

# Transient and Stability Analysis of the Many-Server Heavy-Traffic Fluid Limit for the Overloaded X Call-Center Model

Ohad Perry

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027  
email: op2105@columbia.edu <http://www.columbia.edu/~op2105/>

Ward Whitt

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

We study an ordinary differential equation (ODE) arising as the many-server heavy-traffic fluid limit of an overloaded Markovian queueing model with two customer classes and two service pools, known as the X model in the call-center literature. The system operates under the fixed-queue-ratio-with-thresholds (FQR-T) control, which we proposed in recent papers 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 with FQR-T during an overload. The FQR-T 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., we have a heavy-traffic averaging principle (AP). The AP makes standard ODE theory difficult to apply. Nevertheless, we provide results about the existence and uniqueness of solutions to the ODE, prove that there exists a unique stationary point; and give easily verifiable conditions for the fluid process to converge to its stationary point. Moreover, we show that the convergence to stationarity is exponentially fast. Finally, we provide a numerical algorithm, based on the matrix-geometric method, for solving the ODE.

*Key words:* many-server queues; averaging principle; state-space collapse; heavy-traffic fluid limit; fluid stationarity; ordinary differential equations; nonlinear control systems; Lyapunov stability; matrix-geometric method

*MSC2000 Subject Classification:* Primary: 60K25, ; Secondary: 60K30, 60F17, 90B15, 37C75, 93D05

*OR/MS subject classification:* Primary: Queues ; Secondary: Limit Theorems, Transient Results, Algorithms

---

**1. Introduction.** In this paper we study an *ordinary differential equations* (ODE) that arises as the *many-server heavy-traffic* (MS-HT) fluid limit of an overloaded Markovian X service-system model under the *fixed-queue-ratio-with-thresholds* (FQR-T) control. Specifically, we consider the fluid-limit approximation of a system comprised 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. We assume that the time of the change and the values of the new arrival rates are not known when the overload 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.

This paper is the third in a series of four papers. First, in [9] we initiated study of this overload-control problem and proposed the FQR-T control. We used a heuristic stationary fluid approximation to derive the optimal control when a convex holding cost is charged to the two queues. Within that framework, we showed 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 [10] we applied a heavy-traffic *averaging principle* (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 occurred. 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.

In [10] we then showed that the stationary point of this ODE coincides with the stationary point found by flow balance in [9]. We also developed diffusion-process refinements to the fluid model. However, we gave no proofs of convergence. Instead, we used simulation experiments to show that the fluid-model predictions for the means, and the diffusion-process predictions for the steady-state distribution of the stochastic process, are remarkably accurate, even for relatively small systems (e.g., with 25 agents in each pool).

In this third paper we establish mathematical properties of the ODE introduced in [10]. 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 be computed at each instant. Nevertheless, we provide results about the existence and uniqueness of solutions to the ODE, prove that there exists a unique stationary point; and give easily verifiable conditions for the fluid process to converge to its stationary point. Moreover, we show that the convergence to stationarity is exponentially fast. Finally, we provide a numerical algorithm, based on the matrix-geometric method, for solving the ODE.

Finally, the fourth paper [11] establishes MS-HT limit theorems providing important mathematical support for the ODE approximation introduced in [10] and analyzed in detail here. Both fluid limits, corresponding to the functional weak law of large numbers (FWLLN), and diffusion refinements, corresponding to the the functional central limit theorem (FCLT), are established in [11].

The last two papers may seem out of order, because we establish properties of the limit before we prove the convergence to that limit, but the order is appropriate, because the properties we establish here actually play a key role in the proof of the limit theorems in [11]. Convergence to the ODE is established in [11] by the standard two-step procedure, described in Ethier and Kurtz [2]: (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. Characterizing the limit process depends on the results here. The current paper also establishes important properties of the limiting ODE, useful for direct applications of the fluid-model approximation.

The main difficulty in the characterization of the fluid limit is that, in the sequence of spatially-scaled X systems, the control-driving process - the queue-difference process in (6) - is not being scaled. Hence, it does not converge to a deterministic quantity due to the spatial scaling. However, as indicated above, the driving process operates in a different time scale than the fluid-scaled processes, asymptotically achieving a (time-dependent) steady state at each instant of time, yielding the AP.

In addition to complicating the convergence proof, the AP also complicates the analysis of the limiting ODE. First, it requires that the steady state of a *continuous-time Markov chain* (CTMC), whose distribution depends on the solution to the ODE, be computed at every instant of time. This may seem like a cyclical argument, since the solution to the ODE depends on the steady-state distribution of a CTMC, which in turn depends on the solution to the ODE. However, the separation of time scales in the fluid limit makes this problem well posed.

The second complication is that the AP produces a singularity region in the state space, causing the ODE to be discontinuous in its full state space. Hence, classical results for the existence, uniqueness and stability of solutions to the ODE cannot be applied directly. Nevertheless, in this paper we establish the key properties of the ODE outlined above.

**Here is how the rest of this paper is organized:** The next two sections provide background. In §2 we elaborate on the X queueing model, the fluid approximation (in steady state) under the overload, and the FQR-T control for the original X queueing model, which primarily is a review of [9]. In §3 we review the MS-HT scaling and the fluid ODE, arising as the limit of scaled X model systems operating under the FQR-T control, which primarily is elaborating on [10]. This ODE is what we study in subsequent sections. See [9, 10] for a general literature review.

In §4 we establish properties of the fast-time-scale CTMC, which depends on the state of the ODE, and whose steady-state distribution influences the evolution of the ODE. In §5 we define the state space of the ODE, and show that the ODE possesses a unique solution on an interval. We also give conditions

for the existence of a unique global solution. In §6 we rigorously define fluid stationarity, establish the existence of a unique stationary point and give conditions assuring that the fluid solution converges to the stationary point. In §7 we provide conditions for state-space collapse (SSC) and prove that a converging solution converges to stationarity exponentially fast. In §8 we analyze the system with an initial underloaded state. We prove that in that case the approximating fluid models lead to our main ODE in a finite time. In §9 we develop an algorithm to numerically solve the ODE (given an initial condition), based on the theory developed in the previous sections. Finally, in §10 we draw conclusions and mention remaining open problems.

In addition, in the appendix A we elaborate some more on the algorithm, and give another numerical example, showing that the ODE can be solved, even in cases where we could not prove the existence of a global unique solution. The MS-HT limit will be established in [11].

**2. Preliminaries.** This section reviews the highlights of [9], 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  $\lambda_1$  and  $\lambda_2$ . 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  $\theta_i$ ,  $i = 1, 2$ . 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 which we denote by 0, 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, at time 0, the arrival processes change from Poisson with rates  $\lambda_1$  and  $\lambda_2$  to Poisson processes with *unknown* (but fixed) rates  $\tilde{\lambda}_1$  and  $\tilde{\lambda}_2$ , where  $\tilde{\lambda}_i < m_i/\mu_{i,i}$ ,  $i = 1, 2$  (normal loading), but  $\tilde{\lambda}_i > m_i/\mu_{i,i}$  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  $(\lambda_1, \lambda_2)$ .

The two service systems may be designed to operate independently for various reasons. In [9, 10] we considered the common case in which there is no efficiency gain from service by cross-trained agents. Specifically, in [9] we assumed the *strong inefficient sharing condition*

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

Under condition (1), customers are served at a faster rate when served in their own service pool than when they are being served in the other-class pool. However, many results in [9] hold under the weaker *basic inefficient sharing condition*:  $\mu_{1,1}\mu_{2,2} \geq \mu_{1,2}\mu_{2,1}$ .

If (1) holds then it is disadvantageous (from the standard quality-of-service perspective) for customers to be served in the other-class pool, since their service tends to be longer. Indeed, it is shown in [10] that there can be serious performance degradation, even in normal loading, if both pools are allowed to serve the other class. Without customer abandonment, the sharing can cause the system to become unstable, causing the queue lengths to diverge to infinity.

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

face of the overload, including developing an effective control.

It is easy to see that some sharing can be beneficial if one system is overloaded, while the other is underloaded (has some slack), but sharing may not be desirable if both systems are overloaded. In order to motivate the need for sharing when both systems are overloaded, in [9] we considered a convex-cost framework. With that framework, in [9] we showed that sharing may be beneficial, even if it causes the total queue length (queue 1 plus queue 2) to increase.

Let  $Q_i(t)$  be the number of customers in the class- $i$  queue at time  $t$ , and let  $Z_{i,j}(t)$  be the number of class- $i$  customers being served in pool  $j$  at time  $t$ ,  $i, j = 1, 2$ . Given a stationary routing policy, the stochastic process  $\{(Q_i(t), Z_{i,j}(t) : i, j = 1, 2) : t \geq 0\}$  becomes a six-dimensional CTMC. In principle, the optimal control could be found from the theory of Markov decision processes, but that approach seems prohibitively difficult. For a complete analysis, we would need to consider the unknown transient interval over which the overload occurs, and the random initial conditions, depending on the model parameters under normal loading. In summary, there is a genuine need for the simplifying approximation we review next.

**2.2 The Approximating Deterministic Fluid Model in Steady State.** Given the model above, we want to determine an effective control and analyze its performance. In order to (approximately) minimize the expected cost over the overload incident, in [9] we exploited two characteristics of many-server systems: First, an overloaded many-server service system can be well approximated by a fluid model, which is deterministic and relatively easy to analyze; e.g., see [14]. Second, as demonstrated in [9], many-server systems approach steady state relatively quickly. (In this paper, we provide additional mathematical support by showing that the fluid model converges to stationarity exponentially fast.) These properties support restricting attention to steady-state analysis of the fluid model during the overload incident.

A main conclusion of [9] is that for the fluid model in steady state (in overload), *it is possible to minimize the steady-state cost by choosing appropriate queue-ratio functions*, which can be calculated in advance. (The queue-ratio functions can be functions of the arrival rates or of the queue lengths without sharing.) Moreover, as we explain below, it often suffices to use fixed queue ratios (FQR), with one ratio for each direction of sharing. In addition, under the basic inefficient sharing condition  $\mu_{1,1}\mu_{2,2} \geq \mu_{1,2}\mu_{2,1}$ , it is never optimal to simultaneously share in both directions. That property justifies the additional requirement that *at most one service pool is allowed to serve customers from both classes at any time*. In practice, this additional restriction helps prevent unwanted sharing under normal loads. It directly prevents simultaneous sharing in both directions, which clearly is detrimental under condition (1).

Thus, we are lead to consider the deterministic fluid model. Specifically, we approximate the stochastic processes  $Q_i(t)$  and  $Z_{i,j}(t)$  by deterministic and differentiable (thus, continuous) functions, which we call “fluid”. Let  $q_i(t)$  and  $z_{i,j}(t)$ ,  $i, j = 1, 2$ , be the deterministic fluid approximations of  $Q_i(t)$  and  $Z_{i,j}(t)$ , respectively. Then  $(q_i(t), z_{i,j}(t); i, j = 1, 2, t \geq 0)$  is called the “fluid model” (or the “fluid approximation”) of the stochastic system. Let  $q_i^*$  and  $z_{i,j}^*$  be the limits of the fluid functions as  $t \rightarrow \infty$ , assuming these limits exist. Then the vector  $(q_i^*, z_{i,j}^*; i, j = 1, 2) \in \mathbb{R}^6$  is called the steady-state of the fluid model, or alternatively, the stationary point of the fluid model (see §6 for a formal definition).

As indicated above, we assume that queue 1 is overloaded and is receiving help from pool 2, so that  $z_{1,2}^* > 0$ . As mentioned before, this implies that  $z_{2,1}^* = 0$  and  $z_{1,1}^* = m_1$ . If we further assume that pool 2 is overloaded after sharing, we have that  $z_{2,2}^* = m_2 - z_{1,2}^*$ . That is the main case we want to consider. Hence, we need only consider the three-dimensional steady-state vector  $x^* = (q_1^*, q_2^*, z_{1,2}^*)$ . Now, for  $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$  to remain fixed for all  $t$ , the flow into the system must be equal to the flow out of the system. Hence, in steady state, there are  $m_1$  agents processing class-1 fluid in pool 1 at rate  $\mu_{1,1}$ , plus  $z_{1,2}^*$  agents in pool 2, processing at rate  $\mu_{1,2}$ . In addition to the class-1 fluid leaving the system due to the service process, there is also fluid leaving the system due to the abandonment process, with rate  $\theta_1 q_1^*$  in steady state. Similarly, class-2 fluid is served by the remaining  $m_2 - z_{1,2}^*$  servers in pool 2, which process at rate  $\mu_{2,2}$ . All the class-2 arrivals which are not served, abandon at rate  $\theta_2 q_2^*$ . Equating the input to each queue (which is just the arrival rate to this queue) to the output from each queue, we see that

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

from which we get the expressions for the stationary queue lengths

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

This steady-state fluid framework greatly simplifies the control problem, because in the setting above there is only the single decision variable  $z_{1,2}^*$ . The equations in (2) can be used to find the optimal  $z_{1,2}^*$  by solving the simple optimization problem of minimizing the convex-cost function  $C(q_1^*, q_2^*)$  over the constraint  $0 \leq z_{1,2}^* \leq m_2$ .

It follows immediately from (2) that  $q_1^*$  is decreasing with  $z_{1,2}^*$ , while  $q_2^*$  is increasing with  $z_{1,2}^*$ . Consequently, for given arrival rates  $\lambda_1$  and  $\lambda_2$ , the optimal  $z_{1,2}^*$  determines a unique ratio between the steady-state fluid queues,  $r_{1,2}^*(q_1^*, q_2^*) \equiv r_{1,2}^*(\lambda_1, \lambda_2) \equiv q_1^*/q_2^*$ . (Similar analysis holds for  $r_{2,1}^*(\lambda_1, \lambda_2)$  which is used when class 2 is being helped by pool 1.) In general, the optimal ratios are different for different arrival rates. An efficient algorithm to find the optimal ratio-function was developed in [9]; see Proposition 4 and §5.3 there.

However, as explained in [9], the optimal ratios often tend to be approximately the same for all possible overloads;  $r_{i,j}^*(\lambda_1, \lambda_2) \approx r_{i,j}$ , so that it is usually enough to consider only one fixed queue ratio for each direction of sharing. This conclusion is supported mathematically when we impose additional conditions on the convex cost function. Since the actual cost function is difficult to specify, it is natural to consider simple parametric special cases. In particular, it is natural to assume that the holding cost is a separable quadratic function, i.e., of the form  $C(q_1, q_2) = C_1(q_1) + C_2(q_2)$ , with  $C_i(q_i) = c_i q_i^2 + b_i q_i + a_i$ ,  $i = 1, 2$ . In that case, the optimal queue-ratio function has a relatively simple explicit form, in particular, we can translate each of the state-dependent queue ratios to a fixed ratio shifted by a constant. More specifically, the optimal relation that should hold between the two queues is  $q_1^* - r_{i,j}^* q_2^* = \kappa_{i,j}$ ,  $i, j = 1, 2$ , where  $\kappa_{i,j}$  and  $r_{i,j}^*$  are fixed constants for all possible overloads. If, in addition,  $b_i = a_i = 0$  so that  $C_i(q_i) = c_i q_i^2$ , then  $\kappa_{i,j} = 0$ , and the optimal relation between the queues should be a fixed queue ratio, i.e.,  $r_{i,j}^*(\lambda_1, \lambda_2) \equiv r_{i,j}^*$ . Thus there is a theoretical basis for using FQR once sharing has been activated. However, we also consider shifted FQR, which is the optimal control for all separable quadratic cost functions.

**2.3 The FQR-T Control for the Original Queueing Model.** Having found the optimal steady-state fluid levels for the fluid model, we are ready to construct the corresponding control for the stochastic model. For the original stochastic model, there are additional issues. First, even in large systems, it is hard to distinguish between genuine overloads, caused by higher-than-expected arrival rates, and momentarily heavy loads which are due to stochastic fluctuations. The difficulty is primarily due to the abandonments, which keep the system stable whatever the arrival rates are. Moreover, even if an overload has been established, it is hard to determine the values of the new arrival rates, and thus the optimal sharing policy. The purpose of FQR-T is to automatically detect overloads, immediately when they occur, and maintain the optimal ratio between the two queues when the system is overloaded.

The FQR-T control is based on two positive thresholds,  $k_{1,2}$  and  $k_{2,1}$ , and the two queue-ratio parameters,  $r_{1,2}$  and  $r_{2,1}$ . We define two queue-difference stochastic processes  $\tilde{D}_{1,2}(t) \equiv Q_1(t) - r_{1,2}Q_2(t)$  and  $\tilde{D}_{2,1} \equiv r_{2,1}Q_2(t) - Q_1(t)$ .

As long as  $\tilde{D}_{1,2}(t) < k_{1,2}$  and  $\tilde{D}_{2,1}(t) < k_{2,1}$  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  $\tilde{D}_{1,2}(t) \geq k_{1,2}$ , 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  $\tilde{D}_{1,2}(t) \geq 0$ . 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.

With the assumptions on the X system and the FQR-T control, the six-dimensional stochastic process  $(Q_i(t), Z_{i,j}(t); i, j = 1, 2)$  is a CTMC. Once sharing is initialized, the control keeps the two queues at

approximately the target ratio, e.g., if queue 1 is being helped, then  $Q_1(t) \approx r_{1,2}Q_2(t)$ . If sharing is done in the opposite direction, then  $r_{2,1}Q_2(t) \approx Q_1(t)$  for all  $t \geq 0$ . That is substantiated by simulation experiments, some of which are reported in [9, 10].

It is clear that the thresholds should be small enough so that overloads are detected quickly. On the other hand, the thresholds should be large enough to prevent unwanted sharing when the system is not overloaded. The way to choose the thresholds is discussed at length in §2 of [10]. The main idea is to exploit MS-HT scaling; see §3.1.

