψ is locally Lipschitz continuous on Ω_1 within Ω_2 for the specified reference set Ω_2. Thus, for every initial point in $\mathbb{S}$, we can apply the Picard-Lindelöf theorem to deduce the desired existence and uniqueness of solutions to the IVP (4.3) over an interval $[0, \delta)$ for $\delta > 0$; see Theorem 2.2 of Teschl [21]. That solution will be right differentiable at $t = 0$ and differentiable in the open interval $(0, \delta)$. In particular, this reasoning applies in the case $\mathbb{A}$; in that cases, as discussed after Definition 7.1, the set $\mathbb{A}$ is regarded as a two-dimensional subset of the Banach space $\mathbb{S}^2$. As a consequence of this reasoning, we deduce that if the initial point is in $\mathbb{A}$, then there will be a solution that remains entirely in $\mathbb{A}$ over an initial interval $[0, \delta)$.

So far, this construction only yields a solution over the initial time interval $[0, \delta)$ for some $\delta > 0$. However, if $x(t)$ is the value of a solution to the ODE at time t and that value is within $\mathbb{S}$, then the same construction applies over

an interval $[t, t+\delta)$. Hence, provided that we can establish claim (ii), we obtain the full statement in part (i). □

7.4. *Global existence and uniqueness.* The remainder of section is devoted to completing the proof of Theorem 5.2 by establishing global existence and uniqueness. We first observe that, in general, one overall *differentiable* solution to the ODE over $[0,\infty)$ may not exist. From $\mathbb{S}^+$ or $\mathbb{S}^-$, the solution can eventually move to anywhere in $\mathbb{S}^b$. That movement can produce a discontinuity in $\pi_{1,2}(x)$ and thus in Ψ. For example, in $\mathbb{S}^+$ we necessarily have $\pi_{1,2}(x) = 1$. However, in general there is nothing preventing $x(t) \to x(t_b)$ as $t \uparrow t_b$, where $x(t) \in \mathbb{S}^+$ with $\pi_{1,2}(x(t)) = 1$ but also $\delta_+(x(t)) < 0 < \delta_-(x(t))$ while $x(t_b) \in \mathbb{A}$, necessarily with $\delta_+(x(t_b)) < 0 < \delta_-(x(t_b))$. The probability $\pi_{1,2}(x(t))$ jumps instantaneously from 1 to some value strictly between 0 and 1 when $\mathbb{A}$ is hit. A numerical example is given in the appendix.

Nevertheless, we can treat time points at which $\pi_{1,2}$ and thus Ψ are discontinuous by starting a new ODE at each hitting time of $\mathbb{A}$ from $\mathbb{S}^+$ or $\mathbb{S}^-$. That makes the solution continuous and differentiable a.e. We justify that claim in the remainder of this section.

Before doing so, we observe that the only difficulty occurs when the solution goes from $\mathbb{S}^+$ or $\mathbb{S}^-$ to $\mathbb{S}^b$. As a consequence of Theorem 5.2 (i), a solution starting in $\mathbb{A}$ can only leave $\mathbb{A}$ via one of the sets $\mathbb{A}^+$ or $\mathbb{A}^-$. To understand what can happen, let $d(x(t)) \equiv q_1(t) - rq_2(t)$ and $d'(x(t)) \equiv \dot{q}_1(t) - r\dot{q}_2(t)$, from (4.2), where we regard $d'(x(t))$ as a right derivative. In $\mathbb{S}^b$ we have $d(x(t)) = 0$ and in $\mathbb{A}$ we have $d'(x(t)) = 0$. On $\mathbb{A}^+$ and $\mathbb{A}^-$, the possibilities can be determined from the following lemma.

LEMMA 7.1. *On $\mathbb{S}^b$, if $\pi_{1,2}(x) = 1$, then $d'(x) = \delta_+(x)$; if $\pi_{1,2}(x) = 0$, then $d'(x) = \delta_-(x)$. Hence, on $\mathbb{A}^+$, $d'(x) \geq 0$, while on $\mathbb{A}^-$, $d'(x) \leq 0$.*

PROOF. Substitute the appropriate values of $\pi_{1,2}(x(t))$ into (4.2) and compute $\delta_\pm(x)$ from (6.1)–(6.4), recalling that $r = j/k$. □

We next separate equality from strict inequality for the weak inequalities in Lemma 7.1. For that purpose, we decompose the sets $\mathbb{A}^+$ and $\mathbb{A}^-$ by letting

$$\begin{aligned} \mathbb{A}^+_+ &\equiv \{x \in \mathbb{A}^+ \mid \delta_+(x) > 0\}, \quad \mathbb{A}^+_0 \equiv \{x \in \mathbb{A}^+ \mid \delta_+(x) = 0\}, \\ \mathbb{A}^-_- &\equiv \{x \in \mathbb{A}^- \mid \delta_-(x) < 0\}, \quad \mathbb{A}^-_0 \equiv \{x \in \mathbb{A}^- \mid \delta_-(x) = 0\}. \end{aligned} \tag{7.4}$$

LEMMA 7.2. *Consider a solution to the ODE over a sufficiently small interval starting at $x(0)$. If $x(0) \in \mathbb{A}^+_+$, then $x(t) \in \mathbb{S}^+$ for all $t > 0$ sufficiently*

small; if $x(0) \in \mathbb{A}_0^+$, *then* $x(t) \in \mathbb{S} - \mathbb{S}^- - \mathbb{A}^-$ *for all* $t > 0$ *sufficiently small; if* $x(0) \in \mathbb{A}_-^-$, *then* $x(t) \in \mathbb{S}^-$ *for all* $t > 0$ *sufficiently small; if* $x(0) \in \mathbb{A}_0^-$, *then* $x(t) \in \mathbb{S} - \mathbb{S}^+ - \mathbb{A}^+$ *for all* $t > 0$ *sufficiently small.*

PROOF. We only treat the first two cases, because the reasoning for the last two is the same. Recall the critical role played by the drifts of the FTSP, as first defined in (6.9). Their critical role in positive recurrence is established in (6.10). The inequality in (6.13) implies that there are only three possible cases. If $x(0) \in \mathbb{A}_+^+$, then $d(x(0)) = 0$ and $d'(x(0)) > 0$ by Lemma 7.1. That implies that $d(x(t)) > 0$ for all $t > 0$, which in turn implies that $x(t) \in \mathbb{S}^+$ for all $t > 0$, the claimed result. If $x(0) \in \mathbb{A}_0^+$, then $d'(x(0)) = 0$ by Lemma 7.1. To see why we cannot have $x(t) \in \mathbb{A}^- \cup \mathbb{S}^-$ for all $t > 0$ sufficiently small, note that then $\pi_{1,2}(x)$ would jump from 1 to 0, which would cause a jump in $d'(x)$ because of Lemma 7.1 and the inequality in (6.13). □

7.5. *Boundedness.* We now show that the possible values of a solution to the ODE are contained in a compact subset of $\mathbb{S}$, provided that the initial values of the queue lengths are constrained. That is accomplished by proving that a solution to the IVP (4.3) is bounded. We use the notation: $a \vee b \equiv \max\{a, b\}$.

THEOREM 7.2 (boundedness). *Every solution to the* IVP (4.3) *is bounded. In particular, the following upper bounds for the fluid queues hold:*

$$q_i(t) \le q_i^{bd} \equiv q_i(0) \vee \lambda_i/\theta_i \quad t \ge 0, \quad i = 1, 2. \tag{7.5}$$

PROOF. Since $0 \le z_{1,2} \le m_2$ and $q_i \ge 0$ in $\mathbb{S}$, we only need to establish the upper bounds for the queue contents in (7.5). To do so, it suffices to consider the bounding function describing the queue-length process of each queue in a modified system with no service processes, so that all the fluid output is due to abandonment, which produces a simple one-dimensional ODE for each queue; for the remaining details, see §D in the appendix. □

7.6. *Proof of Theorem 5.2 (ii).*

PROOF. It follows from Theorem 5.2 (*i*) established above, and Theorems 7.1 and 7.2, that any solution x on $[0, \delta)$ can be extended to an interval $[0, \delta')$, $\delta' > \delta$ (even $\delta' = \infty$), with the solution $\{x(t) : t \in [0, \delta')\}$ again being unique, provided that that the solution x makes no transitions from $\mathbb{S} - \mathbb{S}^b$ to $\mathbb{S}^b$, causing a discontinuity in $\pi_{1,2}(x)$ and thus Ψ in (4.2). (See Theorem 3.3 in [10] and its proof for supporting details.)

Moreover, the solution in $\mathbb{S}^+$ or $\mathbb{S}^-$ necessarily has a left limit at the time it hits $\mathbb{S}^b$. The left limit exists because, by Theorem 7.2, the solution is bounded, and because the derivative in either $\mathbb{S}^+$ or $\mathbb{S}^-$ is bounded as well, by virtue of (4.2). At each such hitting time, a new ODE is constructed starting in $\mathbb{S}^b$. That ensures the overall continuity of x.

In general, there can be accumulation points of such hitting times of the set $\mathbb{S}^b$ from $\mathbb{S}-\mathbb{S}^b$; i.e., there could exist sequences $\{t_n^i : n \geq 1\}$, $i = 1, 2$ with $x(t_n^1) \in \mathbb{S}-\mathbb{S}^b$ and $x(t_n^2) \in \mathbb{S}^b$ for all n with $t_n^i \uparrow t < \infty$ and $x(t_n^i) \to x(t) \in \mathbb{S}^b$ as $n \to \infty$ for $i = 1, 2$. However, we claim that, at any such accumulation time t, $x(t)$ must be in either $\mathbb{A}^+$ or $\mathbb{A}^-$. We now show by contradiction that there cannot be an accumulation point in $\mathbb{A}$.

Starting from $x(t)$ in $\mathbb{A}$, it necessarily will take a given positive time for the solution to move through the set $\mathbb{A}$ until it reaches $\mathbb{A}^+ \cup \mathbb{A}^-$, after which it must transition to $\mathbb{S} - \mathbb{S}^b$, after which it may again later hit $\mathbb{A}$. We can apply the continuity of the solution within the set $\mathbb{A}$ as a function of the initial value within $\mathbb{A}$, due to the Lipschitz continuity, see §2.4 of [21]. Since $t_n^2 \to t$ and $x(t_n^2) \to x(t) \in \mathbb{A}$ as $n \to \infty$, the solution of the ODE over an initial interval starting at $x(t_n^2)$ in $\mathbb{A}$ at time t_n^2 must converge to the solution of the ODE over an initial interval starting at $x(t)$ in $\mathbb{A}$ at time t. Thus, we can conclude that the time to reach $\mathbb{A}^+ \cup \mathbb{A}^-$ from $x(t_n^2)$ at time t_n^2 must be bounded below by a strictly positive number for all n sufficiently large. Thus, such an accumulation point $x(t)$ cannot be in $\mathbb{A}$. All such accumulation points must be in either $\mathbb{A}^+$ or $\mathbb{A}^-$.

Finally, by Theorem 7.1, the function Ψ is locally Lipschitz continuous at each point in $\mathbb{A}^+ \cup \mathbb{A}^-$. Hence, the ODE is well defined there. First, a new ODE can be constructed starting at the accumulation point in $\mathbb{A}^+ \cup \mathbb{A}^-$. However, since the ODE is well defined at each accumulation time, the solution x must actually be differentiable at each of these accumulation times of hitting times. As a consequence, x is continuous and differentiable almost everywhere throughout $[0, \infty)$. □

We remark that we have not yet ruled out the possibility of infinitely many discontinuity points of $\pi_{1,2}(x(t))$ and thus $\Psi(x(t))$; i.e., we have not shown that the solution to the ODE necessarily only makes finitely many transitions from $\mathbb{S} - \mathbb{S}^b$ to $\mathbb{A}$ over the entire positive halfline $[0, \infty)$. In later sections we obtain conditions guaranteeing that does not happen. In the proof of Theorem 5.2 (*ii*), just completed, we have established the following result.