As indicated above, in general (if the convex cost function is not separable and quadratic) the two optimal ratios depend on the arrival rates to the system, which are assumed to be unknown. In that case we can use the *queue-ratio-with-thresholds control* (QR-T), proposed in [9], which uses the state-dependent queue ratios at each decision epoch. However, even if QR-T is used, then after a short period of time the system should stabilize at a fixed ratio  $r_{i,j}^*$ , which is optimal for the specific (unknown) arrival rates; i.e., QR-T will automatically “discover” the optimal ratio. Once the queue-ratio stabilizes at a fixed ratio, the control is the same as FQR-T.

If the optimal relation between the queues is  $q_1^* = r_{1,2}^*q_2^* + \kappa_{i,j}$  for some  $\kappa_{i,j} \in \mathbb{R}$ , as is the case when the holding cost is separable and quadratic with non-zero constant and linear terms, then we use the *shifted FQR-T control*. Shifted FQR-T centers about  $\kappa_{i,j}$  instead at about zero. For example, if class 1 is overloaded, then every server takes his new customer from the head of queue 1 if  $\tilde{D}_{i,j}(t) \geq \kappa_{1,2}$ . Otherwise, it takes the new customer from the head of its own class queue. We call that control *shifted FQR-T* since it keeps the two queues at a fixed ratio, but shifted by the constant  $\kappa_{i,j}$ . We can think of FQR-T as the special case of shifted FQR-T with  $\kappa_{i,j} = 0$ .

Our analysis so far relies on the assumption that FQR-T and shifted FQR-T achieve their purpose, i.e., that they keep the the two queues approximately in fixed relation. In the stochastic system this means that the two-dimensional vector  $(Q_1(t), Q_2(t))$  should tend to evolve approximately as a one-dimensional process. In the fluid model this approximation becomes exact; We no longer need to consider the three-dimensional process  $x(t) \equiv (q_1(t), q_2(t), z_{1,2}(t))$ , since it is enough to consider  $z_{1,2}(t)$  together with only one of the queues. The other queue is determined by the first via the *state-space collapse* (SSC) equation  $q_1(t) = r_{i,j}q_2(t) + \kappa_{i,j}$ , depending on which way the sharing is performed. In [10] SSC is substantiated via simulation; in [11] it will be shown to hold asymptotically in the MS-HT limit, which we describe in the next section.

In [10] we suggest the fluid approximation  $\{x(t) : t \geq 0\}$ , which is characterized by a three-dimensional ODE involving the AP. In order to develop this approximation, we considered the fluid as a limit of a properly scaled sequence of stochastic X models operating under (shifted) FQR-T. We then argued that the transient fluid model has a stationary point, which agrees with the optimal solution derived heuristically before. However, none of the claims were proved, and were only verified using simulation experiments.

Unlike the steady-state fluid approximation, there appears to be no simple heuristic derivation of the transient ODE without considering the original stochastic system. To see why, assume that FQR-T is employed with a ratio  $r_{1,2}$ . If FQR-T indeed keeps the ratio between the two queues fixed, then  $q_1(t) = r_{1,2}q_2(t)$  for each  $t$ . But then  $\tilde{D}_{1,2}(t) \equiv q_1(t) - r_{1,2}q_2(t) = 0$  for each  $t$ , which implies that every newly available server takes his next customer from the head of queue 1 *at any time t*. Obviously, this heuristic approximation is meaningless. Hence, a more careful treatment of the difference-processes  $\tilde{D}_{i,j}$  is needed; we somehow need to capture the fact that, in the fluid model, fluid is flowing from queue 1 to both service pools at every time  $t$ . To do that, we evidently must consider the fluid model as a limit of stochastic X models.

**3. The Many-Server Heavy-Traffic Fluid Limit.** In this section we describe the convergence of the sequence of stochastic systems to the fluid limit, as was conjectured in [10] and will be established in [11]. Without loss of generality *we assume that class 1 is overloaded, and 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.)

**3.1 Many-Server Heavy-Traffic (MS-HT) Scaling.** To develop the fluid limit, 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.,

$$\bar{\lambda}_i^n \equiv \frac{\lambda_i^n}{n} \rightarrow \lambda_i \quad \text{and} \quad \bar{m}_i^n \equiv \frac{m_i^n}{n} \rightarrow m_i \quad \text{as } n \rightarrow \infty, \quad (3)$$

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

$$\bar{Q}_i^n(t) \equiv \frac{Q_i^n(t)}{n} \quad \text{and} \quad \bar{Z}_{i,j}^n(t) \equiv \frac{Z_{i,j}^n(t)}{n}, \quad i, j = 1, 2, \quad t \geq 0. \quad (4)$$

For each system  $n$ , there are threshold  $k_{1,2}^n$  and  $k_{2,1}^n$ , scaled as suggested in [9, 10]:

$$\frac{k_{i,j}^n}{n} \rightarrow 0 \quad \text{and} \quad \frac{k_{i,j}^n}{\sqrt{n}} \rightarrow \infty \quad \text{as } n \rightarrow \infty, \quad i, j = 1, 2. \quad (5)$$

The first scaling by  $n$  is chosen to make the thresholds asymptotically negligible in MS-HT fluid scaling, so they have no asymptotic impact on the steady-state cost. The second scaling by  $\sqrt{n}$  is chosen to make the thresholds asymptotically infinite in MS-HT diffusion scaling, so that asymptotically the thresholds will not be exceeded under normal loading. It is significant that MS-HT scaling shows that we should be able to simultaneously satisfy both conflicting objectives in large systems. There are also the shifting thresholds  $\kappa_{i,j}^n$ , arising from consideration of separable quadratic cost functions; see §2.3, but we do not specify their scale.

We let time zero be the time at which  $Q_1^n(0) = k_{1,2}^n$ , and sharing is 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  $\tilde{D}_{1,2}^n(t) \equiv Q_1^n(t) - r_{1,2}^* Q_2^n(t)$ . However, if  $\kappa_{1,2}^n \rightarrow \infty$ , then  $\tilde{D}_{1,2}^n \rightarrow \infty$  as  $n \rightarrow \infty$ . Hence, we redefine the difference process. Let

$$D^n(t) \equiv (Q_1^n(t) - \kappa^n) - r Q_2^n(t), \quad t \geq 0, \quad (6)$$

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

With this definition, we allow  $\kappa^n$  to be of any order less than or equal to  $O(n)$ ; in particular, we assume that  $\kappa^n/n \rightarrow \kappa$  for  $0 \leq \kappa < \infty$ . There are two principle cases:  $\kappa = 0$  and  $\kappa > 0$ . The first case produces FQR; the second case produces shifted FQR. (Since the overload has already begun, the original thresholds  $k_{i,j}^n$  no longer play a role.)

We now apply FQR using the process  $D^n$  in (6): if  $D^n(t) \geq 0$ , then every newly available agent (in either pool) takes his new customer from the head of the class-1 queue. If  $D^n(t) < 0$ , then every newly available agent takes his new customer from the head of his own queue.

**3.2 Representation.** In order to understand why the ODE takes the form it does, it is helpful to see the representation used in the first step in establishing the MS-HT limit. Following common practice, as reviewed in §2 of [8], we represent all the processes of interest in terms of mutually independent random-time-changed rate-1 Poisson processes: Let  $N_i^a$ ,  $N_{i,2}^s$  and  $N_i^u$  for  $i = 1, 2$  be six mutually independent rate-1 Poisson processes.

For simplicity, we restrict attention to the main case, which can be shown to be asymptotically equivalent to the actual system: We assume that all agents are busy all the time and no class-2 customers are being served at service-pool 1. Thus, we have  $Z_{2,1}^n(t) = 0$ ,  $Z_{1,1}^n(t) = m_1^n$  and  $Z_{2,2}^n(t) = m_2^n - Z_{1,2}^n(t)$ , for all  $t \geq 0$ , so that we need only consider  $Z_{1,2}^n$ .

We thus obtain the following representation for the three processes  $Q_1^n$ ,  $Q_2^n$  and  $Z_{1,2}^n$  in terms of the

queue-difference process  $D^n$  in (6):

$$\begin{aligned}
Z_{1,2}^n(t) &\equiv Z_{1,2}^n(0) + N_{2,2}^s \left( \mu_{2,2} \int_0^t 1_{\{D^n(s) \geq 0\}} (m_2^n - Z_{1,2}^n(s)) ds \right) \\
&\quad - N_{1,2}^s \left( \mu_{1,2} \int_0^t 1_{\{D^n(s) < 0\}} Z_{1,2}^n(s) ds \right), \quad t \geq 0. \\
Q_1^n(t) &\equiv Q_1^n(0) + N_{1,1}^a(\lambda_1^n t) - N_{1,1}^s(m_1^n \mu_{1,1} t) - N_{1,2}^s \left( \mu_{1,2} \int_0^t 1_{\{D^n(s) \geq 0\}} Z_{1,2}^n(s) ds \right) \\
&\quad - N_{2,2}^s \left( \mu_{2,2} \int_0^t 1_{\{D^n(s) \geq 0\}} (m_2^n - Z_{1,2}^n(s)) ds \right) - N_1^u \left( \theta_1 \int_0^t Q_1^n(s) ds \right), \quad t \geq 0. \\
Q_2^n(t) &\equiv Q_2^n(0) + N_{2,2}^a(\lambda_2^n t) - N_{2,2}^s \left( \mu_{2,2} \int_0^t 1_{\{D^n(s) < 0\}} (m_2^n - Z_{1,2}^n(s)) ds \right) \\
&\quad - N_{1,2}^s \left( \mu_{1,2} \int_0^t 1_{\{D^n(s) < 0\}} Z_{1,2}^n(s) ds \right) - N_2^u \left( \theta_2 \int_0^t Q_2^n(s) ds \right), \quad t \geq 0.
\end{aligned} \tag{7}$$

We then construct the usual martingale processes by subtracting the stochastic intensities, letting  $M_i^{n,a}(t) \equiv N_i^a(\lambda_i^n t) - \lambda_i^n t$ ,  $M_i^{n,u} \equiv N_i^u \left( \theta_i \int_0^t Q_i^n(s) ds \right) - \theta_i \int_0^t Q_i^n(s) ds$  and  $M_{i,2}^{n,s}(t) \equiv N_{i,2}^s(I_{i,2}^n(t)) - I_{i,2}^n(t)$ , where  $I_{i,2}^n(t)$  is the stochastic intensity used with the Poisson-process  $N_{i,2}^s(t)$ , e.g.,  $I_{1,2}^n(t) \equiv \mu_{1,2} \int_0^t 1_{\{D^n(s) < 0\}} Z_{1,2}^n(s) ds$ .

The fluid limit is a FWLLN. To express it, let  $D$  be the usual function space of right-continuous functions on the interval  $[0, \infty)$  with left limits in  $(0, \infty)$ , endowed with the usual topology and let  $\Rightarrow$  denote convergence in distribution; see [2, 13].

We next rewrite the equations in (7) by subtracting and adding the stochastic intensities, and then dividing each equation by  $n$ . It can be shown that  $M_i^{n,a}/n \Rightarrow 0$ ,  $M_i^{n,u}/n \Rightarrow 0$  and  $M_{i,2}^{n,s}/n \Rightarrow 0$  in  $D$  as  $n \rightarrow \infty$ , (where 0 here stands for the zero function). Hence, we replace these processes by an  $o_p(1)$  term, where a sequence  $\{Y^n : n \geq 1\}$  of processes in  $D$  satisfies  $Y^n = o_p(1)$  if  $Y^n \Rightarrow 0$  in  $D$  as  $n \rightarrow \infty$ . We thus have the associated representation for the fluid-scaled queueing processes:

$$\begin{aligned}
\bar{Z}_{1,2}^n &\equiv \bar{Z}_{1,2}^n(0) + \mu_{2,2} \int_0^t 1_{\{D^n(s) \geq 0\}} (\bar{m}_2^n - \bar{Z}_{1,2}^n(s)) ds - \mu_{1,2} \int_0^t 1_{\{D^n(s) < 0\}} \bar{Z}_{1,2}^n(s) ds + o_p(1), \quad t \geq 0, \\
\bar{Q}_1^n(t) &\equiv \bar{Q}_1^n(0) + \bar{\lambda}_1^n t - \bar{m}_1^n t - \mu_{1,2} \int_0^t 1_{\{D^n(s) \geq 0\}} \bar{Z}_{1,2}^n(s) ds - \mu_{2,2} \int_0^t 1_{\{D^n(s) \geq 0\}} (\bar{m}_2^n - \bar{Z}_{1,2}^n(s)) ds \\
&\quad - \theta_1 \int_0^t \bar{Q}_1^n(s) ds + o_p(1), \quad t \geq 0, \\
\bar{Q}_2^n(t) &\equiv \bar{Q}_2^n(0) + \bar{\lambda}_2^n t - \mu_{2,2} \int_0^t 1_{\{D^n(s) < 0\}} (\bar{m}_2^n - \bar{Z}_{1,2}^n(s)) ds - \mu_{1,2} \int_0^t 1_{\{D^n(s) < 0\}} \bar{Z}_{1,2}^n(s) ds \\
&\quad - \theta_2 \int_0^t \bar{Q}_2^n(s) ds + o_p(1), \quad t \geq 0.
\end{aligned} \tag{8}$$

The ODE we study is an approximation for the three-dimensional fluid-scaled process  $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$  with components defined in (8).

**3.3 A Heuristic View of the AP.** In fact, the ODE we study is the limit of the three-dimensional fluid-scaled process  $\bar{X}^n \equiv (\bar{Q}_1^n, \bar{Q}_2^n, \bar{Z}_{1,2}^n)$  with components defined in (8); i.e., in [11] we show that  $\bar{X}^n \Rightarrow x$  in  $D^3$  as  $n \rightarrow \infty$ , where  $x \equiv (q_1, q_2, z_{1,2})$  is a deterministic limit satisfying the ODE. The resulting ODE can be seen directly from the differential form of the integral representation in (8), provided that we invoke the AP discussed below. As a result of the AP, the indicator functions  $1_{\{D^n(s) \geq 0\}}$  and  $1_{\{D^n(s) < 0\}}$ , appearing in the integrands, are replaced by deterministic functions, denoted by  $\pi_{1,2}(x(s))$  and  $1 - \pi_{1,2}(x(s))$ , respectively (in addition to replacing  $\bar{X}^n$  by  $x$ ).

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. In that situation, it can be shown that the queue-difference process  $D^n$  in (6) 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 (4) and (8). A heuristic explanation is that, with the MS-HT scaling in (3), in order for the two queues to change significantly (in a relative sense), which is captured by the scaling in (4), 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. It transitions above and below 0 of order  $O(n)$  times in any finite interval. Thus, over short time intervals in which  $X^n$  remains nearly unchanged (for large  $n$ ), the process  $D^n$  moves frequently in its state space, nearly achieving a local steady state rapidly with respect to  $\bar{X}^n$ . 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-incremented* difference processes: for each  $n \geq 1$  and  $t \geq 0$ , let

$$D_t^n(s) \equiv D^n(t + s/n) - D^n(t), \quad s \geq 0. \quad (9)$$

Dividing  $s$  by  $n$  in (9) 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. A main result in [11] is that, for each  $t \geq 0$ ,

$$D_t^n \equiv \{D_t^n(s) : s \geq 0\} \Rightarrow D_t \equiv \{D_t(s) : s \geq 0\} \quad \text{as } n \rightarrow \infty \quad \text{in } D, \quad (10)$$

where the limit  $D_t \equiv \{D_t(s) : s \geq 0\}$  is a *pure-jump continuous-time Markov process* with state space  $\{k + rj : k \in \mathbb{Z}, j \in \mathbb{Z}\}$ . We call  $D_t$  the *fast-time-scale-process* (FTSP). This limit is easy to understand by examining the transition rates of the process  $D_t^n$  defined in (9), which depend on the CTMC  $\bar{X}^n(t)$ .

The deterministic function  $\pi_{1,2}$ , mentioned in the first paragraph of this section, is the steady-state probability of the set  $[0, \infty)$  for the FTSP, i.e.,

$$\pi_{1,2}(x(t)) \equiv \lim_{s \rightarrow \infty} P(D_t(s) \geq 0) = \lim_{u \rightarrow \infty} \frac{1}{u} \int_0^u 1_{\{D_t(s) \geq 0\}} ds, \quad (11)$$

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

To actually establish convergence for  $\bar{X}^n$  in (8), we go further in [11] and prove local uniform convergence in  $t$ , which implies that for any  $\epsilon > 0$ , there exist  $n_0$  and  $\eta > 0$  such that, for any  $n \geq n_0$ ,

$$\left| \frac{1}{\eta} \int_t^{t+\eta} 1_{\{D_t^n(s) \geq 0\}} ds - \pi_{1,2}(x(t)) \right| < \epsilon. \quad (12)$$

The local uniform convergence allows us to replace the indicator functions in the integrals in (8) with the  $\pi_{1,2}$  functions in the fluid limit.

**3.4 The Fluid-Limit ODE** The discussion in §§3.2 and 3.3 above is an outline of the convergence result to appear in [11]. A different approach appeared §4.2 of [10], where the ODE was developed directly, assuming that the fluid limit exists, and is differentiable. The convergence  $\bar{X}^n \Rightarrow x$  established in [11] based on the representation (8) together with the AP in (9)-(12) lead to the same ODE as in [10]. We now specify the ODE, which is the main subject of this paper.

The general form of an ODE is  $\dot{x}(t) = \Psi(x(t), t)$  for a function  $\Psi$ , where  $\dot{x}(t)$  is the derivative evaluated at  $t$ . In addition, our ODE is *autonomous* (or *time invariant*) because  $\Psi(x(t), t) \equiv \Psi(x(t))$ . An autonomous ODE does not depend explicitly on the time-argument  $t$ , and its behavior is invariant to shifts in the time origin.