COROLLARY 7.1 (extension to a global solution). *Let x be the unique differentiable solution to the* IVP (4.3) *on an interval $[0, \delta)$, established in §7.1. If it is known that the solution can never transition from $\mathbb{S}^+$ or $\mathbb{S}^-$ to*

$\mathbb{S}^b$, *which is implied by remaining in the set* $\mathbb{A}$ *for all time, then there exists a unique differentiable solution to the* IVP (4.3) *on* $[0, \infty)$.

We give conditions for the solution to the IVP (4.3) to lie entirely in $\mathbb{A}$ in §10.

8. The existence of a unique stationary point. We now indicate what is meant by a stationary point for an autonomous ODE. Then we show that there exists a unique stationary point for the autonomous ODE in (4.1) and (4.2). We then give conditions under which the fluid solution $x \equiv \{x(t) : t \geq 0\}$ converges to stationarity as $t \to \infty$. In §9 we show that it does so exponentially fast.

DEFINITION 8.1 (stationary point for an autonomous ODE). We say that x^* is a stationary point for an autonomous ODE $\dot{x}(t) = \Psi(x(t))$ if $\Psi(x^*) = 0$, i.e., if $x(t) = x^*$ for all $t \geq 0$ is a solution to the ODE. If a solution x of the ODE is a constant function, then we say that the solution is stationary, or in steady state.

Definition 8.1 actually contains two different definitions of a stationary point of an ODE, but they are equivalent for an autonomous ODE; see pp. 20-21 of [21]. Note that an autonomous ODE can have a stationary point without all solutions to the ODE being constant functions. For example, the ODE $\dot{x}(t) = 2\sqrt{x}$ has a stationary point $x^* = 0$, but it also has the non-constant solution $x(t) = t^2$, $t \geq 0$. For this example, the Lipschitz continuity required for the Picard-Lindelof Theorem is violated, so that that non-uniqueness is possible, and occurs.

8.1. *Characterization of the stationary point.* By definition, a stationary point $x^* \in \mathbb{S}$ satisfies $\Psi(x^*) = 0$. From (4.2), we see that this gives a system of three equations with three unknowns, namely, q_1^*, q_2^* and $z_{1,2}^*$. The apparent fourth variable $\pi_{1,2}^* \equiv \pi_{1,2}(x^*)$ is a function of the other three variables and its value is determined by x^*. In principle, the three equations in $\Psi(x) = 0$ can be solved directly to find all the roots of Ψ. However, $\pi_{1,2}^*$ is a complicated function of x^* having the complicated closed-form expression in (6.14) and (6.17).

Theorem 8.1 below states that, if there exists a stationary point for the fluid ODE (4.2), then this point is unique, and must have the specified form. The uniqueness of x^* is proved by treating $\pi_{1,2}^*$ as a fourth variable, and adding a fourth equation to the three equations $\Psi(x) = 0$. However, it does not prove that a stationary point exists. In general, the solution $\pi_{1,2}^*$

we get from the system of four equations may not equal to $\pi_{1,2}(x^*)$, for the function $\pi_{1,2}$ defined in (3.7). The existence of a stationary point is proved in the next section.

The proof of existence is immediate from the proof of uniqueness when $\pi_{1,2}(x^*)$ is known in advance to be 0 or 1, with the value determined. That occurs everywhere except the region $\mathbb{A}$; it occurs in the two regions $\mathbb{S}^+$ and $\mathbb{S}^-$, but it also occurs in $\mathbb{S}^b - \mathbb{A}$. Since the QBD is not positive recurrent in $\mathbb{S}^b - \mathbb{A}$, it follows that $\pi_{1,2}(x^*)$ can only assume one of the values, 0 or 1, achieving the same value as in the neighboring region $\mathbb{S}^+$ or $\mathbb{S}^-$. (We omit detailed demonstration.) But we will have to work harder in $\mathbb{A}$.

We now focus on uniqueness. Although $\pi^*_{1,2}$ is treated as a variable, we still impose conditions on it so that it can be a legitimate solution to (3.7). In particular, if $q_1^* - rq_2^* > \kappa$ then we let $\pi^*_{1,2} = 1$; if $q_1^* - rq_2^* < \kappa$, then we let $\pi^*_{1,2} = 0$. Equation (8.3) below shows that $0 \le \pi^*_{1,2} \le 1$ whenever $q_1^* - rq_2^* = \kappa$, i.e., whenever $x^* \in \mathbb{S}^b$.

For $a, b \in \mathbb{R}$, recall that $a \vee b \equiv \max\{a, b\}$ and let $a \wedge b \equiv \min\{a, b\}$. Let

$$z \equiv \frac{\theta_2(\lambda_1 - m_1\mu_{1,1}) - r\theta_1(\lambda_2 - m_2\mu_{2,2}) - \theta_1\theta_2\kappa}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}}. \tag{8.1}$$

THEOREM 8.1 (uniqueness of the stationary point). *There can be at most one stationary point $x^* \equiv (q_1^*, q_2^*, z^*_{1,2})$ for the ODE (4.1), which must take the form*

$$z^*_{1,2} = 0 \vee z \wedge m_2, \quad q_1^* = \frac{\lambda_1 - m_1\mu_{1,1} - \mu_{1,2}z^*_{1,2}}{\theta_1}, \quad q_2^* = \frac{\lambda_2 - \mu_{2,2}(m_2 - z^*_{1,2})}{\theta_2}, \tag{8.2}$$

for z in (8.1). Moreover,

$$\pi^*_{1,2} = \frac{\mu_{1,2}z^*_{1,2}}{\mu_{1,2}z^*_{1,2} + (m_2 - z^*_{1,2})\mu_{2,2}}. \tag{8.3}$$

PROOF. We start with (8.3). This expression is easily derived from the third equation in (4.2), by equating $\dot{z}_{1,2}(t)$ to zero. Observe that if $z^*_{1,2} = m_2$ then $\pi^*_{1,2}$ in (8.3) is equal to 1, and if $z^*_{1,2} = 0$ then $\pi^*_{1,2} = 0$. Now, by plugging the value of $\pi^*_{1,2}$ in the ODE's for $\dot{q}_1(t)$ and $\dot{q}_2(t)$ in (4.2) we get the expressions of q_1^* and q_2^* in (8.2). We now have the two equations for the stationary queues, but there are three unknowns: $z^*_{1,2}$, q_1^* and q_2^*. We introduce a third equation to resolve this difficulty.

Consider the following three equations with the three unknowns: $z, q_1(z)$ and $q_2(z)$. (here q_1 and q_2 are treated as functions of the variable z, not to be

confused with the fluid solution which is a function of the time argument t.)

$$q_1(z) = \frac{\lambda_1 - \mu_{1,1}m_1 - \mu_{1,2}z}{\theta_1}, \quad q_2(z) = \frac{\lambda_2 - \mu_{2,2}(m_2 - z)}{\theta_2}, \tag{8.4}$$
$$\kappa = q_1(z) - rq_2(z).$$

Notice that $q_1(z)$ is decreasing with z, whereas $q_2(z)$ is increasing with z. Thus, there exists a unique solution to these three equations, which has z as in (8.1). We can recover x^* from the solution to (8.4), and by doing so show that x^* is unique and is always in one of the three regions $\mathbb{S}^-$, $\mathbb{S}^+$ or $\mathbb{S}^b$ (so that $x^* \in \mathbb{S}$).

Let $(q_1(z), q_2(z), z)$ be the unique solution to (8.4). First assume that $z > m_2$, which implies that $q_2(z) > 0$, and, by the third equation, $q_1(z) > \kappa$. By replacing z with m_2, $q_1(\cdot)$ is increased and $q_2(\cdot)$ is decreased (but is still positive), so that $q_1(m_2) - rq_2(m_2) > \kappa$ (and, trivially, $q_1(m_2) > \kappa$, $q_2(m_2) > 0$). This implies that $x^* \equiv (q_1(m_2), q_2(m_2), m_2) \in \mathbb{S}^+$ and, if it is indeed a solution to $\Psi(x) = 0$, then x^* is the unique stationary point for the ODE.

Now assume that the unique solution to (8.4) has $z < 0$. By replacing z with 0 we have $q_1(0) < q_1(z)$ and $q_2(0) > q_2(z)$, which imply that $q_1(0) - rq_2(0) < \kappa$. Now, since $q_1(0) = q_1^a$ we have that $q_1(0) \geq \kappa$ by Assumption A. This implies that $q_1(z) > \kappa$, which further implies that $rq_2(z) = q_1(z) - \kappa > 0$, so that $rq_2(0) > rq_2(z) > 0$. Taking $x^* \equiv (q_1(0), q_2(0), 0)$, we see that $x^* \in \mathbb{S}^-$, and if x^* is indeed a solution to $\Psi(x) = 0$, then x^* is the unique stationary point for the ODE.

Finally, assume that the solution $x(z) \equiv (q_1(z), q_2(z), z)$ to (8.4) has $0 \leq z \leq m_2$. To conclude that $x(z)$ is in $\mathbb{S}^b$ we need to show that $q(z), q_2(z) \geq 0$, so that $q_1^* = q_1(z)$ and $q_2^* = q_2(z)$ are legitimate queue-length solutions. We now show that is the case under Assumption A.

Let $S_2^a \equiv m_2 - \lambda_2/\mu_{2,2}$. Note that, if $S_2^a \geq 0$, then $S_2^a = s_2^a$, for s_2^a in (5.1). We start by rewriting $q_1(z)$ and $q_2(z)$ in (8.4) as

$$q_1(z) = q_1^a - \frac{\mu_{1,2}}{\theta_1}z, \quad q_2(z) = \frac{\mu_{2,2}}{\theta_2}(z - S_2^a). \tag{8.5}$$

Now, it follows from Assumption A that

$$\kappa \leq q_1^a - \frac{\mu_{1,2}}{\theta_1}s_2^a \leq q_1^a - \frac{\mu_{1,2}}{\theta_1}S_2^a, \tag{8.6}$$

where the second inequality follows trivially, since $S_2^a \leq s_2^a$. From the third equation of (8.4), $\kappa = q_1(z) - rq_2(z)$. Combining this with (8.5), we see that

$$\kappa = q_1(z) - rq_2(z) = q_1^a - \frac{\mu_{1,2}}{\theta_1}z - r\frac{\mu_{2,2}}{\theta_2}(z - S_2^a). \tag{8.7}$$

Combining (8.6) and (8.7), we get

$$q_1^a - \frac{\mu_{1,2}}{\theta_1} z - r\frac{\mu_{2,2}}{\theta_2}(z - S_2^a) \le q_1^a - \frac{\mu_{1,2}}{\theta_1} S_2^a,$$

which is equivalent to

$$0 \le \left(\frac{\mu_{1,2}}{\theta_1} + r\frac{\mu_{2,2}}{\theta_2}\right)(z - S_2^a).$$

This, together with the fact that the solution has $z \ge 0$, implies that $z \ge \max\{0, S_2^a\} = s_2^a$. It follows from (8.5) that $q_2(z) \ge 0$ and, by using the third equation in (8.4) again, $q_1(z) = rq_2(z) + \kappa \ge \kappa \ge 0$. □

An immediate consequence of the proof of Theorem 8.1 is that, in order to find the candidate stationary point x^*, one has to solve the three equations in (8.4). The next corollary summarizes the values x^* may take, depending on its region; the proof appears in the appendix.

COROLLARY 8.1. *Let $x^* = (q_1^*, q_2^*, z_{1,2}^*)$ be the point defined in Theorem 8.1.*

(i) If $x^* \in \mathbb{S}^b$, then, for z defined in (8.1),

$$z_{1,2}^* = z = \frac{\theta_1\theta_2(q_1^a - \kappa) - r\theta_1(\lambda_2 - \mu_{2,2}m_2)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}}$$
$$= \begin{cases} \frac{\theta_1\theta_2(q_1^a - rq_2^a - \kappa)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}}, & \text{if } \ q_2^a \ge 0,\ s_2^a = 0. \\ \frac{\theta_1\theta_2(q_1^a + r\mu_{2,2}s_2^a/\theta_2 - \kappa)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}}, & \text{if } \ q_2^a = 0,\ s_2^a > 0. \end{cases}$$
$$q_1^* = \frac{\lambda_1 - m_1\mu_{1,1} - z_{1,2}^*\mu_{1,2}}{\theta_1}, \qquad q_2^* = \frac{\lambda_2 - (m_2 - z_{1,2}^*)\mu_{2,2}}{\theta_2}.$$

(ii) If $x^* = \mathbb{S}^+$, then

$$z_{1,2}^* = m_2, \qquad q_1^* = \frac{\lambda_1 - m_1\mu_{1,1} - m_2\mu_{1,2}}{\theta_1}, \qquad q_2^* = \frac{\lambda_2}{\theta_2}.$$

(iii) If $x^* \in \mathbb{S}^-$, then

$$z_{1,2}^* = 0, \qquad q_1^* = \frac{\lambda_1 - m_1\mu_{1,1}}{\theta_1}, \qquad q_2^* = \frac{\lambda_2 - m_2\mu_{2,2}}{\theta_2}.$$

If $x^* \in \mathbb{S}^+$, as in (ii), then the system does not have enough service capacity to keep the weighted difference between the two queues at κ, even when all agents are working with class 1. In this case, the only output from queue 2 is due to abandonment, since no class-2 fluid is being served (in steady state). Queue 2 is then equivalent to the fluid approximation for an $M/M/\infty$ system with service rate θ_2 and arrival rate λ_2. On the other hand, queue 1 is equivalent to an overloaded inverted-V model: a system in which one class, having one queue, is served by two different service pools.

The next corollary gives necessary and sufficient conditions for x^* to be in each region. It shows that the region of x^* can be determined from rate considerations alone. We give the proof in the appendix.

COROLLARY 8.2. *Let x^* be as in (8.2). Then*

(i) $x^* \in \mathbb{S}^b$ if and only if

$$\frac{\mu_{1,2}s_2^a}{\theta_1} \vee rq_2^a \;\le\; q_1^a - \kappa \;\le\; \frac{r\lambda_2}{\theta_2} + \frac{\mu_{1,2}m_2}{\theta_1}; \tag{8.8}$$

$x^* \in \mathbb{A}$ if and only if both inequalities are strict.

(ii) $x^* \in \mathbb{S}^+$ if and only if $q_1^a - \kappa > \frac{r\lambda_2}{\theta_2} + \frac{\mu_{1,2}m_2}{\theta_1}$.

(iii) $x^* \in \mathbb{S}^-$ if and only if $rq_2^a > q_1^a - \kappa$.

REMARK 8.1 (most likely region in applications). It follows from Corollary 8.2 that, in applications, $\mathbb{A}$ is the most likely region for the stationary point when the system is overloaded, provided that the arrival rates are about $10 - 50\%$ larger than planned during an overload incident. Typically, a much higher overload is needed in order for the stationary point to be in $\mathbb{S}^+$. As an example, consider the canonical example from [16]: There are 100 servers in each pool, serving their own class at rates $\mu_{1,1} = \mu_{2,2} = 1$. Type-2 servers serve class-1 customers at rate $\mu_{1,2} = 0.8$. Also, $\theta_1 = \theta_2 = 0.3$, $r = 0.8$ and $\kappa = 0$. Suppose that class 2 is not overloaded with $\lambda_2 = 90$. Then, for the stationary point to be in $\mathbb{S}^+$, we need to have $\lambda_1 > \mu_{1,1}m_1 + \mu_{1,2}m_2 + \theta_1 r\lambda_2/\theta_2 = 252$, i.e., the class-1 arrival rate is 252% larger than the total service rate of pool 1. If $\lambda_2 > 90$, especially if pool 2 is also overloaded, then λ_1 needs to be even larger than that.

8.2. *Existence of a stationary point.* We have just established uniqueness of the stationary point in $\mathbb{S}$, and characterized it. In the process, we have also established existence in $\mathbb{S} - \mathbb{A}$, because the form of $\pi_{1,2}(x)$ is then known in advance. Now we will establish existence of the stationary point in $\mathbb{A}$. First, we calculate the drift rates at $x^* \in \mathbb{A}$.

LEMMA 8.1 (the drift rates at x^*). *For x^* in Corollary 8.1 (i), where $0 < z^*_{1,2} < m_2$,*

$$\delta_+(x^*) = -(j+k)\mu_{2,2}(m_2 - z^*_{1,2}) < 0, \quad \delta_-(x^*) = +(j+k)\mu_{1,2}z^*_{1,2} > 0. \tag{8.9}$$

PROOF. Substitute x^* in Corollary 8.1 (i) into (6.9), using (6.1)-(6.4). □

We now are ready to prove existence.

THEOREM 8.2 (existence). *If the model parameters produce $x^* \in \mathbb{A}$, i.e., as in Corollary 8.1 (i), where $0 < z^*_{1,2} < m_2$, then x^* is the unique stationary point.*

PROOF. We will prove that there must exist at least one stationary point. Given that result, by Theorem 8.1 and Corollary 8.1, there must be exactly one fixed point and that must be the x^* given there. To establish existence, we will apply the Brouwer fixed point theorem. It concludes that a continuous function mapping a convex compact subset of Euclidean space $\mathbb{R}^k$ into itself has at least one fixed point. We will let our domain be the set

$$C(\eta) \equiv \{x \in \mathbb{A} \cap \mathbb{B} : \delta_+(x) \le -\eta \quad \text{and} \quad \delta_-(x) \ge \eta\} \tag{8.10}$$

for an appropriate small positive η, where $\mathbb{B} \equiv [0, q_1^{bd}] \times [0, q_2^{bd}] \times [0, m^2]$ with q_i^{bd} being the bound on q_i from Theorem 7.2. Choose η sufficiently small that $x^* \in C(\eta)$; that is easily ensured by Lemma 8.1. Since the rates in (6.1)–(6.4) and the drift in (6.9) are linear functions of x, we see that $C(\eta)$ is a convex subset of $\mathbb{A}$ for each $\eta > 0$. Since the inequalities in (8.10) are weak, the set is closed. The intersection with $\mathbb{B}$ guarantees that the set $C(\eta)$ is also bounded. Hence, $C(\eta)$ is compact.

By Theorem 5.2, for any $x(0) \in C(\eta)$, there exists a unique solution to the ODE over $[0, \delta]$ for some positive δ. Hence, for any t with $0 < t < \delta$, the map from $x(0)$ to $x(t)$ is continuous; see §2.4 of [21]. Let $x^*_L \equiv (q^*_{1,L}, q^*_{2,L}, z^*_{1,2,L})$ and $x^*_U \equiv (q^*_{1,U}, q^*_{2,U}, z^*_{1,2,U})$, where $q^*_{1,L} \equiv q^*_1 - \epsilon$, $q^*_{2,L} \equiv q^*_2 - \epsilon$, $z^*_{1,2,L} \equiv z^*_{1,2} - \epsilon$, $q^*_{1,U} \equiv q^*_1 + \epsilon$, $q^*_{2,U} \equiv q^*_2 + \epsilon$ and $z^*_{1,2,U} \equiv z^*_{1,2} + \epsilon$. Let $\phi_t : C(\eta) \to C(\eta)$ be the continuous function defined by $\phi_t(x(0)) \equiv (q_{1,t}, q_{2,t}, z_{1,2,t})$, where

$$q_{i,t} \equiv q_i(t) \vee q^*_{i,L} \wedge q^*_{i,U} \quad \text{and} \quad z_{1,2,t} \equiv z_{1,2,t} \vee z^*_{1,2,L} \wedge z^*_{1,2,U}, \tag{8.11}$$

for $i = 1, 2$. We can choose $\eta > 0$ and $\epsilon > 0$ sufficiently small so that, first, $x^* \in C(\eta)$ and, second, that $x_{i,t} \in C(\eta)$ for each $x(0) \in C(\eta)$. Hence, the pair $(C(\eta), \phi_t)$ satisfies the conditions for the Brouwer fixed point theorem. Hence, there exists $x(0) \in C(\eta)$ such that $x(t) = x(0)$.

Now let $\{t_n : n \geq 1\}$ be a sequence of time points decreasing toward 0. We can apply the argument above to deduce that, for each n, there exists $x_n(0)$ in $C(\eta)$ such that $x_n(t_n) = x_n(0)$. However, from the ODE, we have the relation $|x(t) - x(0) - \Psi(x(0))t| \leq Mt^2$ for all sufficiently small t. Since $\{x_n(0) : n \geq 1\}$ is bounded, there exists a convergent subsequence. Let $x(0)$ be the limit of that convergent subsequence. For that limit, we necessarily have $\Psi(x(0)) = 0$. Hence, that $x(0)$ must be a stationary point for the ODE. By Theorem 8.1, we must have $x(0) = x^*$. □

8.3. *Global asymptotic stability.* Having a unique stationary point of an autonomous ODE does not imply that the solution necessarily converges to that point as $t \to \infty$. It does not even guarantee that a solution to the IVP (4.3) is asymptotically stable in the sense that, if $\|x(0) - x^*\| < \epsilon$, then $x(t) \to x^*$ as $t \to \infty$, no matter how small ϵ is. In fact, there is not even a guarantee that $x(t)$ will remain in the ϵ-neighborhood of x^* for all $t \geq 0$. We will establish all of these properties in Theorem 8.3 below by showing that x^* in §8.1 is globally asymptotically stable, as defined below:

Definition 8.2 (global asymptotic stability). A point x^* is said to be globally asymptotically stable for an autonomous ODE if it is a stationary point and if, for any initial state $x(0)$ and any $\epsilon > 0$, there exists a time $T \equiv T(x(0), \epsilon) \geq 0$ such that $\|x(t) - x^*\| < \epsilon$ for all $t \geq T$.

Global asymptotic stability goes beyond simple convergence by also requiring that the limit be a stationary point.

Theorem 8.3 (global asymptotic stability of x^*). *The unique stationary point x^* of the autonomous ODE in (4.1) is globally asymptotically stable.*

Our proof of Theorem 8.3 relies on Lyapunov stability theory for deterministic dynamical systems; see Chapter 4 of Khalil [10]. Let E be an open and connected subset of $\mathbb{R}^n$ containing the origin. We use standard vector notation to denote the inner product of vectors $a, b \in \mathbb{R}^n$, i.e., $a \cdot b = \sum_{i=1}^n a_i b_i$.

Definition 8.3 (Lie derivative). For a continuously differentiable function $V : E \to \mathbb{R}$, and a function $\Psi : E \to \mathbb{R}^n$, the Lie derivative of V along Ψ is defined by

$$\dot{V}(x) \equiv \frac{\partial V}{\partial x}\Psi(x) = \nabla V \cdot \Psi(x),$$

where $\nabla V \equiv (\frac{\partial V}{\partial x_1}, \ldots, \frac{\partial V}{\partial x_n})$ is the gradient of V.