We consider the autonomous ODE

$$\dot{x}(t) \equiv (\dot{q}_1(t), \dot{q}_2(t), \dot{z}_{1,2}(t)) = \Psi(x(t)) \equiv \Psi(q_1(t), q_2(t), z_{1,2}(t)), \quad t \geq 0, \quad (13)$$

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

$$\begin{aligned} \dot{q}_1(t) &\equiv \lambda_1 - m_1 \mu_{1,1} - \pi_{1,2}(x(t)) [z_{1,2}(t) \mu_{1,2} + z_{2,2}(t) \mu_{2,2}] - \theta_1 q_1(t) \\ \dot{q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(x(t))) [z_{2,2}(t) \mu_{2,2} + z_{1,2}(t) \mu_{1,2}] - \theta_2 q_2(t) \\ \dot{z}_{1,2}(t) &\equiv \pi_{1,2}(x(t)) z_{2,2}(t) \mu_{2,2} - (1 - \pi_{1,2}(x(t))) z_{1,2}(t) \mu_{1,2}, \end{aligned} \quad (14)$$

with  $\pi_{1,2} : [0, \infty)^2 \times [0, m_2] \rightarrow [0, 1]$  defined in (11).

Some of the results in this paper depend on the initial value of the ODE. In that case, we consider the *initial value problem* (IVP)

$$\dot{x}(t) = \Psi(x(t)), \quad x(0) = w_0 \quad (15)$$

for  $\Psi(x)$  in (13) - (14).

We remark that specifying the IVP in (13)-(15) does not fully characterize the limit of  $\bar{X}^n$ , given convergence of the initial conditions  $\bar{X}^n(0) \rightarrow w_0$  w.p. 1, where  $w_0 \geq 0$  is deterministic, as required in [11]. First, it is not initially evident that a solution to the ODE exists. Second, even if a solution does exist, this solution must be unique as well in order for it to characterize the limit of  $\bar{X}^n$ , because in the proof of convergence the ODE initially appears only as the limit of a converging subsequence. In general, different subsequences may converge to different limits. Thus, our first task here is to prove the existence of a unique solution to the IVP in (15).

The proof of existence and uniqueness of a solution to (15), is tied to the characterization of  $\pi_{1,2}$  in (14) and (11), 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.

**4. The Fast-Time-Scale Process.** Recall that the FTSP  $D_t$  is the limit of  $D_t^n$  without any scaling (see (10)), where  $D_t^n$  is the time-incremented difference process defined in (9) in terms of the queue-difference stochastic process  $D^n \equiv (Q_1^n - \kappa^n) - rQ_2^n$  in (6). 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 (6) that  $D^n$  has state space  $\{\pm j \pm kr - \kappa^n : j, k \in \mathbb{Z}\}$ . Next, because of the subtraction in (9),  $D_t^n$  has state space  $\{\pm j \pm kr : j, k \in \mathbb{Z}\}$ . Finally, because of the convergence in (10), the FTSP  $D_t$  has this same state space.

**4.1 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,  $D_t \equiv \{D_t(s) : s \geq 0\}$  becomes a continuous-time Markov chain (CTMC), which we refer to as the *fast-time-scale Markov chain* (FTSMC).

Let  $\lambda_+^{(j)}(x(t))$ ,  $\lambda_+^{(k)}(x(t))$ ,  $\mu_+^{(j)}(x(t))$  and  $\mu_+^{(k)}(x(t))$  be the transition rates of the FTSMC  $D_t$  for transitions of  $+j$ ,  $+k$ ,  $-j$  and  $-k$ , respectively when  $D_t \geq 0$ . Similarly, we define the transitions when  $D_t < 0$ :  $\lambda_-^{(j)}(x(t))$ ,  $\lambda_-^{(k)}(x(t))$ ,  $\mu_-^{(j)}(x(t))$  and  $\mu_-^{(k)}(x(t))$ . These rates are the limits of the rates of  $D_t^n$  as  $n \rightarrow \infty$  with  $\bar{X}^n(t) \Rightarrow x(t)$ ; convergence will be proved in [11].

First, for  $j \in (-\infty, 0)$ , the upward rates are

$$\lambda_-^{(j)}(x(t)) = \lambda_1, \quad \text{and} \quad \lambda_-^{(k)}(x(t)) = \mu_{1,2}z_{1,2}(t) + \mu_{2,2}z_{2,2}(t) + \theta_2q_2(t), \quad (16)$$

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

$$\mu_-^{(j)}(x(t)) = \mu_{1,1}z_{1,1}(t) + \theta_1q_1(t) \quad \text{and} \quad \mu_-^{(k)}(x(t)) = \lambda_2, \quad (17)$$

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  $j \in [0, \infty)$ , we have upward rates

$$\lambda_+^{(j)}(x(t)) = \lambda_1 \quad \text{and} \quad \lambda_+^{(k)}(x(t)) = \theta_2q_2(t), \quad (18)$$

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

$$\mu_+^{(j)}(x(t)) = \mu_{1,1}z_{1,1}(t) + \mu_{1,2}z_{1,2}(t) + \mu_{2,2}z_{2,2}(t) + \theta_1q_1(t) \quad \text{and} \quad \mu_+^{(k)}(x(t)) = \lambda_2, \quad (19)$$

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.

**4.2 Representing the FTSMC  $D_t$  as a QBD.** Further analysis is simplified by exploiting matrix geometric methods, as in [6]. In particular, we represent the integer-valued FTSMC  $D_t \equiv \{D_t(s) : s \geq 0\}$  just constructed as a homogeneous continuous-time QBD, as in Definition 1.3.1 and §6.4 of [6]. To do so, 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 [6] (imitating the shape of a generator matrix of a birth-and-death process). In particular, we write

$$Q \equiv \begin{pmatrix} B & A_0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (20)$$

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.,

$$B \equiv \begin{pmatrix} A_1^+ & B_\mu \\ B_\lambda & A_1^- \end{pmatrix} \quad \text{and} \quad A_i \equiv \begin{pmatrix} A_i^+ & 0 \\ 0 & A_i^- \end{pmatrix} \quad (21)$$

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

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

$$\phi(2nm + i) \equiv nm + i \quad \text{and} \quad \phi((2n + 1)m + i) \equiv -nm - i + 1, \quad 1 \leq i \leq m. \quad (22)$$

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 (20) 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  $r = 0.4$ , so that we can choose  $j = 2$  and  $k = 5$ , yielding  $m = 5$ . The states are ordered in levels as follows

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

Then the submatrices  $B_\mu, B_\lambda, A_i^+$  and  $A_i^-$ , which form the block matrices  $B$  and  $A_i$ ,  $i = 0, 1, 2$ , have the following form:

$$\begin{aligned}
B_\mu &= \begin{pmatrix} 0 & \mu_+^{(2)} & 0 & 0 & \mu_+^{(5)} \\ \mu_+^{(2)} & 0 & 0 & \mu_+^{(5)} & 0 \\ 0 & 0 & \mu_+^{(5)} & 0 & 0 \\ 0 & \mu_+^{(5)} & 0 & 0 & 0 \\ \mu_+^{(5)} & 0 & 0 & 0 & 0 \end{pmatrix} & B_\lambda &= \begin{pmatrix} 0 & \lambda_-^{(2)} & 0 & 0 & \lambda_-^{(5)} \\ \lambda_-^{(2)} & 0 & 0 & \lambda_-^{(5)} & 0 \\ 0 & 0 & \lambda_-^{(5)} & 0 & 0 \\ 0 & \lambda_-^{(5)} & 0 & 0 & 0 \\ \lambda_-^{(5)} & 0 & 0 & 0 & 0 \end{pmatrix} \\
A_0^+ &= \begin{pmatrix} \lambda_+^{(5)} & 0 & 0 & 0 & 0 \\ 0 & \lambda_+^{(5)} & 0 & 0 & 0 \\ 0 & 0 & \lambda_+^{(5)} & 0 & 0 \\ \lambda_+^{(2)} & 0 & 0 & \lambda_+^{(5)} & 0 \\ 0 & \lambda_+^{(2)} & 0 & 0 & \lambda_+^{(5)} \end{pmatrix} & A_0^- &= \begin{pmatrix} \mu_-^{(5)} & 0 & 0 & 0 & 0 \\ 0 & \mu_-^{(5)} & 0 & 0 & 0 \\ 0 & 0 & \mu_-^{(5)} & 0 & 0 \\ \mu_-^{(2)} & 0 & 0 & \mu_-^{(5)} & 0 \\ 0 & \mu_-^{(2)} & 0 & 0 & \mu_-^{(5)} \end{pmatrix} \\
A_1^+ &= \begin{pmatrix} -\sigma_+ & 0 & \lambda_+^{(2)} & 0 & 0 \\ 0 & -\sigma_+ & 0 & \lambda_+^{(2)} & 0 \\ \mu_+^{(2)} & 0 & -\sigma_+ & 0 & \lambda_+^{(2)} \\ 0 & \mu_+^{(2)} & 0 & -\sigma_+ & 0 \\ 0 & 0 & \mu_+^{(2)} & 0 & -\sigma_+ \end{pmatrix} & A_1^- &= \begin{pmatrix} -\sigma_- & 0 & \mu_-^{(2)} & 0 & 0 \\ 0 & -\sigma_- & 0 & \mu_-^{(2)} & 0 \\ \lambda_-^{(2)} & 0 & -\sigma_- & 0 & \mu_-^{(2)} \\ 0 & \lambda_-^{(2)} & 0 & -\sigma_- & 0 \\ 0 & 0 & \lambda_-^{(2)} & 0 & -\sigma_- \end{pmatrix} \\
A_2^+ &= \begin{pmatrix} \mu_+^{(5)} & 0 & 0 & \mu_+^{(2)} & 0 \\ 0 & \mu_+^{(5)} & 0 & 0 & \mu_+^{(2)} \\ 0 & 0 & \mu_+^{(5)} & 0 & 0 \\ 0 & 0 & 0 & \mu_+^{(5)} & 0 \\ 0 & 0 & 0 & 0 & \mu_+^{(5)} \end{pmatrix} & A_2^- &= \begin{pmatrix} \lambda_-^{(5)} & 0 & 0 & \lambda_-^{(2)} & 0 \\ 0 & \lambda_-^{(5)} & 0 & 0 & \lambda_-^{(2)} \\ 0 & 0 & \lambda_-^{(5)} & 0 & 0 \\ 0 & 0 & 0 & \lambda_-^{(5)} & 0 \\ 0 & 0 & 0 & 0 & \lambda_-^{(5)} \end{pmatrix}
\end{aligned} \tag{23}$$

where

$$\sigma_+ = \lambda_+^{(5)} + \lambda_+^{(2)} + \mu_+^{(5)} + \mu_+^{(2)} \quad \text{and} \quad \sigma_- = \lambda_-^{(5)} + \lambda_-^{(2)} + \mu_-^{(5)} + \mu_-^{(2)}. \tag{24}$$

(We solve a full numerical example with these matrices in §9.3.)

Henceforth, we refer to  $D_t$  as the QBD, because this is the only QBD under consideration. To summarize, both the FTSMC and the QBD are alternative representations of the original FTSP (exploiting the assumption that  $r = j/k$  for positive integers  $j$  and  $k$  without common factor).

**4.3 Positive Recurrence.** We now determine when the FTSP  $D_t$  is positive recurrent, so that the AP holds. For that purpose, we employ the theory in §7 of [6], modified to the continuous-time QBD. To apply the theory, we 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 first 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  $\nu^+$  and  $\nu^-$  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

$$\nu^+ A_0^+ \mathbf{1} < \nu^+ A_2^+ \mathbf{1} \quad \text{and} \quad \nu^- A_0^- \mathbf{1} < \nu^- A_2^- \mathbf{1}. \tag{25}$$

In our application it is easy to see that both  $\nu^+$  and  $\nu^-$  are the uniform probability vector, attaching probability  $1/m$  to each of the  $m$  states.

Let  $\delta_+$  and  $\delta_-$  be the drift in the positive and negative region, respectively; i.e., let

$$\begin{aligned}
\delta_+(x(t)) &\equiv j \left( \lambda_+^{(j)}(x(t)) - \mu_+^{(j)}(x(t)) \right) + k \left( \lambda_+^{(k)}(x(t)) - \mu_+^{(k)}(x(t)) \right) \\
\delta_-(x(t)) &\equiv j \left( \lambda_-^{(j)}(x(t)) - \mu_-^{(j)}(x(t)) \right) + k \left( \lambda_-^{(k)}(x(t)) - \mu_-^{(k)}(x(t)) \right).
\end{aligned} \tag{26}$$

By our construction of the rates above, we always have  $\delta_-(x(t)) > \delta_+(x(t))$ . We immediately deduce a simple criterion for the QBD  $D_t$  to be positive recurrent from (25):

**THEOREM 4.1** *The QBD  $D_t$  is positive recurrent if and only if*

$$\delta_-(x(t)) > 0 > \delta_+(x(t)). \tag{27}$$

If the QBD  $D_t$  is positive recurrent, then the AP takes place, and  $\pi_{1,2}(x(t))$  can be computed, as shown in §4.4 below. If, instead, we have net upward drift, i.e., if  $\delta_-(x(t)) > \delta_+(x(t)) \geq 0$ , then the CTMC is either null-recurrent or transient; in either case,  $\pi_{1,2}(x(t)) = 1$ . If, instead, we have net downward drift, i.e., if  $0 \geq \delta_-(x(t)) > \delta_+(x(t))$ , then the CTMC is again either null-recurrent or transient; in either case,  $\pi_{1,2}(x(t)) = 0$ .

**4.4 Computing  $\pi_{1,2}$ .** In this framework, 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  $\alpha_n^+$  and  $\alpha_n^-$  both being  $1 \times m$  vectors. Then the desired probability  $\pi_{1,2}$  can be expressed as

$$\pi_{1,2} = \sum_{n=0}^{\infty} \sum_{j=1}^m \alpha_{n,j}^+ = \sum_{n=0}^{\infty} \alpha_n^+ \mathbf{1} = \sum_{n=0}^{\infty} \alpha_n \mathbf{1}_+, \quad (28)$$

where  $\mathbf{1}$  denotes a  $m \times 1$  column vector with all entries 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 [6], the steady-state distribution has the matrix-geometric form

$$\alpha_n = \alpha_0 R^n, \quad (29)$$

where  $R$  is the  $2m \times 2m$  rate matrix. Since the spectral radius of the rate matrix  $R$  is strictly less than 1 (Corollary 6.2.4 of [6]), we have

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

Also, by Lemma 6.3.1 of [6], the boundary probability vector  $\alpha_0$  is the unique solution to the system

$$\alpha_0(B + RA_2) = 0 \quad \text{and} \quad \pi \mathbf{1} = \alpha_0(I - R)^{-1} \mathbf{1} = 1. \quad (30)$$

Finally, given the above, and using (28), we see that the desired quantity  $\pi_{1,2}$  can be represented as

$$\pi_{1,2} = \alpha_0(I - R)^{-1} \mathbf{1}_+, \quad (31)$$

where  $R$  is the  $2m \times 2m$  rate matrix and  $\alpha_0$  is the  $1 \times 2m$  vector of stationary boundary probabilities. The rate-matrix  $R$  is the minimal nonnegative solutions to the quadratic matrix equation

$$A_0 + RA_1 + R^2A_2 = 0,$$

and can be found efficiently by existing algorithms, as in [6] (see §9). In addition, important topological properties of  $R$  are known, and will be shown to hold in our case.

With the QBD representation, we can determine when the FTSP  $D_t$  is positive recurrent, for a given  $x(t)$ , (using (27)) and then numerically calculate  $\pi_{1,2}$ . This allows us to numerically solve the ODE (13), as in §9. Moreover, we will use the representation (31), and results about the rate matrix  $R$ , to conclude topological properties of  $\pi_{1,2}$ .

**5. Existence and Uniqueness of Solutions.** We now start to analyze the ODE and IVP introduced in §3.4. In this section we show that a unique solution exists to the IVP (15) for every initial point in the state space, at least on some initial interval. In subsequent sections we extend this result, and give sufficient conditions for a unique solution to exist for all  $t \geq 0$ . To apply existence and uniqueness results from ODE theory, we need the function  $\Psi$  in (14) to be (locally) Lipschitz continuous. However,  $\Psi$  is not even continuous on the full state-space  $\mathbb{S} \equiv [0, \infty)^2 \times [0, m_2]$  with elements  $x \equiv (q_1, q_2, z_{1,2})$ . (Here  $x$  denotes a possible value of the function  $x$ ; we use the notation interchangeably; it should be clear from the context. Recall that the ODE is autonomous, so that there is no time argument, i.e.,  $\Psi(x(t), t) = \Psi(x(t))$ .) To overcome this difficulty, we divide the state-space  $\mathbb{S}$  into three regions, and show that  $\Psi$  is indeed locally Lipschitz continuous in each of these regions.

**5.1 Properties of  $\Psi$ .** The ODE inherits essential structure from the queueing system with the FQR control. For the queueing systems, the instantaneous sharing is in a different direction when the (centered) queue-difference process  $D^n(t)$  in (6) is above 0 or below 0. The ODE has similar structure, but a special role is played by the boundary (where equality holds), which is where all averaging takes

place. In particular, the ODE has different behavior when the (fluid-scale, un-centered) queue difference  $q_1 - rq_2$  is above  $\kappa$ , equal to  $\kappa$  or below  $\kappa$ . We refer to the middle region as the *boundary*.

Thus we divide the state space  $\mathbb{S} \equiv [0, \infty)^2 \times [0, m_2] \equiv \{(q_1, q_2, z_{1,2})\}$  of the ODE into three regions:

$$\mathbb{S}^b \equiv \{q_1 - rq_2 = \kappa\}, \quad \mathbb{S}^+ \equiv \{q_1 - rq_2 > \kappa\}, \quad \mathbb{S}^- \equiv \{q_1 - rq_2 < \kappa\}, \quad (32)$$

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

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

The region  $\mathbb{S}^+$  above the boundary is an open subset of  $\mathbb{S}$ . For all  $x \in \mathbb{S}^+$ ,  $\pi_{1,2}(x) = 1$ . The region  $\mathbb{S}^-$  below the boundary is also an open subset of  $\mathbb{S}$ . For all  $x \in \mathbb{S}^-$ ,  $\pi_{1,2}(x) = 0$ . It is important to keep in mind that, in order for  $\mathbb{S}^-$  to be a proper subspace of  $\mathbb{S}$ , both service pools must be constantly full (in the fluid limit). 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).