DEFINITION 8.4 (Lyapunov-function candidate). A continuously differentiable function $V : E \to \mathbb{R}$ is a Lyapunov-function candidate if:

(*i*) $V(0) = 0$

(*ii*) $V(x) > 0$ for all x in $E - \{0\}$

In proving Theorem 8.3 we use the following theorem, which is Theorem 4.2 pg. 124 in [10]:

THEOREM 8.4 (global asymptotic stability for nonlinear ODE). *Let $x = 0$ be a stationary point of $\dot{x} = \Psi(x)$, $\Psi : E \to \mathbb{R}^n$, and let $V : \mathbb{R}^n_+ \to \mathbb{R}$ be a Lyapunov-function candidate. If*

(*i*) $V(x) \to \infty$ *as* $||x|| \to \infty$ *and*

(*ii*) $\dot{V}(x) < 0$ *for all* $x \neq 0$,

then $x = 0$ is globally asymptotically stable as in Definition 8.2.

Notice that, under the conditions of Theorem 8.4, the Lyapunov-function candidate V provides a form of monotonicity: We necessarily have $V(0) = 0$ and $V(x(t))$ strictly decreasing in t for $x(t) \neq 0$. To elaborate, we introduce the notion of a V-ball. We say that $\beta_V(\alpha)$ is the α V-ball with center at x^* and radius α if

$$\beta_V(\alpha) \equiv \{x \in \mathbb{R}^n : \|V(x) - V(x^*)\| \leq \alpha\}. \tag{8.12}$$

If $x(t_0) \in \beta_V(\alpha)$ for some $\alpha \geq 0$ and $t_0 \geq 0$, then $x(t) \in \beta_V(\alpha)$ for all $t \geq t_0$.

PROOF OF THEOREM 8.3. Theorem 8.4 applies directly only within a single region, starting at a point in $\mathbb{S}^+$, $\mathbb{S}^-$, $\mathbb{A}$, $\mathbb{A}^-$ or $\mathbb{A}^+$. However, we will show that the same Lyapunov function V can be used in all regions, leading to global decrease of V as x^* is being approached.

Let x be the unique solution to (4.3). Let $x^* \equiv (q_1^*, q_2^*, z_{1,2}^*)$ be the stationary point for the system (4.1). Without loss of generality, we perform a change of variables and define a new system whose unique stationary point is $x = 0$. To this end, let $y = x - x^*$ so that $\dot{y} = \dot{x} = \Psi(x)$. Hence, $\Psi(x) = \Psi(y + x^*) \equiv g(y)$ and we have that $g(0) = \Psi(0 + x^*) = \Psi(x^*) = 0$. That is, if x^* is a stationary point for the original system $\dot{x} = \Psi(x)$, then the stationary point for the new system, $\dot{y} = g(y)$, is $y^* = 0$. We distinguish between two cases: (*i*) $\mu_{1,2} > \mu_{2,2}$ and (*ii*) $\mu_{1,2} \leq \mu_{2,2}$.

(*i*) First, if $\mu_{1,2} > \mu_{2,2}$, then choose $V_1(x) \equiv x_1 + x_2$ and apply its Lie derivative along $g(y) = \Psi(y+x^*)$ where $y+x^* = (q_1(t)+q_1^*, q_2(t)+q_2^*, z_{1,2}(t)+$

$z^*_{1,2}$) and x^* is given in (8.2). By the definition of the Lie derivative, $\dot{V}_1(y)$ is equal to the inner product

$$\dot{V}_1(y) = (1,1,0) \cdot (\dot{q}_1(t), \dot{q}_2(t), \dot{z}_{1,2}(t))' = \dot{q}_1(t) + \dot{q}_2(t),$$

for $\dot{q}_1$, $\dot{q}_2$ and $\dot{z}_{1,2}$ in (4.2), after the change of variables. Let $\tilde{z}_{1,2}(t) \equiv z_{1,2}(t) + z^*$. Then, for $x^* = (q^*_1, q^*_2, z^*_{1,2})$ as in (8.2)

$$\begin{aligned}
\dot{V}_1(y) &= \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(y(t))[\tilde{z}_{1,2}(t)\mu_{1,2} + (m_2 - \tilde{z}_{1,2}(t))\mu_{2,2}] \\
&\quad - \theta_1(q_1(t) + q^*_1) - (1 - \pi_{1,2}(y(t)))[(m_2 - \tilde{z}_{1,2}(t))\mu_{2,2} + \tilde{z}_{1,2}(t)\mu_{1,2}] \\
&\quad + \lambda_2 - \theta_2(q_2(t) + q^*) \\
&= \lambda_1 + \lambda_2 - m_1\mu_{1,1} - m_2\mu_{2,2} + z_{1,2}(t)\mu_{2,2} + z^*\mu_{2,2} - z_{1,2}(t)\mu_{1,2} \\
&\quad - z^*_{1,2}\mu_{1,2} - \theta_1 q_1(t) - \theta_1 q^*_1 - \theta_2 q_2(t) - \theta_2 q^*_2 \\
&= -\theta_1 q_1(t) - \theta_2 q_2(t) - z_{1,2}(t)(\mu_{1,2} - \mu_{2,2}).
\end{aligned}$$

Thus, $\dot{V}_1(y) < 0$ for all $y \in \mathbb{R}^3$ unless $y = 0$.

(*ii*) When $\mu_{1,2} \leq \mu_{2,2}$, there exists a $B \geq 1$ such that $\mu_{2,2} = B\mu_{1,2}$. We next show that for any $C > B$ the candidate-function $V_2(x) \equiv Cx_1 + x_2 + (C-1)x_3$ is a Lyapunov function. The Lie derivative of $V_2(x)$ for the modified system $g(y)$ is

$$\dot{V}_2(y) = (C, 1, C-1) \cdot (\dot{q}_1(t), \dot{q}_2(t), \dot{z}_{1,2}(t)) = C\dot{q}_1(t) + \dot{q}_2(t) + (C-1)\dot{z}_{1,2}(t).$$

Hence,

$$\begin{aligned}
\dot{V}_2(y) &= C\left[\lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(y(t))(\tilde{z}_{1,2}(t)\mu_{1,2} + (m_2 - \tilde{z}_{1,2}(t))\mu_{2,2})\right] \\
&\quad - \theta_1(q_1(t) + q^*_1) + \lambda_2 - \theta_2(q_2(t) + q^*_2) \\
&\quad - (1 - \pi_{1,2}(y(t)))(\tilde{z}_{1,2}(t)\mu_{1,2} + (m_2 - \tilde{z}_{1,2}(t)\mu_{2,2})) \\
&\quad + (C-1)\left[\pi_{1,2}(y(t))(m_2 - \tilde{z}_{1,2}(t))\mu_{2,2} - (1 - \pi_{1,2}(y(t)))\tilde{z}_{1,2}(t)\mu_{1,2}\right] \\
&= -C\theta_1 q_1(t) - \theta_2 q_2(t) - z_{1,2}(t)(C\mu_{1,2} - \mu_{2,2}),
\end{aligned}$$

so that $\dot{V}_2(y) < 0$ for all $y \neq 0$.

By Theorem 8.4, $y^* = 0$ is globally asymptotically stable for the modified system $g(y)$. Hence, x^* is globally asymptotically stable for the original system $\Psi(x)$. That is, for every initial value $x(0)$ we have that $x(t) \to x^*$. □

REMARK 8.2 (eliminating $\pi_{1,2}$ from the argument). As often occurs with Lyapunov functions, our choice of Lyapunov functions in the two cases (i) and (ii) of the proof of Theorem 8.3 above simplifies the argument. We have chosen the two Lyapunov functions so that we can eliminate $\pi_{1,2}$ from the

analysis. This suggests that there is some some flexibility in the control for achieving the stationary point x^*. That is consistent with Theorem 8.1, which characterizes the form of the unique stationary point. In both cases, we do not need to analyze the FTSP.

8.4. *Staying in* $\mathbb{S}$. We also use the Lyapunov argument to prove Theorem 5.1, i.e., to show that the solution to the ODE can never leave $\mathbb{S}$.

PROOF OF THEOREM 5.1. To simplify the writing of the proof, we use the following notation (with an abuse of conventions): For a function f of a real variable that is continuous at t, let $f(t+) \geq c$ mean that $f(s) \geq c$ for all $s \in (t, t+\epsilon]$, for all $\epsilon > 0$ sufficiently small (and similarly for other inequalities, e.g., $> c$). If f is not continuous at t, then $f(t+)$ has the usual meaning, i.e., the right limit at the point t. We apply this definition below to functions that are derivatives, e.g., $f = \dot{z}_{1,2}$.

By Assumption A, $x(0) \in \mathbb{S}$ and we must show that, for all $t > 0$, $z_{1,2}(t) \in [0, m_2]$, $q_1(t) \in [\kappa, \infty)$ and $q_2(t) \in [0, \infty)$. Consider $t \geq 0$. It is easy to see that if $z_{1,2}(t) = 0$, then $\dot{z}_{1,2}(t) \geq 0$. If $\dot{z}_{1,2}(t) > 0$, then $z_{1,2}(t)$ is increasing and, in particular, $z_{1,2}(t+) > 0$. However, if $\dot{z}_{1,2}(t) = 0$, we must rule out the case $z_{1,2}(t+) < 0$. (Observe that $\dot{z}_{1,2}(t) = z_{1,2}(t) = 0$ implies $\pi_{1,2}(x(t)) = 0$.) To do that, we take the contradictory assumption, namely we assume that $z_{1,2}(t+) < 0$. Hence, for any $s > t$ in a small-enough neighborhood of t, we can find $\varepsilon > 0$, such that $z_{1,2}(s) = -\varepsilon$. Plugging that value of $z_{1,2}(s)$ in the ODE (4.2), we see that, regardless of the value of $\pi_{1,2}(x(s))$,

$$\dot{z}_{1,2}(s) = \pi_{1,2}(x(s))(m_2 + \varepsilon)\mu_{2,2} + (1 - \pi_{1,2}(x(s)))\varepsilon\mu_{1,2} > 0, \tag{8.13}$$

implying that $z_{1,2}(s)$ is strictly increasing at each $s > t$ in a small-enough neighborhood of t. This further implies that no value $-\varepsilon$ can be reached by $z_{1,2}$ because, by continuity, $z_{1,2}$ must first attain all the values in $(-\varepsilon, 0)$. But $z_{1,2}(s)$ is almost-everywhere differentiable by Theorem 5.2, and for all regular points s for which $z_{1,2}(s) < 0$, (8.13) holds. (Where s is a regular point if $z_{1,2}(s)$ is differentiable at s.) In other words, $z_{1,2}$ cannot decrease towards $-\varepsilon$, and this is true for all $-\varepsilon < 0$.

We now treat the case $z_{1,2}(t) = m_2$. It is easy to see that in that case $\dot{z}_{1,2}(t) \leq 0$. If $\dot{z}_{1,2}(t) < 0$, then $z_{1,2}(t)$ is strictly decreasing, so that $z_{1,2}(t+) < m_2$. Once again, we need to show that when $\dot{z}_{1,2}(t) = 0$ we cannot have that $z_{1,2}(t+) > m_2$. (Observe that $z_{1,2}(s) = m_2$ and $\dot{z}_{1,2}(s) = 0$ implies $\pi_{1,2}(x(s)) = 1$.) We prove the desired result similarly as above, by taking the contradictory assumption that $z_{1,2}(t+) > m_2$, so that for any $s > t$ in a small-enough neighborhood of t, we can find $\epsilon > 0$ such that $z_{1,2}(s) =$

$m_2 + \varepsilon$. Plugging that value of $z_{1,2}(s)$ in the ODE of $z_{1,2}$ in (4.2) we see that, regardless of the value of $\pi_{1,2}(x(s))$,

$$\dot{z}_{1,2}(x(s)) = \pi_{1,2}(x(s))(-\varepsilon)\mu_{2,2} - (1 - \pi_{1,2}(x(s)))(m_2 + \varepsilon)\mu_{1,2} < 0,$$

which implies that $z_{1,2}(s)$ is strictly decreasing at each $s > t$ in a small-enough neighborhood of t. As before, that implies that $z_{1,2}(t+) \leq m_2$.