It is immediate that the function  $\Psi$  in (14) 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  $\Psi$  is linear in each region. However,  $\Psi$  is not linear on  $\mathbb{S}^b$ , so we must work harder there.

To analyze  $\Psi$  on  $\mathbb{S}^b$ , we exploit properties of the QBD introduced in §4. First observe that, if  $0 < \pi_{1,2}(x(t)) < 1$  for  $s \leq t \leq u$ , then  $x(t) \in \mathbb{S}^b$  for  $t \in [s, u]$ , i.e., SSC holds on  $[s, u]$ . Recall that, for  $x \in \mathbb{S}$ ,  $\delta_+(x)$  and  $\delta_-(x)$  are the QBD drift rates in (26). Let  $\mathbb{A}$  be the set of all  $x \in \mathbb{S}^b$  for which the QBD is positive recurrent, as given in (27); i.e., let

$$\mathbb{A} \equiv \{x \in \mathbb{S}^b \mid \delta_-(x) > 0 > \delta_+(x)\}. \quad (33)$$

From the continuity of the QBD drift-rates in (26), it follows that  $\mathbb{A}$  is an open and connected subset of  $\mathbb{S}^b$ . Hence,  $\mathbb{A}$  can be regarded as an open connected subset of  $\mathbb{R}_+^2$  (since  $\mathbb{S}^b$  is homoeomorphic to  $\mathbb{R}_+ \times [0, m_2]$ ).

If  $x(t) \in \mathbb{A}$  for  $t \in [s, u]$ , then we say that *strong SSC* holds on that interval. If  $x(t) \in \mathbb{A}$  for all  $t \geq 0$ , then we say that strong SSC holds globally.

**DEFINITION 5.1** (local Lipschitz continuity) *A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is locally Lipschitz continuous if for every  $v_0 \in \mathbb{R}^n$  there exists a neighborhood  $U$  of  $v_0$  such that  $f$  restricted to  $U$  is Lipschitz continuous; i.e., there exists a constant  $K \equiv K(U)$  such that  $\|f(v_1) - f(v_2)\| \leq K\|v_1 - v_2\|$  for every  $v_1, v_2 \in U$ .*

**THEOREM 5.1** *The function  $\Psi$  in (14) is locally Lipschitz continuous on  $\mathbb{A}$ .*

**PROOF.** The key component of the function  $\Psi$  is  $\pi_{1,2}$ . We will look at  $\pi_{1,2}$ , and thus the QBD, as a function of the variable  $x \in \mathbb{A}$ . By the definition of the matrices  $A_0$ ,  $A_1$  and  $A_2$  in (21) (see also the example in §4.2) and the definitions of the rates in (16)-(19), the matrices  $A_i$ ,  $i = 0, 1, 2$ , are twice differentiable (as functions of  $x$ ) at each  $x \in \mathbb{A}$ . It follows from Theorem 2.3 in He [4] that the rate-matrix  $R$  in (29), which is the minimal nonnegative solution to the quadratic matrix equation  $A_0 + RA_1 + R^2A_2 = 0$ , is also twice differentiable at each  $x \in \mathbb{A}$ . In particular, the derivative  $R'$  exists and is continuous in  $\mathbb{A}$ . It follows from the normalizing expression in (30) and the differentiability of  $R$ , that  $\alpha_0$  is also differentiable. Hence, from (31), we see that  $\pi_{1,2}$  is differentiable at each  $x \in \mathbb{A}$ , with

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

By differentiating (30), we have

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

so that  $\alpha'_0$  is continuous. The continuity of  $R'$  and  $\alpha'_0$  implies that the derivative  $\pi'_{1,2}$  is continuous on  $\mathbb{A}$ , which in turn implies that the derivative  $\Psi'$  is continuous on  $\mathbb{A}$ . That in turn implies that  $\Psi$  is locally Lipschitz continuous on  $\mathbb{A}$ , as claimed. For this last step, we use the fact that a function mapping a convex compact subset of  $\mathbb{R}^m$  to  $\mathbb{R}^n$  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 [5]. Finally, since a continuous function on a compact set is bounded,  $\Psi$  satisfies this property. Hence  $\Psi$  is indeed locally Lipschitz continuous.  $\square$

**5.2 Solution to the ODE.** The local Lipschitz continuity of  $\Psi$  allows us to apply the classical Picard-Lindelöf theorem (extended to locally Lipschitz functions) to deduce the desired existence and uniqueness of solutions to the IVP (15); e.g., see Theorem 2.2 of Teschl [12].

**THEOREM 5.2** (local existence and uniqueness) *If  $w_0 \in \mathbb{A}$ , then there exists a unique solution  $x : [0, \delta) \rightarrow \mathbb{A}$  to the IVP (15) for some  $\delta > 0$ .*

**PROOF.** By the classical Picard-Lindelöf theorem, Theorem 2.2 of Teschl [12] or Theorem 3.1 in [5], and Theorem 5.1, there exists  $\delta_1 > 0$  such that there exists a unique solution to the ODE on the interval  $[0, \delta_1)$ , provided that  $x(t) \in \mathbb{A}$  for  $t \in [0, \delta_1)$ . Since  $w_0$  is contained in the open set  $\mathbb{A}$  and the function  $x$  and the drifts  $\delta_-$  and  $\delta_+$  are continuous functions, there necessarily exists  $\delta$  with  $0 < \delta \leq \delta_1$  such that  $x(t) \in \mathbb{A}$  for all  $t \in [0, \delta)$ .  $\square$

We now give sufficient conditions for the existence of a unique solution to the IVP (15) over the entire halfline  $[0, \infty)$ . There are two issues: (i) extending the existence and uniqueness result above, given that the solution falls in  $\mathbb{A}$ , and (ii) showing that a solution necessarily stays within  $\mathbb{A}$ . To address the first issue, we exploit boundedness. In particular, we prove that a solution to the IVP (15) is bounded, so that every fluid solution is contained in a compact subset of  $\mathbb{S}$ . We use the following notation:  $a \vee b \equiv \max\{a, b\}$ .

**THEOREM 5.3** (boundedness) *Every solution to the IVP (15) is bounded. In particular, the following upper bounds for the fluid queues hold:*

$$q_i(t) \leq q_i(0) \vee \lambda_i / \theta_i \quad t \geq 0, \quad i = 1, 2. \quad (34)$$

**PROOF.** For the boundedness, it is clear that  $0 \leq z_{1,2} \leq m_2$  and  $q_i \geq 0$  in  $\mathbb{S}$ . Hence, we only need to prove the upper bounds (34). For  $i = 1, 2$ , let  $u_i(t)$  be the function describing the queue-length process (of queue  $i$ ) in a modified system with no service processes (so that all the fluid output is due to abandonment). The queue-length process in the modified system evolves according to the ODE

$$\dot{u}_i(t) = \lambda_i - \theta_i u_i(t), \quad t \geq 0,$$

whose solution is

$$u_i(t) = \frac{\lambda_i}{\theta_i} + \left( u_i(0) - \frac{\lambda_i}{\theta_i} \right) e^{-\theta_i t}, \quad t \geq 0.$$

It follows that  $u_i(t) \leq u_i(0) \vee \lambda_i / \theta_i$  and, when  $u_i(0) = q_i(0)$ , the the right-hand side in (34) is an upper bound for  $u_i(t)$ . We now show that this is also a bound for  $q_i(t)$ . For that purpose, define the auxiliary function  $f_i(t) \equiv q_i(t) - u_i(t)$ ,  $t \geq 0$ , and observe that  $f_i(0) = 0$  and  $\dot{f}_i(0) < 0$ . Hence,  $f$  is decreasing at 0 with  $f(t) < f(0)$  for all  $t \in [0, \delta)$  for some  $\delta > 0$ . This implies that  $q_i(t) < u_i(t)$  for all  $t \in [0, \delta)$ .

We now want to show that  $q_i(t) \leq u_i(t)$  for all  $t \geq 0$ . For a proof by contradiction, assume that there exists some  $t_0 > 0$  such that  $q_i(t_0) > u_i(t_0)$ , and let

$$t_1 \equiv \sup\{t < t_0 : q_i(t) = u_i(t)\}, \quad t_2 \equiv \inf\{t > t_0 : q_i(t) = u_i(t)\}.$$

By the contradictory assumption and the continuity of  $q$  and  $u$ , we have  $0 < t_1 < t_0 < t_2$ . ( $t_2$  may be infinite.) Then

$$q_i(t) > u_i(t) \quad \text{for all } t_1 < t < t_2. \quad (35)$$

It follows from the mean-value theorem that there exists some  $t_3 \in (t_1, t_0)$  such that

$$\dot{f}_i(t_3) = \frac{f(t_0) - f(t_1)}{t_0 - t_1} = \frac{f(t_0)}{t_0 - t_1} > 0.$$

Hence,  $\dot{q}_i(t_3) > \dot{u}_i(t_3)$ . For  $i = 1$ , this translates to

$$\lambda_1 - \mu_{1,1}m_1 - \pi_{1,2}(x(t_3)) [z_{1,2}(t_3)\mu_{1,2} + z_{2,2}(t_3)\mu_{2,2}] - \theta_1 q_1(t_3) > \lambda_1 - \theta_1 u_1(t_3).$$

Thus,

$$\theta_1 (q_1(t_3) - u_1(t_3)) < -\mu_{1,1}m_1 - \pi_{1,2}(x(t_3)) [z_{1,2}(t_3)\mu_{1,2} + z_{2,2}(t_3)\mu_{2,2}] < 0,$$

so that  $q_1(t_3) < u_1(t_3)$ , contradicting (35). A similar argument holds for  $q_2$ .  $\square$

**THEOREM 5.4** (global existence and uniqueness) *Let  $x$  be the unique solution to the IVP (15) on an interval  $[0, \delta)$ , established by Theorem 5.2. If  $x(\delta) \in \mathbb{A}$ , then the solution can be extended to an interval  $[0, \delta')$ ,  $\delta' > \delta$ , with the solution again being unique. If it is known that the solution can never leave  $\mathbb{A}$ , then  $\delta' = \infty$ ; i.e., there exists a unique solution to the IVP (15) on  $[0, \infty)$ .*

In the proof of Theorem 5.4 we make use of the next lemma. For its proof see Theorem 3.3 in [5].

**LEMMA 5.1** *Consider an ODE  $\dot{x} = f(x)$  in a domain  $U$  in  $\mathbb{R}^n$ , where  $f$  is locally Lipschitz. Let  $K$  be a compact subset of  $U$ . If every solution of the ODE is contained in  $K$ , then there exists a unique solution to the ODE on the entire halfline  $[0, \infty)$ .*

**PROOF OF THEOREM 5.4.** By Theorem 5.1,  $\Psi$  is locally Lipschitz continuous, and by Theorem 5.3, a solution to the IVP (15) is bounded. It follows from Lemma 5.1 that there exists a unique solution to (15) for all  $t \geq 0$ .  $\square$

In Section §7 we give sufficient conditions for the solution of the IVP (15) to lie entirely in  $\mathbb{A}$ , which by Theorem 5.4 will imply existence and uniqueness of a solution over the entire halfline  $[0, \infty)$ . We also go further to provide an a posteriori demonstration of existence and uniqueness of a solution over the entire halfline  $[0, \infty)$  when these sufficient conditions do not hold: In §7.2, we show how being contained in  $\mathbb{A}$  for all  $t > 0$  can be inferred from the *initial behavior* of the solution, which is what we can achieve numerically. We then can apply Theorem 5.2 to conclude that there exists a unique solution to the IVP (15) for all  $t \geq 0$ .

**REMARK 5.1** Theorems 5.1-5.4 also hold for solutions to the IVP (15) in  $\mathbb{S}^-$  and  $\mathbb{S}^+$ . Indeed, they are elementary, because  $\Psi$  is Lipschitz continuous, since  $\pi_{1,2}$  is constant in these regions. The boundedness used in the proof of Theorem 5.4, and proved in Theorem 5.3, applies in these two regions as well.

**6. Fluid Stationarity.** As reviewed in §1 and §2.2, we did our initial analysis of the overloaded X model in [9] using a steady state (or stationary) fluid analysis. That is, we assumed that there exists a unique stationary point  $x^*$  and that  $x(t) \rightarrow x^*$  as  $t \rightarrow \infty$  for all initial states  $x(0)$ , and gave a heuristic derivation of the limit  $x^*$ . In this section we provide mathematical justification. We first give a formal definition of fluid stationarity and prove the existence and uniqueness of a stationary point  $x^*$  for the ODE (14). We then give conditions under which the fluid solution  $x \equiv \{x(t) : t \geq 0\}$  converges to stationarity as  $t \rightarrow \infty$ . In §7, we show that it does so exponentially fast.

**DEFINITION 6.1** (stationary point for the fluid) *We say that  $x^*$  is a stationary point for the ODE (or fluid model) if  $x(t) = x^*$  for all  $t \geq 0$  when  $x(0) = x^*$ . That is,  $x^*$  is a stationary point if  $\Psi(x^*) = 0$  for  $\Psi$  in (13) and (14). If  $x(t) = x^*$ , then we say that the fluid solution is in steady state at time  $t$ .*

We now make some important assumptions, which we will use to show that there exists a unique stationary point for the ODE. 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  $\lambda_i$ , service-rate  $\mu_{i,i}$  and abandonment-rate  $\theta_i$ ; see Theorem 2.3 in [14], especially equation (2.19), and §5.1 in [9]. In particular,

$$q_i^a \equiv \frac{(\lambda_i - \mu_{i,i} m_i)^+}{\theta_i} \quad \text{and} \quad s_i^a \equiv \left( m_i - \frac{\lambda_i}{\mu_{i,i}} \right)^+, \quad i = 1, 2, \quad (36)$$

where  $(x)^+ \equiv \max\{x, 0\}$ . It is easy to see that  $q_i^a s_i^a = 0$ ,  $i = 1, 2$ .

A sufficient condition for the ODE (14) to be well defined (so that the solution is in  $\mathbb{S}$ , possibly after an initial transient) is to have  $s_1^a = s_2^a = 0$ , i.e., there is no spare service capacity in either pool in their individual steady states. However, if  $s_2^a > 0$ , the solution can still be in  $\mathbb{S}$  after an initial transient, if 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  $q_1^a > \kappa$ . 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 that flows to the system, 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  $\kappa$  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 §8.)

From the above, we see that in order to have both service pools full all the time, we must have either  $s_1^a = s_2^a = 0$ , or, if  $s_2^a > 0$ ,  $\theta_1(q_1^a - \kappa) \geq \mu_{1,2}s_2^a$ . We summarize these conditions in the next assumption **which is assumed to hold henceforth**.

ASSUMPTION A.

(I) If  $q_1^a > \kappa$ , then  $\theta_1(q_1^a - \kappa) \geq \mu_{1,2}s_2^a$ .

(II) If  $q_1^a \leq \kappa$ , then  $s_1^a = s_2^a = 0$ .

In words, Condition (I) of the Assumption A guarantees that if there is spare service capacity in pool 2, then there is enough class-1 fluid to have both service pools full. Condition (II) guarantees that when there is no sharing of customers, both pools are full (with their own class fluid only), due to the arrival rates being larger than the total service capacity of each class. If Condition (II) holds, then FQR-T prevents sharing, and the two classes are independent. In this case, we can decompose the system into two independent Erlang-A models (operating in the ED regime), and analyze them separately, as was done in [14].

It is significant that Assumption A involves only the parameters of the system, and requires no knowledge on the specific solution to the IVP (15). We will show that when this assumption holds, there exists a unique stationary point in  $\mathbb{S}$  for every solution to (14).

**6.1 Uniqueness of the Stationary Point.** By definition, a stationary point  $x^* \in \mathbb{S}$  is such that  $\Psi(x^*) = 0$ . From (14), 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  $\Psi$ . However,  $\pi_{1,2}^*$  is a complicated function of  $x^*$  having the complicated closed-form expression in (28) and (31).

Theorem 6.1 below states that *if there exists a stationary point for the fluid ODE* (14), 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 (11). The existence of a stationary point is more involved, and is proved later; See Corollary 6.3.

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 (11). In particular, if  $q_1^* - r q_2^* > \kappa$  then we let  $\pi_{1,2}^* = 1$ ; if  $q_1^* - r q_2^* < \kappa$ , then we let  $\pi_{1,2}^* = 0$ . Equation (39) below shows that  $0 \leq \pi_{1,2}^* \leq 1$  whenever  $q_1^* - r q_2^* = \kappa$ , 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

$$z \equiv \frac{\theta_2(\lambda_1 - m_1\mu_{1,1}) - r\theta_1(\lambda_2 - m_2\mu_{2,2}) - \theta_1\theta_2\kappa}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}}. \quad (37)$$

**THEOREM 6.1** (uniqueness of the stationary point) *There can be at most one stationary point  $x^* \equiv (q_1^*, q_2^*, z_{1,2}^*)$  for the IVP (15), which for  $z$  in (37) must take the form*

$$z_{1,2}^* = 0 \vee z \wedge m_2, \quad q_1^* = \frac{\lambda_1 - m_1 \mu_{1,1} - \mu_{1,2} z_{1,2}^*}{\theta_1}, \quad q_2^* = \frac{\lambda_2 - \mu_{2,2}(m_2 - z_{1,2}^*)}{\theta_2}. \quad (38)$$

Moreover,

$$\pi_{1,2}^* = \frac{\mu_{1,2} z_{1,2}^*}{\mu_{1,2} z_{1,2}^* + (m_2 - z_{1,2}^*) \mu_{2,2}}. \quad (39)$$

**PROOF.** We start with (39). This expression is easily derived from the third equation in (14), by equating  $\dot{z}_{1,2}(t)$  to zero. Observe that if  $z_{1,2}^* = m_2$  then  $\pi_{1,2}^*$  in (39) 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 (14) we get the expressions of  $q_1^*$  and  $q_2^*$  in (38). 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 time.)

$$q_1(z) = \frac{\lambda_1 - \mu_{1,1} m_1 - \mu_{1,2} z}{\theta_1}, \quad q_2(z) = \frac{\lambda_2 - \mu_{2,2}(m_2 - z)}{\theta_2}, \quad \kappa = q_1(z) - r q_2(z). \quad (40)$$

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 (37). We can recover  $x^*$  from the solution to (40), 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 (40). First assume that  $z > m_2$ , which implies that  $q_2(z) > 0$ , and, by the third equation,  $q_1(z) > \kappa \geq 0$ . By replacing  $z$  with  $m_2$ ,  $q_1(\cdot)$  is increased and  $q_2(\cdot)$  is decreased (but is still positive), so that  $q_1(m_2) - r q_2(m_2) > \kappa$  (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 (40) 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  $q_1(0) - r q_2(0) < \kappa$ . In that case there is no sharing, and by Condition (II) of Assumption A, the point  $x^* \equiv (q_1(0), q_2(0), 0)$  is in  $\mathbb{S}^-$ . Once again, if  $x^*$  is indeed a solution to  $\Psi(x) = 0$ , then  $x^*$  is the unique stationary point.