Turning to the queues, note that to leave $\mathbb{S}$ just after time t, we must have $q_1(t) = \kappa$ or $q_2(t) = 0$ (or both). If $q_1(t) = \kappa$ and $q_2(t) > 0$, then $x(t) \in \mathbb{S}^-$ so that $\pi_{1,2}(x(t)) = 0$. Plugging this value of $\pi_{1,2}(x(t))$ in the ODE for $q_1(t)$ in (4.2), we see that $\dot{q}_1(t) \geq \lambda_1 - \mu_{1,1}m_1 - \theta_1\kappa \geq 0$ by Assumption A. Hence, $q_1(t)$ is nondecreasing. If $q_1(t) > \kappa$ and $q_2(t) = 0$, then $x(t) \in \mathbb{S}^+$ and $\pi_{1,2}(x(t)) = 1$, which gives $\dot{q}_2(t) = \lambda_2 > 0$. Hence q_2 is increasing at time t.

Now consider the case $q_1(t) = \kappa$ and $q_2(t) = 0$, so that $x(t) \in \mathbb{S}^b$. For one of the queues to become negative at time $t+$, we need to have its derivative be negative at time t. We will consider various subcases.

First assume that $\dot{q}_1(t) < 0$ and $\dot{q}_2(t) \geq 0$. In that case $(q_2 - q_1)(t+) > 0$, so that $\pi_{1,2}(x(t+)) = 0$. Plugging this value of $\pi_{1,2}(x(t+))$ in the ODE (4.2), together with $q_1(t+) = \kappa$, we see that $\dot{q}_1(t+) > 0$ by Assumption A. Next assume that $\dot{q}_1(t) \geq 0$ and $\dot{q}_2(t) < 0$. Then $(q_1 - q_2)(t+)$, so that $\pi_{1,2}(x(t+)) = 1$. Plugging this value of $\pi_{1,2}(x(t+))$, together with $q_2(t+) = 0$, we see that $\dot{q}_2(t+) > 0$.

We finally consider the remaining more challenging subcase: $\dot{q}_1(t) < 0$ and $\dot{q}_2(t) < 0$. We will show that this subcase is not possible. To that end, we further divide this case into three subcases: $x(t) \in \mathbb{A}^+$, $x(t) \in \mathbb{A}^-$ and $x(t) \in \mathbb{A}$. (Recall that $\mathbb{S}^b = \mathbb{A} \cup \mathbb{A}^+ \cup \mathbb{A}^-$.) However, $x(t)$ cannot be in $\mathbb{A}^-$, since then $\pi_{1,2}(x(t)) = 0$, which implies that $q_1(t)$ is nondecreasing (plug $\pi_{1,2}(x(t)) = 0$ and $q_1(t) = \kappa$ into the ODE (4.2)). Moreover, $x(t)$ cannot be in $\mathbb{A}^+$, since then $\pi_{1,2}(x(t)) = 1$, which implies that $q_2(t)$ is strictly increasing.

Now assume the remaining possibility, $x(t) \in \mathbb{A}$, and recall that Ψ is Lipschitz continuous in $\mathbb{A}$, so that the Lyapunov argument holds over $[t, t + \eta)$, for some $\eta > 0$. Specifically, the Lyapunov function V is monotone increasing in $x(t)$, because $x^* > 0$. (The inequality holds componentwise.) If $\mu_{1,2} > \mu_{2,2}$, then we take the Lyapunov function $V_1(x(t)) = q_1(t) + q_2(t)$. The monotonicity of V_1 at $x(t)$ implies that at least one of the queues must be increasing, which contradicts the assumption that the derivative of both queues is negative at t. If $\mu_{1,2} \leq \mu_{2,2}$, then we take the Lyapunov function $V_2(x(t)) = Cq_1(t) + q_2(t) + (C - 1)z_{1,2}(t)$. We then choose $C = 1 + \epsilon$ with ϵ small enough, such that $\dot{V}_2(x(t)) < 0$ (assuming the derivatives of both queues are strictly negative at t). Once again, this contradicts the positive

monotonicity of V at $x(t)$. This concludes the proof. □

9. Exponential stability.

DEFINITION 9.1 (exponential stability). A stationary point x^* is said to be exponentially stable if there exist two real constants ϑ, $\beta > 0$ such that

$$\|x(t) - x^*\| \leq \vartheta \|x(0) - x^*\| e^{-\beta t},$$

for all $t \geq 0$ and for all $x(0)$, where $\|\cdot\|$ is a norm on $\mathbb{R}^n$.

We use Theorem 3.4 on p. 82 of Marquez [12], stated below.

THEOREM 9.1 (exponential stability of the origin). *Suppose that all the conditions of Theorem 8.4 are satisfied. In addition, assume that there exist positive constants K_1, K_2, K_3 and p such that*

$$K_1 \|x\|^p \leq V(x) \leq K_2 \|x\|^p \quad \textit{and} \quad \dot{V}(x) \leq -K_3 \|x\|^p.$$

Then the origin is exponentially stable, and

$$\|x(t)\| \leq \|x(0)\| \, (K_2/K_1)^{1/p} \, e^{-(K_3/2K_2)t} \quad \textit{for all} \quad t \quad \textit{and} \quad x(0).$$

We use the L_1 norm: $\|x\| = |x_1| + |x_2| + |x_3|$ for $x \in \mathbb{R}^3$.

THEOREM 9.2 (exponential stability of x^*). *Each x^* in $\mathbb{S}$ is exponentially stable.*

(i) If $\mu_{1,2} > \mu_{2,2}$, then

$$\|x(t) - x^*\| \leq \|x(0) - x^*\| e^{-(K_3/2)t} \quad \textit{for all} \quad t \quad \textit{and} \quad x(0)$$

for all $x(0) \in \mathbb{S}$ and $t \geq 0$, where $K_3 \equiv \max\{\theta_1, \theta_2, \mu_{1,2} - \mu_{2,2}\}$.

(ii) If $\mu_{2,2} = B\mu_{1,2}$, $B \geq 1$, then

$$\|x(t) - x^*\| \leq \|x(0) - x^*\| (C/K_1) e^{-(K_4/2)t}$$

for all $x(0) \in \mathbb{S}$, $t \geq 0$ and $C > B$, where $K_1 \equiv \min\{1, C - 1\}$ and $K_4 \equiv \max\{C\theta_1, \theta_2, (C\mu_{1,2} - \mu_{2,2})\}$.

PROOF. As in the proof of Theorem 8.3, Theorem 9.2 applies directly only within one region, starting at a point in $\mathbb{S}^+$, $\mathbb{S}^-$, $\mathbb{A}$, $\mathbb{A}^-$ or $\mathbb{A}^+$. However, again, the same Lyapunov function V can be used in all regions.

We consider the two cases in turn: (i) In the proof of Theorem 8.4, $V_1(x) \equiv x_1 + x_2$ was shown to be a Lyapunov function with a strictly negative Lie derivative. Since $x \geq 0$, we can take $K_1 = K_2 = 1$ and $p = 1$. Since $\dot{V}_1(x) = -\theta_1 q_1(t) - \theta_2 q_2(t) - (\mu_{1,2} - \mu_{2,2}) z_{1,2}(t)$, we can take K_3 specified above, and the result follows from Theorem 9.1.

(ii) We use the Lyapunov function $V_2(x) = Cx_1 + x_2 + (C-1)x_3$. Then $K_1\|x\| \leq V_2(x) < C\|x\|$ for $K_1 \equiv \min\{1, C-1\}$. From the proof of Theorem 8.4, we know that $\dot{V}_2(x) = -C\theta_1 q_1(t) - \theta_2 q_2(t) - (C\mu_{1,2} - \mu_{2,2}) z_{1,2}(t)$, so that $\dot{V}_2(x) \leq -K_4\|x\|$. □

10. Conditions for state-space collapse. In this section we give ways of verifying that x lies entirely in $\mathbb{A}$, given that $x(0)$ and x^* are both in $\mathbb{A}$. In the appendix we provide conditions for the solution to eventually reach $\mathbb{A}$ after an initial transient. The results here are intended to apply after this initial transient period has concluded.

THEOREM 10.1 (sufficient conditions for global SSC). *Let $\nu \equiv \mu_{1,2} \wedge \mu_{2,2}$, and suppose that $x(0) \in \mathbb{A}$. Also assume that*

$$q_2(0) \leq \lambda_2/\theta_2 \quad \textit{and} \quad q_1(0) \leq (\lambda_1 - m_1\mu_{1,1})/\theta_1. \tag{10.1}$$

If, in addition, the following inequalities are satisfied, then the solution to the IVP (4.3) is in $\mathbb{A}$ for all t:

$$(i) \quad \lambda_1 < \nu m_2 + m_1\mu_{1,1} \quad \textit{and} \quad (ii) \quad \lambda_2 > \nu m_2. \tag{10.2}$$

PROOF. We start by showing, under Condition (i), that $\delta_+(x(t))$ in (6.9) is strictly negative for each t. For a fixed t,

$$\delta_+(x(t)) \equiv j\left(\lambda_+^{(j)}(t) - \mu_+^{(j)}(t)\right) + k\left(\lambda_+^{(k)}(t) - \mu_+^{(k)}(t)\right) < 0$$

if and only if

$$(\mu_{2,2} - \mu_{1,2}) z_{1,2}(t) - m_2\mu_{2,2} < -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2 q_2(t)) + \theta_1 q_1(t). \tag{10.3}$$

If $\mu_{2,2} > \mu_{1,2}$, then the left-hand side (LHS) of (10.3) is maximized at $z_{1,2}(t) = m_2$, and is equal to $-\mu_{1,2}m_2$. If $\mu_{2,2} < \mu_{1,2}$, the the LHS is maximized at $z_{1,2}(t) = 0$, and is equal to $-\mu_{2,2}m_2$. When $\mu_{2,2} = \mu_{1,2}$ the LHS is equal to $-\mu_{2,2}m_2 = -\mu_{1,2}m_2$. Overall, the LHS of (10.3) is smaller than or equal to $-\nu m_2$.

Since $q_2(0) \leq \lambda_2/\theta_2$, we conclude, using the bound in (7.5), that $\theta_2 q_2(t) \leq \lambda_2$ for all $t \geq 0$. This, together with the fact that $q_1(t) \geq 0$ for all t, implies that the RHS of (10.3) is larger than or equal to $-(\lambda_1 - m_1\mu_{1,1})$, so that

$$\begin{aligned}(\mu_{2,2} - \mu_{1,2})z_{1,2}(t) - \mu_{2,2}m_2 &\leq -\nu m_2 < -(\lambda_1 - m_1\mu_{1,1}) \\ &\leq -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2 q_2(t)) + \theta_1 q_1(t)\end{aligned}$$

where the second inequality is due to condition (i).