Finally, assume that the solution  $x(z) \equiv (q_1(z), q_2(z), z)$  to (40) 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 (36). We start by rewriting  $q_1(z)$  and  $q_2(z)$  in (40) as

$$q_1(z) = q_1^a - \frac{\mu_{1,2}}{\theta_1} z, \quad q_2(z) = \frac{\mu_{2,2}}{\theta_2} (z - S_2^a). \quad (41)$$

Now, it follows from Assumption A that

$$\kappa \leq q_1^a - \frac{\mu_{1,2}}{\theta_1} s_2^a \leq q_1^a - \frac{\mu_{1,2}}{\theta_1} S_2^a, \quad (42)$$

where the second inequality follows trivially, since  $S_2^a \leq s_2^a$ . From the third equation of (40),  $\kappa = q_1(z) - r q_2(z)$ . Combining this with (41), we see that

$$\kappa = q_1(z) - r q_2(z) = q_1^a - \frac{\mu_{1,2}}{\theta_1} z - r \frac{\mu_{2,2}}{\theta_2} (z - S_2^a). \quad (43)$$

Combining (42) and (43), we get

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

which is equivalent to

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

This, together with the fact that the solution has  $z \geq 0$ , implies that  $z \geq \max\{0, S_2^a\} = s_2^a$ . It follows from (41) that  $q_2(z) \geq 0$  and, by using the third equation in (40) again,  $q_1(z) = rq_2(z) + \kappa \geq \kappa \geq 0$ .  $\square$

An immediate consequence of the proof of Theorem 6.1 is that, in order to find the candidate stationary point  $x^*$ , one has to solve the three equations in (40). If the (unique) solution has  $z < 0$ , then  $x^* \in \mathbb{S}^-$  and  $z_{1,2}^* = 0$ . If  $z > m_2$  then  $x^* \in \mathbb{S}^+$  and  $z_{1,2}^* = m_2$ . Otherwise,  $x^* \in \mathbb{S}^b$  with  $0 \leq z_{1,2}^* \leq m_2$ . The queue lengths have always the same expressions, and their values depend only on the value of  $z$ . The next corollary summarizes the values  $x^*$  may take, depending on its region.

**COROLLARY 6.1** *Let  $x^* = (q_1^*, q_2^*, z_{1,2}^*)$  be the point defined in Theorem 6.1.*

(i) *If  $x^* \in \mathbb{S}^b$ , then, for  $z$  defined in (37),*

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

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

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

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

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

**PROOF.** If  $x^* \in \mathbb{S}^b$ , then the solution to (40) will have  $0 \leq z \leq m_2$ , where the exact value of  $x^*$  is readily seen to be the one in (i). If  $x^* \in \mathbb{S}^+$ , then  $q_1^* - r q_2^* > \kappa$ , so that  $\pi_{1,2}^* = 1$ . Plugging  $\pi_{1,2}^* = 1$  in the ODE for  $z_{1,2}(t)$  in (14), we get  $\dot{z}_{1,2}(t) = z_{2,2}(t) \mu_{2,2}$ . Since at stationarity  $\dot{z}_{1,2}(t) = 0$ , it follows that  $z_{2,2}^* = 0$ , which implies that  $z_{1,2}^* = m_2$ . Plugging this value of  $z_{1,2}^*$ , together with  $\pi_{1,2}^* = 1$  when  $\dot{q}_i(t) = 0$ ,  $i = 1, 2$ , we get the values of  $q_1^*$  and  $q_2^*$  as in (ii).

Finally, if  $x^* \in \mathbb{S}^-$ , i.e., if  $q_1^* - r q_2^* < \kappa$ , then  $\pi_{1,2}^* = 0$ , so that, by plugging this value of  $\pi_{1,2}^*$  in the ODE for  $z_{1,2}(t)$  in (14), we see that  $\dot{z}_{1,2}(t) = \mu_{1,2} z_{1,2}(t)$ . Equating to zero, to get the value at stationarity, we see that  $z_{1,2}^* = 0$ . Plugging  $\pi_{1,2}^* = 0$  and  $z_{1,2}^* = 0$  in the ODE for  $q_1(t)$  and  $q_2(t)$ , and equating these to zero, we get the values in (iii).  $\square$

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  $\kappa$ , 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 an  $M/M/\infty$  system with service rate  $\theta_2$  and arrival rate  $\lambda_2$ . 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.

As we remarked at the beginning of this subsection, from the proofs of Theorem 6.1 and Corollary 6.1, and from the expression of  $\pi^*$  in (39), it is clear that  $x^*$  is a stationary point for the ODE (14) when  $x^*$  is in  $\mathbb{S}^+$  or  $\mathbb{S}^-$ . In that case  $\pi_{1,2}(x^*) = \pi_{1,2}^*$  (equals 1 in  $\mathbb{S}^+$  and equals 0 in  $\mathbb{S}^-$ ). That same conclusion applies when  $x^*$  is in  $\mathbb{S}^b - \mathbb{A}$ , once we have verified that  $\pi_{1,2}(x^*) = \pi_{1,2}^*$ . In these cases,  $x^*$  is the unique stationary point to the ODE. The problem of existence is only when the suspected stationary-point  $x^*$  is in  $\mathbb{A}$ .

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.

**COROLLARY 6.2** *Let  $x^*$  be as in (38). Then*

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

$$\frac{\mu_{1,2}s_2^a}{\theta_1} \vee rq_2^a \leq q_1^a - \kappa \leq \frac{r\lambda_2}{\theta_2} + \frac{\mu_{1,2}m_2}{\theta_1}; \quad (44)$$

$x^* \in \mathbb{A}$  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  $rq_2^a > q_1^a - \kappa$ .

PROOF. We prove (i) only. The proofs for (ii) and (iii) are similar. First assume that  $x^* \in \mathbb{S}^b$ . Since  $z_{1,2}^* \geq 0$ , It follows from the expression for  $z_{1,2}^*$  in (i) of Corollary 6.1 that if  $q_2^a \geq 0$  then  $q_1^a - \kappa \geq rq_2^a$ . If  $s_2^a > 0$  then  $q_1^a - \kappa \geq \mu_{1,2}s_2^a/\theta_1$  by Assumption A. For the other inequality we use the fact that

$$z_{1,2}^* = \frac{\theta_1\theta_2(q_1^a - \kappa) - r\theta_1(\lambda_2 - \mu_{2,2}m_2)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} \leq m_2,$$

which implies the right-hand inequality in (44).

Now Assume that (44) holds. It follows from the right-hand-side (RHS) inequality and the expression of  $z$  in (37) that

$$\begin{aligned} z &\equiv \frac{\theta_1\theta_2(q_1^a - \kappa) - r\theta_1(\lambda_2 - \mu_{2,2}m_2)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} \\ &\leq \frac{\theta_1\theta_2(r\lambda_2/\theta_2 + \mu_{1,2}m_2/\theta_1) - r\theta_1(\lambda_2 - \mu_{2,2}m_2)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} = m_2. \end{aligned}$$

From the left-hand inequality in (44), we see that, if  $s_2^a = 0$  (and necessarily  $q_2^a \geq 0 = s_2^a$ ), then

$$z \geq \frac{\theta_1\theta_2rq_2^a - r\theta_1(\lambda_2 - \mu_{2,2}m_2)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} = 0.$$

If  $s_2^a > 0$  (and  $q_2^a = 0$ ), then

$$z \geq \frac{\theta_2\mu_{1,2}s_2^a - r\theta_1(\lambda_2 - \mu_{2,2}\lambda_2)}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} = \frac{\theta_2\mu_{1,2}s_2^a + r\theta_1\mu_{2,2}s_2^a}{r\theta_1\mu_{2,2} + \theta_2\mu_{1,2}} = s_2^a.$$

Thus, if (44) holds, then  $s_2^a \leq z \leq m_2$ . This was shown to imply that  $x^* \in \mathbb{S}^b$  in the proof of Theorem 6.1. (In fact, we have a stronger result, since we have  $z \geq s_2^a$ . This is due to the requirement that  $q_1^a - \kappa \geq \mu_{1,2}s_2^a/\theta_1$ , which is exactly Condition (I) in Assumption A.)

We can show that the inequalities in (44) are strict if and only if  $x^* \in \mathbb{A}$  by first observing that the inequalities are strict if and only if  $0 < z^* < m_2$ , and then directly calculate the QBD drift rates at the point  $x^*$ . This is done in §7.2; see (50). It then follows that (27) holds at  $x^*$  if and only if  $0 < z^* < m_2$ .

Alternatively, in Corollary (6.3) we show that  $\pi_{1,2}^*$  in (39) is indeed the value of (11) at the point  $x^*$ . It is easy to see that  $0 < \pi_{1,2}^* < 1$  in (39) if and only if  $0 < z^* < m_2$ .  $\square$

REMARK 6.1 It follows from Corollary 6.2 that in applications  $\mathbb{A}$ , is the most likely region for the stationary point when the system is overloaded. This is because we expect the arrival rates to be 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 [9, 10]: 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  $\lambda_1 > \mu_{1,1}m_1 + \mu_{1,2}m_2 + \theta_1r\lambda_2/\theta_2 = 252$ , 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  $\lambda_1$  needs to be even larger than that.

**6.2 Existence of a Stationary Point and Stability.** We have just established uniqueness of the stationary point in  $\mathbb{S}$ , and characterized it. In the process, we have also established existence in  $\mathbb{S} - \mathbb{A}$ . Now we will establish existence of the stationary point in  $\mathbb{A}$ . However, we want to do more. Having a unique stationary point does not imply that a fluid solution necessarily converges to this point as  $t \rightarrow \infty$ .

It does not even guarantee that a solution to the IVP (15) is asymptotically stable in the sense that, if  $\|x(0) - x^*\| < \epsilon$ , then  $x(t) \rightarrow x^*$  as  $t \rightarrow \infty$ , no matter how small  $\epsilon$  is. In fact, there is not even a guarantee that  $x(t)$  will remain in the  $\epsilon$ -neighborhood of  $x^*$  for all  $t \geq 0$ . We will establish all of these properties in Theorem 6.2 below by showing that  $x^*$  in §6.1 is globally asymptotically stable, as defined below:

**DEFINITION 6.2** (global asymptotic stability) *A point  $x^*$  is said to be globally asymptotically stable 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*

$$\|x(t) - x^*\| < \epsilon, \quad \text{for all } t \geq T,$$

Note that our definition of global asymptotic stability goes beyond simple convergence by also requiring that the limit be a stationary point. (In general, it is possible to have convergence without the limit being a stationary point.)

The next theorem concludes that, if  $x(0)$  and  $x^*$  in (38) are both in one of the regions  $\mathbb{S}^-$ ,  $\mathbb{S}^+$  or  $\mathbb{A}$ , and if the fluid solution  $x$  lies entirely in that same region, then  $x^*$  is a globally asymptotically stable point for the ODE (14); i.e.,  $x^*$  is a stationary point and  $x(t) \rightarrow x^*$  as  $t \rightarrow \infty$ . (So far, We are unable to establish global asymptotic stability for  $x^*$  in the boundary region  $\mathbb{S}^b - \mathbb{A}$ .)

**THEOREM 6.2** (global asymptotic stability of  $x^*$ ) *If the solution to (15) lies entirely in one of the regions  $\mathbb{S}^+$ ,  $\mathbb{S}^-$  or  $\mathbb{A}$ , then  $x^*$  in Theorem 6.1 is globally asymptotically stable.*

The proof of Theorem 6.2 relies on results from nonlinear-control theory for deterministic dynamical systems, specifically, Lyapunov stability theory; for background, see Chapter 4 of Khalil [5]. 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 6.3** (Lie derivative) *For a continuously differentiable function  $V : E \rightarrow \mathbb{R}$ , and a function  $\Psi : E \rightarrow \mathbb{R}^n$ , the Lie derivative of  $V$  along  $\Psi$  is defined by*

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

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

**DEFINITION 6.4** (Lyapunov-function candidate) *A continuously differentiable function  $V : E \rightarrow \mathbb{R}$  is a Lyapunov-function candidate if:*

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

In proving Theorem 6.2 we use the following theorem, which is Theorem 4.2 pg. 124 in [5]:

**THEOREM 6.3** (global asymptotic stability for nonlinear ODE) *Let  $x = 0$  be a stationary point of  $\dot{x} = \Psi(x)$ ,  $\Psi : E \rightarrow \mathbb{R}^n$ , and let  $V : \mathbb{R}_+^n \rightarrow \mathbb{R}$  be a Lyapunov-function candidate. If*

- (i)  $V(x) \rightarrow \infty$  as  $\|x\| \rightarrow \infty$  and
- (ii)  $\dot{V}(x) < 0$  for all  $x \neq 0$ ,

then  $x = 0$  is globally asymptotically stable as in Definition 6.2.

Notice that, under the conditions of Theorem 6.3, 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, which we will apply further in §7.2. We say that  $\beta_V(\alpha)$  is the  $\alpha$   $V$ -ball with center at  $x^*$  and radius  $\alpha$  if

$$\beta_V(\alpha) \equiv \{x \in \mathbb{R}^n : \|V(x) - V(x^*)\| \leq \alpha\}. \quad (45)$$

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  $t \geq t_0$ . Thus, with the Lyapunov-function approach, we show *both* that  $x^*$  is a stationary point and that there is convergence  $x(t) \rightarrow x^*$  as  $t \rightarrow \infty$  for all initial values  $x(0)$ . We also establish this stronger “V-monotonicity.”

**PROOF OF THEOREM 6.2.** Let  $x \equiv \{x(t) : t \geq 0\}$  be the unique solution to (15), and assume that  $x$  lies entirely in only one of the regions  $\mathbb{S}^-$ ,  $\mathbb{S}^+$  or  $\mathbb{A}$ . Let  $x^* \equiv (q_1^*, q_2^*, z_{1,2}^*)$  be the stationary point for the system (13), and assume that  $x^*$  is in the same region as  $x$ . Since  $x^* \neq 0$ , 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 (38). 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 (14), 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 (38)

$$\begin{aligned} \dot{V}_1(y) &= \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(y(t))[\tilde{z}_{1,2}(t)\mu_{1,2} + (m_2 - \tilde{z}_{1,2}(t))\mu_{2,2}] - \theta_1(q_1(t) + q_1^*) \\ &\quad + \lambda_2 - (1 - \pi_{1,2}(y(t)))[(m_2 - \tilde{z}_{1,2}(t))\mu_{2,2} + \tilde{z}_{1,2}(t)\mu_{1,2}] - \theta_2(q_2(t) + q_2^*) \\ &= \lambda_1 + \lambda_2 - m_1\mu_{1,1} - m_2\mu_{2,2} + z_{1,2}(t)\mu_{2,2} + z^*\mu_{2,2} - z_{1,2}(t)\mu_{1,2} - z_{1,2}^*\mu_{1,2} \\ &\quad - \theta_1q_1(t) - \theta_1q_1^* - \theta_2q_2(t) - \theta_2q_2^* \\ &= -\theta_1q_1(t) - \theta_2q_2(t) - z_{1,2}(t)(\mu_{1,2} - \mu_{2,2}). \end{aligned}$$

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,

$$\begin{aligned} \dot{V}_2(y) &= C[\lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(y(t))(\tilde{z}_{1,2}(t)\mu_{1,2} + (m_2 - \tilde{z}_{1,2}(t))\mu_{2,2})] - \theta_1(q_1(t) + q_1^*) \\ &\quad + \lambda_2 - (1 - \pi_{1,2}(y(t)))(\tilde{z}_{1,2}(t)\mu_{1,2} + (m_2 - \tilde{z}_{1,2}(t))\mu_{2,2}) - \theta_2(q_2(t) + q_2^*) \\ &\quad + (C - 1)[\pi_{1,2}(y(t))(m_2 - \tilde{z}_{1,2}(t))\mu_{2,2} - (1 - \pi_{1,2}(y(t)))\tilde{z}_{1,2}(t)\mu_{1,2}] \\ &= -C\theta_1q_1(t) - \theta_2q_2(t) - z_{1,2}(t)(C\mu_{1,2} - \mu_{2,2}), \end{aligned}$$

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

By Theorem 6.3,  $y^* = 0$  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  $x(t) \rightarrow x^*$ , provided that  $x$  is in the same region ( $\mathbb{S}^+$ ,  $\mathbb{S}^-$  or  $\mathbb{A}$ ) for all  $t \geq 0$ .  $\square$

We summarize the existence and uniqueness result of the stationary point in the next corollary.