To show that condition (ii) is sufficient to have $\delta_-(x(t)) > 0$ for all t, fix $t \geq 0$ and note that, for $\delta_-(x(t))$ in (6.9), we have

$$\delta_-(x(t)) \equiv j\left(\lambda_-^{(j)}(t) - \mu_-^{(j)}(t)\right) + k\left(\lambda_-^{(k)}(t) - \mu_-^{(k)}(t)\right) > 0$$

if and only if

(10.4)
$$r(\mu_{1,2} - \mu_{2,2})z_{1,2}(t) + r\mu_{2,2}m_2 > -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2 q_2(t)) + \theta_1 q_1(t).$$

It is easy to see that the LHS of (10.4) has a minimum value of $r(\mu_{1,2} \wedge \mu_{2,2})m_2 \equiv r\nu m_2$. By essentially the same arguments as in Theorem 7.2 we can show that $q_1(t) \leq q_1(0) \vee (\lambda_1 - m_1\mu_{1,1})/\theta_1$. Since we assume that $q_1(0) \leq (\lambda_1 - m_1\mu_{1,1})/\theta_1$, we have the bound $q_1(t) \leq (\lambda_1 - m_1\mu_{1,1})/\theta_1$ for all $t \geq 0$. With this bound, we see that the RHS of (10.4) is smaller than or equal to $r\lambda_2$. Overall, we have

$$\begin{aligned}r(\mu_{1,2} - \mu_{2,2})z_{1,2}(t) + r\mu_{2,2}m_2 &\geq r\nu m_2 > r\lambda_2 \\ &\geq -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2 q_2(t)) + \theta_1 q_1(t),\end{aligned}$$

where the second inequality is due to Condition (ii). Since (6.10) holds for all $t \geq 0$, we also have $0 < \pi_{1,2}(t) < 1$ for all t. Hence, every solution to the IVP in (4.3) must lie entirely in $\mathbb{A}$. □

For $x^* \in \mathbb{A}$, we will now show that there exist $\alpha > 0$ and $T \equiv T(\alpha)$, such that global SSC can be inferred once $\|x(T) - x^*\| < \alpha$. We exploit the drift rates at stationarity, defined by $\delta_+^* \equiv \delta_+(x^*)$ and $\delta_-^* \equiv \delta_-(x^*)$. It follows from the expressions in (6.9) that

(10.5)
$$\delta_+^* \equiv \delta_+(x^*) = -\mu_{2,2}(r+1)(m_2 - z_{1,2}^*), \quad \delta_-^* \equiv \delta_-(x^*) = \mu_{1,2}(r+1)z_{1,2}^*.$$

Thus, if $0 < z_{1,2}^* < m_2$, then the positive recurrence condition (6.10) holds at the stationary point x^*. (This agrees with (8.3) which has $0 < \pi_{1,2}^* < 1$ if and only if $0 < z_{1,2}^* < m_2$.)

In the next theorem we give explicit expressions for α. For reasonable rates, such as will hold in applications, α is quite large. In fact, in the numerical example considered in §11.3 we show that, typically in applications, α is so large, that we can infer that x lies entirely in $\mathbb{A}$ without even solving the IVP; i.e., $x(0) \in \beta_V(\alpha)$.

THEOREM 10.2. *Suppose that $x^* \in \mathbb{A}$ and let $\xi \equiv \min\{|\delta_+^*|, \delta_-^*\}$.*

(i) If $\mu_{2,2} \geq \mu_{1,2}$, then let $\alpha = \xi / r\theta_2$

(ii) If $\mu_{2,2} < \mu_{1,2}$, then let $\alpha = \xi/\varsigma$, where $\varsigma \equiv \mu_{1,2} - \mu_{2,2} + \theta_1 + r\theta_2 > 0$. In both cases, if there exists $T \geq 0$ such that $x(T) \in \beta_V(\alpha)$, then $\{x(t) : t \geq T\}$ lies entirely in $\mathbb{A}$.

PROOF. We use $\beta_V(\alpha)$, the α V-ball with center at x^* and radius α, in (8.12). To find a proper α for the V-ball $\beta_V(\alpha)$, we once again use the conditions (10.3) and (10.4). We first show how to find α for the case $\mu_{2,2} = B\mu_{1,2}$ for some $B \geq 1$, i.e., when $\mu_{1,2} \leq \mu_{2,2}$. Recall (proof of Theorem 8.3) that in this case, $V_2(x) = Cx_1 + x_2 + (C-1)x_3$ is a Lyapunov function for any $C > B$. Also, the Lyapunov function was defined for the modified system in which the origin was the stationary point.

Let $x^* = (q_1^*, q_2^*, z_{1,2}^*)$ be the stationary point in $\mathbb{A}$. First assume that, at some time T, $V_2(x(T)) = \epsilon_1$, i.e., $Cq_1(T) + q_2(T) + (C-1)z_{1,2}(T) = \epsilon_1$. If $x(t) \in \beta_{V_2}(\epsilon_1)$ for all $t > T$, then it must hold that

(10.6)
$$q_1^* - \frac{\epsilon_1}{C} < q_1(t) < q_1 + \frac{\epsilon_1}{C}, \qquad q_2^* - \epsilon_1 < q_2(t) < q_2^* + \epsilon_1 \quad \text{and}$$
$$z_{1,2}^* - \frac{\epsilon_1}{C-1} < z_{1,2}(t) < z_{1,2}^* + \frac{\epsilon_1}{C-1}, \quad t \geq T.$$

To make sure $\delta_+(x(t)) < 0$, we use (10.3), reorganizing the terms. We need to have

$$(\mu_{2,2} - \mu_{1,2})z_{1,2}(t) + r\theta_2 q_2(t) - \theta_1 q_1(t) < -(\lambda_1 - \mu_{1,1}m_1) + r\lambda_2 + \mu_{2,2}m_2.$$

By (10.6), the above inequality holds if

$$(\mu_{2,2} - \mu_{1,2})\left(z_{1,2}^* + \frac{\epsilon_1}{C-1}\right) + r\theta_2(q_2^* + \epsilon_1) - \theta_1\left(q_1^* - \frac{\epsilon_1}{C}\right)$$
$$< -(\lambda_1 - \mu_{1,1}m_1) + r\lambda_2 + \mu_{2,2}m_2.$$

Plugging in the expressions for q_1^*, q_2^* and $z_{1,2}^*$, we see that we need to find an $\epsilon_1 > 0$ such that

$$(\mu_{2,2} - \mu_{1,2})\frac{\epsilon_1}{C-1} + r\theta_2\epsilon_1 + \theta_1\frac{\epsilon_1}{C} < \mu_{2,2}(r+1)(m_2 - z_{1,2}^*).$$

We can take C as large as needed, so that the only term that matters on the LHS is $r\theta_2\epsilon_1$. Hence, we need to have

$$\epsilon_1 < \frac{\mu_{2,2}(r+1)(m_2 - z^*_{1,2})}{r\theta_2} = \frac{|\delta^*_+|}{r\theta_2}.$$

Similarly, to make sure that $\delta_-(x(t)) > 0$, we use (10.4), reorganizing the terms. We need to have

$$\begin{aligned} r(\mu_{1,2} - \mu_{2,2})z_{1,2}(t) + r\theta_2 q_2(t) - \theta_1 q_1(t) \\ > -(\lambda_1 - \mu_{1,1}m_1) + r(\lambda_2 - \mu_{2,2}m_2). \end{aligned}$$

Using (10.6) again (with a different ϵ_2), we see that it suffices to show that

$$\begin{aligned} r(\mu_{1,2} - \mu_{2,2})\left(z^*_{1,2} + \frac{\epsilon_2}{C-1}\right) + r\theta_2(q^*_2 - \epsilon_2) - \theta_1\left(q^*_1 + \frac{\epsilon_2}{C}\right) \\ > -(\lambda_1 - \mu_{1,1}m_1) + r(\lambda_2 - \mu_{2,2}m_2). \end{aligned}$$

Once again, plugging in the values of q^*_1, q^*_2 and $z^*_{1,2}$, and taking C as large as needed, we can choose $\epsilon_2 > 0$ such that

$$\epsilon_2 < \frac{\mu_{1,2}(r+1)z^*_{1,2}}{r\theta_2} = \frac{\delta^*_-}{r\theta_2}.$$

Hence, we can take α as in (i).

For the second case, when $\mu_{1,2} > \mu_{2,2}$, we use the Lyapunov function $V_1(x) = x_1 + x_2$. Using similar reasoning as above, we get

$$\epsilon_1 < \frac{\mu_{2,2}(r+1)(m_2 - z^*_{1,2})}{\mu_{1,2} - \mu_{2,2} + \theta_1 + r\theta_2} = \frac{|\delta^*_+|}{\varsigma} \quad \text{and} \quad \epsilon_2 < \frac{\mu_{1,2}(r+1)z^*_{1,2}}{\mu_{1,2} - \mu_{2,2} + \theta_1 + r\theta_2} = \frac{\delta^*_-}{\varsigma}.$$

Hence, in this case we can take α in (ii). □

11. A numerical algorithm to solve the IVP.

11.1. *Computing $\pi_{1,2}(x)$ at a point x.* The QBD structure in §6.2 allows us to use established efficient numerical algorithms from [11] to solve for the steady state of the QBD to compute $\pi_{1,2}(x)$, for any given $x \equiv x(t) \in \mathbb{A}$.

We start by computing the rate matrix $R \equiv R(x)$. (To simplify notation, we drop the argument x, with the understanding that all matrices, are functions of x.) By Proposition 6.4.2 of [11], R is related to matrices G and U via

$$G = (-U)^{-1}A_2, \quad U = A_1 + A_0G \quad \text{and} \quad R = A_0(-U)^{-1}. \tag{11.1}$$

In addition, the matrices G and R are the minimal nonnegative solutions to the quadratic matrix equations

$$A_2 + A_1G + A_0G^2 = 0 \quad \text{and} \quad A_0 + RA_1 + R^2A_2 = 0. \tag{11.2}$$

Hence, if can compute the matrix G, then the rate matrix R can be found via (11.1). Once R is known, we use (6.16) to compute α_0. With α_0 and R in hand, $\pi_{1,2}(x)$ is easily computed via (6.17).

It remains to compute the matrix G. We use the *logarithmic reduction* (LR) *algorithm* in §8.4 of [11], modified to the continuous case, as in §8.7 of [11]. The LR algorithm is quadratically convergent and is numerically well behaved. These two properties are important, because the matrix $R(x)$ needs to be computed for many values of x when we numerically solve the IVP (4.3). From our experience with this algorithm, it takes fewer than ten iterations to achieve a 10^{-6} precision (when calculating G).

11.2. *Computing the solution x.* To compute the solution x, we combine the forward Euler method for solving an ODE with the algorithm to solve for $\pi_{1,2}(x(t))$ described above. Specifically, we start with a specified initial value $x(0)$, a step-size h and number of iterations n, such that $nh = T$. First, assume that $z_{1,1}(0) = m_1$ and $z_{1,2}(0) + z_{2,2}(0) = m_2$, so that $x(0) \in \mathbb{S}$. If $\bar{D}(0) \equiv (q_1(0) - \kappa) - rq_2(0) > 0$ then $\pi_{1,2}(x(0)) = 1$. If $\bar{D}(0) < 0$ then $\pi_{1,2}(x(0)) = 0$ and, if $\bar{D}(0) = 0$, then we check to see whether (6.10) holds. If it does, then $x(0) \in \mathbb{A}$ and we calculate $\pi_{1,2}(x(0))$ as described above. If $x(0) \in \mathbb{S}^b - \mathbb{A}$ then we can still determine the value of $\pi_{1,2}(x(0))$ in the following way: If $\delta_-(x(t)) = 0 > \delta_+(x(t))$, then we let $\pi_{1,2}(x(t)) = 0$; if instead $\delta_-(x(t)) > 0 = \delta_+(x(t))$, then we let $\pi_{1,2}(x(t)) = 1$.

Given $x(0)$ and $\pi_{1,2}(x(0))$ we can calculate $\Psi(x(0))$ explicitly, and perform the Euler step $x(h) = x(0) + h\Psi(x(0))$. We then repeat the procedure for each k, $0 \le k \le n-1$, i.e.,

$$x((k+1)h) = x(kh) + h\Psi(x(kh)), \quad 0 \le k \le n, \tag{11.3}$$

where $x(kh)$ is given from the previous iteration, and $\Psi(x(kh))$ can be computed once $\pi_{1,2}(x(kh))$ is found.