**COROLLARY 6.3** (existence and uniqueness of a stationary point) *Under Assumption A, there exists a unique stationary point  $x^*$  in  $\mathbb{S}$  for the ODE in (13) and (14), with  $x^*$  defined in (38). As a consequence, we have  $\pi_{1,2}(x^*) = \pi_{1,2}^*$  for  $\pi_{1,2}$  in (11) and  $\pi_{1,2}^*$  in (39).*

**PROOF.** Uniqueness of a stationary point for the ODE in (14) was fully treated in §6.1, so it suffices to consider only existence. We already observed after the proof of Corollary 6.1 that both existence and uniqueness are immediate if  $x^*$  is in  $\mathbb{S} - \mathbb{A}$ . The existence of the stationary point  $x^* \in \mathbb{A}$  follows from Theorem 6.2 provided that there exists a solution lying entirely in  $\mathbb{A}$ . However, we can choose to take  $x(0) = x^*$  in  $\mathbb{A}$ , in which case,  $x(t) = x^*$  for all  $t \geq 0$ , so that extra condition is satisfied.  $\square$

**7. Conditions for State-Space Collapse.** Both our result establishing global existence and uniqueness of a solution  $x$  to the IVP (15) (Theorem 5.4) and our result establishing global asymptotic stability of the stationary point  $x^*$  to the ODE (13) (Theorem 6.2) require that the solution  $x$  lies in the same region for all  $t \geq 0$ . As before, we are mostly interested in region  $\mathbb{A}$ , where the AP is operating, and which is the most likely region for the stationary point  $x^*$  to be (during overloads). 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 §8 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 has concluded. (It is then reasonable to consider  $x(0)$  as well as  $x^*$  as being in  $\mathbb{A}$ .)

We start by giving sufficient conditions for global strong SSC, i.e., having  $x \in \mathbb{A}$  on  $[0, \infty)$ . Afterwards, for the cases in which these sufficient conditions do not hold, we provide a method to infer strong SSC by solving the ODE (14) up to some finite time  $T$  (which is shown to be not very large).

**7.1 Sufficient Conditions for Strong SSC.** We now give sufficient conditions for global strong SSC. These conditions depend only on the initial point  $x(0)$  and the basic parameters of the system.

**THEOREM 7.1** (sufficient conditions for global strong SSC) *Let  $\nu \equiv \mu_{1,2} \wedge \mu_{2,2}$ , and suppose that  $x(0) \in \mathbb{A}$ . Also assume that*

$$q_2(0) \leq \lambda_2/\theta_2 \quad \text{and} \quad q_1(0) \leq (\lambda_1 - m_1\mu_{1,1})/\theta_1. \quad (46)$$

*If, in addition, the following inequalities are satisfied, then the solution to the IVP (15) is in  $\mathbb{A}$  for all  $t$ :*

$$\begin{aligned} (i) \quad & \lambda_1 < \nu m_2 + m_1\mu_{1,1} \quad \text{and} \\ (ii) \quad & \lambda_2 < \nu m_2 \end{aligned} \quad (47)$$

**REMARK 7.1** The rate conditions in (47) are intuitive, at least when  $\mu_{1,2} = \mu_{2,2}$ . Under condition (i), there is enough service capacity in both service pools to serve all of the class-1 input. Thus, a situation in which  $q_1 - rq_2 > \kappa$  can not be sustained for long, since if  $q_1$  grows above the boundary, pool 2 can allocate more service capacity in order to “pull” queue 1 back to the boundary. Similarly, under condition (ii), there is enough service capacity in pool 2 (which is the only one serving class 2 in our settings) to “pull” queue 2 back to the boundary whenever it grows above it, so that  $q_1 - rq_2 < \kappa$  is not sustainable either. Observe that Condition (i) is relatively weak, since it allows  $\lambda_1$  to be quite large compared to the total service capacity of pool 1, i.e., class 1 can be highly overloaded. On the other hand, Condition (ii) is more restrictive, and when  $\mu_{2,2} > \mu_{1,2}$  is likely not to hold in applications. However, if  $\mu_{2,2} \leq \mu_{1,2}$  (equality of the rates is often assumed), then Condition (ii) simply states that class 2 is not overloaded.

**PROOF OF THEOREM 7.1.** We start by showing, under Condition (i), that  $\delta_+(x(t))$  in (26) 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

$$(\mu_{2,2} - \mu_{1,2})z_{1,2}(t) - m_2\mu_{2,2} < -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2q_2(t)) + \theta_1q_1(t). \quad (48)$$

If  $\mu_{2,2} > \mu_{1,2}$ , then the left-hand side (LHS) of (48) 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 (48) is smaller than or equal to  $-\nu m_2$ .

Since  $q_2(0) \leq \lambda_2/\theta_2$ , we conclude, using the bound in (34), that  $\theta_2q_2(t) \leq \lambda_2$  for all  $t \geq 0$ . This, together with the fact that  $q_1(t) \geq 0$  for all  $t$ , implies that the RHS of (48) is larger than or equal to  $-(\lambda_1 - m_1\mu_{1,1})$ , so that

$$(\mu_{2,2} - \mu_{1,2})z_{1,2}(t) - \mu_{2,2}m_2 \leq -\nu m_2 < -(\lambda_1 - m_1\mu_{1,1}) \leq -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2q_2(t)) + \theta_1q_1(t)$$

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 (26), 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

$$r(\mu_{1,2} - \mu_{2,2})z_{1,2}(t) + r\mu_{2,2}m_2 > -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2q_2(t)) + \theta_1q_1(t). \quad (49)$$

It is easy to see that the LHS of (49) 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 5.3 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 (49) is smaller than or equal to  $r\lambda_2$ . Overall, we have

$$r(\mu_{1,2} - \mu_{2,2})z_{1,2}(t) + r\mu_{2,2}m_2 \geq r\nu m_2 > r\lambda_2 \geq -(\lambda_1 - m_1\mu_{1,1}) + r(\lambda_2 - \theta_2q_2(t)) + \theta_1q_1(t),$$

where the second inequality is due to Condition (ii).

Since (27) 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 (15) must lie entirely in  $\mathbb{A}$ .  $\square$

Combining Theorems 5.4, 6.2 and 7.1, we have the following corollary providing sufficient conditions for all good results discussed so far:

**COROLLARY 7.1** *If (44) holds with strict inequalities,  $x(0) \in \mathbb{A}$  and the four inequalities in Theorem 7.1 hold, then (i) there exists a unique solution  $x$  to the IVP (15) which lies entirely in  $\mathbb{A}$  and (ii) there exists a unique stationary point  $x^*$  to the ODE (14) which is globally asymptotically stable. That stationary point  $x^*$  is given in Corollary 6.2.*

**7.2 Verifying Eventual Convergence to Stationarity.** It is reasonable to assume that, if we look at the system after an initial transient over  $[0, T]$ , then  $x(T)$  and the unique stationary point  $x^*$  will be in the same region, and the fluid solution  $x(t)$  will converge to  $x^*$  as  $t \rightarrow \infty$ . Even if  $x$  leaves the region for some period of time, we expect that, after some transient period, it will return to the region where  $x^*$  is, stay there and converge to  $x^*$ . However, it remains to prove in full generality that there necessarily exists a time  $T$  after which the solution will never leave a region.

However, for every individual IVP, we may be able to infer that  $x(t)$  will converge to  $x^*$  by numerically solving the IVP over an initial interval  $[0, T]$  and observing that, after some initial transient (which has passed),  $x(t)$  is indeed in the set  $\mathbb{A}$  and is close to  $x^*$ . Specifically, we will show that there exist  $\alpha > 0$  and  $T \equiv T(\alpha)$ , such that global strong SSC can be inferred once  $\|x(T) - x^*\| < \alpha$ .

To achieve that goal, we make use of the Lyapunov function  $V$  and, more specifically,  $\beta_V(\alpha)$ , the  $\alpha$   $V$ -ball with center at  $x^*$  and radius  $\alpha$  in (45). We will exploit the fact that the solution  $x$  cannot leave a  $V$ -ball once it enters it. Thus we seek an  $\alpha > 0$  such that  $\beta_V(\alpha) \subseteq \mathbb{A}$ . Once  $x$  enters this  $\beta_V(\alpha)$ , it can never leave, so the function  $x$  remains in  $\mathbb{A}$  thereafter.

To find an appropriate radius  $\alpha$ , we introduce the drift rates at stationarity,  $\delta_+^* \equiv \delta_+(x^*)$  and  $\delta_-^* \equiv \delta_-(x^*)$ . It follows from the expressions in (26) that

$$\delta_+^* \equiv \delta_+(x^*) = -\mu_{2,2}(r+1)(m_2 - z_{1,2}^*) \quad \text{and} \quad \delta_-^* \equiv \delta_-(x^*) = \mu_{1,2}(r+1)z_{1,2}^*. \quad (50)$$

Thus, if  $0 < z_{1,2}^* < m_2$ , then the positive recurrence condition (27) holds at the stationary point  $x^*$ . (This agrees with (39) which has  $0 < \pi_{1,2}^* < 1$  if and only if  $0 < z_{1,2}^* < 1$ .)

In the next theorem we give explicit expressions for  $\alpha$ . Observe that for reasonable rates, such as will hold in applications,  $\alpha$  is quite large (which is what we want, because we will then be able to infer that  $x$  lies entirely in  $\mathbb{A}$  with only modest computation). In fact, in the numerical example considered in §9.3 we show that, typically in applications,  $\alpha$  is so large, that we can infer that  $x$  lies entirely in  $\mathbb{A}$  without even solving the IVP! That is, the initial condition is already in the  $V$ -ball  $\beta_V(\alpha)$ .

**THEOREM 7.2** *Suppose that  $x^* \in \mathbb{A}$  and let  $\xi \equiv \min\{|\delta_+^*|, \delta_-^*\}$ .*

- (i) *When  $\mu_{2,2} \geq \mu_{1,2}$ , let  $\alpha = \xi/r\theta_2$*
- (ii) *When  $\mu_{2,2} < \mu_{1,2}$ , 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  $\mathbb{A}$ , so that  $x^*$  in (i) of Corollary 6.1 is a globally asymptotically stable stationary point.*

PROOF. To find a proper  $\alpha$  for the  $V$ -ball  $\beta_V(\alpha)$ , we once again use the conditions (48) and (49). We first show how to find  $\alpha$  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 6.2) 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

$$\begin{aligned} q_1^* - \frac{\epsilon_1}{C} < q_1(t) < q_1 + \frac{\epsilon_1}{C}, \quad q_2^* - \epsilon_1 < q_2(t) < q_2^* + \epsilon_1 \quad \text{and} \\ z_{1,2}^* - \frac{\epsilon_1}{C-1} < z_{1,2}(t) < z_{1,2}^* + \frac{\epsilon_1}{C-1}, \quad t \geq T. \end{aligned} \quad (51)$$

To make sure  $\delta_+(x(t)) < 0$ , we use (48), reorganizing the terms. We need to have

$$(\mu_{2,2} - \mu_{1,2})z_{1,2}(t) + r\theta_2q_2(t) - \theta_1q_1(t) < -(\lambda_1 - \mu_{1,1}m_1) + r\lambda_2 + \mu_{2,2}m_2.$$

By (51), 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 (49), reorganizing the terms. We need to have

$$r(\mu_{1,2} - \mu_{2,2})z_{1,2}(t) + r\theta_2q_2(t) - \theta_1q_1(t) > -(\lambda_1 - \mu_{1,1}m_1) + r(\lambda_2 - \mu_{2,2}m_2).$$

Using (51) again (with a different  $\epsilon_2$ ), we see that it suffices to show that

$$r(\mu_{1,2} - \mu_{2,2}) \left( z_{1,2}^* + \frac{\epsilon_2}{C-1} \right) + r\theta_2(q_2^* - \epsilon_2) - \theta_1 \left( q_1^* + \frac{\epsilon_2}{C} \right) > -(\lambda_1 - \mu_{1,1}m_1) + r(\lambda_2 - \mu_{2,2}m_2).$$

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  $\alpha$  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  $\alpha$  in (ii). □

**7.3 Exponential Stability.** In this section we will establish exponential stability, i.e., we will show that the solution converges to the stationary point exponentially fast. We do this for two reasons: first, to help justify using the stationary point for performance approximations and, second, to show that it should not require a lengthy calculation to verify that the solution will remain within the set  $\mathbb{A}$  and converge to the stationary point  $x^*$ .

In the previous section, we have shown that for a system with a steady state  $x^*$  in  $\mathbb{A}$ , we can run the algorithm, starting at an arbitrary initial point  $x(0)$ , until  $x \equiv (q_1, q_2, z_{1,2})$  falls in the  $V$ -ball  $\beta_V(\alpha)$  in (45) for an  $\alpha$  identified in Theorem 7.2. It is easy to see that if  $z_{1,2}^*$  is not too close to 0 or  $m_2$ , then  $\alpha$  is relatively large, so that numerical issues do not arise. However, we want to know that the time  $T$  at which the solution enters this  $\alpha$ -neighborhood of  $x^*$  should not be too large.

DEFINITION 7.1 (exponential stability) *A stationary point  $x^*$  is said to be (globally) exponentially stable if there exist two real constants  $\vartheta, \beta > 0$  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  $\mathbb{R}^n$ .

To show that  $x^*$  in (38) is exponentially stable, we use Theorem 3.4 on p. 82 of Marquez [7], which we state here for completeness.

THEOREM 7.3 (exponential stability of the origin) *Suppose that all the conditions of Theorem 6.3 are satisfied. In addition, assume that there exist positive constants  $K_1, K_2, K_3$  and  $p$  such that*

$$\begin{aligned} K_1 \|x\|^p &\leq V(x) \leq K_2 \|x\|^p \\ \dot{V}(x) &\leq -K_3 \|x\|^p. \end{aligned}$$

Then the origin is exponentially stable, and

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

We now state our application of the general theorem. We will use the  $L_1$  norm:  $\|x\| = |x_1| + |x_2| + |x_3|$  for  $x \in \mathbb{R}^3$ .

THEOREM 7.4 (exponential stability of  $x^*$ ) *If the entire trajectory of the solution to the IVP (15) is in  $\mathbb{A}$ , then  $x^*$  in (38) is exponentially stable, and the following hold:*

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

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

where

$$K_3 \equiv \max\{\theta_1, \theta_2, \mu_{1,2} - \mu_{2,2}\}. \quad (52)$$

(ii) *If  $\mu_{2,2} = B\mu_{1,2}$ ,  $B \geq 1$ , then for any  $C > B$*

$$\|x(t) - x^*\| \leq \|x(0) - x^*\| (C/K_1) e^{-(K_4/2)t} \quad \text{for all } t \text{ and } x(0),$$

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. We consider the two cases in turn:

(i) If  $\mu_{1,2} > \mu_{2,2}$ , then  $V_1(x) \equiv x_1 + x_2$ ,  $x \geq 0$ , was shown to be a Lyapunov function in Theorem 6.3 with a strictly negative Lie derivative. Thus, since  $x \geq 0$ , we can take  $K_1 = K_2 = 1$  and  $p = 1$ . As  $\dot{V}_1(x) = -\theta_1 q_1(t) - \theta_2 q_2(t) - (\mu_{1,2} - \mu_{2,2}) z_{1,2}(t)$ , we can take  $K_3$  in (52), and the result follows from Theorem 7.3.

(ii) If  $\mu_{1,2} \leq \mu_{2,2}$ , then 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 Theorem 6.3 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  $\dot{V}_2(x) \leq -K_4 \|x\|$ .  $\square$

If  $x(0)$  and  $x^*$  are in  $\mathbb{S}^-$  or  $\mathbb{S}^+$ , then the same methods can be applied to verify whether  $x$  lies entirely in the same region, and thus converges to  $x^*$ . These methods, together with the fast rate of convergence, suggest that if  $x(0)$  and  $x^*$  are both in the same region, then  $x$  will converge to  $x^*$ , and will do so exponentially fast. As mentioned in the beginning of the subsection, we cannot prove this in full generality. There should be convergence for any initial state, even outside  $\mathbb{S}$ , but that requires formulating ODE's for other regions, which we turn to next. In fact, as we explain in Remark 8.1 in the next section, we need to add another feature to make it possible to have convergence to the stationary point for all initial conditions.

**8. Transient Behavior Before Hitting  $\mathbb{S}$ .** Recall that our model is designed to respond to unexpected overloads. We assume that the two classes operate independently until a time at which the arrival rates change, and the system becomes overloaded. Let 0 be the time that the arrival rates change. We thus think of a system in steady state at time 0 when the arrival rates change, with

$$q_1(0) = q_2(0) = z_{1,2}(0) = z_{2,1}(0) = 0. \quad (53)$$

In particular,  $q_1(0) \leq \kappa$ , and no sharing is taking place. A well-operated system tends to have a critically loaded fluid limit, yielding steady-state values  $z_{1,1}(0) = m_1$  and  $z_{2,2}(0) = m_2$ , but we could also have an underloaded steady state, with  $z_{1,1}(0) < m_1$  and/or  $z_{2,2}(0) < m_2$  as well.

The ODE in (13)-(14) can be regarded as the fluid limit of a sequence of overloaded queueing models. Class 1 was assumed to be overloaded due to the arrival rate being larger than the total service rate of service pool 1, while class 2 was overloaded either because its arrival rate was also too large (but less so than class 1), or because pool 2 was helping class-1 customers. For the ODE, the system overload assumption translates into having  $z_{1,1}(t) = m_1$  and  $z_{1,2}(t) + z_{2,2}(t) = m_2$  for all  $t$ , so that the state space for the fluid limit was taken to be  $\mathbb{S}$ . (The space  $\mathbb{S}$  was defined in (32) in §5, but the assumption that the service pools are both full was introduced at the beginning of §3.2.) However, if either  $z_{1,1}(0) < m_1$  or  $z_{2,2}(0) < m_2$ , then the initial state is not in  $\mathbb{S}$ , so we cannot use the ODE (13) to describe the system. There is a transient period  $[0, t_{\mathbb{S}})$  during which the two service pools fill up, but the system is not yet overloaded.