If $z_{1,1}(0) < m_1$ or $z_{1,2}(0) + z_{2,2}(0) < m_2$, so that $x(0) \notin \mathbb{S}$, we use the appropriate fluid model for the alternative region, as specified in the appendix, where at each Euler step we check to see which fluid model should be applied.

The forward Euler algorithm is known to have an error proportional to the step size h, and to be relatively numerically unstable at times, but it was found to be adequate. It would be easy to apply more sophisticated algorithms, such as general linear methods, which have a smaller error, and can be more numerically stable. The only adjustment required is to replace the Euler step in (11.3) by the alternative method.

In the numerical example in §11.3 below we let the ratio be $r = 0.8 = 4/5$, so that all the matrices, used in the computations for $\pi_{1,2}$, are of size 10×10. It took less than 10 seconds for the algorithm to terminate (using a relatively slow, 1 GB memory, laptop). The same example, but with $r = 20/25$, so that the matrices are now 50×50, the algorithm took less than a minute to terminate. Moreover, the answers to both trials were exactly the same, up to the $7th$ digit. In both cases, we performed 5000 Euler steps (each of size $h = 0.01$, so that the termination time is $T = 50$). It is easily seen that $\pi_{1,2}$ had to be calculated for over 4500 different points, starting at the time $\pi_{1,2}$ becomes positive (see Figure 2 in the following example).

The validity of the solution can be verified by comparing it to simulation results, as in the example below and others in [16, 19]. There are two other ways to verify the validity: First, we can check that the solution converges to the stationary point x^*, which can be computed explicitly using (8.2). Second, within $\mathbb{A}$ we can see that the two queues keep at the target ratio r, even though this relation between the two queues is not forced explicitly by the algorithm.

11.3. *A numerical example.* We now provide a numerical example of the algorithm for solving the ODE in (4.1). In addition, we added the sample paths of the stochastic processes Q_1^n and $Z_{1,2}^n$, after scaling as in (3.2), on top of the trajectories of the solution to their fluid counterparts q_1 and $z_{1,2}$.

The model has the same target ratio $r = 0.8$ as in the example in §6.2 with component rate matrices in (6.12). We chose a large queueing system with scaling factor $n = 1000$, so that the stochastic fluctuations do not to hide the general structure of the simulated sample paths. We let the ODE model parameters be $m_1 = m_2 = 1$, $\lambda_1 = 1.3$, $\lambda_2 = 0.9$, $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$, $\theta_1 = \theta_2 = 0.3$ and $\kappa = 0$. The associated queueing model has the same parameters $\mu_{i,j}$ and θ_i, but the other parameters are multiplied by n. The plots are shown without dividing by n.

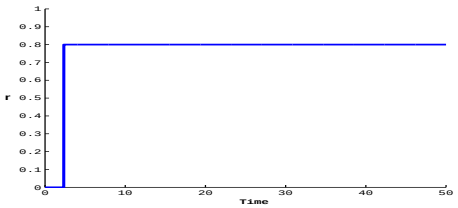

FIG 1. *Ratio between the queues.*

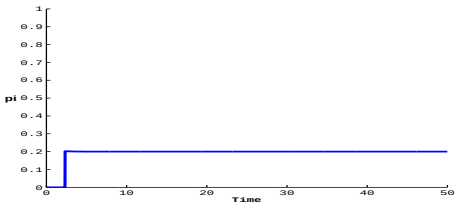

FIG 2. $\pi_{1,2}$ *calculated at each iteration.*

We ran the algorithm and the simulation for 50 time units. We used an Euler step of size $h = 0.01$, so we performed 5000 Euler iterations. In each Euler iteration we performed several iterations to calculate the matrix G in (11.1), which is used to calculate the instantaneous steady-state probability $\pi_{1,2}$.

Figures 1–4 show $q_1(t)/q_2(t)$, $\pi_{1,2}(x(t))$, $q_1(t)$ and $z_{1,2}(t)$ as functions of time t for a system initialized empty. After a short period, the pools fill up. Then $q_1(t)$ starts to grow, and immediately then fluid (customers) starts flowing to pool 2, causing $z_{1,2}(t)$ to grow. Figures 1–4 show that, for practical purposes, steady state is achieved for $t \in [10, 20]$.

In Figure 1 we see that once $\mathbb{S}^b$ is hit, the ratio between the queues is kept at the target ratio 0.8. This is an evidence for the validity of the numerical solution, and a strong demonstration of the AP. In Figure 2 we see that initially, while $q_1 = 0$, $\pi_{1,2} = 0$. This lasts until $z_{2,2}(t) + z_{1,2}(t) = m_2$, at which time the space $\mathbb{S}$ is hit, specifically $\mathbb{S}^b$), and the averaging begins. Once $\mathbb{S}^b$ is hit, $\pi_{1,2}$ becomes almost constant, even before the system reaches steady state. Thus the functions q_1, q_2 and $z_{1,2}$ have exponential form, supporting the results of §9.

We got $x(t_n) \equiv (q_1(t_n), q_2(t_n), z_{1,2}(t_n)) = (0.3639, 0.4550, 0.2385)$ and $\pi_{1,2}(t_n) = 0.2$ when the algorithm terminated. From (8.2), $x^* \equiv (q_1^*, q_2^*, z_{1,2}^*) = (0.3667, 0.4595, 0.2375)$. From (8.3), we get $\pi_{1,2}^* = 0.2$.

Before solving the ODE, we can apply Theorem 10.2 to conclude that the solution will remain in $\mathbb{A}$ after it first hits $\mathbb{A}$.

12. Proof of Theorem 7.1. It is immediate that the function Ψ in (4.1) and (4.2) is Lipschitz continuous on $\mathbb{S}^+$ and $\mathbb{S}^-$, because $\pi_{1,2}(x) = 1$ when $x \in \mathbb{S}^+$ and $\pi_{1,2}(x) = 0$ when $x \in \mathbb{S}^-$, so that Ψ is linear in each of these regions. However, Ψ is not linear on $\mathbb{S}^b$, because Ψ involves $\pi_{1,2}(x)$, which is a nonlinear function of the state x determined by the FTSP in §6.

We now prove the three conclusions involving $\mathbb{A}$, $\mathbb{A}^+$ and $\mathbb{A}^-$. We will use the fact that a function mapping a convex compact subset of $\mathbb{R}^m$ to $\mathbb{R}^n$

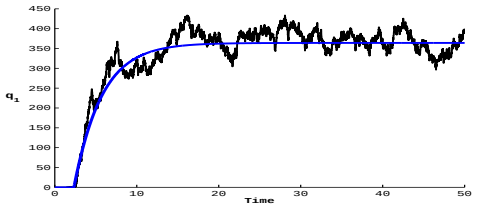

FIG 3. *Trajectory of q_1 together with a simulated sample path of the stochastic process Q_1 in a system initializing empty.*

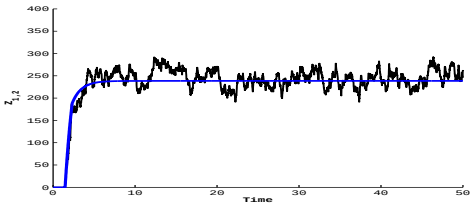

FIG 4. *Trajectory of $z_{1,2}$ together with a simulated sample path of the stochastic process $Z_{1,2}$ in a system initializing empty.*

is Lipschitz on that domain if it has a bounded derivative. Since we can always work with balls in $\mathbb{R}^m$ (which are convex with compact closure), that in turn implies that a function mapping an open subset of $\mathbb{R}^m$ to $\mathbb{R}^n$ is locally Lipschitz whenever it has a bounded derivative on each ball in the domain; e.g., see Lemma 3.2 of [10]. The three sets $\mathbb{A}$, $\mathbb{A}^+$ and $\mathbb{A}^-$ are convex. The key is what happens in $\mathbb{A}$.

For understanding, it is helpful to first verify this theorem in the special case $r = 1$, where the QBD process reduces to a BD process. Thus we first give a proof for that special case.

Proof for the special case $r = 1$. Since $\mathbb{S}^b$ is homeomorphic to a closed subset of $\mathbb{R}^2$, we can (and do) regard $\mathbb{A}$ as an open connected convex subset of $\mathbb{R}^2$. The key component of the function Ψ in $\mathbb{A}$ is $\pi_{1,2}$. We exploit the explicit representations in (6.18) and (6.21). From (6.1)–(6.4), the partial derivatives of $\lambda^{\pm}(x)$ and $\mu^{\pm}(x)$ with respect to the three components of x, i.e., q_1, q_2 and $z_{1,2}$, are constants. From (6.18) and (6.21), we see that the partial derivatives of $\pi_{1,2}(x)$ with respect to each of the three components of x exist, are finite and continuous. That takes care of $\mathbb{A}$.

We elaborate: We can obtain an explicit expression for the derivatives by applying the chain rule. Suppose that we consider the partial derivative with respect component i of $x = (q_1, q_2, z_{1,2})$ (e.g, $x_1 = q_1$). By (6.1)–(6.4) and (6.9), the partial derivatives of $\delta_+(x)$ and $\delta_-(x)$ with respect to x_i are constant as functions of x; let these constants be denoted by $\dot{\delta}_+$ and $\dot{\delta}_-$. Let $\dot{\pi}_{1,2}(x)$ denote the partial derivative of $\pi_{1,2}(x)$ with respect to x_i. Then, from (6.18), we obtain the explicit representation

$$\dot{\pi}_{1,2}(x) = \frac{\dot{\delta}_+\delta_-(x) - \dot{\delta}_-\delta_+(x)}{(\delta_-(x) - \delta_+(x))^2} \quad \text{in} \quad \mathbb{A}. \tag{12.1}$$

By (7.2), the denominator in (12.1) is strictly positive in $\mathbb{A}$. Since the functions $\delta_+(x)$ and $\delta_-(x)$ are linear in this case, they are continuous. Thus,

from (12.1), we see that indeed the partial derivative $\dot{\pi}_{1,2}(x)$ is well defined and continuous on $\mathbb{A}$.

We next consider $\mathbb{A}^-$ and $\mathbb{A}^+$; the reasoning for these two cases is essentially the same, with the representations in (6.13), (6.18), and (12.1) making it quite elementary. The relation (6.13) implies that the denominator in (12.1) is uniformly bounded below in $\mathbb{A}$. Thus, $\pi_{1,2}(x) \to 0$ and these partial derivatives approach finite limits as $x \to x_b \in \mathbb{A}^-$ for $x \in \mathbb{A}$, while $\pi_{1,2}(x) \to 1$ and these partial derivatives approach finite limits as $x \to x_b \in \mathbb{A}^+$ for $x \in \mathbb{A}$. In both cases we have a conventional heavy-traffic limit: $\rho^{\pm}(x) \uparrow 1$ as $x \to x_b$. Hence, the partial derivatives of $\pi_{1,2}(x)$ are continuous and bounded on $\mathbb{S}^b$. As a consequence, for any ϵ-ball in $\mathbb{S}-\mathbb{S}^-$ about x in $\mathbb{A}^+$, there exists a constant K such that $|\pi_{1,2}(x_1)-\pi_{1,2}(x_2)| \le K\|x_1-x_2\|_3$ for all x_1 and x_2 in the ϵ-ball, where $\|\cdot\|_3$ is the maximum norm on $\mathbb{R}^3$. A similar statement applies to $\mathbb{A}^-$. Hence we have completed the proof for $r = 1$.

We make two concluding remarks. We first note that the derivative depends on the neighborhood of x that we consider. At a point x in $\mathbb{A}^+$ (and similarly for $\mathbb{A}^-$), if we take a sequence of points $x_n : n \ge 1$ with $x_n \to x$ as $n \to \infty$, where $x_n \in \mathbb{S}^+$ for all $n \ge 1$, then $\pi_{1,2}(x_n) = 1$ for all n, so that the derivative is 0. On the other hand, the derivative approaching $x \in \mathbb{A}^+$ through A need not be 0. However, by the reasoning above that derivative is finite, That is sufficient for the required local Lipschitz continuity.