If sharing is eventually going to take place (i.e., if  $x^*$  is in either  $\mathbb{A}$  or  $\mathbb{S}^+$ ), then with initial conditions as in (53), we should certainly hit  $\mathbb{S}^b$ . Sharing will begin only at a time  $T$  such that  $q_1(T) - r q_2(T) = \kappa$ . In this section we show that, if indeed  $x^* \in \mathbb{A} \cup \mathbb{S}^+$ , then  $T < \infty$ , where

$$T \equiv \inf\{t \geq 0 : x(t) \in \mathbb{S}^b\}. \quad (54)$$

The transient period of the fluid system can be divided into two distinct periods: The first transient period, on the interval  $[0, T)$ , lasts until the fluid limit hits  $\mathbb{S}^b$ . The second transient period is the one starting at the hitting time  $T$ , and is described by the ODE (14). This period was analyzed in the previous sections. The first transient period is described by different ODE's, depending on the state of the system. These ODE's, for the initial condition in (53), are given in the proof of Theorem 8.1 below.

We shall prove that  $T < \infty$  under the extra assumption that at no time during  $[0, T)$  is  $z_{2,1} > 0$ . The assumption can be verified directly by solving the fluid model of the first transient period. We discuss this condition after the proof of Theorem 8.1.

**THEOREM 8.1** *If  $x^* \in \mathbb{A} \cup \mathbb{S}^+$ , if (53) holds and if  $z_{2,1}(t) \equiv 0$  for all  $t \geq 0$ , then  $T < \infty$ , for  $T$  in (54).*

**PROOF.** We start by developing the ODE to describe the system before hitting  $\mathbb{S}$ . As before, we do not consider the original queueing model and prove convergence to the appropriate fluid limit, but instead we develop the ODE directly. We first consider the case in  $s_2^a > 0$  (so that  $q_2^a = 0$ ), i.e., class 2 experiences no overload by itself (before pool 2 starts serving class-1 fluid). First, there is an initial period in which the pools are being filled with fluid. It is easy to see that as long as neither pool is full, the pool-content functions  $z_{i,i}(t)$  behave as the fluid approximations for the number in system at time  $t$  in an  $M/M/\infty$  queueing model with arrival rate  $\lambda_i$  and service rate  $\mu_{i,i}$ ,  $i = 1, 2$ ; e.g., see [8] (where it assumed that  $\lambda = \mu$ , so that  $\lambda/\mu = 1$ ). Therefore, the system evolution is described by the pair of ODE's

$$\begin{aligned} \dot{z}_{1,1}(t) &= \lambda_1 - \mu_{1,1} z_{1,1}(t), & z_{1,1}(0) &= \zeta_1 \\ \dot{z}_{2,2}(t) &= \lambda_2 - \mu_{2,2} z_{2,2}(t), & z_{2,2}(0) &= \zeta_2, \end{aligned}$$

and the unique solution to each ODE is

$$z_{i,i}(t) = \frac{\lambda_i}{\mu_{i,i}} + \left( \zeta_i - \frac{\lambda_i}{\mu_{i,i}} \right) e^{-\mu_{i,i} t}, \quad t \geq 0, \quad i = 1, 2.$$

These ODE's describe the dynamics of the two classes until one of the pools is full, i.e., until the time

$$t_1 \equiv \min_{i=1,2} \inf\{t \geq 0 : z_{i,i}(t) = m_i\}. \quad (55)$$

Since we assume that  $s_2^a > 0$ ,  $t_1$  is the time at which  $z_{1,1}(t) = m_1$ , and at this time we need to start considering  $q_1$ . Clearly,  $q_1$  evolves independently of class 2 until  $q_1(t) = \kappa$  (when sharing is initialized). Let

$$t_2 \equiv \inf\{t \geq t_1 : q_1(t) = \kappa\}. \quad (56)$$

Recall that  $\kappa$  may be equal to 0, in which case  $t_1 = t_2$ . If  $t_2 > t_1$ , then  $q_1(t)$ ,  $t \in [t_1, t_2)$ , evolves as the fluid approximation for the queue-length process in an Erlang-A model operating in the ED MS-HT regime, as in [14]. The ODE describing the evolution of  $q_1$  is

$$\dot{q}_1(t) = \lambda_1 - \mu_{1,1}m_1 - \theta_1q_1(t), \quad t_1 \leq t < t_2, \quad \text{with } q_1(t_1) = 0, \quad (57)$$

and its unique solution is

$$q_1(t) = \frac{\lambda_1 - \mu_{1,1}m_1}{\theta_1} \left(1 - e^{-\theta_1(t-t_1)}\right), \quad t_1 \leq t < t_2.$$

Now, since  $q_1(t_2) = \kappa$  and  $q_2(t_2) = 0$ , class-1 fluid starts flowing to service pool 2, so that  $z_{1,2}$  starts increasing. There is a time  $t_3$  such that, for  $t \in [t_2, t_3)$ ,  $q_1(t) = \kappa$ ,  $q_2(t) = 0$  and all the excess class-1 fluid, that is not lost due to abandonment, is flowing to pool 2. Hence,  $z_{1,2}$  satisfies the ODE

$$\dot{z}_{1,2}(t) = (\lambda_1 - \mu_{1,1}m_1 - \theta_1\kappa) - \mu_{1,2}z_{1,2}(t), \quad t_2 \leq t < t_3, \quad \text{with } z_{1,2}(t_2) = 0,$$

whose unique solution is

$$z_{1,2}(t) = \frac{\lambda_1 - \mu_{1,1}m_1 - \theta_1\kappa}{\mu_{1,2}} \left(1 - e^{-\mu_{1,2}(t-t_2)}\right), \quad t_2 \leq t < t_3.$$

Hence,  $t_3 \equiv \inf\{t \geq t_2 : z_{1,2}(t) + z_{2,2}(t) = m_2\}$ , so that at time  $t_3$  both service pools are full, with  $q_1(t_3) = \kappa$ ,  $q_2(t_3) = 0$  and  $q_1(t_3) - rq_2(t_3) = \kappa$ . It follows that  $t_3$  is the time at which the fluid model hits the space  $\mathbb{S}^b$ , and the first transient period is over, i.e.,  $t_3 = T$  for  $T$  in (54).

Now we consider the second case in which  $q_2^a > 0$ . In this case there are different scenarios: In the first scenario, pool 2 can be filled before pool 1, so that  $t_1 = \inf\{t \geq 0 : z_{2,2} = m_2\}$ , for  $t_1$  in (55). In that case  $q_2$  begins to increase at time  $t_1$ , evolving according to the ODE of the overloaded Erlang-A model

$$\dot{q}_2(t) = \lambda_2 - \mu_{2,2}m_2 - \theta_2q_2(t).$$

However, by the assumption of the theorem, we have ruled out the case in which  $q_1(t) - r_{2,1}q_2(t) = \kappa_{2,1}$ , so that no class-2 fluid will flow to pool 1. Hence, from the beginning (time 0),  $z_{1,1}$  increases until time  $t'_1 \geq t_1$  at which  $z_{1,1} = m_1$ . Then  $q_1$  increases, satisfying (57) with  $q_1(t'_1) = 0$ . By the assumption on  $x^*$ , and following Corollary 6.2, there exists a time  $T < \infty$  such that  $q_1(T) - rq_2(T) = \kappa$ . This is because  $rq_2(t) \leq rq_2^a < q_1^a - \kappa$  for all  $t \leq T$ . On the other hand, it follows trivially from the solution to (57), that  $q_1^a$  is the globally asymptotically stable point of (57). Hence, for every  $\epsilon > 0$ , there exists  $t_\epsilon$  such that  $q_1(t) > q_1^a - \epsilon$  for all  $t \geq t_\epsilon$ . (This is because, by the initial conditions,  $q_1(t) \leq q_1^a$  for all  $t$ ). Thus, we can find  $\epsilon > 0$  such that

$$rq_2^a < q_1^a - \epsilon - \kappa < q_1(t) - \kappa \text{ for all } t \geq t_\epsilon. \quad (58)$$

The second scenario of the second case has pool 1 filled first at time  $t_1$ , so that  $q_1$  starts increasing according to (57). If  $q_1$  reaches  $\kappa$  before  $q_2$  starts increasing, then we have the same behavior as when  $s_2^a > 0$ . However, if at time  $t_2$  in (56)  $q_2 > 0$ , then the two queues will continue increasing independently until time  $T$ . Once again, (58) can be shown to hold, so that  $T < \infty$ .  $\square$

We can easily calculate the exact value of  $x(T)$  and use it to calculate the QBD drift rates  $\delta_+(x(T))$  and  $\delta_-(x(T))$  to find whether the positive-recurrence condition (27) holds at  $T$ , so that  $x(T) \in \mathbb{A}$ .

**REMARK 8.1** (*sharing in the wrong direction*) In Theorem 8.1 we assumed that we never have  $z_{2,1} > 0$ . The reason is that, if  $z_{2,1}$  ever does become positive, then the fluid  $x$  never hits the region  $\mathbb{S}$ . To see that this is so, suppose that for some time  $t_4$  sharing is initialized, with class-2 fluid flowing to service pool 1. Then  $z_{2,1}$  is increasing until a time  $t_5$  at which  $q_1(t_5) - rq_2(t_5) = \kappa$ , and the AP begins to operate. At that time,  $z_{2,1}$  will start decreasing according to the ODE

$$\dot{z}_{2,1}(t) = -\mu_{2,1}z_{2,1}(t), \quad t \geq t_5,$$

whose unique solution is

$$z_{2,1}(t) = z_{2,1}(t_5)e^{-\mu_{2,1}(t-t_5)}, \quad t \geq t_5. \quad (59)$$

Hence  $z_{2,1}$  remains strictly positive for all  $t \geq t_5$ , and  $\mathbb{S}$  is never hit.

Of course, the fluid state should be approaching a state in  $\mathbb{S}$  as  $t$  increases. However, if there is such a limit point, then that limit point itself typically will *not* be a stationary point, because if  $x$  did start at that limit point, then it will have to continue to move toward the final stationary point  $x^*$ .

More generally, the failure of  $z_{2,1}$  to actually reach 0 in finite time has practical implications for the FQR-T control in the original queueing system. It suggests that it should be beneficial to introduce lower positive thresholds for  $z_{1,2}$  and  $z_{2,1}$ , below which we relax the one-way sharing restriction. It remains to examine the system performance in response to such more complex transient behavior.

For the cases covered by Theorem 8.1, the system evolution over the entire halfline  $[0, \infty)$  is a continuous “soldering” of the different ODE’s, but at the soldering points  $t_i$ , the functions under consideration are typically not differentiable. Hence, there is no single ODE that captures the full dynamics of the system. To see why, consider the case in which  $s_2^a > 0$  and  $\kappa > 0$ . Then, for  $t < t_1$ ,  $q_1(t) = 0$  and  $\dot{q}_1 = 0$ , but for  $t_1 \leq t < t_2$ ,  $q_1(t)$  evolves according to (57), which typically has a strictly positive derivative at  $t_1$ . Thus the left and right derivatives at  $t_1$  are not equal. Similar arguments hold for all the other soldering points.

We observe that all the fluid approximations used in the proof of Theorem 8.1 can be shown to hold as fluid limits of a sequence of scaled queueing processes. In fact, these MS-HT fluid limits are much easier to establish than the MS-HT convergence to the fluid limit described by (13), since they do not include the AP. As a consequence, their limiting ODE’s are continuous in their full state spaces. In addition, the ODE’s describing the fluid limits have unique closed-form solutions.

**9. A Numerical Algorithm to Solve the IVP.** In this section we provide a numerical algorithm for solving the IVP (15). To the best of our knowledge, there are no other algorithms available to solve such an IVP. The difficulty, of course, is that the ODE is driven by the stochastic FTSC process  $D_t$ . Having an efficient algorithm for solving the IVP clearly is vital for having the fluid approximation be a useful tool for applications, but the algorithm is also important for other reasons. First, establishing convergence by the method in §7.2 (when the sufficient conditions for global stability in §7.1 do not hold) depends on calculating the solution up to a finite time  $T$ , where we can observe that the solution is close enough to the stationary point  $x^*$ , for which an explicit expression is given in §6. Second, the ability to solve the IVP provides a powerful demonstration of the AP, and a verification of its correctness, because we can compare it to simulation results. The close agreement with simulation also shows that the overall approximation is effective; see the numerical example below and the comparisons between the fluid solutions to simulation results in [10].

**9.1 Computing  $\pi_{1,2}(x)$  at a point  $x$ .** In §4.2 we saw that our representation of the FTSP  $D_t$  as a QBD was very helpful for characterizing positive recurrence and the set  $\mathbb{A}$  where the AP prevails. This QBD structure also plays a key role in our numerical algorithm. The QBD structure allows us to use established efficient numerical algorithms 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 with a given  $x \in \mathbb{A}$ , so that averaging is taking place. As before, we assume that class 1 is overloaded, and that service pool 2 is helping class 1. From (31) it is clear that we must start with computing the rate matrix  $R \equiv R(x)$ . (To simplify notation, we drop the argument  $x$  with the understanding that all matrices, and the vector  $\alpha_0$  are functions of  $x$ .)

We exploit the well-developed theory for QBD processes in Latouche and Ramaswami [6]. By Proposition 6.4.2 of [6], the matrix  $R$  is related to two other matrices,  $G$  and  $U$ , via

$$G = (-U)^{-1}A_2, \quad U = A_1 + A_0G \quad \text{and} \quad R = A_0(-U)^{-1}. \quad (60)$$

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

$$A_2 + A_1G + A_0G^2 = 0 \quad \text{and} \quad A_0 + RA_1 + R^2A_2 = 0. \quad (61)$$

Hence, if can compute the matrix  $G$ , then the rate matrix  $R$  can be found via (60). Once  $R$  is known, we use (30) to compute  $\alpha_0$ . With  $\alpha_0$  and  $R$  in hand,  $\pi_{1,2}(x)$  is easily computed via (31).

It remains to compute the matrix  $G$ . In §8 of [6], three different numerical algorithms to calculate  $G$  are provided. We chose to use the *logarithmic reduction algorithm* in §8.4, modified to the continuous case, as in §8.7, in [6]. As reviewed there, this algorithm is quadratically convergent (as opposed to the linear rate of convergence of the other two algorithms), and is numerically well behaved. These two properties are important for us, since we need to compute the matrix  $R(x)$  for thousands of points  $x$  when we numerically solve the IVP (15). From our experience with this algorithm, it takes fewer than ten iterations to achieve a  $10^{-6}$  precision (when calculating  $G$ ).

**9.2 Computing the Solution  $x$ .** To compute the solution  $\{x(t) : 0 \leq t \leq T\}$ , 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 (27) 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  $\pi_{1,2}(x(t)) = 1$ . As long as we the calculated solution remains within one of the regions  $\mathbb{A}$ ,  $\mathbb{S}^+$  or  $\mathbb{S}^-$ , we know that we are calculating the unique solution to the IVP, by virtue of Theorem 5.4 and Remark 5.1. We do not yet have such a supporting theoretical result in  $\mathbb{S}^b - \mathbb{A}$ , but numerical experience indicates that this method is effective.

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 use the same procedure to find  $x(2h)$ ,  $x(3h)$ ,  $\dots$ ,  $x(nh)$ , i.e., for each  $k = 0, 1, 2, \dots, n-1$ ,

$$x((k+1)h) = x(kh) + h\Psi(x(kh)), \quad 0 \leq k \leq n, \quad (62)$$

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 §8, where at each Euler step we check to see which fluid model should be applied.

We have chosen to use the forward Euler algorithm, although it is known to have an error proportional to the step size  $h$ , and to be relatively numerically unstable at times. We have two reasons for doing so: First, the Euler method is the simplest numerical method for solving ODE's. Thus, one can immediately observe the main structure of the algorithm. It is also very easy to see how to apply more sophisticated algorithms, such as general linear methods, which have a smaller error, and can be more numerically stable. The only adjustment needed, is to replace the Euler step in (62) by the different method. At any iteration,  $\pi_{1,2}$  is computed as in §9.1. Moreover, as can be seen the numerical example below,  $\pi_{1,2}$  is almost constant throughout (starting at the time  $x$  hits the set  $\mathbb{A}$ ). This suggests that the solution behaves very much like a simple exponential function (strengthening the result of §7.3), which is very smooth and stable. Hence, we have no problem with numerical stability with the Euler method.

In the numerical example in §9 we took the ratio  $r = 0.8 = 4/5$ , so that all the matrices, used in the computations for  $\pi_{1,2}$ , are of size  $10 \times 10$ . 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  $50 \times 50$ , 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. See the example below. See also [9] for comprehensive verifications via simulation experiments. However, there are two features of the numerical solution itself that strongly suggest its validity. First, we can check whether the solution converges to the stationary point  $x^*$ , which can be computed explicitly using (38). An even stronger verification of the solution's correctness is the fact that the two queues keep at the ratio  $r$ , even though this relation between the two queues is not forced explicitly by the algorithm (it is only used to

calculate  $\pi_{1,2}$ . Hence it appears implicitly in the ODE via the expression for  $\pi_{1,2}$ ). Specifically, the fact that the SSC equation,  $q_1(t) - rq_2(t) = \kappa$ , holds for all  $t$  from the moment the solution hits  $\mathbb{S}$ , is a strong evidence that  $\pi_{1,2}(t)$  (and, consequently,  $x(t)$ ) is computed correctly; See Figure 1.

**9.3 A Numerical Example.** Below are figures produced by a Matlab code implementing the algorithm above. In addition, we added the sample paths of the stochastic processes  $Q_1$  and  $Z_{1,2}$ , on top of the trajectories of the solution to their fluid counterparts  $q_1$  and  $z_{1,2}$ . These sample paths were created by a single simulation run. The model is the same one introduced in §4.2 with component rate matrices in (23). The model parameters are  $m_1 = m_2 = 1000$ ,  $\lambda_1 = 1300$ ,  $\lambda_2 = 900$ ,  $\mu_{1,1} = \mu_{2,2} = 1$ ,  $\mu_{1,2} = \mu_{2,1} = 0.8$  and  $\theta_1 = \theta_2 = 0.3$ . We take  $\kappa = 0$  and  $r = 0.8$ . We chose to take a relatively large system ( $n = 1000$ ), so that the stochastic fluctuations do not to hide the general structure of the simulated sample paths. The time-dependent mean values follow the fluid solutions very closely, as can be confirmed by considering multiple replications; see [10] for more comparisons with simulations. There it is shown that even for surprisingly small systems (e.g., with 25 agents in each pool) the mean values are well approximated by the fluid.

We ran the algorithm and the simulation for 50 time units. Since we used an Euler step of size  $h = 0.01$ , we performed 5000 Euler iterations, but in each Euler iteration we performed several iterations to calculate the matrix  $G$  in (60), which is used to calculate the instantaneous steady-state probability  $\pi_{1,2}$ . The QBD matrices for this example with  $r = 0.8$  appear in (23).

Figures 1-4 show the curves of the ratio between the queues (as a function of  $t$ , i.e., the actual ratio between the queues through time),  $\pi_{1,2}$ ,  $q_1$  together with  $Q_1$ , and  $z_{1,2}$  together with  $Z_{1,2}$ , for a system initializing empty. After a short period in which the pools fill up,  $q_1(t)$  starts to grow, and immediately then fluid (customers) starts flowing to pool 2, causing  $z_{1,2}(t)$  to grow. At this initial time period, the stochastic processes and their fluid approximations are almost indistinguishable.

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. As discussed before, 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. It is interesting that once  $\mathbb{S}^b$  is hit,  $\pi_{1,2}$  becomes almost a constant, even before the system reaches steady state. This explains why the curves of  $q_1$ ,  $q_2$  and  $z_{1,2}$  resemble the curves of exponential functions, and strengthens the results of §7.3. (Observe that if  $\pi_{1,2}(x(t))$  is replaced by a constant in the ode (14), then its solution is easily seen to be an exponential function.)

When the algorithm terminated, the value of  $x(t_n)$  was  $q_1(t_n) = 363.9$ ,  $q_2(t_n) = 455.0$  and  $z_{1,2}(t_n) = 238.5$ . Also,  $\pi_{1,2}(t_n) = 0.2$ . Calculating the value of  $x^* = (q_1^*, q_2^*, z_{1,2}^*)$  (using (38)) we have  $x^* = (366.7, 459.5, 237.5)$ . Plugging  $z_{1,2}^*$  in (39), we get  $\pi_{1,2}^* = 0.2$ . As we mentioned before, these steady-state values also suggest that the algorithm is achieving the correct solution to the ODE.

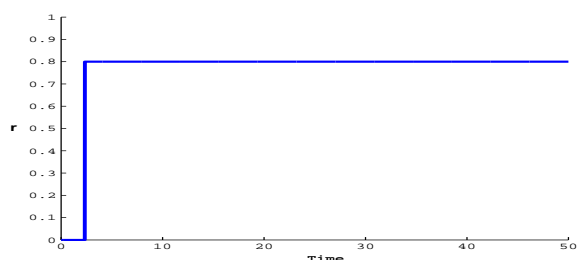


Figure 1: ratio between the queues.

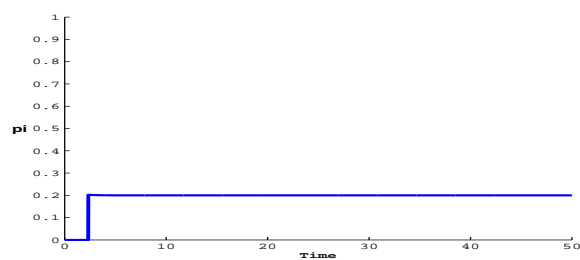


Figure 2:  $\pi_{1,2}$  calculated at each iteration.

Note that in this example, the sufficient conditions for strong SSC in §7.1 do not hold. Specifically, condition (ii) in Theorem 7.1 does not hold since  $\lambda_2 = 900 > \nu m_2 = 800$ , for  $\nu \equiv \mu_{1,2} \wedge \mu_{2,2}$ . Observe that Condition (i) in that theorem does hold, since  $\lambda_1 = 1300 < \nu m_2 + \mu_{1,1} m_1 = 1800$ ; See Remark 7.1.

However, this example shows how useful the results of §7.2 are. By Theorem 7.2 we have  $\alpha = \xi / r \theta_2$ , where  $\xi \equiv |\delta_+^*| \wedge \delta_-^*$ . With the value of  $z_{1,2}^*$  computed above, it follows that  $\xi = \delta_-^* = 342$ , so that  $\alpha = 1425$ . This means that  $x(t)$ ,  $t \geq T$ , where  $T$  is the time the solution hits  $\mathbb{A}$ , is known to lie

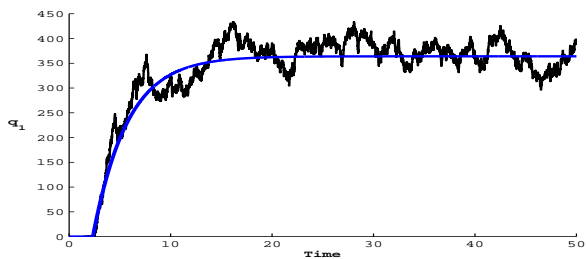


Figure 3: trajectory of  $q_1$  together with a simulated sample path of the stochastic process  $Q_1$  in a system initializing empty.

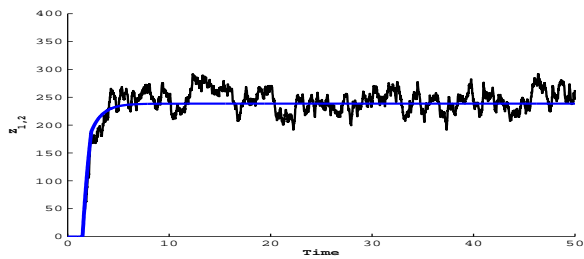


Figure 4: trajectory of  $z_{1,2}$  together with a simulated sample path of the stochastic process  $Z_{1,2}$  in a system initializing empty.

entirely in  $\mathbb{A}$  without even solving the algorithm. That is because  $x(T) = (0, 0, 100) \in \beta_V(\alpha)$ , and  $\beta_V(\alpha) \subset \mathbb{A}$ . (Recall that the solution hits  $\mathbb{S}^b$  when  $z_{1,2} + z_{2,2} = m_2$ . In our example it is easy to see that  $z_{2,2}(T) = \lambda_2 = 900$ , so that  $z_{1,2}(T) = 100$ . Since  $\kappa = 0$ , we also have  $q_1(T) = q_2(T) = 0$ . We can calculate  $\delta_-(x(T))$  and  $\delta_+(x(T))$ , to conclude that  $x(T) \in \mathbb{A}$ .)

**10. Conclusions and Further Research.** In this paper we analyzed the deterministic ODE (13)-(14), arising as the MS-HT fluid limit of the overloaded X call-center model operating under the FQR-T control. In addition to being an interesting mathematical object in its own right, the ODE analyzed in this paper is a vital link between our past research in [9, 10] and our future research in [11].

We showed that the existence of a unique solution to the IVP (15) depends heavily on the characterization of the function  $\Psi$  in (13) and its topological properties. These properties, in turn, depend on the state space of  $\Psi$ , and the regions of the state space in which  $\Psi$  is continuous. These regions are further characterized by the probabilistic properties of the family of FTSC processes  $\{D_t : t \geq 0\}$ . The existence of a global unique solution further depends on other properties of the solution, specifically, its stability. Since the proof of convergence depends on the uniqueness of the solution to the IVP, this paper prepares the way for the future work in [11].

The connection to the past research is clear: First, we prove that the stationary point  $x^*$ , which was developed heuristically in [9] using flow-balance arguments, and was claimed to be the stationary point of (14) in [10], using reasonings similar to those in §6, is indeed the unique stationary point for the fluid. Moreover, we provided mild conditions assuring the convergence of the solution to  $x^*$ . We also showed that the convergence to  $x^*$  is exponentially fast, further justifying the steady-state analysis in [9].

To fully connect to the model considered in our previous papers, in §8 we considered the system at the time when the arrival rates change. At that time, denoted by 0, the system will typically be underloaded, so that the state space should not be  $\mathbb{S}$ . After the change, we assume that the arrival rates are larger than the total service rate of the two pools. Specifically, we assumed Assumption A in §6. We then considered the first transient period  $[0, T)$ , where  $T$  is the time at which  $\mathbb{S}^b$  is hit. Using alternative fluid models (ODE's), we showed that  $T < \infty$ , under the conditions of Theorem 8.1. The solutions to the fluid models during the first transient period are all exponential functions, so that this period also passes exponentially fast.

Finally, we developed an efficient algorithm to solve the IVP (15), based on the matrix geometric method. This algorithm solves the different fluid models described in §8, and combines these solutions with the solution to (14) once the set  $\mathbb{A}$ , where the AP takes place, is hit.

Our main results in this paper were based on classical results from ODE theory, specifically the Picard-Lindelöf theorem establishing the existence and uniqueness of solutions to IVP's, and the theory of QBD processes. Since the function  $\Psi$  appearing in (13) is not continuous in  $\mathbb{S}$ , and not Lipschitz continuous in  $\mathbb{S}^b - \mathbb{A}$ , we could not apply this theorem for solutions that are not known to be confined to one region. We do not yet have a proof that a global solution to the IVP exists in general, or that a solution passing through  $\mathbb{S}^b - \mathbb{A}$  is unique in that region.

It also remains to generalize Theorem 8.1, and include the case in which  $z_{2,1}$  becomes positive during

the first transient period. We do make the following conjecture:

**CONJECTURE 10.1** *Make Assumption A as usual and introduce lower thresholds as in Remark 8.1. If the appropriate ODE is defined for each relevant region, as in the proof of Theorem 8.1, then  $x(t) \rightarrow x^*$  as  $t \rightarrow \infty$ , where  $x^* \in \mathbb{S}$ , for any initial state  $x(0)$ , in  $\mathbb{S}$  or not.*

It also remains to consider more complicated dynamics than provided by a single change in the arrival rates. The numerical algorithm applies more generally, but it remains to establish mathematical results and examine the performance. For example, it remains to consider a second overload incident happening before the system has recovered from the first one.

**Acknowledgments** This research was supported by NSF grant DMI-0457095.

## References

- [1] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [2] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.
- [3] I. Gurvich and W. Whitt, *Queue-and-idleness-ratio controls in many-server service systems*. Math. Oper. Res. **34** (2) (2009), 363–396.
- [4] Q. He, *Differentiability of the matrices  $R$  and  $G$  in the matrix analytic method*. Stochastic Models **11** (1) (1995), 123–132.
- [5] H. K. Khalil, *Nonlinear Systems*. Prentice Hall, New Jersey, 2002.
- [6] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, Siam and ASA, Philadelphia, 1999.
- [7] M. J. Horacio, *Nonlinear Control Systems*. Wiley, New Jersey, 2003.
- [8] G. Pang, R. Talreja and W. Whitt, *Martingale proofs of many-server heavy-traffic limits for Markovian queues*. Probability Surveys. **4** (2007), 193–267.
- [9] O. Perry and W. Whitt, *Responding to unexpected overloads in large-scale service systems*. Management Sci., forthcoming.
- [10] O. Perry and W. Whitt, *A fluid approximation for service systems responding to unexpected overloads*. Submitted to Oper. Res.
- [11] O. Perry and W. Whitt, *Heavy-traffic limits for the overloaded X model via an averaging principle*. in preparation.
- [12] G. Teschl, *Ordinary Differential Equations and Dynamical Systems*, Universität Wien, 2009. Available online: <http://www.mat.univie.ac.at/~gerald/ftp/book-ode/ode.pdf>
- [13] W. Whitt, *Stochastic-Process Limits*, New York, Springer, 2002.
- [14] W. Whitt, *Efficiency-driven heavy-traffic approximations for many-server queues with abandonments*. Management Sci. **50** (10) (2004), 1449–1461.

**Appendix A. More on the Algorithm.** In this appendix we elaborate further on the algorithm introduced in §9. Let  $\{t_m : m = 0, 1, 2, \dots, n\}$  be the Euler steps, with  $t_{m+1} - t_m = h$ . In our experiments we found  $h = 0.01$  to be a good candidate for the step size since it is small enough to minimize numerical errors, while the number of iterations needed for the ODE to reach its stationary point, is just a few thousands. Hence the algorithm takes only a few seconds to terminate.

Let  $\bar{D}(t) \equiv q_1(t) - rq_2(t)$ , denote the weighted difference between the two fluid queues. The discretization of the ODE in the numerical algorithm means that if, at step  $k - 1$ ,  $\bar{D}(t_{k-1}) \notin \mathbb{S}^b$  but is close to it, then  $\bar{D}(t_k)$  may miss the boundary, even though the (continuous) ODE is at the boundary at time  $t_k$ . For that reason, if  $\kappa - h < \bar{D}(t_k) < \kappa + h$ , then we force  $x(t_k)$  to be in  $\mathbb{S}^b$ , by taking  $\bar{D}(t_k) = \kappa$ . Once we have  $\bar{D}(t_k) = \kappa$  we decide whether to keep staying on the boundary for the next Euler step, by checking whether (27) holds. According to the relation between the QBD drift rates at time  $t_k$ , we decide whether we should apply the AP, in order to find  $\pi_{1,2}(t_k)$ , or rather set  $\pi_{1,2}(t_k)$  to zero or one.

At any step in the algorithm, we must also decide which ODE to use. That depends on the state of the system at each time, as described in §8. If the fluid state is not in  $\mathbb{S}$ , as in the initial period of the example in §9 and the example below, then we use the appropriate fluid model, as given in the proof of Theorem 8.1.

**A.1 An Example with  $x^* \in \mathbb{S}^+$ .** We now consider the same example as in §9.3, except now we increase the arrival rate for class 1 substantially, so that  $x^* \in \mathbb{S}^+$ . In particular, we let  $\lambda_1 = 3000$  instead of 1300. Once again, the system is initialized empty. That means that the fluid solution in  $\mathbb{S}$  is moving between the two regions  $\mathbb{S}^b$  and  $\mathbb{S}^+$ . In particular, the solution first hits  $\mathbb{S}^b$  (specifically,  $\mathbb{S}^b - \mathbb{A}$ ), as was proved in Theorem 8.1, but it stays there for a short amount of time, and then crosses to  $\mathbb{S}^+$ .

We see how  $z_{2,2}$  starts increasing up to the time  $T$  in which  $z_{1,2}(T) + z_{2,2}(T) = m_2$ . At this time  $z_{2,2}(T)$  starts decreasing, and is replaced by class-1 fluid. Since no class-2 fluid is flowing to either of the service pool, all the class-2 fluid output is due to abandonment. We can also observe that  $z_{2,2}$  eventually hits 0, even though  $z_{2,2}$  satisfies the equation (59). This is due to the numerical errors, as described in §8.

In steady-state we have  $q_2^* = \lambda_2/\theta_2 = 900/0.3 = 3000$  and  $q_1^* = (\lambda_1 - m_1\mu_{1,1} - m_2\mu_{1,2})/\theta_2 = 4000$ , as in Corollary 6.1 (ii).

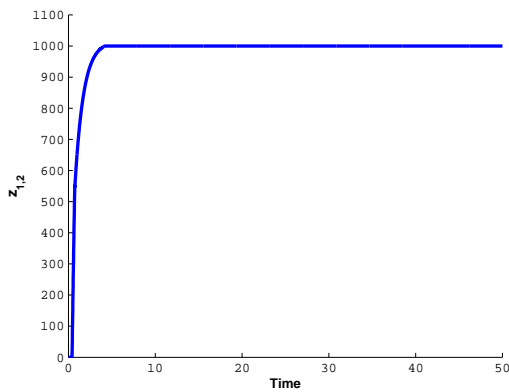


Figure 5:  $z_{1,2}$  when  $\lambda_1$  exceeds the system's capacity.

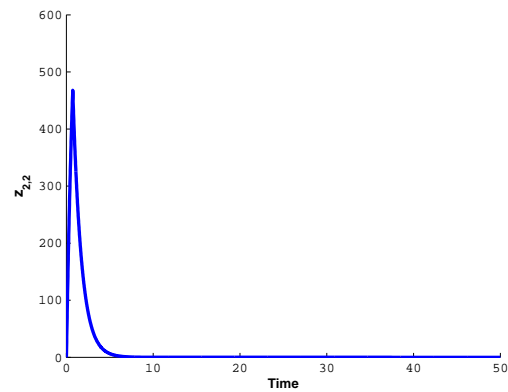


Figure 6:  $z_{2,2}$  when  $\lambda_1$  exceeds the system's capacity.

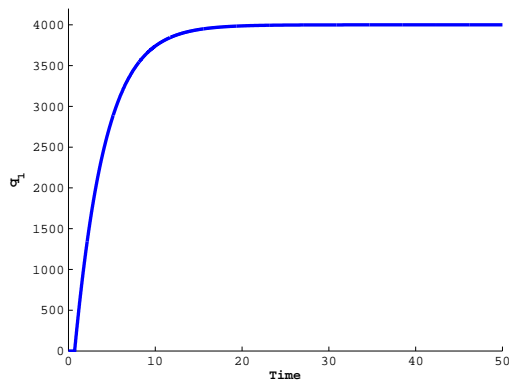


Figure 7:  $q_2$  when  $\lambda_1$  exceeds the system's capacity.

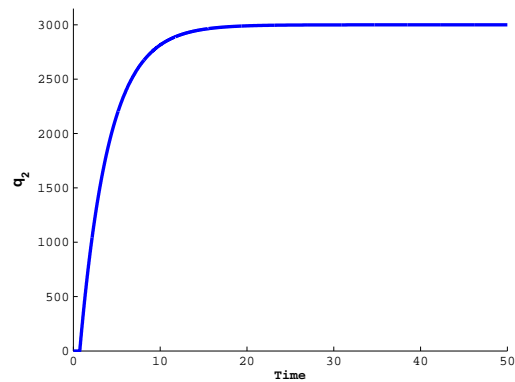


Figure 8:  $q_1$  when  $\lambda_1$  exceeds the system's capacity.