Second, we observe that we cannot conclude that $\pi_{1,2}(x)$ is even continuous on all of $\mathbb{S}$, because for $x \in \mathbb{A}$ we can have a sequence $\{x_n : n \ge 1\}$ with $x_n \in \mathbb{S}^+$ for all n (or $x_n \in \mathbb{S}^-$ for all n), with $x_n \to x$ as $n \to \infty$, $\pi_{1,2}(x_n) = 1$ for all n (or $= 0$), while $0 < \pi_{1,2}(x) < 1$.

We now treat the general case.

Proof of Theorem 7.1 in the general case. We first consider $\mathbb{A}$. As in the case $r = 1$, we are regarding $\mathbb{A}$ as an open connected convex subset of $\mathbb{R}^2$. We will look at $\pi_{1,2}$, and thus the QBD, as a function of the variable $x \in \mathbb{A}$, which is an element of $\mathbb{R}^3$. By the definition of the matrices A_0, A_1 and A_2 in (6.6) (see also the example in §6.2), these matrices are twice differentiable with respect to any of their elements. By the definition of the rates in (6.1)-(6.4), which are the elements of the matrices A_0, A_1 and A_2, these matrix elements in turn have constant partial derivatives with respect to each of the three real components of x at each $x \in \mathbb{A}$, i.e., with respect to q_1, q_2 and $z_{1,2}$. It follows from Theorem 2.3 in He [8] that the rate matrix R in (6.15), which is the minimal nonnegative solution to the quadratic matrix equation $A_0 + RA_1 + R^2A_2 = 0$, is also twice differentiable with respect to the matrix elements of A_0, A_1 and A_2, and thus also with respect to the three real components of x at each $x \in \mathbb{A}$.

It thus suffices to look at the derivatives with respect to one of the elements of the matrices A_0, A_1 or A_2. It follows from the normalizing expression in (6.16) and the differentiability of R, that α_0 is also differentiable. Hence, from (6.17), we see that $\pi_{1,2}$ is differentiable at each $x \in \mathbb{A}$, with

$$\pi'_{1,2} = \alpha'_0(I-R)^{-1}\mathbf{1}_+ + \alpha_0(I-R)^{-1}R'(I-R)^{-1}\mathbf{1}_+. \tag{12.2}$$

By differentiating (6.16), we have

$$\alpha'_0(I-R)^{-1}\mathbf{1} + \alpha_0(I-R)^{-1}R'(I-R)^{-1}\mathbf{1} = 0, \tag{12.3}$$

so that α'_0 is continuous. The continuity of R' and α'_0 with respect to one of the elements of the matrices A_0, A_1 or A_2 implies that the derivative $\pi'_{1,2}$ with respect to one of the elements of the matrices A_0, A_1 or A_2 is finite and continuous on $\mathbb{A}$, which in turn implies that the partial derivatives with respect to the three real components of x at each $x \in \mathbb{A}$ are finite and continuous as well. Hence, Ψ is locally Lipschitz continuous on $\mathbb{A}$, as claimed.

We next show that $\pi_{1,2}$ and thus Ψ are locally Lipschitz continuous in neighborhoods of points in $\mathbb{A}^+$ within $\mathbb{S} - \mathbb{S}^-$ and of points in $\mathbb{A}^-$ within $\mathbb{S} - \mathbb{S}^+$. We will only consider $\mathbb{A}^+$, because the two cases are essentially the same. In both cases, the situation is complicated starting from (12.2) because the entries of $\alpha_0(x)$ become negligible, while the entries of $(I-R)^{-1}(x)$ explode as $x \to x_b$. However, the two different limits cancel their effect. We exploit (6.19). The representation in (6.19) is convenient because now $\alpha_0(x) \to \alpha_0(x_b)$ as $x \to x_b$, where $\alpha_0(x_b)$ is finite. All key asymptotics take place in R^+.

Since the crucial asymptotics involves only $\mathbb{R}^+$, we see that we only need carefully consider one of the two regions, in this case the upper one. To obtain results about $\mathbb{R}^+$, from a process perspective, it suffices to replace the given QBD by a new QBD with the upper region and reflection at the lower boundary. The new QBD model involving only $\mathbb{R}^+$ is equivalent to a relatively simple single-server queue. The net input is a linear combination of four Poisson processes, and so has stationary and independent increments. The queue length process in the revised model is an elementary $MAP/MSP/1$ queue, as in §4 of [1], which has as QBD representation with rate matrix R^+.

For the asymptotics, the key quantities are the spectral radii of the matrices $R^+(x)$ and $R^-(x)$, say $\eta^+(x)$ and $\eta^-(x)$, and the way that these depend on the drifts $\delta_+(x)$ and $\delta_-(x)$ as $x \to x_b$. The spectral radius $\eta^+(x)$ is the unique root in the interval $(0,1)$ of the equation $det[A_0^+(x) + A_1^+(x)\eta + A_2^+(x)\eta^2] = 0$, and similarly for $\eta^-(x)$; see (39) on p. 241 of [13], the

Appendix of [14] and §4 of [1]. We see that $\eta^+(x) \to \eta^+(x_b) = 1$ and $\eta^-(x) \to \eta^-(x_b) < 1$ as $x \to x_b \in \mathbb{A}^+$. In general, we can represent powers of the matrix R (and similarly for R^+ and R^-) asymptotically as

$$R^n = vu\eta^n + o(\eta^n) \quad \text{as} \quad n \to \infty, \tag{12.4}$$

where u and v are the left and right eigenvectors of the eigenvalue η, respectively, normalized so that $u\mathbf{1} = \mathbf{1}$ and $uv = 1$. Moreover, as $\eta \to 1$, the matrix inverse $(I - R)^{-1}$ is dominated by these terms.

Hence, we can do a heavy-traffic expansion of $\eta^+(x)$ and the related quantities as $x \to x_b \in \mathbb{A}^+$ with $x \in \mathbb{A}$, as in [3]; see the Appendix of [14]. As $x \to x_b$, all quantities in (6.19) have finite continuous limits as $x \to x_b \in \mathbb{A}^+$ except $(I - R^+(x))^{-1}$. We first have $|\delta_+(x)| \to 0$ and $\delta_-(x) \to \delta_-(x_b)$, where $0 < \delta_-(x_b) < \infty$. We then obtain

$$\begin{aligned} 1 - \eta^+(x) &= c(x_b)|\delta_+(x)| + o(|\delta_+(x)|) \\ (I - R(x)^+)^{-1} &= \frac{v^+(x_b)u^+(x_b)}{1 - \eta^+(x)} + o((1 - \eta^+(x))^{-1}) \\ &= \frac{v^+(x_b)u^+(x_b)}{c(x_b)|\delta_+(x)|} + o(|\delta_+(x)|^{-1}) \end{aligned} \tag{12.5}$$

as $x \to x_b$ and $|\delta_+(x)| \to 0$, where c, v^+ and u^+ are continuous functions of x_b on $\mathbb{A}^+$. The asymptotic relations in (12.5) together with (6.19) imply that

$$|\pi_{1,2}(x) - \pi_{1,2}(x_b)| = |\pi_{1,2}(x) - 1| = |-r(x)/(1 + r(x)|, \tag{12.6}$$

where

$$r(x) \equiv \frac{\alpha_0^-(I - R^-)^{-1}\mathbf{1}}{\alpha_0^+(I - R^+)^{-1}\mathbf{1}} \sim h(x_b)|\delta_+(x)| \tag{12.7}$$

as $x \to x_b$ and $|\delta_+(x)| \to 0$, where h is a continuous function on $\mathbb{A}^+$. Hence, there exist constants K_1 and K_2 such that

$$|\pi_{1,2}(x) - \pi_{1,2}(x_b)| \le K_1|\delta_+(x)| \le K_2\|x - x_b\|_3 \tag{12.8}$$

for all x sufficiently close to x_b. Finally, we can apply the triangle inequality with (12.8) to obtain $|\pi_{1,2}(x_1) - \pi_{1,2}(x_2)| \le 2K_2\|x_1 - x_2\|_3$ for x_1, x_2 in an ϵ ball about x_b in $\mathbb{S} - \mathbb{S}^-$. Hence, $\pi_{1,2}(x)$ and thus Ψ are locally Lipschitz continuous on $\mathbb{A}^+$ within $\mathbb{S} - \mathbb{S}^-$. Hence the proof is complete. □

Acknowledgments. This research began while the first author was completing his Ph.D. in the Department of Industrial Engineering and Operations Research at Columbia University and was completed while he held a postdoctoral fellowship at CWI in Amsterdam. This research was partly supported by NSF grants DMI-0457095 and CMMI 0948190.

REFERENCES

[1] Abate, J., Choudhury, G. L., Whitt, W. (1994). Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models* **10** 99–143. MR1259856

[2] Bassamboo, A., Zeevi, A. (2009). On a data-driven method for staffing large call centers. *Oper. Res.* **57**, 714–726.

[3] Choudhury, G. L., Whitt, W. (1994). Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 queue. *Stochastic Models* **10** 453–498. MR1268561

[4] Coddington, E. A., Levinson, N. (1955). *Theory of Ordinary Differential Equations*, McGraw-Hill, New York. MR0069338

[5] Coffman, E. G., Puhalskii, A. A., Reiman, M. I. (1995). Polling systems with zero switchover times: a heavy-traffic averaging principle. *Annals of Applied Probability* **5** 681–719. MR1359825

[6] Ethier, S. N., Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York. MR0838085

[7] Gurvich, I., Whitt, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* **34** (2) 363–396. MR2554064

[8] He, Q. (1995). Differentiability of the matrices R and G in the matrix analytic method. *Stochastic Models* **11** (1) 123–132. MR1316771

[9] Hunt, P. J., Kurtz, T. G. (1994). Large loss networks. *Stochastic Processes and their Applications* **53** 363–378. MR1302919

[10] Khalil, H. K. (2002). *Nonlinear Systems.* Prentice Hall, New Jersey.

[11] Latouche, G., Ramaswami, V. (1999). *Introduction to Matrix Analystic Methods in Stochastic Modeling,* Siam and ASA, Philadelphia MR1674122

[12] Marquez, H. J. (2003). *Nonlinear Control Systems.* Wiley, New York.

[13] Neuts, M. F. (1986). The caudal characteristic curve of queues. *Adv. Appl. Prob.* **18** 221–254. MR0827337

[14] Neuts, M. F. (1989). *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, Marcel Dekker, New York. MR1010040

[15] Pang, G., Talreja, R., Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4**, 193–267. MR2368951

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

[17] Perry, O., Whitt, W. (2010a). A fluid limit for an overloaded X model via an averaging principle. working paper, Columbia University, NY. Available at: http://www.columbia.edu/~ww2040/allpapers.html

[18] Perry, O., Whitt, W. (2010b). Gaussian approximations for an overloaded X model via an averaging principle. working paper, Columbia University, NY.Available at: http://www.columbia.edu/~ww2040/allpapers.html

[19] Perry, O., Whitt, W. (2011). A fluid approximation for service systems responding to unexpected overloads. *Operations Res.*, forthcoming Available at: http://www.columbia.edu/~ww2040/allpapers.html

[20] Stolyar, A. L., Tezcan, T. (2010). Control of systems with flexible multi-server pools: a shadow routing approach. *Queueing Systems* **66**, 1–51. MR2674107

[21] Teschl, G. (2009). *Ordinary Differential Equations and Dynamical Systems*, Universität Wien. Available online: www.mat.univie.ac.at/~gerald/ftp/book-ode/ode.pdf

[22] Whitt, W. (2002). *Stochastic-Process Limits*, New York, Springer MR1876437

[23] 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